Bioinformatics Advance Access published January 31, 2007

BIOINFORMATICS Original Paper

Structural Bioinformatics

De Novo SVM Classification of Precursor MicroRNAs from Genomic Pseudo Hairpins Using Global and Intrinsic Folding Measures

Stanley NG Kwang Loong[†,‡,*], Santosh K. MISHRA[†,‡]

[†] Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

[‡] NUS Graduate School for Integrative Sciences & Engineering, Centre for Life Sciences, #05-01, 28 Medical Drive, Singapore 117456

Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: MicroRNAs (miRNAs) are small ncRNAs participating in diverse cellular and physiological processes through the post-transcriptional gene regulatory pathway. Critically associated with the miRNAs biogenesis, the hairpin structure is a necessary feature for the computational classification of novel precursor miRNAs (*pre-miRs*). Though many of the abundant genomic inverted repeats (pseudo hairpins) can be filtered computationally, novel specie-specific *pre-miRs* are likely to remain elusive.

Results: *miPred* is a *de novo* Support Vector Machine (SVM) classifier for identifying *pre-miRs* without relying on phylogenetic conservation. To achieve significantly higher sensitivity and specificity than existing (quasi) *de novo* predictors, it employs a Gaussian Radial Basis Function kernel (RBF) as a similarity measure for 29 global and intrinsic hairpin folding attributes. They characterize a *pre-miR* at the dinucleotide sequence, hairpin folding, non-linear statistical thermodynamics, and topological levels. Trained on 200 human *pre-miRs* and 400 pseudo hairpins, *miPred* achieves 93.50% (five-fold cross-validation accuracy) and 0.9833 (area under the ROC). Tested on the remaining 123 human *pre-miRs* and 246 pseudo hairpins, it reports 84.55% (sensitivity), 97.97% (specificity), and 93.50% (accuracy). Validated onto 1,918 *pre-miRs* across 40 non-human species and 3,836 pseudo hairpins, it yields 87.65% (92.08%), 97.75% (97.42%), and 94.38% (95.64%) for the mean (overall) sensitivity, specificity, and accuracy. Notably, *A. mellifera*, *A. Geoffroyi*, *C. familiaris*, *E. Barr*, *H. Simplex virus*, *H. cytomegalovirus*, *O. aries*, *P. patens*, *R. lymphocryptovirus*, *Simian virus*, and *Z. mays* are unambiguously classified with 100.00% (sensitivity) and >93.75% (specificity).

Availability: Datasets, raw statistical results, and source codes are available at http://web.bii.a-star.edu.sg/~stanley/Publications

Contact: stanley@bii.a-star.edu.sg*; santosh@bii.a-star.edu.sg

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

MicroRNAs (miRNAs) constitute an abundant class of small (~21–23-nts), endogenous, and evolutionarily conserved ncRNA molecules that mediate post-transcriptionally the production of intra-cellular proteins in most eukaryotes via sequence-specific target mechanisms (Bartel 2004). The founding members *lin*-4 and *let*-7 miRNAs discovered respectively in 1993 and 2000, are key heterochronic regulators directing temporal aspects of development timing in early larval *C. elegans* (Lee *et al*., 1993; Reinhart *et al*., 2000). Subsequently, thousands of novel miRNA genes have been unraveled across plants, worms, flies, vertebrates, and even viruses; >4000 miRNAs spanning 45 species are listed in miRBase 8.2 (Griffiths-Jones *et al*., 2006). Biologically pivotal and more prevalent genomically than presumed, miRNAs perform key regulatory roles in diverse cellular and physiological events such as apoptosis, proliferation, and fat metabolism in the *D. melanogaster* (Brennecke *et al*., 2003; Xu *et al*., 2003); patterning and developmental specification in plants (Chen 2004; Palatnik *et al*., 2003); genetic diseases including oncogenesis (Calin and Croce 2006; Lu *et al*., 2005).

Previously, novel miRNA genes were identified almost exclusively by directional cloning of endogenous small RNAs and high-throughput sequencing of large numbers of cDNA clones (Lagos-Quintana *et al*., 2001; Lau *et al*., 2001; Lee and Ambros 2001). Conventional forward genetic screening is highly biased towards abundantly and/or ubiquitously expressed miRNAs that usually dominate the cloned products (Lagos-Quintana *et al*., 2003). Evidently, miRNAs expressed constitutively at low levels or in highly constrained tissue- and time-specific patterns are intricate to detect experimentally. Computational prediction techniques have been employed extensively to overcome this technical hurdle (Berezikov *et al*., 2006). The underlying principle revolves around two tenets. First, precursor miRNAs (*pre-miRs*) should possess statistically significant and evolutionarily conserved (a)symmetric RNA hairpin, a structural prerequisite functionally critical for the early stages of the mature miRNA biogenesis (Bartel 2004; Kim 2005); see supplementary "Biogenesis of Mature MicroRNAs" for details. Second, the hairpin feature of *pre-miRs* should be distinct from those of random inverted repeats (termed as pseudo hairpins) that can potentially fold into dysfunctional candidate hairpins e.g., $1.1E^7$ in human (Bentwich *et al*., 2005) and $4.4E^4$ in *C. elegans* (Pervouchine *et al*., 2003). Removing these overwhelming and irrelevant genomic pool of false-positives without sacrificing excessively putative *pre-miRs* is most technically challenging, as they are relatively short in length (~60–80-nts in animal and ~100−400-nts in plants) and have highly diverse base compositions (Zhang *et al*., 2006b). Unlike protein-coding genes, they frequently exhibit seemingly weak or lack detectable statistically significant primary-sequence signals such as the open reading frames (ORFs), promoter motifs, and codon signatures (Berezikov *et al*., 2006).

Earlier attempts in circumventing these difficulties relied on identifying close homologs of published *pre-miRs* e.g., *let-7* (Pasquinelli *et al*., 2000). This can be as straightforward as aligning sequences through NCBI BlastN (McGinnis and Madden 2004) while allowing several mismatches and gaps depending on their inter-phylogenetic distance. False-positives not residing in the orthologous locations are deemed not conserved phylogentically between closely related species, and are consequently masked (Floyd and Bowman 2004; Pasquinelli *et al*., 2000). The candidate orthologues of evolutionary conserved miRNAs genes are then assessed for their capability to potentially fold into hairpin structures with the lowest minimum free energy of folding (MFE), have ≥16-bps involving the first 22-nts of the mature miRNA embedded within one arm of the fold-back precursor, and in the absence of

*To whom correspondence should be addressed.

large internal loops or bulges especially large asymmetric ones (Ambros *et al*., 2003). Apparently, mere application of simple alignment queries and positive-selection rules is likely to overlook novel families lacking clear homologues to published mature miRNAs.

Advanced comparative approaches like MiRscan (Lim *et al*., 2003b; Lim *et al*., 2003a), MIRcheck (Jones-Rhoades and Bartel 2004), miRFinder (Bonnet *et al*., 2004a), miRseeker (Lai *et al*., 2003), findMiRNA (Adai *et al*., 2005), PalGrade (Bentwich *et al*., 2005), and MiRAlign (Wang *et al*., 2005) have systematically exploit the greater availability of sequenced genomes for eliminating the over-represented false-positives. Cross-species sequence conservation based on computationally intensive multiple genome alignments is a powerful approach for genome-wide screening of phylogentically well conserved *pre-miRs* between closely related species. However, it suffers lower sensitivity in divergent evolutionary distance (Berezikov *et al*., 2005; Boffelli *et al*., 2003). Identifying *pre-miRs* that differ significantly or evolve rapidly at the sequence level while retaining their characteristic evolutionary conserved hairpin structures may also be an issue. Another significant drawback is that non-conserved *pre-miRs* with genus-specific patterns are likely to evade detection. Pathogenic viral-encoded *pre-miRs* have been uncovered in *E. Barr virus*, *K. sarcoma-associated herpesvirus*, *M. γ-herpesvirus 68*, *H. Cytomegalovirus*, and *Simian virus 40* that neither share significant sequence homology with known host *pre-miRs* nor among themselves (Cullen 2006; Sarnow *et al*., 2006).

(Table 1) To surmount the technical shortfalls of comparative works for distinguishing species-specific and non-conserved *pre-miRs*, several state-of-the-art *de novo* (or *ab initio*) predictive approaches have been extensively developed. The inaugural and definitive work by Sewer *et al.* (2005) compiled 40 distinctive sequence and structural "markers" from the hairpins that obviates the use of comparative genomics information. The SVM classifier model trained with the experimental domain knowledge and binary-labeled feature vectors, recovered 71% of the positive *pre-miRs* with a remarkably low false-positive rate of ~3%. It also predicted ~50 to 100 novel *pre-miRs* for several species; ~30% of these were previously experimentally validated. The validation rate among the predicted cases that were conserved in ≥1 other species was higher at ~60%; many had not been detected by comparative genomics approaches. The *3SVM* (Xue *et al*., 2005) improved the performances to ~90.00% for human and up to 90.00% in other species. Albeit its methodological simplicity, promising performances, and independence of comparative genomics information, *3SVM* was largely limited to classifying RNA sequences that fold into secondary structures without multiple loops. RNAmicro (Hertel and Stadler 2006) incorporating sequence and structural information as part of its feature vector, reported an incredibly promising efficiency of 91.16% (sensitivity) and 99.47% (specificity). Still, its classification pipeline required computationally expensive multiple sequence alignments for inputs.

ProMiR (Nam *et al*., 2005) took advantage of a probabilistic co-learning model HMM to classify miRNA genes based on their pairwise aligned sequences. ProMiR minimized the false-positive rate to as low as 4.00%, but compromised for a less performing sensitivity of 73.00%. A relatively recent work BayesMIRfinder (Yousef *et al*., 2006) adopted an alternative discriminative machine learning algorithm NBI as its underlying classifier model. Notwithstanding its technical novelty, BayesMIRfinder relied on the comparative analysis of conserved genomics regions for post-processing to yield a considerably higher sensitivity of 97.00% and comparable specificity of 91.00% in mouse to existing algorithms.

Table 1. Existing (quasi) *de novo* classifiers, in chronological order, for distinguishing novel *pre-miRs* from genomic pseudo hairpins.

Works	*Classifiers*	*Num*	*Description of Features*	*Datasets*	*P*	*N*	*SE*	*SP*
Sewer *et al.* (2005)	SVM	40	16 statistics computed from the entire hairpin structure, 10 from the longest symmetrical region of the stem, 11 from the longest relaxed symmetry region, and 3 from the candidate stem-loop.	H	178	5,395	71.00	97.00
ProMiR (Nam *et al*., 2005)	HMM	–	A hairpin structure is represented as a pairwise sequence. Each position of the pairwise sequence has two states, structural and hidden.	H	136	1,000	73.00	96.00
3SVM (Xue *et al*., 2005)	SVM	32	Each hairpin is encoded as a set of 32 triplet elements: a nucleotide type and three local continuous sub-structure-sequence attributes e.g., "A(((" and "G(..".	H H	30 39	1,000 2,444	93.30 92.30	88.10 89.00
BayesMIRfinder (Yousef *et al*., 2006)	NBI	84	62 secondary structural features derived from the foot, mature, and head of a hairpin-loop; 12 sequence features extracted from the candidate sequence.	C.E M	11 22	150 150	83.00 97.00	96.00 91.00
RNAmicro (Hertel and Stadler 2006)	SVM	12	2 lengths of stem and hairpin loop regions; 1 G+C sequence composition; 4 sequence conservation; 4 thermodynamic stability; and 1 structural conservation.	Animal	136	394	91.16	99.47

(*Classifiers*) SVM (Support Vector Machine), NBI (Naïve Bayesian Induction), and HMM (Hidden Markov Model). (*Num*) Number of features. (*Datasets*) H (*H. sapiens*), C.E (*C. elegans*), and M (*M. musculus*). *P* (real *pre-miRs*), *N* (pseudo hairpins), *SE* (Sensitivity), and *SP* (Specificity).

2 MATERIALS AND METHODS

2.1 Biologically relevant datasets

Training, testing, and independent datasets. They are pooled separately from four independent sources (see supplementary "Materials and Methods" for details). To improve the quality of this comprehensive collection, sequences with non-ACG[TU] nucleotides are filtered and no sequence is reused. Entire set of 2,241 *pre-miRs* is obtained from miRBase 8.2 (Griffiths-Jones *et al*., 2006); 8,494 pseudo hairpins from human RefSeq genes (Pruitt and Maglott 2001) without undergoing any known experimentally validated alternative splicing (AS) events. For hyperparameter estimation and training the decision function of *miPred*, binary-class labeled samples consisting of 200 human *pre-miRs* (positives) and 400 pseudo hairpins (negatives) are randomly selected without replacement to avoid the classifier being skewed towards specifically screened training samples. The remaining 123 human *pre-miRs* (positives) and 246 randomly selected pseudo hairpins (negatives) are used for testing. They are denoted as TR-H and TE-H. The comparable ratio of 1:2 ensures that the selected negatives contribute more significantly to the specificity of a classifier than positives, while avoiding the problem of overtraining. Typically, the decision function of SVM converges to a solution where all samples belonging to the smaller class are classified as that of the larger class if the class sizes differ significantly. The performance of *miPred* is evaluated against three datasets IE-NH, IE-NC, and IE-M. They represent the remaining 1,918 *pre-miRs* spanning 40 non-human species (positives) and 3,836 randomly selected pseudo hairpins (negatives); 12,387 functional ncRNAs (negatives) from Rfam 7.0 (Griffiths-Jones *et al*., 2005); and 31 mRNAs (negatives) from GenBank DNA database (Benson *et al*., 2005), respectively.

Four complete viral genomes. They are downloaded from GenBank DNA database (Benson *et al*., 2005), namely *E. Barr virus* (EBV; 171,823-bps; DNA circular; AJ507799.2), *K. sarcoma-associated herpesvirus* (KSHV; 137,508-bps; DNA linear; U75698.1), *M. γ-herpesvirus 68 strain WUMS* (MGHV68; 119,451-bps; DNA linear; U97553.2), and *H. cytomegalovirus strain AD169* (HCMV; 229,354-bps; DNA linear; X17403.1).

2.2 Computational pipeline of *miPred*

Background of SVM. Integral to *miPred* is the Support Vector Machine (SVM), a supervised classification technique derived from the statistical learning theory of structural risk minimization principle (Burges 1998). Given its simplicity to deal easily with multi-dimensional datasets that can be noisy or redundant (non-informative or highly correlated), SVM has been adopted extensively as an invaluable machine learning tool to address diverse bioinformatics problems (Dror *et al*., 2005; Han *et al*., 2004; Liu *et al*., 2006).

Briefly, the primary objective of SVM is to explicitly construct a multi-dimensional hyperplane separating a set of complex feature vectors $\mathbf{x}_i$ into binary labeled classes $y_i \in$ (1 or -1) with the distance between the hyperplane and the closest support vectors (the margin) maximized. In non-linear separable cases, the maximum-margin hyperplane is constructed after transforming uniquely the input variables into a high-dimensional feature space via the Gaussian Radial Basis Function kernel (RBF) $K(\mathbf{x}, \mathbf{x}_i)$ in Eq. (1). Typically, SVM is conducted using three straightforward steps: feature extraction, training the decision function on a set of selected binary-labeled training vectors, and classifying a given test sample $\mathbf{x}_i$ into either positive or negative classes (Burges 1998).

$$K(\mathbf{x},\mathbf{x}_i) = exp(-g\,\|\mathbf{x}-\mathbf{x}_i\|^2), \quad g = 1/2s^2. \tag{1}$$

Extraction of miPred's features. Without relying on phylogenetic conservation, *miPred* undertakes a novel approach that posits the entire hairpin structure of each RNA sequence can be characterized solely into a feature vector $\mathbf{x}_i$ containing **29** RNA global and intrinsic folding measures; see supplementary "Materials and Methods" for details. **17** sequence composition variables: 16 dinucleotide frequencies *%**XY*** (***X***, ***Y*** $\in \sum = [A, C, G, U]$) and 1 aggregate dinucleotide frequency *%G+C* ratio. Dinucleotide is the preferred predicting descriptor to mononucleotide or higher-order frequencies, as it strikes a compromise between the resolution and computation tractability. **6** folding measures: adjusted base pairing propensity *dP* (Schultes *et al*., 1999), adjusted Minimum Free Energy of folding *dG* (Freyhult *et al*., 2005; Seffens and Digby 1999), Minimum Free Energy of folding index 1 $MFEI_1$ (Zhang *et al*., 2006a), adjusted base pair distance *dD* (Freyhult *et al*., 2005; Moulton *et al*., 2000), adjusted shannon entropy *dQ* (Freyhult *et al*., 2005), and Minimum Free Energy of folding index 2 $MFEI_2$. **1** topological descriptor: degree of compactness *dF* (Fera *et al*., 2004; Gan *et al*., 2004). **5** normalized variants of *dP*, *dG*, *dQ*, *dD*, and *dF* i.e., *zP*, *zG*, *zQ*, *zD*, and *zF* derived from dinucleotide shuffling. We compute the 17 sequence composition variables as well as the folding measures *dQ* and *dD* by a custom perl script interfaced to the module RNAlib of Vienna RNA Package 1.4 (Hofacker 2003); *dG* by RNAfold program (Hofacker 2003) that predicts the most favorable RNA secondary structure and the corresponding MFE; the topological descriptors *S* and *dF* by a custom program RNAspectral. After synthesizing the set of random RNA sequences, the normalized variants *zP*, *zG*, *zQ*, *zD*, and *zF* are computed.

Parameter estimation, training, and evaluation of miPred. The libSVM version 2.82 (http://www.csie.ntu.edu.tw/~cjlin/libsvm), a free SVM implementation is used for training and testing *miPred*'s binary classification. Samples are randomly selected without replacement via a custom python script. Foremost, the 29 attributes of *miPred* are rescaled linearly by the *svm-scale* program to the interval [-1.0, 1.0] or [0.0, 1.0] to guard against asymptomatic biasness in the numeric ranges for all the datasets; larger variance may dominate the classification e.g., [6.0, 50.0] *vs.* [-0.5, -0.2]. All *miPred* classifier models are generated with "*svm-train* -b 1 -c 2^C -g γ"; default RBF kernel; "-b 1" option computes the SVM probability estimates (*P*-values) for classification thresholding. As both the penalty parameter *C* (determines the trade-off between training error minimization and margin maximization) and the RBF kernel parameter γ (defines the nonlinear mapping from input space to some high-dimensional feature space) are critical for the performance of SVM (Duan *et al*., 2003), they are optimally calibrated by an exhaustive grid-search strategy. Briefly, at each hyperparameter pair (C, γ) selected from the search space $\log_2 C \in$ [-10, -9, …, 15] and $\log_2 \gamma \in$ [-15, -14, …, 10], we perform a five-fold cross validation. The training dataset is randomly partitioned into approximately five distinct equal-sized subsets. Repeating the validation process five times for each subset i.e., retaining a set for testing and the remaining four sets for training, the average accuracy of the five models gives the five-fold leave-one-out cross-validation (LOOCV) accuracy rate (Duan *et al*., 2003). To avoid over-fitting the generalization, the best combination of hyperparameters (C, γ) maximizing the five-fold LOOCV accuracy rate serve as the default setting for training *miPred*. Finally, the classification is conducted on the testing and independent evaluation datasets with "*svm-predict* -b 1". See supplementary "Materials and Methods" for details on statistical tests and performance evaluation metrics.

3 RESULTS AND DISCUSSION

3.1 Training and classifying human *pre-miRs*

We calibrate *miPred* using TR-H, the optimal hyperparameter pair (C, γ) is (16.0, 0.03125) that maximizes the five-fold cross-validation accuracy rate of 93.50%. A classification score ranging [0.0, 1.0] is assigned by *miPred* to each hairpin, which is designated as a candidate *pre-miR* if its score is beyond a specified threshold. Across the entire spectrum of thresholds, a trade-off generally exists between specificity (greater value at higher threshold) and sensitivity (value increases at lower threshold) (Dror *et al*., 2005; Liu *et al*., 2006). The ROC analysis of *miPred*'s classification model (*figure not shown*) reported that the AUC is approximately unity i.e., 0.9833.

(Figure 1A and Table S1) With the default *miPred*'s threshold predefined at 0.5, the SE (Sensitivity), SP (Specificity), and ACC (Accuracy) reported for TR-H are 88.00%, 97.50%, and 94.33%, respectively. Here, SP > SE is more desirable in screening for novel *pre-miRs* from the entire genomic sequences or cloned small RNAs as abundant dysfunctional hairpins are encoded in the human (Bentwich *et al*., 2005) and *C. elegans* (Pervouchine *et al*., 2003) genomes. An implication of a slightly lower SP than SE will reduce the signal (genuine *pre-miRs*) to background (pseudo hairpins) ratio, inflating significantly the effort and resources demanded in experimental validation of the putative precursor transcripts as biologically functional *pre-miRs*.

(Figure 1B and Table S1) Next, conducting *miPred* onto TE-H obtains comparable performances of 84.55% (SE), 97.97% (SP), and 93.50% (ACC). In all, *miPred* can classify correctly 86.69% (280/323) human *pre-miRs* as positives and 97.68% (631/646) pseudo hairpins as negatives. Three of the human *pre-miRs* designated as negatives receive very low classification scores from *miPred*: *hsa-mir-565* (0.454), *hsa-mir-566* (0.012), and *hsa-mir-594* (0.187). Coincidently, they have been suspected to be falsely annotated as precursor transcripts encoding mature miRNAs on two grounds (Berezikov *et al*., 2006). First, both *hsa-mir-565* and *hsa-mir-594* overlap with tRNA annotations; *hsa-mir-566* overlaps with Alu repeats. Second, none was represented by >1 clone or differentially expressed in a Dicer-deficient cell-line (Cummins *et al*., 2006). Nevertheless, we believe that neither criterion is sufficient to eliminate a candidate as repeat- (Smalheiser and Torvik 2005) and pseudogene-derived miRNAs (Devor 2006) have been discovered, and miRNAs expressed at low levels may be elusive to detection in a Dicer-disrupted mutant (Berezikov *et al*., 2006).

(Table S1) In contrast, *3SVM* based on triplet-encoding scheme (Xue *et al*., 2005) yields slightly poorer results: 86.00%, 97.00%, and 93.33% for TR-H; 73.15%, 95.37%, and 87.96% for TE-H; or 81.49% (251/308) of human *pre-miRs* as positives and 96.43% (594/616) of pseudo hairpins as negatives. The evaluation demonstrates the outstanding and consistent classification performance of *miPred* in partitioning specifically human *pre-miRs* from pseudo hairpins. The improved distinct separation by *miPred* is likely due to its excellent capability in recognizing the specific intrinsic and global features of human *pre-miRs* against those of pseudo hairpins.

3.2 Improved classification of non-human *pre-miRs*

(Figure 1C and Table S1) We next extend the validation of *miPred* to

IE-NH and quantify its mean (overall) SE, SP, and ACC. Here, mean denotes the average performance for all species within IE-NH; overall performance is derived from the entire IE-NH independent of species. In this setting, *miPred* yields excellent and comparable classification performances to those of TR-H and TE-H, with respective SE, SP, and ACC of 87.65% (92.08%; 1,766/1,918 non-human *pre-miRs* as positives), 97.75% (97.42%; 3,737/3,836 pseudo hairpins as negatives), and 94.38% (95.64%). (Table S1) In contrast, *3SVM* reports 80.10% (86.15%; 1,443/1,675 non-human *pre-miRs* as positives), 96.81% (96.27%; 3,225/3,350 pseudo hairpins as negatives), and 91.24% (92.90%). Apparently, these results point to *miPred* as a more credible and consistent classifier distinguishing reliably specie-specific and evolutionary well-conserved *pre-miRs* across diverse organisms covering plants, worms, flies, vertebrates, and viruses listed in miRBase 8.2 (Griffiths-Jones *et al*., 2006).

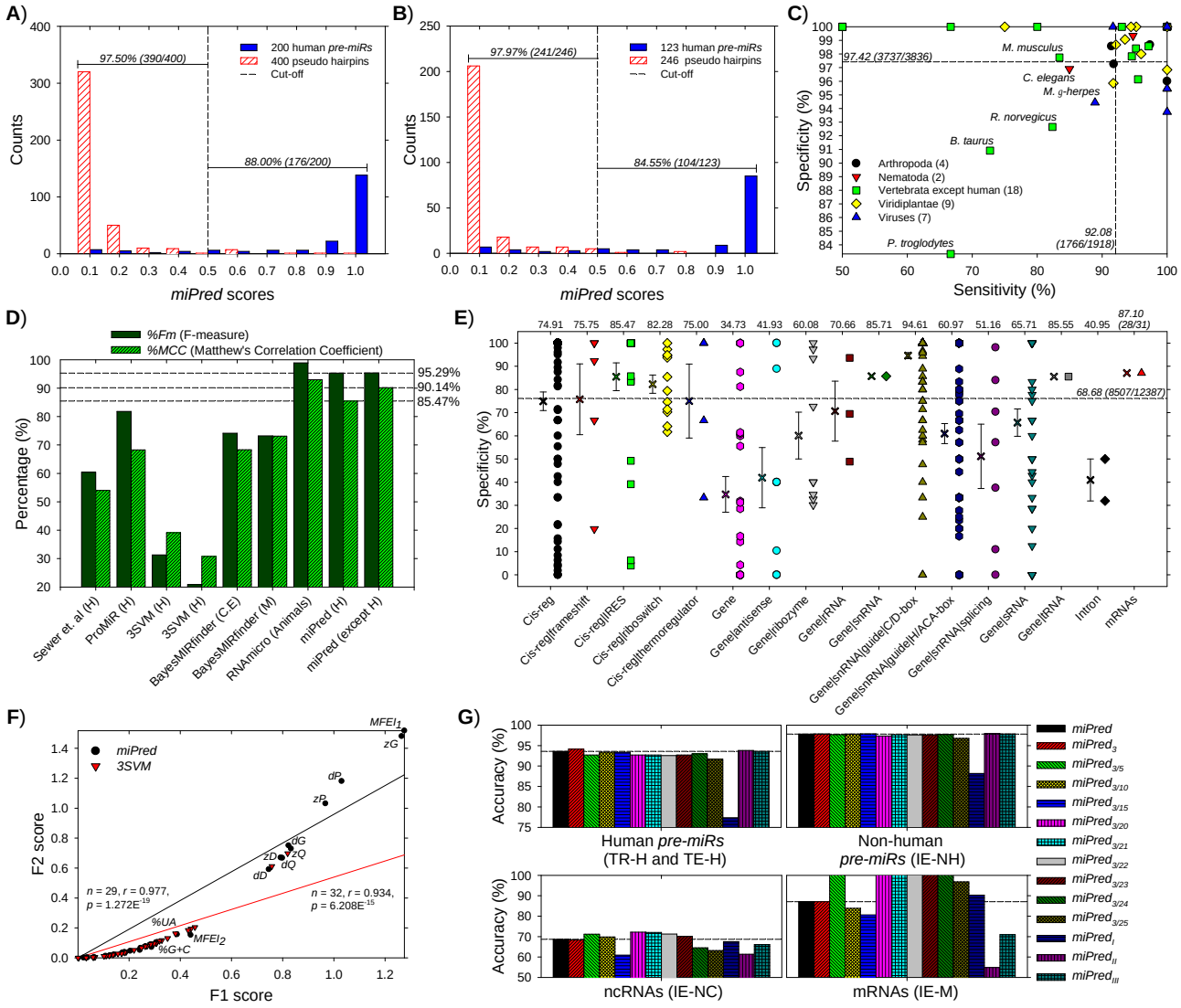

Fig. 1. A–**B**) Distribution of TR-H (200 human *pre-miRs* and 400 pseudo hairpins) and TE-H (remaining 123 human *pre-miRs* and 246 pseudo hairpins) by *miPred* scores. Default *miPred* decision boundary (vertical dash line at 0.5). **C**) Distribution of IE-NH (1,918 *pre-miRs* across 40 non-human species and 3,836 pseudo hairpins) by specificity and sensitivity; details at Table S1. Dash lines denote overall performances. For clarity, only specie names are assigned in left-bottom quarter. **D**) Performance comparison with existing (quasi) *de novo* classifiers (Table 1). H (*H. sapiens*), C.E (*C. elegans*), and M (*M. musculus*). **E**) Distribution of IE-NC (12,387 ncRNAs) and IE-M (31 mRNAs) by specificity; details at Table S3. Dash line denotes overall specificity. **F**) F1 and F2 scores for features of *miPred* and *3SVM*; details at Table S5. For clarity, only top 12 ranking attributes of *miPred* are shown. **G**) Effects of feature selection on *miPred*'s accuracy; details at Table S6. Dash lines denote accuracies of original *miPred*.

Notably, those *pre-miRs* present in the genomes of *A. mellifera*, *A. Geoffroyi*, *C. familiaris*, *E. Barr*, *H. Simplex virus*, *H. cytomegalovirus*, *O. aries*, *P. patens*, *R. lymphocryptovirus*, *Simian virus*, and *Z. mays* are unambiguously identified by *miPred* with 100.00% (SE) and >93.75% (SP). Moreover, *pre-miRs* encoded in *C. briggsae* and *C. elegans* are excellently classified with SE of 94.74% and 84.96%, as well as SP of 99.34% and 96.90%; the remaining two pathogenic viruses *M. γ-herpesvirus* and *K. sarcoma-associated herpesvirus* have SE of 88.89% and 91.67%, as well as SP of 94.44% and 100.00%. Since *miPred* was not trained initially on any specie-specific *pre-miRs* and especially viral-encoded ones, this supporting evidence reinforces the premise that its selected descriptors have successfully captured the intrinsic and global properties characterizing the biologically functional *pre-miRs* spanning across different species including viruses.

(Table S2) An obvious question is how viral-encoded *pre-miRs* can be distinguished by *miPred* so outstandingly, especially when they are known to lack homologs in other viruses or in the host (Cullen 2006; Sarnow *et al*., 2006). As there are few experimental studies elucidating their biological activities and biogenesis (Sullivan *et al*., 2005), we speculate pathogenic viruses do not possess homologous genes that can express functionally similar host miRNA processing proteins e.g., Drosha, Dicer, and RISC. After infecting the human immune cells, they hijack these critical host proteins to regulate viral and host gene expression (Cullen 2006; Sarnow *et al*., 2006). This will facilitate their viral replication and pathogenesis by blocking the innate or adaptive host immune responses or by interfering with the appropriate regulation of apoptosis, cell growth, or DNA replication. Consequently, viral-encoded *pre-miRs* are likely to be recognized and processed identically to the host (i.e., human) *pre-miRs* that *miPred* was trained on.

3.3 Performance comparison with existing predictors

(Figure 1D) By evaluating the published results of existing (quasi) *de novo* classifiers (Table 1), both RNAmicro (Hertel and Stadler 2006) and *miPred* are the highest-scoring predictors in identifying putative *pre-miRs* from a genomic pool of candidate hairpins. RNAmicro displays comparable *%Fm* (F-measure) and *%MCC* (Matthew's Correlation Coefficient) of 98.90% and 92.97% *vs. miPred* of 95.29% and 85.47% (human *pre-miR* datasets; 95.34% and 90.14% for non-human *pre-miR* datasets). The remaining classifiers range 20.85−91.87% (*%Fm*) and 30.80−79.51% (*%MCC*).

Notably, *miPred* benefits two key areas of technical advancements. First, its 29 features are extracted from a single RNA sequence for classifying novel *pre-miRs* against pseudo hairpins in an unequivocal *de novo* manner. This is the primary advantage that *miPred* has over RNAmicro by avoiding costly and occasionally unreliable multiple sequences alignments due to large phylogenetic distant or rapidly evolving *pre-miRs*. RNAmicro relies on computationally expensive comparative genomic alignments for predicting the consensus secondary structures and computing its feature vector (Hertel and Stadler 2006). Moreover, ProMiR (Nam *et al*., 2005) and BayesMIRfinder (Yousef *et al*., 2006) depend on similar phylogenetic/conservation information for not incurring any significant loss of performances. Due to the sequence homologous nature of the genomics datasets being generated, their predictive accuracy may suffer when the cross-species evolutionary distance (e.g., verterbrates *vs.* nematode/urochordate) is too exceptionally diverged in rendering reliable multi-genomes alignment technically difficult or impossible. Second, distinct from classifiers by Sewer *et al*. (2005) and *3SVM* (Xue *et al*., 2005), the 29 attributes from *miPred* represent the global and intrinsic properties of any RNA structure, and not specific regions of it. Besides avoiding the *pars pro toto* fallacy in mistaking part for the entire, *miPred* can handle both hairpin structures as well as RNA sequences that fold with multiple loops.

3.4 Classification of functional ncRNAs and mRNAs

The original intent of *miPred* is to distinguish *pre-miRs* spanning diverse species from genomic pseudo hairpins, according to the classifier model trained solely on human datasets. Since ncRNAs and mRNAs were not included in the initial training, it will be very instructive to assess how well *miPred* can discriminate them as non *pre-miRs* without relying on their specific dinucleotide sequence, structural, and topological characteristics. Moreover, such assessment was lacking or not available from existing (quasi) *de novo* predictors (Table 1). (Figure 1E and Table S3) Evaluating *miPred* and *3SVM* (Xue *et al*., 2005) onto IE-NC and IE-M, the former reports mean (overall) SP of 76.15% (68.68%; 8507/12387 ncRNAs) and 87.10% (27/31 mRNAs). Here, mean or average SP is computed from all ncRNA types within IE-NC; overall SP corresponds to the entire IE-NC independent of ncRNA types. In contrast, *3SVM* yields 90.30% (78.37%; 1,884/2,404 ncRNAs across 155 types) and 0.00% (0/31 mRNAs) for SP (*figure not shown*). Upon scrutiny, its "better" performances are attained at the expense of excluding 9,983 ncRNAs spanning 302 types (IE-NC) and 31 mRNAs (IE-M) that fold into complex structures containing multiple loops. This structural exclusion is a major limitation experienced commonly by most of the existing (quasi) *de novo* classifiers (Table 1) that extract modularized features from predefined RNA sub-structures. The comparison with *3SVM* clearly demonstrates that *miPred* trained solely on human *pre-miRs* and pseudo hairpins, can provide reasonable generalization in identifying unambiguously at least two-thirds of all the samples in IE-NC and IE-M as *bona fide* negatives.

Among the ncRNA samples in IE-NC, tRNAs (Sprinzl and Vassilenko 2005) and snoRNAs (Weinstein and Steitz 1999) are two of the largest classes of small ncRNAs present in the eukaryotic genomes. They are frequently misclassified as *pre-miRs* in most experimental settings, due to the absence of statistical signatures like codon structure and open reading frame (ORF) encoded by protein-coding genes (Sprinzl and Vassilenko 2005; Weinstein and Steitz 1999). The snoRNAs can be divided into C/D snoRNAs or H/ACA snoRNAs acting as guides for site-specific 2'-O-ribose methylation or for pseudouridylation in the post-transcriptional processing of rRNAs (Weinstein and Steitz 1999). (Figure 1E and Table S4) 94.61% C/D snoRNAs, 60.97% H/ACA snoRNAs, and 85.55% tRNAs are identified by *miPred* as genuine non *pre-miRs*. To enhance the quality of *miPred*'s identification, specialized algorithmic tools like snoseeker (Yang *et al*., 2006) and tRNAscan-SE (Lowe and Eddy 1997) can serve as rapid and preprocessing filters in excluding these abundant ncRNAs, except C/D snoRNAs. They have reported SE of 90.00%, 75.00%, and 99.5% for detecting C/D snoRNAs, H/ACA snoRNAs and tRNAs, respectively.

(Figure 1E and Table S4) *miPred* is capable of discriminating correctly 75.75% frameshift, 85.47% IRES, 75.00% thermoregulator, 70.66% rRNA, and 85.71% snRNA as authentic non *pre-miRs*. Interestingly, a novel and abundant class of ncRNAs known as riboswitches (Winkler and Breaker 2003) are correctly classified by *miPred* as non *pre-miRs* with comparable SP of 82.28%. These riboswitches found only in prokaryotes to date, can *cis*-modulate their expressions upon binding to metabolite (e.g., guanine and thiamine pyrophosphate) without involving accessory protein cofactors. Our SVM classifier *miPred* will likely to become an invaluable pre-experimental predictor in the event eukaryotic riboswitches(-like) molecules are identified.

(Figure 1E and Table S4) Several classes of ncRNA are poorly classified by *miPred* as potential *pre-miRs* with SP $\leq$ 60.00%: Antisense, Ribozymes, Spliceosomes like U1−2 and U4−6, and Group I/II intron RNAs. Careful inspection into their sequence, structural, and topological properties reveals no general noticeable trends to explain the evasive detection. This finding prompts us to speculate that the feature vector used by *miPred* may lack specific discriminative components against these elusive classes of functional ncRNAs, or in part that they may possibly be exceedingly mobile or rapidly evolving. To identify and eliminate such ncRNAs will definitely require specialized tools built on the domain knowledge of their characteristic properties.

3.5 Contribution of individual features

We next investigate the essential attributes of *miPred* that contribute substantially to the class distinctions between *pre-miRs* and pseudo hairpins, or whether exclusion of selected feature(s) can further en-

hance/degrade *miPred*'s performances. Elucidating the "contributory quality" of individual attribute within a feature vector reaps the potential benefits of enhancing the predictive performance and computational tractability of the classifier, and gaining deeper insights into the domain problem (Isabelle and Andre 2003). Despite the importance, only *3SVM* (Xue *et al*., 2005) among the existing (quasi) *de novo* classifiers (Table 1) has conducted an analysis (less detailed than ours) on its feature selection.

(Figure 1F and Table S5) We evaluate the F-scores F1 and F2 (see supplementary "Materials and Methods" for definitions) on the class-conditional distributions, which measure the discriminative power of the *miPred*'s 29 attributes. They are strongly and positively correlated, reporting pearson correlation coefficient $r = 0.977$ and $p = 1.272E^{-19}$. As expected, structural features possess the strongest discriminative importance/powers by dominating the 12 highest scoring attributes (ranked according to descending F1 scores): $MFEI_1$, *zG*, *dP*, *zP*, *zQ*, *dG*, *dQ*, *zD*, *dD*, $MFEI_2$, *%AU*, and *%G+C*. They overlap to some degree with RNAmicro's features (Hertel and Stadler 2006) i.e., *%G+C*, $MFEI_1$, *dG* (RNAmicro uses mean MFE of the aligned sequences and MFE of the consensus structure), and *zG* (RNAmicro computes via a regression model). Since the majority of the *pre-miRs* are well-defined and thermodynamically stable stem-loop structures critical for the biogenesis of mature miRNAs (Bonnet *et al*., 2004b), these common features and *miPred*'s top-ranking ones are most probable to be conserved across all species from human to viruses. We believe they are likely to be indispensable for rendering more robustness to the multi-feature capability of *miPred* against erroneous classifications of novel *pre-miRs*.

Generally, the efficiency and reliability of classifiers depend on the size and selection of both the relevant data samples and specific attributes (Isabelle and Andre 2003). We next repeat previous experiments using 10 variants of *miPred* i.e., they contain smaller collection of features and trained in the exact manner as *miPred* with identical samples in TR-H, and their performances are assessed against the remaining datasets (TE-H, IE-NH, IE-NC, and IE-M). $miPred_3$ contains a subset of 26 features from *miPred* that excludes *dQ*, *dD*, and *zD*. When evaluated statistically onto the 2,241 non-redundant *pre-miRs*, three pairs of attributes are strongly and positively correlated (*figure not shown*) with *r* ranging 0.9221−0.9846 and $p < 0.001$: *dQ vs. dD*, *dQ vs. zQ*, and *zQ vs. zD*. *zQ* is selected due to its higher discriminative power (as indicated by both its F1 and F2 scores) than *dQ*, *dD*, and *zD* (Figure 1F). Derived from $miPred_3$, the remaining nine variants represented by $miPred_{3/5}$, $miPred_{3/10}$, …, $miPred_{3/24}$, and $miPred_{3/25}$ include only the top ranking 21, 16, 11, 6, 5, 4, 3, 2, and 1 feature(s), respectively.

(Figure 1G and Table S6) As expected, *miPred* and $miPred_3$ demonstrate consistent and comparable classification accuracies spanning the five datasets. The former containing near perfect correlated features *dQ*, *dD*, and *zD* as part of its larger feature vector is highly resilient to redundancy, since it also relies on SVM. SVM incorporates regularization techniques and is based on the theory of risk minimization, which can provide robust generalization control in accommodating redundant (i.e., strongly correlated) variables (Burges 1998). Removing 5 to 15 low-scoring features, $miPred_{3/5} - miPred_{3/15}$ yield negligible performance differences compared to $miPred_3$ when applied to *pre-miR* datasets; better improvements reported by $miPred_{3/5}$ for ncRNAs and mRNAs datasets. This result suggests that the removed features are likely to contribute in a smaller degree to *miPred* as non-informative attributes and they generally do not degrade the performance of the discriminant method by overfitting the training data. With fewer than seven top-ranking features contained in $miPred_{3/20} - miPred_{3/25}$, their overall classification accuracies degrade slightly for *pre-miR* datasets; generally have better performances for ncRNAs and mRNAs datasets. Both findings indicate that these six highest-scoring attributes $MFEI_1$, *zG*, *dP*, *zP*, *zQ*, and *dG* are likely to be predominantly functioning, in order to contribute significantly to the prediction accuracies of *miPred*.

(Figure 1G and Table S6) Features with weak discriminative power (like those sequence attributes in *miPred* possessing low F-scores) are viewed largely as redundant (i.e., non-informative), as no additional performance is gained by including them (Isabelle and Andre 2003). To affirm this premise, we evaluate another three variants of *miPred*: $miPred_I$ (17 features: 16 dinucleotides frequencies and *%G+C*), $miPred_{II}$ (12 features; $MFEI_1$, $MFEI_2$, *dP*, *dG*, *dQ*, *dD*, *dF*, *zP*, *zG*, *zQ*, *zD*, and *zF*), and $miPred_{III}$ (9 features; a subset of $miPred_{II}$ that excludes *dQ*, *dD*, and *zD*). Apparently, $miPred_I$ performs the worst when identifying *pre-miRs* and degrades moderately for IE-NC, but reports better than expected classification when applying to IE-M. In contrast, the absence of sequence information (i.e., 16 dinucleotide frequencies and *%G+C*) shows no noticeable effect on the performances of $miPred_{II}$ and $miPred_{III}$ for human *pre-miRs* in comparison to *miPred* and $miPred_3$; both classifiers fare slightly inferior to $miPred_I$ for IE-NH and much worst for IE-NC and IE-M. As indicated by both findings, the sequence information do not contribute (significantly or at all) towards discriminating *pre-miRs* from pseudo hairpins. Nevertheless, they are probable to perform a critical or compensatory role in the classification of ncRNAs and mRNAs as non *pre-miRs*.

3.6 Screening viral-encoded miRNA genes

A recent *rna22*-based census suggested that the previous numbers for *pre-miRs* present in several species were gross underestimation, and are likely to range in the tens of thousands (Miranda *et al*., 2006): *C. elegans* (359), *D. melanogaster* (654), *M. musculus* (>25,000) and *H. sapiens* (>25,000). As an illustrative application of *miPred*, we randomly select four complete viral genomes for screening novel *pre-miRs* via a similar methodology (Miranda *et al*., 2006): *E. Barr virus* (EBV), *K. sarcoma-associated herpesvirus* (KSHV), *M. γ-herpesvirus 68 strain WUMS* (MGHV68), and *H. cytomegalovirus strain AD169* (HCMV). To date, miRBase 8.2 (Griffiths-Jones *et al*., 2006) have annotated 23 (EBV; 23 + strands), 13 (KSHV; 12 − and 1 unknown strand), 9 (MGHV68; 9 + strands), and 11 (HCMV; 6 +, 4 −, and 1 unknown strands) viral-encoded *pre-miRs*. The four viral genomic sequences are oriented to the corresponding +/− strands along which the published *pre-miRs* are located, and then scanned with a predefined sliding window (size of 95-nts in 1-nt steps) for potential viral-encoded hairpins. Those genomic regions satisfying the maximum length (≤ 95-nts), minimum size of terminal loop (≥ 3-nts), and MFEs (≤ -25 kcal/mol) are reserved for classification via *miPred*. The three thresholds were empirically determined from available genuine *pre-miRs* encoded in the four pathogenic viruses. The computational approach was described previously by Grad *et al*. (2003) with differences in the parameter settings as mentioned earlier. Briefly, it uses a BLAST-like algorithm to search for short complementary words (stem-like structure) within a specified distance and dynamic programming to determine the complete alignment. MFEs are predicted by the RNAfold program (Hofacker 2003) with default parameters.

(Figure 2A) Roughly, 30.15% (EBV; 60/199), 16.51% (KSHV; 36/218), 10.87% (MGHV68; 20/184), and 27.71% (HCMV; 133/480) of the hairpins are classified as putative *pre-miRs* (positives) at the default *miPred* score cut-off ≥0.5; remaining ones are regarded as negatives. (Table S7) The viral-encoded hairpins are manually mapped to the published *pre-miRs*, 25 true-positives (and 1 false-negative) match 25 published viral-encoded *pre-miRs* (red region) and their mature miRNAs (underlined region): 12 (1) EBV, 6 (0) KSHV, 3 (0) MGHV68, and 4 (0) HCMV. Except ‡*kshv-mir-K12-9* and $^{\Delta}$*kshv-mir-*

K12-9, the remaining true-positive predictions have one or two mature miRNAs embedded exclusively in either arms of their (a)symmetric stem. $^{\Delta}$*kshv-mir-K12-9* is subsequently eliminated as it is a duplicate copy containing the exact sequence of *kshv-mir-K12-9*, and the encoded mature miRNAs overlap the most with its predicted 4-nts (**uaua**) terminal loop. Together, we can identify 44.64% (25/56) of the known *pre-miRs* for the four viruses as hairpins, and recover 96.00% from these hairpins (24/25) as true-positives.

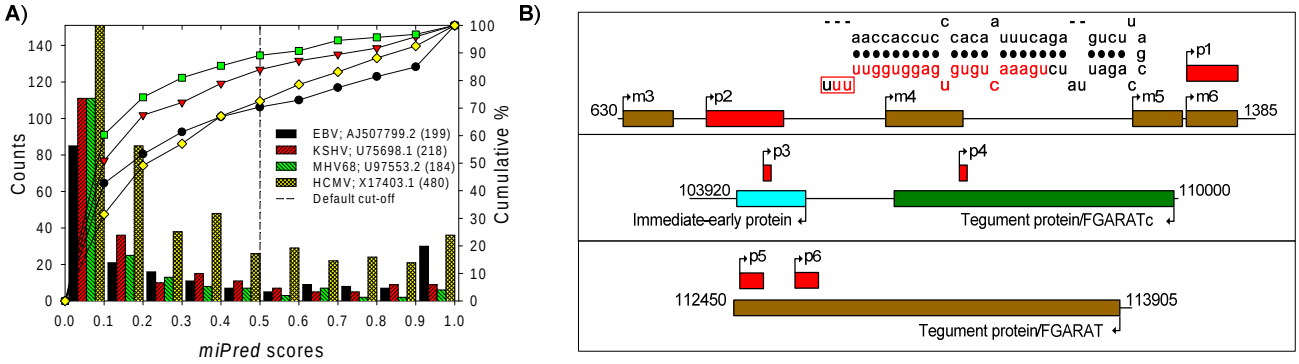

Fig. 2. A) Distribution of viral-encoded hairpins by *miPred* scores. **B**) Genomic map of predicted (pX denotes *mghv-mir-pX*) and published (mX denotes *mghv-mir-M1-X*) MGHV68-encoded *pre-miRs*, drawn *not to scale* by Genepalette 1.2 (Rebeiz and Posakony 2004); RNA structure of m6 (inset; *mghv-mir-M1-6*) as obtained from miRBase 8.2 (Griffiths-Jones *et al*., 2006); red region denotes mature miRNA.

The 25 identified positives report high *miPred* scores ≥0.815 except for two †*ebv-mir-BHRF1-1* (0.437 *miPred* score) and §*mghv-mir-M1-8* (0.658), indicative of the default cut-off at 0.5 was unlikely to be stringent. (Table S7) With the new cut-off set at 0.815, only 92.00% (EBV; 23/35), 60.00% (KSHV; 9/15), 75.00% (MGHV68; 6/8), and 92.73% (HCMV; 51/55) of the previous positives (excluding published *pre-miRs*) survive as novel putatives. Majority has not yet been discovered (more will arise due to innate evolutionary mutations), suggesting previous estimates of viral-encoded *pre-miRs* and miRNAs especially in EBV and HCMV may be grossly understated. (Figure 2B) By mapping carefully the 6 newly found MGHV68-encoded *pre-miRs* to the entire MGHV68 viral genome (the closest relative to human EBV and KSHV (Pfeffer *et al*., 2005)), we observe that p1 overlaps exactly with but is shorter than m6 by 3-nts (UUU) at the 3' termini (see inset for RNA structure). Since the mature miRNA (red region) encoded in m6 was experimentally cloned (Pfeffer *et al*., 2005), p1 is reassigned as a false-positive. p2 resides immediate downstream of m3 and within a known miRNA cluster ~1.5 kb consisting of m1−7 that are transcribed by RNA Polymerase III (Pol-III) (Pfeffer *et al*., 2005), which indicates p2 is likely to be regulated by similar Pol-III promoter. Known host miRNA transcripts are synthesized from intergenic or intronic regions of annotated transcription units (Rodriguez *et al*., 2004) by Pol-II with the hallmarks of 5' m^7G cap structures and 3' poly(A) tails (Cai *et al*., 2004; Lee *et al*., 2004), however, there are emerging evidence of them being transcribed from the exons of protein-coding genes like in *O. sativa* (Sunkar *et al*., 2005). Thus, p3, p4, and p5−6 located in the exons of three proteins may also undergo distinct processing and nuclear export mechanism from the host cell's miRNA maturation machinery.

4 CONCLUSION

In this work, we have proposed a *de novo* SVM classifier model *miPred* to address specifically the challenges in improving the classification accuracy of existing (quasi) *de novo* approaches. Considering that a single criterion to filter pseudo hairpins has not yet been identified, it is trained solely on 29 global and intrinsic attributes derived from human *pre-miRs*, without relying on species-conservation information. The feature set defines distinctively a *pre-miR* at the base (dinucleotide frequencies and *%G+C* ratio), hairpin folding (MFE and base pairing propensity), non-linear statistical thermodynamics (shannon entropy and base pair distance), and topological (Fiedler eigenvalue) levels. By integrating simultaneously the Gaussian Radial Basis Function kernel (RBF) of SVM as a similarity measure into *miPred*'s design, our classifier yields comparable or significantly better performances (in terms of sensitivity and specificity) than existing classifiers for distinguishing non-conserved functional *pre-miRs* (spanning diverse species) from genomic pseudo hairpins and non *pre-miRs* (most classes of ncRNAs and mRNAs) with high discriminative accuracy.

Deployment of *miPred* will likely to translate into considerable saving on precious and scarce experimental resources devoted to validating significantly fewer false-positives, since we are highly assured those precursor transcripts predicted would be experimentally confirmed as functional *pre-miRs*. Recognizing these benefits that underscore *miPred* as a potential and invaluable pre-experimental screening tool, we are currently revamping our research prototype into a user-friendly online predictor. As part of our ongoing research, we are actively identifying novel and clustered *pre-miRs* in human, mouse, and viruses. We believe with the availability of a comprehensive repository of combined *pre-miRs* and mature miRNAs, will then computational mRNA target identification and comprehensive genome annotation be greatly advanced. An expanded repertoire of miRNA genes will definitely signify both a huge opportunity and technical challenge, as we delve into the functional roles of miRNAs interplay with other genetic regulatory networks, biological pathways, and signaling cascades.

ACKNOWLEDGEMENTS

The authors are deeply indebted to the anonymous reviewers (current and ISMB2006) for their generous feedback and constructive suggestions, which has greatly inspired the quality and technical ideas developed in this paper. Sincere appreciation to BII's Clustering Group for their best effort in ensuring the three clusters run smoothly. This work was supported by Bioinformatics Institute. SNKL received Ph.D scholarship funds from Agency for Science, Technology and Research (A*STAR), Singapore.

Authors' contributions: SNKL and SKM conceived the initial ideas. SNKL designed and performed the experiments. SNKL and SKM wrote this manuscript.

Competing interests: none declared.

REFERENCES

Adai,A. *et al.* (2005) Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res.*, **15**, 78-91.

Ambros,V. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277-279.

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

Benson,D.A. *et al.* (2005) GenBank. *Nucleic Acids Res.*, **33**, D34-D38.

Bentwich,I. *et al.* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, **37**, 766-770.

Berezikov,E. *et al.* (2006) Approaches to microRNA discovery. *Nat. Genet.*, **38 Suppl**, S2-S7.

Berezikov,E. *et al.* (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21-24.

Boffelli,D. *et al.* (2003) Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. *Science*, **299**, 1391-1394.

Bonnet,E. *et al.* (2004a) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc. Natl. Acad. Sci. USA*, **101**, 11511-11516.

Bonnet,E. *et al.* (2004b) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911-2917.

Brennecke,J. *et al.* (2003) bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, **113**, 25-36.

Burges,C.J.C. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2**, 121-167.

Cai,X. *et al.* (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**, 1957-1966.

Calin,G.A. and Croce,C.M. (2006) MicroRNA-Cancer Connection: The Beginning of a New Tale. *Cancer Res.*, **66**, 7390-7394.

Chen,X. (2004) A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development. *Science*, **303**, 2022-2025.

Cullen,B.R. (2006) Viruses and microRNAs. *Nat. Genet.*, **38 Suppl**, S25-S30.

Cummins,J.M. *et al.* (2006) The colorectal microRNAome. *Proc. Natl. Acad. Sci. USA*, **103**, 3687-3692.

Devor,E.J. (2006) Primate MicroRNAs miR-220 and miR-492 Lie within Processed Pseudogenes. *J. Hered.*, **97**, 186-190.

Dror,G. *et al.* (2005) Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*, **21**, 897-901.

Duan,K. *et al.* (2003) Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, **51**, 41-59.

Fera,D. *et al.* (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, **5**, 88.

Floyd,S.K. and Bowman,J.L. (2004) Gene regulation Ancient microRNA target sequences in plants. *Nature*, **428**, 485-486.

Freyhult,E. *et al.* (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241.

Gan,H.H. *et al.* (2004) RAG: RNA-As-Graphs database--concepts, analysis, and features. *Bioinformatics*, **20**, 1285-1291.

Griffiths-Jones,S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140-D144.

Griffiths-Jones,S. *et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121-D124.

Han,L.Y. *et al.* (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, **10**, 355-368.

Hertel,J. and Stadler,P.F. (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, e197-e202.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429-3431.

Isabelle,G. and Andre,E. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157-1182.

Jones-Rhoades,M.W. and Bartel,D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787-799.

Kim,V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.*, **6**, 376-385.

Lagos-Quintana,M. *et al.* (2003) New microRNAs from mouse and human. *RNA*, **9**, 175-179.

Lagos-Quintana,M. *et al.* (2001) Identification of Novel Genes Coding for Small Expressed RNAs. *Science*, **294**, 853-858.

Lai,E. *et al.* (2003) Computational identification of Drosophila microRNA genes. *Genome Biol.*, **4**, R42.

Lau,N.C. *et al.* (2001) An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. *Science*, **294**, 858-862.

Lee,R.C. and Ambros,V. (2001) An Extensive Class of Small RNAs in Caenorhabditis elegans. *Science*, **294**, 862-864.

Lee,R.C. *et al.* (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843-854.

Lee,Y. *et al.* (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051-4060.

Lim,L.P. *et al.* (2003a) Vertebrate MicroRNA Genes. *Science*, **299**, 1540.

Lim,L.P. *et al.* (2003b) The microRNAs of Caenorhabditis elegans. *Genes Dev.*, **17**, 991-1008.

Liu,J. *et al.* (2006) Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. *PLoS Genet.*, **2**, e29.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955-964.

Lu,J. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834-838.

McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20-W25.

Miranda,K.C. *et al.* (2006) A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell*, **126**, 1203-1217.

Moulton,V. *et al.* (2000) Metrics on RNA Secondary Structures. *J. Comp. Biol.*, **7**, 277-292.

Nam,J.W. *et al.* (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.*, **33**, 3570-3581.

Palatnik,J.F. *et al.* (2003) Control of leaf morphogenesis by microRNAs. *Nature*, **425**, 257-263.

Pasquinelli,A.E. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86-89.

Pervouchine,D.D. *et al.* (2003) On the normalization of RNA equilibrium free energy to the length of the sequence. *Nucleic Acids Res.*, **31**, e49.

Pfeffer,S. *et al.* (2005) Identification of microRNAs of the herpesvirus family. *Nat. Method*, **2**, 269-276.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137-140.

Rebeiz,M. and Posakony,J.W. (2004) GenePalette: a universal software tool for genome sequence visualization and analysis. *Dev. Biol.*, **271**, 431-438.

Reinhart,B.J. *et al.* (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, **403**, 901-906.

Rodriguez,A. *et al.* (2004) Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Res.*, **14**, 1902-1910.

Sarnow,P. *et al.* (2006) MicroRNAs: expression, avoidance and subversion by vertebrate viruses. *Nat. Rev. Microbiol.*, **4**, 651-659.

Schultes,E.A. *et al.* (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, **49**, 76-83.

Seffens,W. and Digby,D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, **27**, 1578-1584.

Smalheiser,N.R. and Torvik,V.I. (2005) Mammalian microRNAs derived from genomic repeats. *Trends Genet.*, **21**, 322-326.

Sprinzl,M. and Vassilenko,K.S. (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33**, D139-D140.

Sullivan,C.S. *et al.* (2005) SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature*, **435**, 682-686.

Sunkar,R. *et al.* (2005) Cloning and Characterization of MicroRNAs from Rice. *Plant Cell*, **17**, 1397-1411.

Wang,X. *et al.* (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, **21**, 3610-3614.

Weinstein,L.B. and Steitz,J.A. (1999) Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.*, **11**, 378-384.

Winkler,W.C. and Breaker,R.R. (2003) Genetic control by metabolite-binding riboswitches. *Chembiochem.*, **4**, 1024-1032.

Xu,P. *et al.* (2003) The Drosophila MicroRNA Mir-14 Suppresses Cell Death and Is Required for Normal Fat Metabolism. *Curr. Biol.*, **13**, 790-795.

Xue,C. *et al.* (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.

Yang,J.H. *et al.* (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res.* gkl672.

Yousef,M. *et al.* (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, **22**, 1325-1334.

Zhang,B. *et al.* (2006a) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.*, **63**, 246-254.

Zhang,B. *et al.* (2006b) Plant microRNA: A small regulatory molecule with big impact. *Dev. Biol.*, **289**, 3-16.

BIOINFORMATICS Supplementary

pages 1–30
doi:10.1093/bioinformatics/bti283

Structural Bioinformatics

De Novo SVM Classification of Precursor MicroRNAs from Genomic Pseudo Hairpins Using Global and Intrinsic Folding Measures

Stanley NG Kwang Loong[†,‡,*], Santosh K. MISHRA[†,‡]

[†] Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

[‡] NUS Graduate School for Integrative Sciences & Engineering, Centre for Life Sciences, #05-01, 28 Medical Drive, Singapore 117456

Received on December 13, 2006; revised on; accepted on

5 BIOGENESIS OF MATURE MICRORNAS

In the prevailing biogenesis model of miRNAs maturation (Bartel 2004; Kim 2005), >1000-nts primary transcripts (*pri-miRs*) originate from the intergenic or intronic regions of annotated (non-)protein-coding transcription units (Rodriguez *et al*., 2004). The *pri-miRs* are cleaved by the nuclear RNase III endonuclease Drosha/Pasha complex, yielding ~70–120-nts precursor transcripts (*pre-miRs*) with 5' phosphate and ~2-nts 3' overhang. These *pre-miRs* exhibiting characteristic imperfect and extended RNA stem-loop structures are actively exported by the cargo transporter Exportin-5 in a Ran-GTP dependent manner into the cytoplasm (Bohnsack *et al*., 2004; Yi *et al*., 2003; Zeng and Cullen 2004). The *pre-miRs* are further excised by another RNase III endonuclease Dicer into an intermediate duplex *miR*:*miR**, a ~21–23-nts asymmetric mature miRNA duplex. The *miR*:*miR** is recruited by a ribonucleoprotein RNA-Induced Silencing Complex (RISC) (Gregory *et al*., 2005; Maniataki and Mourelatos 2005; Rivas *et al*., 2005; Tang 2005). The strand *miR* with weaker hydrogen binding survives as the mature miRNA, which is preferentially loaded into RISC. miRNA-directed posttranscriptional silencing of target genes occurs by mRNA degradation (Brennecke *et al*., 2005), or translational arrest of protein synthesis (Doench and Sharp 2004), or mRNA deadenylation (Wu *et al*., 2006).

6 MATERIALS AND METHODS

6.1 Biologically relevant datasets

8,494 pseudo hairpin sequences. We analyze 8,494 pseudo hairpin sequences that were extracted from the protein-coding regions (CDSs) according to the UCSC refGene annotation tables (Karolchik *et al*., 2003) and human RefSeq genes (Pruitt and Maglott 2001). As wrongly assumed 'negative samples' can distort the decision boundary of SVM in an unpredictable and/or significant manner, special requirements are imposed on the selection of genomic inverted repeats. First, they originate from genomic regions that do not undergo any known experimentally validated alternative splicing (AS) events, as described previously (Xue *et al*., 2005). This criterion ensures that they do not encode genuine human *pre-miRs*. Second, they are analogous to genuine human *pre-miRs* by displaying similar distribution in terms of their length ~90-nts, hairpin structures with stem ≥8-bps including the GU wobble pairs, and MFEs ≤ -15 kcal/mol. In addition, they fold without multiple loops in their RNA structures as verified by the RNAfold program (Hofacker 2003).

2,241 non-redundant pre-miR sequences. We retrieve 4,028 annotated *pre-miR* sequences spanning across 45 species from miRBase Registry Database release 8.2 (July 2006) (Griffiths-Jones *et al*., 2006). As strong sequence homologies exist among *pre-miRs* both within a single and between different specie(s), the homologs of the training *pre-miRs* must be excluded from the testing and independent evaluation sets. The original dataset is filtered to 90% identity using a greedy incremental clustering algorithm (Li and Godzik 2006). Briefly, all the sequences are first sorted in order of decreasing length and the longest one becomes the representative of the first cluster. Each remaining sequence is compared with the existing representatives and grouped into their cluster if the similarity with any representative is above a given threshold (default value is 0.9), else that sequence becomes the representative of a new cluster. Consequently, 2,241 non-redundant *pre-miRs* spanning 41 species (categorized into arthropoda, nematoda, verterbrata, viridiplantae, and viruses) serve as the reference positive set; none of the sequences from *G. gorilla*, *M. nemestrina* *P. paniscus*, and *P. pygmaeus* is retained.

12,387 functional prokaryotic and eukaryotic ncRNA sequences. We retrieve all curated seed ncRNA sequences from Rfam repository release 7.0 (March 2005) (Griffiths-Jones *et al*., 2005). After removing 46 types of *pre-miRs*, 12,387 functional prokaryotic and eukaryotic ncRNAs spanning 457 types (categorized into 16 classes) serve as the negative non *pre-miR* dataset. They have similar length distribution to the known *pre-miRs*, and can fold with hairpin(s) or stem-loop(s) (Eddy 2001; Storz 2002; Svoboda and Cara 2006).

31 mRNA sequences. We investigate 31 mRNA sequences that tend to fold into complex RNA structures with extremely negative MFEs (Freyhult *et al*., 2005). They are randomly selected from GeneBank DNA database (Benson *et al*., 2005) with the following GenBank accession numbers: NM_001005151.1, NM_001003967.1, NM_177233.4, AY675236.1, NM_001004202.1, NM_178539.2, AB164385.1, AY555511.1, AB189435.1, NM_178307.2, NM_001003966.1, NM_205498.1, NM_013564.3, Z81556.1, NM_131070.2, X56279.1, AK045412.1, AF452886.1, BC049701.1, BC050086.1, NM_172343.1, AY182163.1, BC072691.1, CV127341.1, NC_004671.1, X00910.1, AY226143.1, AJ621386, CV122154.1, X68284, and CV199185.1.

6.2 Feature vector

Adjusted base pairing propensity, dP measures the total number of base pairs present in the RNA secondary structure *S* (Schultes *et al*., 1999) divided by the length *L* in nucleotides. It removes the bias that a long sequence tends to have more base pairs. *dP* ranges [0.0, 0.5], 0.0 for no base pair interactions and 0.5 for maximum of *L*/2 base pairs.

Adjusted minimum free energy of folding, dG measures the thermodynamic stability of RNA structure *S* i.e., the lowest MFE for the most favorable conformation, divided by the length *L* in nucleotides (Freyhult *et al*., 2005). It removes the bias that a long sequence tends to have lower negative MFE (Seffens and Digby 1999).

MFE Index 1, $MFEI_1$ is the ratio of *dG* and %*G+C* content (Zhang *et al*., 2006).

Adjusted shannon entropy, dQ in Eq. (1), characterizes the base pairing probability distribution (BPPD) in a RNA structure *S* as a chaotic dynamical system (Freyhult *et al*., 2005; Huynen *et al*., 1997; Schultes *et al*., 1999). Low values of *dQ* correspond to BPPD that are dominated by single, a few, or by the absence of base pairings. These bases are better predicted than those having multiple alternative states.

$$dQ = -\frac{1}{L}\sum_{i<j} p_{ij}\log_2(p_{ij}), \quad p_{ij} = \sum_{S_\alpha \in \mathbf{S}(\mathbf{x})} P(S_\alpha) d_{ij}^{\alpha}. \tag{1}$$

Here, p_{ij} denotes the probability of bases *i* and *j* pair, computed from the McCaskill's algorithm (J.S.McCaskill 1990); $d_{ij}^{\alpha} = 1$ if *i* and *j* pair, 0 otherwise. RNA molecules exist *in vivo* as an ensemble of secondary structures $S_\alpha \in S(\mathbf{x})$ with the Boltzmann distribution probability $P(S_\alpha)$ (Mathews 2004).

Adjusted base pair distance, dD in Eq. (2), is the base pair distance for all pairs of structures S_α and S_β inferred from sequence **s** (Freyhult *et al*., 2005; Moulton *et al*., 2000).

$$dD = \frac{1}{2L}\sum_{S_a, S_b \in \mathbf{S}(\mathbf{s})} P(S_a)P(S_b)d_{BP}(S_a, S_b) = \frac{1}{L}\sum_{i<j} p_{ij}(1-p_{ij}). \tag{2}$$

Here, the number of base pairs not shared by them is given by $d_{BP}(S_\alpha, S_\beta) = |S_\alpha \cup S_\beta| - |S_\alpha \cap S_\beta| = \sum_{i<j}(d_{ij}^{a} + d_{ij}^{b} - 2d_{ij}^{a}d_{ij}^{b})$. The number of base pairs in S_α is $|S_\alpha| = \sum_{i<j} d_{ij}^{a}$. Definitions of p_{ij} and d_{ij}^{a} follow those of *dQ* in Eq. (1).

Second (or the Fiedler) eigenvalue, dF in Eq. (3), measures the compactness of a tree-graph $G = (V, E)$ (Fera *et al*., 2004; Gan *et al*., 2004). At the coarsest scale, each vertex $v \in V$ represents a bulge loop, hairpin loop, internal loop, the 5' and 3' unpaired termini, or the multi-branch loop; each edge $e \in E$ denotes a RNA stem. *dF* is computed from the Laplacian matrix $\mathbf{L}(G)$, a mathematical representation of the tree-graph *G*. *dF* can be used as a similarity measure among a collection of RNA secondary structures.

$$\mathbf{L}(G)\mathbf{X} = l\,\mathbf{X} \Leftrightarrow dF = FidlerEigen[\mathbf{L}(G)]. \tag{3}$$

MFE Index 2, MFEI$_2$ is the ratio of *dG* and the number of stems *S*, which are structural motifs containing more than three contiguous base pairs.

Normalized feature vectors. The *Z*-score $Z(\mathbf{s}_n)$ in Eq. (4) normalizes the feature $S(\mathbf{s}_n)$ of n^{th} native RNA sequence $\mathbf{s}_n$ in terms of the number of standard deviations by which $S(\mathbf{s}_n)$ differs from the mean of inferred $R = 10^4$ random RNA sequences $\mathbf{r}_n$. $S(\mathbf{s}_n)$ can be *dG*, *dP*, *dQ*, *dD*, and *dF*; corresponding *Z*-scores are denoted as *zG*, *zP*, *zQ*, *zD*, and *zF*.

$$Z(\mathbf{s}_n) = \frac{S(\mathbf{s}_n) - m_n}{s_n}, \quad s_n^2 = \frac{1}{R-1}\sum_{i=1}^{R}[S_i(\mathbf{r}_n) - m_n]^2. \tag{4}$$

Here, $S_i(\mathbf{r}_n)$ is the computed feature for the i^{th} random sequence of $\mathbf{r}_n$, μ_n and σ_n are the sample mean and the standard deviation of the feature $S(\mathbf{s}_n)$. The entire set of *R* random sequences $\mathbf{r}_n$ is synthesized by the "Altschul-Erikson algorithm" (Altschul and Erickson 1985), a form of dinucleotide shuffling. Briefly, it shuffles $\mathbf{s}_n$ while preserving exactly both the mono- and di-nucleotide frequencies. The $\mathbf{r}_n$ shares the same first and last nucleotides as $\mathbf{s}_n$. The order of the shuffled nucleotides is 'less random' due to fewer possible dinucleotide-preserving permutations.

6.3 Statistical tests and performance evaluation

F-scores of features. The "quality" of the i^{th} feature is described commonly by the F-scores F1 (Dror *et al*., 2005) and F2 (Chen and Lin 2006) in Eq. (5). The larger their values for the i^{th} feature, the more likely this feature possesses discriminative importance/power.

$$F1 = \frac{\left|m_i^+ - m_i^-\right|}{\left|s_i^+ + s_i^-\right|}, \quad F2 = \frac{\left(m_i^+ - \bar{m}_i\right)^2 + \left(m_i^- - \bar{m}_i\right)^2}{\left(s_i^+\right)^2 + \left(s_i^-\right)^2}. \tag{5}$$

Here m_i^+ / m_i^- and s_i^+ / s_i^- denote the means and standard deviations of the positive (+) and negative (−) training datasets, correspondingly. The numerator and denominator describe the discrimination between the two classes, and that within each of the two classes.

Performance measures. SE (Sensitivity or recall), *SP* (Specificity), *ACC* (Accuracy), *Fm* (F-measure) (Liu *et al*., 2006), and *MCC* (Matthew's Correlation Coefficient) (Bhasin *et al*., 2006) are defined in Eq. (6). All metrics (except *MCC*) range [0.0, 1.0]; closer to 1.0 implies better scores, and *vice-versa*. *MCC* ranges [-1.0, 1.0]; -1.0, 0.0, and 1.0 indicate worst possible, perfectly random, and best possible classification, respectively. Unlike *ACC*, *Fm* and *MCC* account for the unbalanced datasets.

$$SE = \frac{TP}{TP+FN}, \quad SP = \frac{TN}{TN+FP}, \quad ACC = \frac{TP+TN}{TP+TN+FP+FN},$$
$$Fm = \frac{2(SP \times PPV)}{SP+PPV} \text{ where } PPV \text{ (Positive Predictive Value)} = \frac{TP}{TP+FP}, \tag{6}$$
$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}.$$

Here *TP*, *FN*, *FP*, and *TN* denote the number of true/false *pre-miRs* detected/missed, correspondingly. The "quality" of a binary classification is measured by the area under the Receiver Operating Characteristic curve (ROC) or simply AUC. ROC plots the trade-off between the *SE* and *FPR* (False-positive rate = 1 − *SP*) across all possible thresholds (Lasko *et al*., 2005). AUC ranges [0.5, 1.0]; closer to 0.5 (about the upward diagonal) and to 1.0 (along the left-top boundary) signify a totally random and a perfect classifier (Lasko *et al*., 2005). In this work, AUC is computed by the trapezoidal rule (Kestler 2001).

Benchmarking miPred. Both *3SVM* (Xue *et al*., 2005) and Naïve Bayesian Classifier (*NBC*) serve as independent baseline models to benchmark the performance improvements or deterioration (if any) of *miPred*. The original *3SVM* was previously trained on 163 human *pre-miRs* and 168 pseudo hairpins using the older libSVM 2.36 with the "-b 1" option disabled. Here, *3SVM* is trained on randomly selected 200 human *pre-miRs* and 400 pseudo hairpins using the latest libSVM 2.82 (the "-b 1" option is enabled) and the optimal hyperparameter pair (C, γ). *3SVM* is applied to the testing and independent evaluation datasets with "*svm-predict* -b 1". The Bayes Classifier Induction (*bci*) version 2.14, a free implementation of *NBC* (http://fuzzy.cs.unimagdeburg.de/~borgelt/bayes.html), is used for training and testing with the exact samples and attributes employed by *3SVM* and *miPred*; denoted as *3SVM-NBC* and *miPred-NBC*. For training, "*bci* -L1" yields better classification results than the default "-L0". *NBC* seeks to maximize the probability $P(X|C) = P(f_1, f_2, \ldots f_n)|C)$ such that the sample *X* belongs to one of the binary classes $C = (T, F)$.

REFERENCES

Altschul,S.F. and Erickson,B.W. (1985) Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.*, **2**, 526-538.

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

Benson,D.A. *et al.* (2005) GenBank. *Nucleic Acids Res.*, **33**, D34-D38.

Bhasin,M. *et al*. (2006) Recognition and Classification of Histones Using Support Vector Machine. *J. Comp. Biol.*, **13**, 102-112.

Bohnsack,M.T. *et al*. (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, **10**, 185-191.

Brennecke,J. *et al*. (2005) Principles of MicroRNA-Target Recognition. *PLoS Biol.*, **3**, e85.

Chen,Y.-W. and Lin,C.-J. (2006) Combining SVMs with various feature selection strategies. In Guyon,I., Gunn,S., Nikravesh,M.and Zadeh,L. (eds), *Feature extraction, foundations and applications. Springer*, pp. 315-323.

Doench,J.G. and Sharp,P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.*, **18**, 504-511.

Dror,G. *et al.* (2005) Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*, **21**, 897-901.

Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919-929.

Fera,D. *et al.* (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, **5**, 88.

Freyhult,E. *et al.* (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241.

Gan,H.H. *et al.* (2004) RAG: RNA-As-Graphs database--concepts, analysis, and features. *Bioinformatics*, **20**, 1285-1291.

Gregory,R.I. *et al*. (2005) Human RISC Couples MicroRNA Biogenesis and Posttranscriptional Gene Silencing. *Cell*, **123**, 631-640.

Griffiths-Jones,S. *et al*. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140-D144.

Griffiths-Jones,S. *et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121-D124.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429-3431.

Huynen,M. *et al.* (1997) Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.*, **267**, 1104-1112.

J.S.McCaskill (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105-1119.

Karolchik,D. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51-54.

Kestler,H.A. (2001) ROC with confidence - a Perl program for receiver operator characteristic curves. *Comput. Methods Programs Biomed.*, **64**, 133-136.

Kim,V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell Biol.*, **6**, 376-385.

Lasko,T.A. *et al*. (2005) The use of receiver operating characteristic curves in biomedical informatics. *J. Biomed. Inform.*, **38**, 404-415.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-1659.

Liu,J. *et al*. (2006) Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. *PLoS Genet.*, **2**, e29.

Maniataki,E. and Mourelatos,Z. (2005) A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. *Genes Dev.*, **19**, 2979-2990.

Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178-1190.

Moulton,V. *et al*. (2000) Metrics on RNA Secondary Structures. *J. Comp. Biol.*, **7**, 277-292.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137-140.

Rivas,F.V. *et al*. (2005) Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat. Struct. Mol. Biol.*, **12**, 340-349.

Rodriguez,A. *et al*. (2004) Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Res.*, **14**, 1902-1910.

Schultes,E.A. *et al*. (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, **49**, 76-83.

Seffens,W. and Digby,D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, **27**, 1578-1584.

Storz,G. (2002) An Expanding Universe of Noncoding RNAs. *Science*, **296**, 1260-1263.

Svoboda,P. and Cara,A.D. (2006) Hairpin RNA: a secondary structure of primary importance. *Cell. Mol. Life Sci.*, **63**, 901-908.

Tang,G. (2005) siRNA and miRNA: an insight into RISCs. *Trends Biochem. Sci.*, **30**, 106-114.

Wu,L. *et al*. (2006) MicroRNAs direct rapid deadenylation of mRNA. *Proc. Natl. Acad. Sci. USA*, **103**, 4034-4039.

Xue,C. *et al*. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.

Yi,R. *et al*. (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.*, **17**, 3011-3016.

Zeng,Y. and Cullen,B.R. (2004) Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res.*, **32**, 4776-4785.

Zhang,B. *et al*. (2006) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.*, **63**, 246-254.

Table S2. The prediction performances of *miPred*, *miPred-NBC*, *3SVM*, and *3SVM-NBC* evaluated on the *pre-miR* datasets TR-H (200 human *pre-miRs* and 400 pseudo hairpins), TE-H (remaining 123 human *pre-miRs* and 246 pseudo hairpins), and IE-NH (1,918 *pre-miRs* across 40 non-human species and 3,836 pseudo hairpins).

miPred

Species	*Genus*	*TP*	*FN*	*P*	*FP*	*TN*	*N*	*%SE*	*%SP*	*%FPR*	*%ACC*
Homo sapiens	*Vertebrata*	176	24	200	10	390	400	88.00	97.50	2.50	94.33
Homo sapiens	*Vertebrata*	104	19	123	5	241	246	84.55	97.97	2.03	93.50
Anopheles gambiae	*Arthropoda*	37	1	38	1	75	76	97.37	98.68	1.32	98.25
Apis mellifera	*Arthropoda*	25	0	25	2	48	50	100	96	4	97.33
Arabidopsis thaliana	*Viridiplantae*	101	7	108	2	214	216	93.52	99.07	0.93	97.22
Ateles geoffroyi	*Vertebrata*	2	0	2	0	4	4	100	100	0	100
Bos taurus	*Vertebrata*	8	3	11	2	20	22	72.73	90.91	9.09	84.85
Caenorhabditis briggsae	*Nematoda*	72	4	76	1	151	152	94.74	99.34	0.66	97.81
Caenorhabditis elegans	*Nematoda*	96	17	113	7	219	226	84.96	96.9	3.1	92.92
Canis familiaris	*Vertebrata*	3	0	3	0	6	6	100	100	0	100
Danio rerio	*Vertebrata*	235	11	246	19	473	492	95.53	96.14	3.86	95.94
Drosophila melanogaster	*Arthropoda*	67	6	73	4	142	146	91.78	97.26	2.74	95.43
Drosophila pseudoobscura	*Arthropoda*	32	3	35	1	69	70	91.43	98.57	1.43	96.19
Epstein Barr (EBV)	*Viruses*	22	0	22	2	42	44	100	95.45	4.55	96.97
Fugu rubripes	*Vertebrata*	68	2	70	2	138	140	97.14	98.57	1.43	98.1
Gallus gallus	*Vertebrata*	87	5	92	4	180	184	94.57	97.83	2.17	96.74
Glycine max	*Viridiplantae*	20	1	21	0	42	42	95.24	100	0	98.41
Herpes Simplex (HSV)	*Viruses*	1	0	1	0	2	2	100	100	0	100
Human cytomegalovirus (HCMV)	*Viruses*	11	0	11	1	21	22	100	95.45	4.55	96.97
Kaposi sarcoma-associated herpesvirus (KSHV)	*Viruses*	11	1	12	0	24	24	91.67	100	0	97.22
Lagothrix lagotricha	*Vertebrata*	1	1	2	0	4	4	50	100	0	83.33
Lemur catta	*Vertebrata*	2	1	3	0	6	6	66.67	100	0	88.89
Macaca mulatta	*Vertebrata*	1	1	2	0	4	4	50	100	0	83.33
Medicago truncatula	*Viridiplantae*	17	1	18	0	36	36	94.44	100	0	98.15
Mouse γ-herpesvirus (MGHV68)	*Viruses*	8	1	9	1	17	18	88.89	94.44	5.56	92.59
Mus musculus	*Vertebrata*	166	33	199	9	389	398	83.42	97.74	2.26	92.96
Oryza sativa	*Viridiplantae*	140	12	152	4	300	304	92.11	98.68	1.32	96.49
Ovis aries	*Vertebrata*	2	0	2	0	4	4	100	100	0	100
Pan troglodytes	*Vertebrata*	2	1	3	1	5	6	66.67	83.33	16.67	77.78
Physcomitrella patens	*Viridiplantae*	17	0	17	0	34	34	100	100	0	100
Populus trichocarpa	*Viridiplantae*	144	13	157	13	301	314	91.72	95.86	4.14	94.48
Rattus norvegicus	*Vertebrata*	56	12	68	10	126	136	82.35	92.65	7.35	89.22
Rhesus lymphocryptovirus	*Viruses*	16	0	16	2	30	32	100	93.75	6.25	95.83
Saccharum officinarum	*Viridiplantae*	3	1	4	0	8	8	75	100	0	91.67
Saguinus labiatus	*Vertebrata*	1	1	2	0	4	4	50	100	0	83.33
Simian virus (SV40)	*Viruses*	1	0	1	0	2	2	100	100	0	100
Sorghum bicolor	*Viridiplantae*	48	2	50	2	98	100	96	98	2	97.33
Sus scrofa	*Vertebrata*	1	1	2	0	4	4	50	100	0	83.33
Tetraodon nigroviridis	*Vertebrata*	40	3	43	0	86	86	93.02	100	0	97.67
Xenopus laevis	*Vertebrata*	4	1	5	0	10	10	80	100	0	93.33
Xenopus tropicalis	*Vertebrata*	119	6	125	4	246	250	95.2	98.4	1.6	97.33
Zea mays	*Viridiplantae*	79	0	79	5	153	158	100	96.84	3.16	97.89
Total samples		2046	195	2241	114	4368	4482				

(*Species*) Row 1 (TR-H), row 2 (TE-H), and the remaining rows 3−43 (IE-NH). *TP* (real *pre-miRs* detected), *FN* (real *pre-miRs* missed), *P* (real *pre-miRs*), *FP* (pseudo hairpins detected), *TN* (pseudo hairpins missed), *N* (pseudo hairpins), *%SE* (Sensitivity), *%SP* (Specificity), *%FPR* (False-positive rate), and *%ACC* (Accuracy).

miPred-NBC

Species	*Genus*	*TP*	*FN*	*P*	*FP*	*TN*	*N*	*%SE*	*%SP*	*%FPR*	*%ACC*
Homo sapiens	*Vertebrata*	200	0	200	0	400	400	100.00	100.00	0.00	100.00
Homo sapiens	*Vertebrata*	46	77	123	36	210	246	37.40	85.37	14.63	69.38
Anopheles gambiae	*Arthropoda*	12	26	38	7	69	76	31.58	90.79	9.21	71.05
Apis mellifera	*Arthropoda*	6	19	25	6	44	50	24.00	88.00	12.00	66.67
Arabidopsis thaliana	*Viridiplantae*	20	88	108	27	189	216	18.52	87.50	12.50	64.51
Ateles geoffroyi	*Vertebrata*	0	2	2	1	3	4	0.00	75.00	25.00	50.00
Bos taurus	*Vertebrata*	1	10	11	2	20	22	9.09	90.91	9.09	63.64
Caenorhabditis briggsae	*Nematoda*	27	49	76	20	132	152	35.53	86.84	13.16	69.74
Caenorhabditis elegans	*Nematoda*	51	62	113	26	200	226	45.13	88.50	11.50	74.04
Canis familiaris	*Vertebrata*	0	3	3	1	5	6	0.00	83.33	16.67	55.56
Danio rerio	*Vertebrata*	71	175	246	62	430	492	28.86	87.40	12.60	67.89
Drosophila melanogaster	*Arthropoda*	21	52	73	17	129	146	28.77	88.36	11.64	68.49
Drosophila pseudoobscura	*Arthropoda*	12	23	35	5	65	70	34.29	92.86	7.14	73.33
Epstein Barr (EBV)	*Viruses*	6	16	22	4	40	44	27.27	90.91	9.09	69.70
Fugu rubripes	*Vertebrata*	10	60	70	18	122	140	14.29	87.14	12.86	62.86
Gallus gallus	*Vertebrata*	24	68	92	22	162	184	26.09	88.04	11.96	67.39
Glycine max	*Viridiplantae*	2	19	21	3	39	42	9.52	92.86	7.14	65.08
Herpes Simplex (HSV)	*Viruses*	0	1	1	1	1	2	0.00	50.00	50.00	33.33
Human cytomegalovirus (HCMV)	*Viruses*	0	11	11	5	17	22	0.00	77.27	22.73	51.52
Kaposi sarcoma-associated herpesvirus (KSHV)	*Viruses*	1	11	12	5	19	24	8.33	79.17	20.83	55.56
Lagothrix lagotricha	*Vertebrata*	0	2	2	0	4	4	0.00	100.00	0.00	66.67
Lemur catta	*Vertebrata*	0	3	3	2	4	6	0.00	66.67	33.33	44.44
Macaca mulatta	*Vertebrata*	0	2	2	0	4	4	0.00	100.00	0.00	66.67
Medicago truncatula	*Viridiplantae*	4	14	18	4	32	36	22.22	88.89	11.11	66.67
Mouse γ-herpesvirus (MGHV68)	*Viruses*	2	7	9	3	15	18	22.22	83.33	16.67	62.96
Mus musculus	*Vertebrata*	37	162	199	52	346	398	18.59	86.93	13.07	64.15
Oryza sativa	*Viridiplantae*	35	117	152	37	267	304	23.03	87.83	12.17	66.23
Ovis aries	*Vertebrata*	0	2	2	0	4	4	0.00	100.00	0.00	66.67
Pan troglodytes	*Vertebrata*	2	1	3	1	5	6	66.67	83.33	16.67	77.78
Physcomitrella patens	*Viridiplantae*	3	14	17	3	31	34	17.65	91.18	8.82	66.67
Populus trichocarpa	*Viridiplantae*	33	124	157	41	273	314	21.02	86.94	13.06	64.97
Rattus norvegicus	*Vertebrata*	23	45	68	11	125	136	33.82	91.91	8.09	72.55
Rhesus lymphocryptovirus	*Viruses*	5	11	16	2	30	32	31.25	93.75	6.25	72.92
Saccharum officinarum	*Viridiplantae*	1	3	4	0	8	8	25.00	100.00	0.00	75.00
Saguinus labiatus	*Vertebrata*	0	2	2	1	3	4	0.00	75.00	25.00	50.00
Simian virus (SV40)	*Viruses*	0	1	1	1	1	2	0.00	50.00	50.00	33.33
Sorghum bicolor	*Viridiplantae*	7	43	50	13	87	100	14.00	87.00	13.00	62.67
Sus scrofa	*Vertebrata*	0	2	2	1	3	4	0.00	75.00	25.00	50.00
Tetraodon nigroviridis	*Vertebrata*	9	34	43	9	77	86	20.93	89.53	10.47	66.67
Xenopus laevis	*Vertebrata*	2	3	5	4	6	10	40.00	60.00	40.00	53.33
Xenopus tropicalis	*Vertebrata*	35	90	125	33	217	250	28.00	86.80	13.20	67.20
Zea mays	*Viridiplantae*	16	63	79	18	140	158	20.25	88.61	11.39	65.82
Total samples		724	1517	2241	504	3978	4482				

(*Species*) Row 1 (TR-H), row 2 (TE-H), and the remaining rows 3−43 (IE-NH). *TP* (real *pre-miRs* detected), *FN* (real *pre-miRs* missed), *P* (real *pre-miRs*), *FP* (pseudo hairpins detected), *TN* (pseudo hairpins missed), *N* (pseudo hairpins), *%SE* (Sensitivity), *%SP* (Specificity), *%FPR* (False-positive rate), and *%ACC* (Accuracy).

3SVM†

Species	*Genus*	*TP*	*FN*	*P*	*FP*	*TN*	*N*	*%SE*	*%SP*	*%FPR*	*%ACC*
Homo sapiens	*Vertebrata*	172	28	200	12	388	400	86.00	97.00	3.00	93.33
Homo sapiens	*Vertebrata*	79	29	108	10	206	216	73.15	95.37	4.63	87.96
Anopheles gambiae	*Arthropoda*	33	4	37	1	73	74	89.19	98.65	1.35	95.50
Apis mellifera	*Arthropoda*	23	2	25	1	49	50	92.00	98.00	2.00	96.00
Arabidopsis thaliana	*Viridiplantae*	69	2	71	5	137	142	97.18	96.48	3.52	96.71
Ateles geoffroyi	*Vertebrata*	2	0	2	0	4	4	100.00	100.00	0.00	100.00
Bos taurus	*Vertebrata*	7	1	8	3	13	16	87.50	81.25	18.75	83.33
Caenorhabditis briggsae	*Nematoda*	68	2	70	6	134	140	97.14	95.71	4.29	96.19
Caenorhabditis elegans	*Nematoda*	94	13	107	4	210	214	87.85	98.13	1.87	94.70
Canis familiaris	*Vertebrata*	3	0	3	1	5	6	100.00	83.33	16.67	88.89
Danio rerio	*Vertebrata*	201	32	233	30	436	466	86.27	93.56	6.44	91.13
Drosophila melanogaster	*Arthropoda*	57	9	66	7	125	132	86.36	94.70	5.30	91.92
Drosophila pseudoobscura	*Arthropoda*	28	7	35	1	69	70	80.00	98.57	1.43	92.38
Epstein Barr (EBV)	*Viruses*	19	3	22	0	44	44	86.36	100.00	0.00	95.45
Fugu rubripes	*Vertebrata*	48	16	64	5	123	128	75.00	96.09	3.91	89.06
Gallus gallus	*Vertebrata*	73	14	87	4	170	174	83.91	97.70	2.30	93.10
Glycine max	*Viridiplantae*	16	0	16	0	32	32	100.00	100.00	0.00	100.00
Herpes Simplex (HSV)	*Viruses*	0	1	1	0	2	2	0.00	100.00	0.00	66.67
Human cytomegalovirus (HCMV)	*Viruses*	8	3	11	0	22	22	72.73	100.00	0.00	90.91
Kaposi sarcoma-associated herpesvirus (KSHV)	*Viruses*	4	8	12	0	24	24	33.33	100.00	0.00	77.78
Lagothrix lagotricha	*Vertebrata*	1	0	1	0	2	2	100.00	100.00	0.00	100.00
Lemur catta	*Vertebrata*	2	0	2	0	4	4	100.00	100.00	0.00	100.00
Macaca mulatta	*Vertebrata*	1	0	1	0	2	2	100.00	100.00	0.00	100.00
Medicago truncatula	*Viridiplantae*	15	0	15	2	28	30	100.00	93.33	6.67	95.56
Mouse γ-herpesvirus (MGHV68)	*Viruses*	5	4	9	1	17	18	55.56	94.44	5.56	81.48
Mus musculus	*Vertebrata*	145	41	186	5	367	372	77.96	98.66	1.34	91.76
Oryza sativa	*Viridiplantae*	106	9	115	11	219	230	92.17	95.22	4.78	94.20
Ovis aries	*Vertebrata*	1	0	1	0	2	2	100.00	100.00	0.00	100.00
Pan troglodytes	*Vertebrata*	2	1	3	0	6	6	66.67	100.00	0.00	88.89
Physcomitrella patens	*Viridiplantae*	14	0	14	0	28	28	100.00	100.00	0.00	100.00
Populus trichocarpa	*Viridiplantae*	106	15	121	12	230	242	87.60	95.04	4.96	92.56
Rattus norvegicus	*Vertebrata*	50	12	62	5	119	124	80.65	95.97	4.03	90.86
Rhesus lymphocryptovirus	*Viruses*	16	0	16	1	31	32	100.00	96.88	3.13	97.92
Saccharum officinarum	*Viridiplantae*	0	0	0	0	0	0	NaN	NaN	NaN	NaN
Saguinus labiatus	*Vertebrata*	0	1	1	0	2	2	0.00	100.00	0.00	66.67
Simian virus (SV40)	*Viruses*	1	0	1	0	2	2	100.00	100.00	0.00	100.00
Sorghum bicolor	*Viridiplantae*	33	2	35	2	68	70	94.29	97.14	2.86	96.19
Sus scrofa	*Vertebrata*	0	2	2	0	4	4	0.00	100.00	0.00	66.67
Tetraodon nigroviridis	*Vertebrata*	39	2	41	3	79	82	95.12	96.34	3.66	95.94
Xenopus laevis	*Vertebrata*	2	3	5	1	9	10	40.00	90.00	10.00	73.33
Xenopus tropicalis	*Vertebrata*	101	21	122	7	237	244	82.79	97.13	2.87	92.35
Zea mays	*Viridiplantae*	50	2	52	7	97	104	96.15	93.27	6.73	94.23
Total samples		1694	289	1983	147	3819	3966				

†, *3SVM* model was trained on 200 human ***pre-miRs*** and 400 pseudo hairpins randomly selected using the latest libSVM 2.82 (the "-b 1" option was enabled) and the optimal hyperparameter pair (*C*, *γ*). (***Species***) Row 1 (TR-H), row 2 (TE-H), and the remaining rows 3−43 (IE-NH). ***TP*** (real ***pre-miRs*** detected), ***FN*** (real ***pre-miRs*** missed), ***P*** (real ***pre-miRs***), ***FP*** (pseudo hairpins detected), ***TN*** (pseudo hairpins missed), ***N*** (pseudo hairpins), ***%SE*** (Sensitivity), ***%SP*** (Specificity), ***%FPR*** (False-positive rate), and ***%ACC*** (Accuracy).

3SVM-NBC

Species	*Genus*	*TP*	*FN*	*P*	*FP*	*TN*	*N*	*%SE*	*%SP*	*%FPR*	*%ACC*
Homo sapiens	*Vertebrata*	196	4	200	13	387	400	98.00	96.75	3.25	97.17
Homo sapiens	*Vertebrata*	71	37	108	51	165	216	65.74	76.39	23.61	72.84
Anopheles gambiae	*Arthropoda*	27	10	37	18	56	74	72.97	75.68	24.32	74.77
Apis mellifera	*Arthropoda*	20	5	25	9	41	50	80.00	82.00	18.00	81.33
Arabidopsis thaliana	*Viridiplantae*	44	27	71	30	112	142	61.97	78.87	21.13	73.24
Ateles geoffroyi	*Vertebrata*	1	1	2	1	3	4	50.00	75.00	25.00	66.67
Bos taurus	*Vertebrata*	4	4	8	4	12	16	50.00	75.00	25.00	66.67
Caenorhabditis briggsae	*Nematoda*	52	18	70	23	117	140	74.29	83.57	16.43	80.48
Caenorhabditis elegans	*Nematoda*	87	20	107	39	175	214	81.31	81.78	18.22	81.62
Canis familiaris	*Vertebrata*	3	0	3	2	4	6	100.00	66.67	33.33	77.78
Danio rerio	*Vertebrata*	140	93	233	112	354	466	60.09	75.97	24.03	70.67
Drosophila melanogaster	*Arthropoda*	38	28	66	31	101	132	57.58	76.52	23.48	70.20
Drosophila pseudoobscura	*Arthropoda*	20	15	35	15	55	70	57.14	78.57	21.43	71.43
Epstein Barr (EBV)	*Viruses*	12	10	22	9	35	44	54.55	79.55	20.45	71.21
Fugu rubripes	*Vertebrata*	31	33	64	33	95	128	48.44	74.22	25.78	65.63
Gallus gallus	*Vertebrata*	48	39	87	44	130	174	55.17	74.71	25.29	68.20
Glycine max	*Viridiplantae*	5	11	16	5	27	32	31.25	84.38	15.63	66.67
Herpes Simplex (HSV)	*Viruses*	0	1	1	0	2	2	0.00	100.00	0.00	66.67
Human cytomegalovirus (HCMV)	*Viruses*	3	8	11	1	21	22	27.27	95.45	4.55	72.73
Kaposi sarcoma-associated herpesvirus (KSHV)	*Viruses*	2	10	12	5	19	24	16.67	79.17	20.83	58.33
Lagothrix lagotricha	*Vertebrata*	1	0	1	0	2	2	100.00	100.00	0.00	100.00
Lemur catta	*Vertebrata*	2	0	2	0	4	4	100.00	100.00	0.00	100.00
Macaca mulatta	*Vertebrata*	1	0	1	1	1	2	100.00	50.00	50.00	66.67
Medicago truncatula	*Viridiplantae*	8	7	15	6	24	30	53.33	80.00	20.00	71.11
Mouse γ-herpesvirus (MGHV68)	*Viruses*	4	5	9	7	11	18	44.44	61.11	38.89	55.56
Mus musculus	*Vertebrata*	110	76	186	83	289	372	59.14	77.69	22.31	71.51
Oryza sativa	*Viridiplantae*	73	42	115	56	174	230	63.48	75.65	24.35	71.59
Ovis aries	*Vertebrata*	1	0	1	1	1	2	100.00	50.00	50.00	66.67
Pan troglodytes	*Vertebrata*	2	1	3	1	5	6	66.67	83.33	16.67	77.78
Physcomitrella patens	*Viridiplantae*	7	7	14	3	25	28	50.00	89.29	10.71	76.19
Populus trichocarpa	*Viridiplantae*	73	48	121	52	190	242	60.33	78.51	21.49	72.45
Rattus norvegicus	*Vertebrata*	39	23	62	24	100	124	62.90	80.65	19.35	74.73
Rhesus lymphocryptovirus	*Viruses*	10	6	16	6	26	32	62.50	81.25	18.75	75.00
Saccharum officinarum	*Viridiplantae*	0	0	0	0	0	0	NaN	NaN	NaN	NaN
Saguinus labiatus	*Vertebrata*	0	1	1	0	2	2	0.00	100.00	0.00	66.67
Simian virus (SV40)	*Viruses*	1	0	1	0	2	2	100.00	100.00	0.00	100.00
Sorghum bicolor	*Viridiplantae*	19	16	35	14	56	70	54.29	80.00	20.00	71.43
Sus scrofa	*Vertebrata*	1	1	2	0	4	4	50.00	100.00	0.00	83.33
Tetraodon nigroviridis	*Vertebrata*	18	23	41	17	65	82	43.90	79.27	20.73	67.48
Xenopus laevis	*Vertebrata*	0	5	5	1	9	10	0.00	90.00	10.00	60.00
Xenopus tropicalis	*Vertebrata*	68	54	122	49	195	244	55.74	79.92	20.08	71.86
Zea mays	*Viridiplantae*	18	34	52	30	74	104	34.62	71.15	28.85	58.97
Total samples		1260	723	1983	796	3170	3966				

(*Species*) Row 1 (TR-H), row 2 (TE-H), and the remaining rows 3−43 (IE-NH). *TP* (real *pre-miRs* detected), *FN* (real *pre-miRs* missed), *P* (real *pre-miRs*), *FP* (pseudo hairpins detected), *TN* (pseudo hairpins missed), *N* (pseudo hairpins), *%SE* (Sensitivity), *%SP* (Specificity), *%FPR* (False-positive rate), and *%ACC* (Accuracy).

Table S3. The mean sensitivity and specificity of *miPred*, *miPred-NBC*, *3SVM*, and *3SVM-NBC* evaluated on the non-human *pre-miR* dataset IE-NH (1,918 *pre-miRs* across 40 non-human species and 3,836 pseudo hairpins) categorized by genus of *pre-miRs*.

		miPred		*miPred-NBC*		*3SVM*‡			*3SVM-NBC*	
Genus	*No. of species*	*%SE*	*%SP*	*%SE*	*%SP*	*No. of excluded species*	*%SE*	*%SP*	*%SE*	*%SP*
Arthropoda	4	95.14 ± 2.11	97.63 ± 0.63	29.66 ± 2.20	90.00 ± 1.14	0	86.89 ± 2.57	97.48 ± 0.94	66.92 ± 5.71	78.19 ± 1.41
Viridiplantae	9	93.11 ± 2.47	98.72 ± 0.51	19.02 ± 1.60	90.09 ± 1.40	1	95.92 ± 1.57	96.31 ± 0.93	51.16 ± 4.31	79.73 ± 1.92
Vertebrata†	18	79.29 ± 4.56	97.53 ± 1.05	15.91 ± 4.43	84.83 ± 2.60	0	76.44 ± 7.48	96.11 ± 1.35	61.23 ± 7.22	79.58 ± 3.53
Nematoda	2	89.85 ± 4.89	98.12 ± 1.22	40.33 ± 4.80	87.67 ± 0.83	0	92.50 ± 4.65	96.92 ± 1.21	77.80 ± 3.51	82.68 ± 0.90
Viruses	7	97.22 ± 1.81	97.01 ± 1.08	12.72 ± 5.23	74.92 ± 6.81	0	64.00 ± 14.04	98.76 ± 0.84	43.63 ± 12.49	85.22 ± 5.36

†, *Homo sapiens* is excluded. ‡, *3SVM* model was trained on 200 human *pre-miRs* and 400 pseudo hairpins randomly selected using the latest libSVM 2.82 (the "-b 1" option was enabled) and the optimal hyperparameter pair (C, γ). *%SE* (Sensitivity) and *%SP* (Specificity). Values are expressed as mean ± standard error.

Table S4. The prediction performances of *miPred*, *miPred-NBC*, *3SVM*, and *3SVM-NBC* evaluated on the non *pre-miR* datasets IE-NC (12,387 functional ncRNAs) and IE-M (31 mRNAs).

				miPred		*miPred-NBC*		*3SVM*[‡]			*3SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
RF00001	*5S ribosomal RNA*	*Gene|rRNA*	589	409	69.44	517	87.78	2	2	100.00	1	50.00
RF00002	*5.8S ribosomal RNA*	*Gene|rRNA*	63	59	93.65	59	93.65	1	1	100.00	1	100.00
RF00003	*U1 spliceosomal RNA*	*Gene|snRNA|splicing*	54	38	70.37	45	83.33	0	0	NaN	0	NaN
RF00004	*U2 spliceosomal RNA*	*Gene|snRNA|splicing*	73	8	10.96	53	72.60	0	0	NaN	0	NaN
RF00005	*tRNA*	*Gene|tRNA*	1114	953	85.55	969	86.98	158	150	94.94	142	89.87
RF00006	*Vault RNA*	*Gene*	9	5	55.56	8	88.89	3	3	100.00	1	33.33
RF00007	*U12 minor spliceosomal RNA*	*Gene|snRNA|splicing*	7	4	57.14	7	100.00	0	0	NaN	0	NaN
RF00008	*Hammerhead ribozyme (type III)*	*Gene|ribozyme*	84	61	72.62	68	80.95	1	1	100.00	1	100.00
RF00009	*Nuclear RNase P*	*Gene|ribozyme*	53	16	30.19	50	94.34	0	0	NaN	0	NaN
RF00010	*Bacterial RNase P class A*	*Gene|ribozyme*	236	77	32.63	203	86.02	0	0	NaN	0	NaN
RF00011	*Bacterial RNase P class B*	*Gene|ribozyme*	30	12	40.00	28	93.33	0	0	NaN	0	NaN
RF00012	*U3 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	21	10	47.62	18	85.71	0	0	NaN	0	NaN
RF00013	*6S / SsrS RNA*	*Gene*	7	1	14.29	6	85.71	2	0	0.00	1	50.00
RF00014	*DsrA RNA*	*Gene|sRNA*	3	0	0.00	2	66.67	0	0	NaN	0	NaN
RF00015	*U4 spliceosomal RNA*	*Gene|snRNA|splicing*	25	21	84.00	25	100.00	1	1	100.00	1	100.00
RF00016	*U14 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	18	17	94.44	16	88.89	2	2	100.00	2	100.00
RF00017	*Eukaryotic type signal recognition particle RNA*	*Gene*	70	3	4.29	61	87.14	0	0	NaN	0	NaN
RF00018	*CsrB/RsmB RNA family*	*Gene|sRNA*	9	9	100.00	8	88.89	0	0	NaN	0	NaN
RF00019	*Y RNA*	*Gene*	15	9	60.00	12	80.00	5	5	100.00	2	40.00
RF00020	*U5 spliceosomal RNA*	*Gene|snRNA|splicing*	32	12	37.50	26	81.25	0	0	NaN	0	NaN
RF00021	*Spot 42 RNA*	*Gene|sRNA*	8	0	0.00	8	100.00	0	0	NaN	0	NaN
RF00022	*GcvB RNA*	*Gene|sRNA*	5	3	60.00	5	100.00	0	0	NaN	0	NaN
RF00023	*tmRNA*	*Gene*	87	53	60.92	79	90.80	0	0	NaN	0	NaN
RF00024	*Vertebrate telomerase RNA*	*Gene*	35	10	28.57	31	88.57	0	0	NaN	0	NaN
RF00025	*Ciliate telomerase RNA*	*Gene*	16	13	81.25	12	75.00	0	0	NaN	0	NaN
RF00026	*U6 spliceosomal RNA*	*Gene|snRNA|splicing*	53	52	98.11	48	90.57	0	0	NaN	0	NaN
RF00028	*Group I catalytic intron*	*Intron*	30	15	50.00	29	96.67	0	0	NaN	0	NaN
RF00029	*Group II catalytic intron*	*Intron*	116	37	31.90	89	76.72	0	0	NaN	0	NaN
RF00030	*RNase MRP*	*Gene|ribozyme*	26	9	34.62	25	96.15	0	0	NaN	0	NaN
RF00031	*Selenocysteine insertion sequence*	*Cis-reg*	64	52	81.25	50	78.13	56	56	100.00	50	89.29
RF00032	*Histone 3' UTR stem-loop*	*Cis-reg*	64	64	100.00	57	89.06	26	26	100.00	26	100.00
RF00033	*MicF RNA*	*Gene|antisense*	9	8	88.89	6	66.67	0	0	NaN	0	NaN
RF00034	*RprA RNA*	*Gene|sRNA*	9	7	77.78	9	100.00	0	0	NaN	0	NaN
RF00035	*OxyS RNA*	*Gene|sRNA*	6	4	66.67	6	100.00	0	0	NaN	0	NaN
RF00036	*HIV Rev response element*	*Cis-reg*	65	0	0.00	39	60.00	0	0	NaN	0	NaN
RF00037	*Iron response element*	*Cis-reg*	39	39	100.00	33	84.62	0	0	NaN	0	NaN
RF00038	*PrfA thermoregulator UTR*	*Cis-reg|thermoregulator*	11	11	100.00	11	100.00	5	5	100.00	5	100.00
RF00039	*DicF RNA*	*Gene|antisense*	5	5	100.00	5	100.00	2	2	100.00	2	100.00
RF00040	*RNase E 5' UTR element*	*Cis-reg*	7	5	71.43	7	100.00	0	0	NaN	0	NaN
RF00041	*Enteroviral 3' UTR element*	*Cis-reg*	60	49	81.67	45	75.00	0	0	NaN	0	NaN
RF00042	*CopA-like RNA*	*Gene|antisense*	17	0	0.00	11	64.71	0	0	NaN	0	NaN
RF00043	*R1162-like plasmid antisense RNA*	*Gene|antisense*	6	6	100.00	5	83.33	0	0	NaN	0	NaN
RF00044	*Bacteriophage pRNA*	*Gene*	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00045	*U17/E1 small nucleolar RNA*	*Gene|snRNA|guide|H/ACA-box*	23	16	69.57	18	78.26	0	0	NaN	0	NaN
RF00046	*Small nucleolar RNA R30/Z108*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	2	33.33	0	0	NaN	0	NaN
RF00048	*Enterovirus cis-acting replication element*	*Cis-reg*	56	31	55.36	35	62.50	56	30	53.57	23	41.07
RF00049	*U36/R47/Z100 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	20	20	100.00	19	95.00	3	3	100.00	2	66.67
RF00050	*FMN riboswitch (RFN element)*	*Cis-reg|riboswitch*	48	41	85.42	45	93.75	0	0	NaN	0	NaN
RF00054	*U25 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	8	8	100.00	7	87.50	2	2	100.00	1	50.00
RF00055	*Small nucleolar RNA Z37*	*Gene|snRNA|guide|C/D-box*	8	8	100.00	5	62.50	0	0	NaN	0	NaN
RF00056	*U71 small nucleolar RNA*	*Gene|snRNA|guide|H/ACA-box*	15	10	66.67	11	73.33	0	0	NaN	0	NaN
RF00057	*RyhB RNA*	*Gene|sRNA*	9	9	100.00	6	66.67	0	0	NaN	0	NaN
RF00058	*HgcF RNA*	*Gene*	4	0	0.00	4	100.00	0	0	NaN	0	NaN
RF00059	*TPP riboswitch (THI element)*	*Cis-reg|riboswitch*	236	223	94.49	201	85.17	4	4	100.00	4	100.00

				miPred		miPred-NBC		3SVM‡			3SVM-NBC	
Accession	Type†	Class	N	TN	%SP	TN	%SP	N	TN	%SP	TN	%SP
RF00060	HgcE RNA	Gene	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00061	Hepatitis C virus IRES	Cis-reg|IRES	786	658	83.72	674	85.75	1	0	0.00	0	0.00
RF00062	HgcC family RNA	Gene	22	7	31.82	22	100.00	0	0	NaN	0	NaN
RF00063	SscA RNA	Gene	5	5	100.00	3	60.00	0	0	NaN	0	NaN
RF00064	HgcG RNA	Gene	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00065	snoR9 / snoR19 family	Gene|snRNA|guide|C/D-box	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00066	U7 small nuclear RNA	Gene|snRNA	28	24	85.71	24	85.71	7	7	100.00	7	100.00
RF00067	U15 small nucleolar RNA	Gene|snRNA|guide|C/D-box	18	16	88.89	15	83.33	2	1	50.00	2	100.00
RF00068	U21 small nucleolar RNA	Gene|snRNA|guide|C/D-box	5	5	100.00	4	80.00	3	3	100.00	1	33.33
RF00069	U24/Z20/U76 small nucleolar RNA	Gene|snRNA|guide|C/D-box	14	14	100.00	10	71.43	3	3	100.00	3	100.00
RF00070	Small nucleolar RNA U29	Gene|snRNA|guide|C/D-box	10	10	100.00	6	60.00	2	2	100.00	2	100.00
RF00071	U73 small nucleolar RNA	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	2	2	100.00	2	100.00
RF00072	U23 small nucleolar RNA	Gene|snRNA|guide|H/ACA-box	6	2	33.33	3	50.00	0	0	NaN	0	NaN
RF00077	SraB RNA	Gene|sRNA	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00078	SraD RNA	Gene|sRNA	5	5	100.00	4	80.00	0	0	NaN	0	NaN
RF00079	SraE/RygA/RygB family RNA	Gene|sRNA	6	5	83.33	4	66.67	0	0	NaN	0	NaN
RF00080	yybP-ykoY element	Cis-reg|riboswitch	74	52	70.27	70	94.59	2	2	100.00	1	50.00
RF00081	SraH RNA	Gene|sRNA	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00082	SraG RNA	Gene|sRNA	5	4	80.00	5	100.00	0	0	NaN	0	NaN
RF00083	SraJ RNA	Gene|sRNA	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00084	CsrC RNA family	Gene|sRNA	5	1	20.00	4	80.00	0	0	NaN	0	NaN
RF00085	U28 small nucleolar RNA	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	2	2	100.00	1	50.00
RF00086	U27/Z191/snR74/Z4 small nucleolar RNA	Gene|snRNA|guide|C/D-box	10	10	100.00	7	70.00	0	0	NaN	0	NaN
RF00087	U26 small nucleolar RNA	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	2	0	0.00	2	100.00
RF00088	U30 small nucleolar RNA	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	2	1	50.00	1	50.00
RF00089	U31 small nucleolar RNA	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	4	4	100.00	2	50.00
RF00090	U19 small nucleolar RNA	Gene|snRNA|guide|H/ACA-box	3	0	0.00	2	66.67	0	0	NaN	0	NaN
RF00091	Small nucleolar RNA E2/ACA6/M2/MBI-136	Gene|snRNA|guide|H/ACA-box	10	2	20.00	8	80.00	0	0	NaN	0	NaN
RF00092	E3 small nucleolar RNA	Gene|snRNA|guide|H/ACA-box	9	4	44.44	9	100.00	0	0	NaN	0	NaN
RF00093	U18 small nucleolar RNA	Gene|snRNA|guide|C/D-box	16	16	100.00	14	87.50	10	9	90.00	8	80.00
RF00094	Hepatitis delta virus ribozyme	Gene|ribozyme	15	14	93.33	15	100.00	0	0	NaN	0	NaN
RF00095	Pyrococcus C/D box small nucleolar RNA	Gene|snRNA|guide|C/D-box	38	38	100.00	37	97.37	18	18	100.00	17	94.44
RF00096	U8 small nucleolar RNA	Gene|snRNA|guide|C/D-box	5	2	40.00	3	60.00	0	0	NaN	0	NaN
RF00097	Plant small nucleolar RNA R71	Gene|snRNA|guide|C/D-box	21	18	85.71	21	100.00	0	0	NaN	0	NaN
RF00098	Snake H/ACA box small nucleolar RNA	Gene|snRNA|guide|H/ACA-box	22	22	100.00	20	90.91	0	0	NaN	0	NaN
RF00099	U22 small nucleolar RNA	Gene|snRNA|guide|C/D-box	3	2	66.67	2	66.67	0	0	NaN	0	NaN
RF00100	7SK RNA	Gene	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00101	SraC/RyeA RNA	Gene|sRNA	7	3	42.86	7	100.00	0	0	NaN	0	NaN
RF00102	VA RNA	Gene	23	0	0.00	22	95.65	0	0	NaN	0	NaN
RF00105	HBII-52 small nucleolar RNA	Gene|snRNA|guide|C/D-box	23	23	100.00	14	60.87	1	1	100.00	1	100.00
RF00106	RNAI	Gene|antisense	10	0	0.00	6	60.00	0	0	NaN	0	NaN
RF00107	FinP	Gene	6	0	0.00	6	100.00	0	0	NaN	0	NaN
RF00108	HBII-85 small nucleolar RNA	Gene|snRNA|guide|C/D-box	7	7	100.00	7	100.00	0	0	NaN	0	NaN
RF00109	Vimentin 3' UTR protein-binding region	Cis-reg	12	12	100.00	11	91.67	2	2	100.00	2	100.00
RF00110	RybB RNA	Gene|sRNA	4	2	50.00	4	100.00	2	2	100.00	0	0.00
RF00111	RyeB RNA	Gene|sRNA	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00112	RyeE RNA	Gene|sRNA	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00113	QUAD RNA	Gene|sRNA	15	6	40.00	15	100.00	0	0	NaN	0	NaN
RF00114	Ribosomal S15 leader	Cis-reg	11	11	100.00	7	63.64	0	0	NaN	0	NaN
RF00115	IS061 RNA	Gene|sRNA	5	5	100.00	2	40.00	0	0	NaN	0	NaN
RF00116	C0465 RNA	Gene|sRNA	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00117	C0719 RNA	Gene|sRNA	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00118	rydB RNA	Gene|sRNA	5	5	100.00	4	80.00	5	5	100.00	0	0.00
RF00119	C0299 RNA	Gene|sRNA	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00120	C0343 RNA	Gene|sRNA	4	4	100.00	1	25.00	0	0	NaN	0	NaN
RF00121	MicC RNA	Gene|sRNA	4	3	75.00	3	75.00	0	0	NaN	0	NaN
RF00122	GadY	Gene|sRNA	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00124	IS102 RNA	Gene|sRNA	8	1	12.50	8	100.00	0	0	NaN	0	NaN
RF00125	IS128 RNA	Gene|sRNA	3	1	33.33	3	100.00	0	0	NaN	0	NaN

				miPred		*miPred-NBC*			*3SVM*[‡]		*3SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
RF00126	*ryfA RNA*	*Gene|sRNA*	6	0	0.00	6	100.00	0	0	NaN	0	NaN
RF00127	*t44 RNA*	*Gene|sRNA*	9	4	44.44	9	100.00	0	0	NaN	0	NaN
RF00128	*tke1 RNA*	*Gene|sRNA*	7	2	28.57	5	71.43	0	0	NaN	0	NaN
RF00132	*Small nucleolar RNA R24*	*Gene|snRNA|guide|C/D-box*	12	7	58.33	9	75.00	4	2	50.00	1	25.00
RF00133	*Small nucleolar RNA Z195*	*Gene|snRNA|guide|C/D-box*	8	8	100.00	8	100.00	0	0	NaN	0	NaN
RF00134	*Small nucleolar RNA Z196*	*Gene|snRNA|guide|C/D-box*	7	7	100.00	3	42.86	0	0	NaN	0	NaN
RF00135	*Small nucleolar RNA Z223*	*Gene|snRNA|guide|C/D-box*	5	4	80.00	4	80.00	2	2	100.00	0	0.00
RF00136	*U81 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00137	*U83/U84 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	7	7	100.00	7	100.00	0	0	NaN	0	NaN
RF00138	*U16 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	5	3	60.00	5	100.00	3	2	66.67	2	66.67
RF00139	*U72 small nucleolar RNA*	*Gene|snRNA|guide|H/ACA-box*	7	4	57.14	7	100.00	0	0	NaN	0	NaN
RF00140	*Alpha operon ribosome binding site*	*Cis-reg*	9	3	33.33	8	88.89	0	0	NaN	0	NaN
RF00141	*Small nucleolar RNA R39/R59*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00142	*Small nucleolar RNA Z118/Z121/Z120*	*Gene|snRNA|guide|C/D-box*	7	4	57.14	4	57.14	2	2	100.00	1	50.00
RF00145	*Small nucleolar RNA Z105*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	5	100.00	1	1	100.00	1	100.00
RF00146	*Small nucleolar RNA U33*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00147	*Small nucleolar RNA U34*	*Gene|snRNA|guide|C/D-box*	9	9	100.00	8	88.89	4	3	75.00	3	75.00
RF00149	*Small nucleolar RNA Z103*	*Gene|snRNA|guide|C/D-box*	9	9	100.00	7	77.78	0	0	NaN	0	NaN
RF00150	*Small nucleolar RNA U42*	*Gene|snRNA|guide|C/D-box*	7	7	100.00	6	85.71	0	0	NaN	0	NaN
RF00151	*Small nucleolar RNA U58*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00152	*Small nucleolar RNA U79/Z22*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	6	100.00	3	3	100.00	0	0.00
RF00153	*Small nucleolar RNA U62*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	1	1	100.00	1	100.00
RF00154	*Small nucleolar RNA U63*	*Gene|snRNA|guide|C/D-box*	2	2	100.00	1	50.00	1	1	100.00	1	100.00
RF00155	*Small nucleolar RNA U66*	*Gene|snRNA|guide|H/ACA-box*	3	1	33.33	2	66.67	0	0	NaN	0	NaN
RF00156	*Small nucleolar RNA U70*	*Gene|snRNA|guide|H/ACA-box*	14	11	78.57	13	92.86	0	0	NaN	0	NaN
RF00157	*Small nucleolar RNA U39/U55*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00158	*Small nucleolar RNA U82/Z25*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	1	1	100.00	1	100.00
RF00159	*Small nucleolar RNA Z168/Z174*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00160	*Small nucleolar RNA Z159/U59*	*Gene|snRNA|guide|C/D-box*	10	10	100.00	9	90.00	0	0	NaN	0	NaN
RF00161	*Nanos 3' UTR translation control element*	*Cis-reg*	2	1	50.00	1	50.00	0	0	NaN	0	NaN
RF00162	*SAM riboswitch (S box leader)*	*Cis-reg|riboswitch*	71	53	74.65	60	84.51	1	1	100.00	0	0.00
RF00163	*Hammerhead ribozyme (type I)*	*Gene|ribozyme*	74	72	97.30	67	90.54	39	36	92.31	26	66.67
RF00164	*Coronavirus 3' stem-loop II-like motif (s2m)*	*Cis-reg*	37	37	100.00	37	100.00	33	33	100.00	32	96.97
RF00165	*Coronavirus 3' UTR pseudoknot*	*Cis-reg*	14	14	100.00	13	92.86	0	0	NaN	0	NaN
RF00166	*PrrB/RsmZ RNA family*	*Gene|sRNA*	6	6	100.00	5	83.33	0	0	NaN	0	NaN
RF00167	*Purine riboswitch*	*Cis-reg|riboswitch*	37	36	97.30	23	62.16	0	0	NaN	0	NaN
RF00168	*Lysine riboswitch*	*Cis-reg|riboswitch*	60	37	61.67	54	90.00	0	0	NaN	0	NaN
RF00169	*Bacterial signal recognition particle RNA*	*Gene*	70	43	61.43	52	74.29	58	55	94.83	32	55.17
RF00170	*Retron msr RNA*	*Gene*	8	7	87.50	6	75.00	2	2	100.00	1	50.00
RF00171	*Tombusvirus 5' UTR*	*Cis-reg*	9	9	100.00	9	100.00	1	0	0.00	1	100.00
RF00172	*ctgf/hcs24 CAESAR*	*Cis-reg*	9	9	100.00	8	88.89	0	0	NaN	0	NaN
RF00173	*Hairpin ribozyme*	*Gene|ribozyme*	3	3	100.00	3	100.00	1	1	100.00	0	0.00
RF00174	*Cobalamin riboswitch*	*Cis-reg|riboswitch*	170	135	79.41	157	92.35	0	0	NaN	0	NaN
RF00175	*Retroviral Psi packaging element*	*Cis-reg*	168	168	100.00	156	92.86	0	0	NaN	0	NaN
RF00176	*Tombusvirus 3' UTR region IV*	*Cis-reg*	18	18	100.00	17	94.44	0	0	NaN	0	NaN
RF00177	*Small subunit ribosomal RNA, 5' domain*	*Gene|rRNA*	358	175	48.88	325	90.78	0	0	NaN	0	NaN
RF00179	*GAIT element*	*Cis-reg*	8	8	100.00	8	100.00	4	4	100.00	4	100.00
RF00180	*Renin stability regulatory element (REN-SRE)*	*Cis-reg*	13	13	100.00	13	100.00	0	0	NaN	0	NaN
RF00181	*C/D box small nucleolar RNA 14q(I)/14q(II)*	*Gene|snRNA|guide|C/D-box*	59	57	96.61	50	84.75	36	36	100.00	33	91.67
RF00182	*Coronavirus packaging signal*	*Cis-reg*	15	10	66.67	15	100.00	15	5	33.33	5	33.33
RF00183	*G-CSF factor stem-loop destabilising element (SLDE)*	*Cis-reg*	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00184	*Potato virus X cis-acting regulatory element*	*Cis-reg*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00185	*Flavivirus 3' UTR pseudoknot*	*Cis-reg*	14	3	21.43	11	78.57	0	0	NaN	0	NaN
RF00186	*Small nucleolar RNA U101*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	3	3	100.00	3	100.00
RF00187	*Small nucleolar RNA U102*	*Gene|snRNA|guide|C/D-box*	2	2	100.00	2	100.00	2	2	100.00	2	100.00
RF00188	*Small nucleolar RNA U103*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00189	*Small nucleolar RNA U95*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	4	80.00	5	5	100.00	5	100.00
RF00190	*U98 small nucleolar RNA*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	2	66.67	0	0	NaN	0	NaN

				miPred		miPred-NBC		3SVM‡			3SVM-NBC	
Accession	Type†	Class	N	TN	%SP	TN	%SP	N	TN	%SP	TN	%SP
RF00191	U99 small nucleolar RNA	Gene|snRNA|guide|H/ACA-box	3	1	33.33	3	100.00	0	0	NaN	0	NaN
RF00192	Bovine leukaemia virus RNA packaging signal	Cis-reg	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00193	Citrus tristeza virus replication signal	Cis-reg	9	9	100.00	9	100.00	0	0	NaN	0	NaN
RF00194	Rubella virus 3' cis-acting element	Cis-reg	9	9	100.00	9	100.00	0	0	NaN	0	NaN
RF00195	RsmY RNA family	Gene|sRNA	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00196	Alfalfa mosaic virus RNA 1 5' UTR stem-loop	Cis-reg	4	2	50.00	0	0.00	2	2	100.00	0	0.00
RF00197	rbcL 5' UTR RNA stabilising element	Cis-reg	3	2	66.67	3	100.00	0	0	NaN	0	NaN
RF00198	SL1 RNA	Gene	28	0	0.00	24	85.71	0	0	NaN	0	NaN
RF00199	SL2 RNA	Gene	32	10	31.25	24	75.00	0	0	NaN	0	NaN
RF00200	Small nucleolar RNA Z199	Gene|snRNA|guide|C/D-box	8	8	100.00	7	87.50	6	6	100.00	4	66.67
RF00201	Small nucleolar RNA Z278	Gene|snRNA|guide|C/D-box	7	5	71.43	7	100.00	7	7	100.00	5	71.43
RF00202	Small nucleolar RNA R66	Gene|snRNA|guide|C/D-box	6	6	100.00	6	100.00	1	1	100.00	1	100.00
RF00203	Small nucleolar RNA R160	Gene|snRNA|guide|C/D-box	9	9	100.00	9	100.00	4	4	100.00	4	100.00
RF00204	Small nucleolar RNA R12	Gene|snRNA|guide|C/D-box	9	9	100.00	8	88.89	2	2	100.00	2	100.00
RF00205	Small nucleolar RNA R41	Gene|snRNA|guide|C/D-box	7	7	100.00	6	85.71	7	7	100.00	1	14.29
RF00206	Small nucleolar RNA U54	Gene|snRNA|guide|C/D-box	13	13	100.00	11	84.62	1	1	100.00	1	100.00
RF00207	K10 transport/localisation element (TLS)	Cis-reg	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00208	Small nucleolar RNA R72	Gene|snRNA|guide|C/D-box	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00209	Pestivirus IRES	Cis-reg|IRES	25	1	4.00	20	80.00	0	0	NaN	0	NaN
RF00210	Aphthovirus IRES	Cis-reg|IRES	32	2	6.25	29	90.63	0	0	NaN	0	NaN
RF00211	Small nucleolar RNA U35	Gene|snRNA|guide|C/D-box	8	8	100.00	5	62.50	1	1	100.00	0	0.00
RF00212	U38 small nucleolar RNA	Gene|snRNA|guide|C/D-box	7	7	100.00	6	85.71	3	3	100.00	2	66.67
RF00213	Small nucleolar RNA R38	Gene|snRNA|guide|C/D-box	12	10	83.33	11	91.67	6	6	100.00	3	50.00
RF00214	Retrovirus direct repeat 1 (dr1)	Cis-reg	25	24	96.00	21	84.00	1	0	0.00	1	100.00
RF00215	Tombus virus defective interfering (DI) RNA region 3	Cis-reg	48	48	100.00	34	70.83	6	6	100.00	6	100.00
RF00216	c-myc IRES	Cis-reg|IRES	23	23	100.00	21	91.30	0	0	NaN	0	NaN
RF00217	Small nucleolar RNA U20	Gene|snRNA|guide|C/D-box	4	4	100.00	3	75.00	4	4	100.00	3	75.00
RF00218	Small nucleolar RNA U40	Gene|snRNA|guide|C/D-box	9	9	100.00	9	100.00	8	8	100.00	4	50.00
RF00219	Small nucleolar RNA U32	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00220	Human rhinovirus internal cis-acting regulatory element	Cis-reg	12	12	100.00	12	100.00	10	10	100.00	10	100.00
RF00221	Small nucleolar RNA U43	Gene|snRNA|guide|C/D-box	6	5	83.33	3	50.00	3	2	66.67	3	100.00
RF00222	Bag-1 IRES	Cis-reg|IRES	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00223	bip IRES	Cis-reg|IRES	4	4	100.00	4	100.00	2	2	100.00	2	100.00
RF00224	FGF-2 IRES	Cis-reg|IRES	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00225	Tobamovirus IRES	Cis-reg|IRES	7	7	100.00	7	100.00	0	0	NaN	0	NaN
RF00226	n-myc IRES	Cis-reg|IRES	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00227	FIE3 (ftz instability element 3') element	Cis-reg	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00228	Hepatitis A virus IRES	Cis-reg|IRES	23	9	39.13	22	95.65	0	0	NaN	0	NaN
RF00229	Picornavirus IRES	Cis-reg|IRES	195	96	49.23	180	92.31	0	0	NaN	0	NaN
RF00230	T-box leader	Cis-reg	66	28	42.42	60	90.91	0	0	NaN	0	NaN
RF00231	U93 small nucleolar RNA	Gene|snRNA|guide|H/ACA-box	3	1	33.33	3	100.00	0	0	NaN	0	NaN
RF00232	Spi-1 (PU.1) 5' UTR regulatory element	Cis-reg	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00233	Tymovirus/Pomovirus tRNA-like 3' UTR element	Cis-reg	27	27	100.00	23	85.19	0	0	NaN	0	NaN
RF00234	glmS glucosamine-6-phosphate activated ribozyme	Cis-reg|riboswitch	14	10	71.43	11	78.57	0	0	NaN	0	NaN
RF00235	Plasmid RNAIII	Gene	7	0	0.00	7	100.00	0	0	NaN	0	NaN
RF00236	ctRNA	Gene|antisense	17	0	0.00	16	94.12	0	0	NaN	0	NaN
RF00238	ctRNA	Gene|antisense	48	5	10.42	44	91.67	0	0	NaN	0	NaN
RF00240	RNA-OUT	Gene	7	0	0.00	3	42.86	7	2	28.57	3	42.86
RF00242	ctRNA	Gene|antisense	15	6	40.00	10	66.67	0	0	NaN	0	NaN
RF00243	traJ 5' UTR	Cis-reg	6	2	33.33	6	100.00	0	0	NaN	0	NaN
RF00250	Trans-activation response element (TAR)	Cis-reg	416	26	6.25	370	88.94	412	49	11.89	221	53.64
RF00252	Alfalfa mosaic virus coat protein binding (CPB) RNA	Cis-reg	18	2	11.11	18	100.00	0	0	NaN	0	NaN
RF00259	Interferon gamma 5' UTR regulatory element	Cis-reg	5	5	100.00	2	40.00	0	0	NaN	0	NaN
RF00260	Hepatitis C virus (HCV) cis-acting replication element	Cis-reg	52	52	100.00	52	100.00	52	52	100.00	46	88.46
RF00261	L-myc IRES	Cis-reg|IRES	2	2	100.00	2	100.00	0	0	NaN	0	NaN
RF00262	sar RNA	Gene	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00263	U68 small nucleolar RNA	Gene|snRNA|guide|H/ACA-box	4	3	75.00	3	75.00	0	0	NaN	0	NaN
RF00264	Small nucleolar RNA U64	Gene|snRNA|guide|H/ACA-box	3	1	33.33	3	100.00	0	0	NaN	0	NaN
RF00265	Small nucleolar RNA U69	Gene|snRNA|guide|H/ACA-box	3	1	33.33	2	66.67	0	0	NaN	0	NaN

				miPred		*miPred-NBC*			*3SVM*[‡]		*3SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
RF00266	*Small nucleolar RNA Z17*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	2	50.00	0	0	NaN	0	NaN
RF00267	*Small nucleolar RNA R64*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	0	0.00	0	0	NaN	0	NaN
RF00268	*Small nucleolar RNA snoZ7/snoR77*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00270	*U61 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	2	2	100.00	2	100.00
RF00271	*U60 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00272	*U67 small nucleolar RNA*	*Gene|snRNA|guide|H/ACA-box*	10	10	100.00	8	80.00	0	0	NaN	0	NaN
RF00273	*U59 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	2	2	100.00	2	100.00
RF00274	*U57 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	1	1	100.00	1	100.00
RF00275	*U56 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	7	7	100.00	7	100.00	1	1	100.00	0	0.00
RF00276	*U52 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	3	75.00	3	3	100.00	2	66.67
RF00277	*U49 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	1	1	100.00	0	0.00
RF00278	*U50 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	6	100.00	1	1	100.00	1	100.00
RF00279	*U45 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	11	11	100.00	10	90.91	7	7	100.00	7	100.00
RF00280	*U51 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	1	0	0.00	1	100.00
RF00281	*U47 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00282	*U48 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	1	1	100.00	1	100.00
RF00283	*U91 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00284	*Z18 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	2	2	100.00	2	100.00
RF00285	*Z6 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	1	1	100.00	1	100.00
RF00286	*U92 small nucleolar RNA*	*Gene|snRNA|guide|H/ACA-box*	3	1	33.33	2	66.67	0	0	NaN	0	NaN
RF00287	*U44 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	2	66.67	1	1	100.00	1	100.00
RF00288	*Z30 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	3	75.00	4	4	100.00	1	25.00
RF00289	*Z12 small nucleolar RNA*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	2	2	100.00	2	100.00
RF00290	*Bamboo mosaic potexvirus (BaMV) CE*	*Cis-reg*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00291	*Small nucleolar RNA snoR639/H1*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00292	*Small nucleolar RNA TBR5*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00293	*Small nucleolar RNA snoM1*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00294	*Small nucleolar RNA TBR17*	*Gene|snRNA|guide|C/D-box*	4	3	75.00	4	100.00	0	0	NaN	0	NaN
RF00295	*Small nucleolar RNA TBR7*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	5	83.33	1	1	100.00	1	100.00
RF00296	*Small nucleolar RNA R16*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	5	83.33	2	2	100.00	2	100.00
RF00297	*Small nucleolar RNA Z177*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00299	*Small nucleolar RNA Z200*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00300	*Small nucleolar RNA Z221*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	2	66.67	2	2	100.00	1	50.00
RF00301	*Small nucleolar RNA Z256*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	1	33.33	0	0	NaN	0	NaN
RF00302	*Small nucleolar RNA U65*	*Gene|snRNA|guide|H/ACA-box*	4	0	0.00	4	100.00	0	0	NaN	0	NaN
RF00303	*Small nucleolar RNA snoR86*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	1	33.33	0	0	NaN	0	NaN
RF00304	*Small nucleolar RNA Z279*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00305	*Small nucleolar RNA Z248*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00306	*Small nucleolar RNA Z178*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00307	*Small nucleolar RNA snoR98*	*Gene|snRNA|guide|H/ACA-box*	5	5	100.00	5	100.00	1	1	100.00	1	100.00
RF00308	*Small nucleolar RNA Z268*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	2	2	100.00	1	50.00
RF00309	*Small nucleolar RNA snR60/Z15/Z230/Z193/J17*	*Gene|snRNA|guide|C/D-box*	24	23	95.83	21	87.50	5	4	80.00	1	20.00
RF00310	*Small nucleolar RNA Z165*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	1	33.33	3	3	100.00	0	0.00
RF00311	*Small nucleolar RNA Z188*	*Gene|snRNA|guide|C/D-box*	4	1	25.00	4	100.00	3	0	0.00	3	100.00
RF00312	*Small nucleolar RNA Z206*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00313	*Small nucleolar RNA Z173*	*Gene|snRNA|guide|C/D-box*	3	1	33.33	3	100.00	0	0	NaN	0	NaN
RF00314	*Small nucleolar RNA Z182*	*Gene|snRNA|guide|C/D-box*	7	7	100.00	7	100.00	4	4	100.00	0	0.00
RF00315	*Small nucleolar RNA J33*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	3	60.00	2	2	100.00	0	0.00
RF00316	*Small nucleolar RNA R43*	*Gene|snRNA|guide|C/D-box*	16	16	100.00	16	100.00	6	6	100.00	5	83.33
RF00317	*Small nucleolar RNA Z163*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00318	*Small nucleolar RNA Z175*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00319	*Small nucleolar RNA MBI-1*	*Gene|snRNA|guide|H/ACA-box*	4	2	50.00	4	100.00	0	0	NaN	0	NaN
RF00320	*Small nucleolar RNA Z185*	*Gene|snRNA|guide|C/D-box*	3	2	66.67	2	66.67	1	1	100.00	0	0.00
RF00321	*Small nucleolar RNA Z247*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00322	*Small nucleolar RNA MBI-161*	*Gene|snRNA|guide|H/ACA-box*	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00323	*Small nucleolar RNA R79*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00324	*Small nucleolar RNA MBII-202*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	4	80.00	0	0	NaN	0	NaN
RF00325	*Small nucleolar RNA U53*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	3	3	100.00	3	100.00

				miPred		miPred-NBC		3SVM‡			3SVM-NBC	
Accession	Type†	Class	N	TN	%SP	TN	%SP	N	TN	%SP	TN	%SP
RF00326	Small nucleolar RNA Z155	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00327	Small nucleolar RNA Z194	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00328	Small nucleolar RNA Z161/Z228	Gene|snRNA|guide|C/D-box	7	7	100.00	5	71.43	2	2	100.00	2	100.00
RF00329	Small nucleolar RNA Z162	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00330	Small nucleolar RNA Z43	Gene|snRNA|guide|C/D-box	4	4	100.00	3	75.00	2	2	100.00	1	50.00
RF00331	Small nucleolar RNA Z169	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	2	2	100.00	2	100.00
RF00332	Small nucleolar RNA Z266	Gene|snRNA|guide|C/D-box	4	4	100.00	2	50.00	2	2	100.00	0	0.00
RF00333	Small nucleolar RNA Z157/R69/R10	Gene|snRNA|guide|C/D-box	10	8	80.00	5	50.00	2	2	100.00	0	0.00
RF00334	Small nucleolar RNA MBI-28	Gene|snRNA|guide|H/ACA-box	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00335	Small nucleolar RNA Z13/snr52	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	3	3	100.00	3	100.00
RF00336	Small nucleolar RNA J26	Gene|snRNA|guide|C/D-box	5	5	100.00	3	60.00	0	0	NaN	0	NaN
RF00337	Small nucleolar RNA Z112	Gene|snRNA|guide|C/D-box	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00338	Small nucleolar RNA snR53	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	1	0	0.00	1	100.00
RF00339	Small nucleolar RNA snoR60	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00340	Small nucleolar RNA snoMBI-87	Gene|snRNA|guide|H/ACA-box	6	0	0.00	4	66.67	0	0	NaN	0	NaN
RF00341	Small nucleolar RNA Z39	Gene|snRNA|guide|C/D-box	5	5	100.00	5	100.00	4	4	100.00	4	100.00
RF00342	Small nucleolar RNA Z40	Gene|snRNA|guide|C/D-box	5	5	100.00	3	60.00	5	5	100.00	5	100.00
RF00343	Small nucleolar RNA Z122	Gene|snRNA|guide|C/D-box	3	3	100.00	0	0.00	1	1	100.00	1	100.00
RF00344	Small nucleolar RNA Z267	Gene|snRNA|guide|C/D-box	5	5	100.00	3	60.00	2	2	100.00	2	100.00
RF00345	Small nucleolar RNA snoR1	Gene|snRNA|guide|C/D-box	7	7	100.00	6	85.71	3	3	100.00	1	33.33
RF00346	Small nucleolar RNA snoZ1	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00347	Small nucleolar RNA Z50	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	1	1	100.00	0	0.00
RF00348	Small nucleolar RNA snoR9	Gene|snRNA|guide|C/D-box	7	7	100.00	6	85.71	3	3	100.00	3	100.00
RF00349	Small nucleolar RNA R11/Z151	Gene|snRNA|guide|C/D-box	5	5	100.00	3	60.00	0	0	NaN	0	NaN
RF00350	Small nucleolar RNA Z152/R70/R12/	Gene|snRNA|guide|C/D-box	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00351	Small nucleolar RNA R20	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	2	2	100.00	0	0.00
RF00352	Small nucleolar RNA R21	Gene|snRNA|guide|C/D-box	4	4	100.00	2	50.00	0	0	NaN	0	NaN
RF00353	Small nucleolar RNA snoR31/Z110/Z27	Gene|snRNA|guide|C/D-box	8	5	62.50	2	25.00	0	0	NaN	0	NaN
RF00355	Small nucleolar RNA snoR28	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00356	Small nucleolar RNA R32/R81/Z41	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	4	4	100.00	1	25.00
RF00357	Small nucleolar RNA R44/J54	Gene|snRNA|guide|C/D-box	4	4	100.00	3	75.00	3	3	100.00	3	100.00
RF00358	Small nucleolar RNA Z101	Gene|snRNA|guide|C/D-box	3	3	100.00	2	66.67	2	2	100.00	0	0.00
RF00359	Small nucleolar RNA Z102/R77	Gene|snRNA|guide|C/D-box	4	4	100.00	4	100.00	2	2	100.00	2	100.00
RF00360	Small nucleolar RNA Z107/R87	Gene|snRNA|guide|C/D-box	6	5	83.33	6	100.00	0	0	NaN	0	NaN
RF00361	Small nucleolar RNA Z119	Gene|snRNA|guide|C/D-box	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00362	Pospiviroid RY motif stem loop	Cis-reg	16	14	87.50	15	93.75	11	11	100.00	3	27.27
RF00368	sroB RNA	Gene|sRNA	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00369	sroC RNA	Gene|sRNA	5	0	0.00	4	80.00	0	0	NaN	0	NaN
RF00370	sroD RNA	Gene|sRNA	3	2	66.67	3	100.00	0	0	NaN	0	NaN
RF00371	sroE RNA	Gene|sRNA	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00372	sroH RNA	Gene|sRNA	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00373	Archaeal RNase P	Gene|ribozyme	40	16	40.00	33	82.50	0	0	NaN	0	NaN
RF00374	Gammaretrovirus core encapsidation signal	Cis-reg	23	11	47.83	23	100.00	0	0	NaN	0	NaN
RF00375	HIV primer binding site (PBS)	Cis-reg	373	265	71.05	334	89.54	0	0	NaN	0	NaN
RF00376	HIV gag stem loop 3 (GSL3)	Cis-reg	1374	1371	99.78	1200	87.34	9	9	100.00	4	44.44
RF00377	Small nucleolar RNA U6-53/MBII-28	Gene|snRNA|guide|C/D-box	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00378	Qrr RNA	Gene|sRNA	14	7	50.00	9	64.29	0	0	NaN	0	NaN
RF00379	ydaO/yuaA element	Cis-reg|riboswitch	35	35	100.00	32	91.43	0	0	NaN	0	NaN
RF00380	ykoK element	Cis-reg|riboswitch	39	25	64.10	32	82.05	0	0	NaN	0	NaN
RF00381	Antizyme RNA frameshifting stimulation element	Cis-reg|frameshift	13	12	92.31	12	92.31	10	10	100.00	7	70.00
RF00382	DnaX ribosomal frameshifting element	Cis-reg|frameshift	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00383	Insertion sequence IS1222 ribosomal frameshifting element	Cis-reg|frameshift	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00384	Poxvirus AX element late mRNA CE	Cis-reg	7	7	100.00	7	100.00	0	0	NaN	0	NaN
RF00385	Infectious bronchitis virus D-RNA	Cis-reg	10	6	60.00	10	100.00	10	8	80.00	6	60.00
RF00386	Enterovirus 5' cloverleaf cis-acting replication element	Cis-reg	60	5	8.33	52	86.67	0	0	NaN	0	NaN
RF00387	FGF-1 IRES	Cis-reg|IRES	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00388	Qa RNA	Gene|antisense	5	2	40.00	3	60.00	0	0	NaN	0	NaN
RF00389	Bamboo mosaic virus satellite RNA CE	Cis-reg	42	42	100.00	41	97.62	0	0	NaN	0	NaN
RF00390	UPSK RNA	Cis-reg	4	4	100.00	4	100.00	0	0	NaN	0	NaN

				miPred		*miPred-NBC*			*3SVM*[‡]		*3SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
RF00391	*RtT RNA*	*Cis-reg*	19	16	84.21	18	94.74	0	0	NaN	0	NaN
RF00392	*Small nucleolar RNA ACA5*	*Gene|snRNA|guide|H/ACA-box*	6	6	100.00	4	66.67	0	0	NaN	0	NaN
RF00393	*Small nucleolar RNA ACA8*	*Gene|snRNA|guide|H/ACA-box*	5	4	80.00	4	80.00	0	0	NaN	0	NaN
RF00394	*Small nucleolar RNA ACA4*	*Gene|snRNA|guide|H/ACA-box*	7	4	57.14	7	100.00	0	0	NaN	0	NaN
RF00395	*Small nucleolar RNA ACA10*	*Gene|snRNA|guide|H/ACA-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00396	*Small nucleolar RNA ACA13*	*Gene|snRNA|guide|H/ACA-box*	3	0	0.00	1	33.33	0	0	NaN	0	NaN
RF00397	*Small nucleolar RNA ACA14*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00398	*Small nucleolar RNA ACA15*	*Gene|snRNA|guide|H/ACA-box*	4	2	50.00	4	100.00	0	0	NaN	0	NaN
RF00399	*Small nucleolar RNA ACA24*	*Gene|snRNA|guide|H/ACA-box*	5	5	100.00	4	80.00	0	0	NaN	0	NaN
RF00400	*Small nucleolar RNA ACA28*	*Gene|snRNA|guide|H/ACA-box*	3	2	66.67	3	100.00	0	0	NaN	0	NaN
RF00401	*Small nucleolar RNA ACA20*	*Gene|snRNA|guide|H/ACA-box*	17	4	23.53	14	82.35	0	0	NaN	0	NaN
RF00402	*Small nucleolar RNA ACA25*	*Gene|snRNA|guide|H/ACA-box*	9	7	77.78	8	88.89	0	0	NaN	0	NaN
RF00403	*Small nucleolar RNA ACA41*	*Gene|snRNA|guide|H/ACA-box*	6	1	16.67	6	100.00	0	0	NaN	0	NaN
RF00404	*Small nucleolar RNA ACA46*	*Gene|snRNA|guide|H/ACA-box*	3	1	33.33	2	66.67	0	0	NaN	0	NaN
RF00405	*Small nucleolar RNA ACA44*	*Gene|snRNA|guide|H/ACA-box*	6	6	100.00	6	100.00	1	1	100.00	1	100.00
RF00406	*Small nucleolar RNA ACA42*	*Gene|snRNA|guide|H/ACA-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00407	*Small nucleolar RNA ACA50*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00408	*Small nucleolar RNA ACA1*	*Gene|snRNA|guide|H/ACA-box*	6	5	83.33	5	83.33	0	0	NaN	0	NaN
RF00409	*Small nucleolar RNA ACA7*	*Gene|snRNA|guide|H/ACA-box*	8	8	100.00	6	75.00	1	1	100.00	1	100.00
RF00410	*Small nucleolar RNA ACA2/ACA34*	*Gene|snRNA|guide|H/ACA-box*	18	5	27.78	16	88.89	0	0	NaN	0	NaN
RF00411	*Small nucleolar RNA ACA9*	*Gene|snRNA|guide|H/ACA-box*	6	3	50.00	5	83.33	0	0	NaN	0	NaN
RF00412	*Small nucleolar RNA ACA21*	*Gene|snRNA|guide|H/ACA-box*	5	1	20.00	3	60.00	0	0	NaN	0	NaN
RF00413	*Small nucleolar RNA ACA19*	*Gene|snRNA|guide|H/ACA-box*	4	1	25.00	3	75.00	0	0	NaN	0	NaN
RF00414	*Small nucleolar RNA ACA22*	*Gene|snRNA|guide|H/ACA-box*	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00415	*Small nucleolar RNA ACA30/ACA37/MBI-26*	*Gene|snRNA|guide|H/ACA-box*	6	6	100.00	6	100.00	0	0	NaN	0	NaN
RF00416	*Small nucleolar RNA ACA43*	*Gene|snRNA|guide|H/ACA-box*	7	7	100.00	6	85.71	0	0	NaN	0	NaN
RF00417	*Small nucleolar RNA ACA56*	*Gene|snRNA|guide|H/ACA-box*	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00418	*Small nucleolar RNA ACA52*	*Gene|snRNA|guide|H/ACA-box*	4	0	0.00	3	75.00	0	0	NaN	0	NaN
RF00419	*Small nucleolar RNA ACA52*	*Gene|snRNA|guide|H/ACA-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00420	*Small nucleolar RNA ACA61*	*Gene|snRNA|guide|H/ACA-box*	4	3	75.00	3	75.00	0	0	NaN	0	NaN
RF00421	*Small nucleolar RNA ACA32*	*Gene|snRNA|guide|H/ACA-box*	9	6	66.67	6	66.67	0	0	NaN	0	NaN
RF00422	*Small nucleolar RNA ACA12*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00423	*Small nucleolar RNA ACA26*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00424	*Small nucleolar RNA ACA47*	*Gene|snRNA|guide|H/ACA-box*	6	2	33.33	4	66.67	0	0	NaN	0	NaN
RF00425	*Small nucleolar RNA ACA18*	*Gene|snRNA|guide|H/ACA-box*	6	3	50.00	3	50.00	0	0	NaN	0	NaN
RF00426	*Small nucleolar RNA ACA45*	*Gene|snRNA|guide|H/ACA-box*	3	1	33.33	3	100.00	0	0	NaN	0	NaN
RF00427	*Small nucleolar RNA ACA11*	*Gene|snRNA|guide|H/ACA-box*	3	1	33.33	3	100.00	0	0	NaN	0	NaN
RF00428	*Small nucleolar RNA ACA38*	*Gene|snRNA|guide|H/ACA-box*	5	4	80.00	5	100.00	0	0	NaN	0	NaN
RF00429	*Small nucleolar RNA ACA29*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00430	*Small nucleolar RNA ACA54*	*Gene|snRNA|guide|H/ACA-box*	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00431	*Small nucleolar RNA ACA55*	*Gene|snRNA|guide|H/ACA-box*	3	0	0.00	3	100.00	0	0	NaN	0	NaN
RF00432	*Small nucleolar RNA ACA51*	*Gene|snRNA|guide|H/ACA-box*	9	8	88.89	9	100.00	0	0	NaN	0	NaN
RF00433	*Hsp90 CE*	*Cis-reg|thermoregulator*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00434	*Luteovirus cap-independent translation element (BTE)*	*Cis-reg*	17	17	100.00	13	76.47	0	0	NaN	0	NaN
RF00435	*Repression of heat shock gene expression (ROSE) element*	*Cis-reg|thermoregulator*	3	2	66.67	2	66.67	0	0	NaN	0	NaN
RF00436	*UnaL2 line 3' element*	*Cis-reg*	144	141	97.92	113	78.47	50	49	98.00	13	26.00
RF00437	*Hairy RNA localisation element (HLE)*	*Cis-reg*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00438	*Small nucleolar RNA ACA33*	*Gene|snRNA|guide|H/ACA-box*	5	5	100.00	4	80.00	0	0	NaN	0	NaN
RF00439	*Small nucleolar RNA U87*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	3	75.00	0	0	NaN	0	NaN
RF00440	*Small nucleolar RNA U37*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	3	3	100.00	3	100.00
RF00441	*Small nucleolar RNA Z242*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00442	*ykkC-yxkD element*	*Cis-reg|riboswitch*	16	15	93.75	14	87.50	0	0	NaN	0	NaN
RF00443	*Small nucleolar RNA ACA27*	*Gene|snRNA|guide|H/ACA-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00444	*PrrF RNA*	*Gene|sRNA*	7	2	28.57	7	100.00	0	0	NaN	0	NaN
RF00447	*Voltage-gated potassium-channel Kv1.4 IRES*	*Cis-reg|IRES*	6	5	83.33	6	100.00	0	0	NaN	0	NaN
RF00448	*Epstein-Barr virus nuclear antigen (EBNA) IRES*	*Cis-reg|IRES*	8	8	100.00	8	100.00	0	0	NaN	0	NaN
RF00449	*HIF-1 alpha IRES*	*Cis-reg|IRES*	7	7	100.00	7	100.00	0	0	NaN	0	NaN
RF00450	*Small nucleolar RNA R105/R108*	*Gene|snRNA|guide|C/D-box*	4	3	75.00	4	100.00	0	0	NaN	0	NaN

				miPred		*miPred-NBC*			*3SVM*‡		*3SVM-NBC*	
Accession	*Type*†	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
RF00453	*Cardiovirus cis-acting replication element*	*Cis-reg*	12	11	91.67	9	75.00	2	2	100.00	2	100.00
RF00454	*p27 CE*	*Cis-reg*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00457	*Mnt IRES*	*Cis-reg|IRES*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00458	*Cripavirus IRES*	*Cis-reg|IRES*	7	6	85.71	6	85.71	0	0	NaN	0	NaN
RF00459	*Mason-Pfizer monkey virus packaging signal*	*Cis-reg*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
RF00460	*U1A polyadenylation inhibition element (PIE)*	*Cis-reg*	6	6	100.00	6	100.00	3	3	100.00	3	100.00
RF00461	*Vascular endothelial growth factor (VEGF) IRES A*	*Cis-reg|IRES*	7	7	100.00	7	100.00	0	0	NaN	0	NaN
RF00462	*APC IRES*	*Cis-reg|IRES*	6	6	100.00	2	33.33	0	0	NaN	0	NaN
RF00463	*Apolipoprotein B (apoB) 5' UTR CE*	*Cis-reg*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00465	*Japanese encephalitis virus (JEV) hairpin structure*	*Cis-reg*	20	19	95.00	19	95.00	12	12	100.00	5	41.67
RF00466	*Agrobacterium tumefaciens ROSE element*	*Cis-reg|thermoregulator*	3	1	33.33	3	100.00	0	0	NaN	0	NaN
RF00467	*Rous sarcoma virus (RSV) primer binding site (PBS)*	*Cis-reg*	23	1	4.35	21	91.30	22	13	59.09	18	81.82
RF00468	*Hepatitis C stem-loop VII*	*Cis-reg*	63	9	14.29	32	50.79	63	45	71.43	63	100.00
RF00469	*Hepatitis C stem-loop IV*	*Cis-reg*	109	2	1.83	109	100.00	109	109	100.00	61	55.96
RF00470	*Togavirus 5' plus strand CE*	*Cis-reg*	32	5	15.63	29	90.63	0	0	NaN	0	NaN
RF00471	*Small nucleolar RNA snR48*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	5	83.33	1	1	100.00	0	0.00
RF00472	*Small nucleolar RNA snR55/Z10*	*Gene|snRNA|guide|C/D-box*	7	7	100.00	4	57.14	0	0	NaN	0	NaN
RF00473	*Small nucleolar RNA snR54*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00474	*Small nucleolar RNA snR57*	*Gene|snRNA|guide|C/D-box*	6	6	100.00	5	83.33	2	2	100.00	0	0.00
RF00475	*Small nucleolar RNA snR69*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00476	*Small nucleolar RNA snR61/Z11*	*Gene|snRNA|guide|C/D-box*	9	9	100.00	8	88.89	0	0	NaN	0	NaN
RF00477	*Small nucleolar RNA snR66*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	5	100.00	0	0	NaN	0	NaN
RF00478	*Small nucleolar RNA U88*	*Gene|snRNA|guide|C/D-box*	4	0	0.00	3	75.00	0	0	NaN	0	NaN
RF00479	*Small nucleolar RNA snR71*	*Gene|snRNA|guide|C/D-box*	5	5	100.00	3	60.00	0	0	NaN	0	NaN
RF00480	*HIV Ribosomal frameshift signal*	*Cis-reg|frameshift*	768	152	19.79	704	91.67	765	719	93.99	107	13.99
RF00481	*Hepatitis C virus 3'X element*	*Cis-reg*	22	0	0.00	13	59.09	0	0	NaN	0	NaN
RF00482	*Small nucleolar RNA F1/F2/snoR5a*	*Gene|snRNA|guide|H/ACA-box*	8	5	62.50	6	75.00	0	0	NaN	0	NaN
RF00483	*Insulin-like growth factor II IRES*	*Cis-reg|IRES*	8	8	100.00	7	87.50	0	0	NaN	0	NaN
RF00484	*Connexin-32 IRES*	*Cis-reg|IRES*	6	6	100.00	5	83.33	0	0	NaN	0	NaN
RF00485	*Potassium channel RNA editing signal*	*Cis-reg*	85	76	89.41	69	81.18	13	10	76.92	7	53.85
RF00487	*Connexin-43 IRES*	*Cis-reg|IRES*	13	13	100.00	12	92.31	0	0	NaN	0	NaN
RF00488	*Yeast U1 spliceosomal RNA*	*Gene|snRNA|splicing*	6	0	0.00	5	83.33	0	0	NaN	0	NaN
RF00489	*ctRNA*	*Gene|antisense*	15	6	40.00	14	93.33	10	8	80.00	7	70.00
RF00490	*S-element*	*Cis-reg*	13	13	100.00	9	69.23	3	3	100.00	3	100.00
RF00491	*Simian virus 40 late polyadenylation signal (SVLPA)*	*Cis-reg*	3	3	100.00	2	66.67	0	0	NaN	0	NaN
RF00492	*Small nucleolar RNA U12-22*	*Gene|snRNA|guide|C/D-box*	7	7	100.00	6	85.71	3	3	100.00	3	100.00
RF00493	*Small nucleolar RNA U2-30*	*Gene|snRNA|guide|C/D-box*	3	3	100.00	3	100.00	0	0	NaN	0	NaN
RF00494	*Small nucleolar RNA U2-19*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	1	1	100.00	1	100.00
RF00495	*Heat shock protein 70 (Hsp70) IRES*	*Cis-reg|IRES*	13	13	100.00	13	100.00	0	0	NaN	0	NaN
RF00496	*Coronavirus SL-III cis-acting replication element*	*Cis-reg*	5	5	100.00	5	100.00	3	3	100.00	1	33.33
RF00497	*Dengue virus 3'-SL cis-acting replication element*	*Cis-reg*	23	5	21.74	21	91.30	0	0	NaN	0	NaN
RF00498	*Equine arteritis virus leader TRS hairpin (LTH)*	*Cis-reg*	4	4	100.00	4	100.00	4	4	100.00	4	100.00
RF00499	*Human parechovirus 1 (HPeV1) cis regulatory element*	*Cis-reg*	5	2	40.00	5	100.00	0	0	NaN	0	NaN
RF00500	*Turnip crinkle virus (TCV) repressor of minus strand synthesis H5*	*Cis-reg*	3	2	66.67	3	100.00	3	3	100.00	2	66.67
RF00501	*Rotavirus cis-acting replication element*	*Cis-reg*	14	14	100.00	8	57.14	4	4	100.00	1	25.00
RF00502	*Turnip crinkle virus (TCV) core promoter hairpin (Pr)*	*Cis-reg*	4	4	100.00	4	100.00	4	4	100.00	2	50.00
RF00503	*RNAIII*	*Gene*	12	2	16.67	12	100.00	0	0	NaN	0	NaN
RF00504	*gcvT element*	*Cis-reg|riboswitch*	117	111	94.87	102	87.18	3	3	100.00	2	66.67
RF00505	*RydC RNA*	*Gene|sRNA*	3	3	100.00	3	100.00	2	2	100.00	2	100.00
RF00506	*Threonine operon leader*	*Cis-reg*	27	1	3.70	25	92.59	0	0	NaN	0	NaN
RF00507	*Coronavirus frameshifting stimulation element*	*Cis-reg|frameshift*	18	12	66.67	15	83.33	0	0	NaN	0	NaN
RF00509	*Small nucleolar RNA snR64*	*Gene|snRNA|guide|C/D-box*	4	4	100.00	4	100.00	0	0	NaN	0	NaN
–	mRNAs	–	31	27	87.10	27	87.10	0	0	NaN	0	NaN
Total ncRNA samples (exclude mRNAs)		–	12387	8507		10771		2404	1884		1199	

†, cis-regulatory element (CE); internal ribosome entry site (IRES). *N* (non *pre-miRs*), *TN* (non *pre-miRs* missed), and *%SP* (Specificity). ‡, *3SVM* model was trained on 200 human *pre-miRs* and 400 pseudo hairpins randomly selected using the latest libSVM 2.82 (the "-b 1" option was enabled) and the optimal hyperparameter pair (C, γ).

Table S5. The mean specificity of *miPred*, *miPred-NBC*, *3SVM*, and *3SVM-NBC* evaluated on the non *pre-miR* dataset IE-NC (12,387 functional ncRNAs) categorized by classes of ncRNAs.

		miPred	*miPred-NBC*		*3SVM*†	*3SVM-NBC*
Classes of ncRNAs	*No. of types*	*%SP*	*%SP*	*No. of excluded types*	*%SP*	*%SP*
Cis-reg	77	74.91 ± 4.03	87.99 ± 2.03	46	83.36 ± 5.60	69.96 ± 5.61
Cis-reg|frameshift	5	75.75 ± 15.27	86.80 ± 5.68	3	96.99 ± 3.01	42.00 ± 28.01
Cis-reg|IRES	24	85.47 ± 6.02	91.02 ± 3.06	22	50.00 ± 50.00	50.00 ± 50.00
Cis-reg|riboswitch	12	82.28 ± 3.96	85.77 ± 2.56	8	100.00 ± 0.00	54.17 ± 20.83
Cis-reg|thermoregulator	4	75.00 ± 15.96	91.67 ± 8.33	3	100.00 ± 0.00	100.00 ± 0.00
Gene	24	34.73 ± 7.71	86.65 ± 3.03	18	70.57 ± 18.19	45.23 ± 3.26
Gene|antisense	10	41.93 ± 13.01	78.05 ± 5.03	8	90.00 ± 10.00	85.00 ± 15.00
Gene|ribozyme	9	60.08 ± 10.10	91.54 ± 2.36	6	97.44 ± 2.56	55.56 ± 29.40
Gene|rRNA	3	70.66 ± 12.94	90.74 ± 1.70	1	100.00 ± 0.00	75.00 ± 25.00
Gene|snRNA	1	85.71 ± 0.00	85.71 ± 0.00	0	100.00 ± 0.00	100.00 ± 0.00
Gene|snRNA|guide|C/D-box	165	94.61 ± 1.28	84.59 ± 1.58	72	92.78 ± 2.32	68.60 ± 4.06
Gene|snRNA|guide|H/ACA-box	71	60.97 ± 4.33	84.97 ± 2.04	68	100.00 ± 0.00	100.00 ± 0.00
Gene|snRNA|splicing	7	51.16 ± 13.89	87.30 ± 3.83	6	100.00 ± 0.00	100.00 ± 0.00
Gene|sRNA	42	65.71 ± 5.90	87.53 ± 2.81	39	100.00 ± 0.00	33.33 ± 33.33
Gene|tRNA	1	85.55 ± 0.00	86.98 ± 0.00	0	94.94 ± 0.00	89.87 ± 0.00
Intron	2	40.95 ± 9.05	86.70 ± 9.98	2	NaN	NaN

†, *3SVM* model was trained on 200 human *pre-miRs* and 400 pseudo hairpins randomly selected using the latest libSVM 2.82 (the "-b 1" option was enabled) and the optimal hyperparameter pair (C, γ). *%SP* (Specificity). Values are expressed as mean ± standard error.

Table S6. F1 and F2 scores for features of *miPred* and *3SVM*, sorted by descending F1 scores.

		miPred			*3SVM*†			
Rank	*Features*	*F1 score*	*F2 score*	*ΔF = F1 − F2*	*Features*	*F1 score*	*F2 score*	*ΔF = F1 − F2*
01	$MFEI_1$	1.28	1.52	$-2.42E^{-01}$	*A(((*	$8.20E^{-01}$	$6.97E^{-01}$	$1.22E^{-01}$
02	*zG*	1.27	1.48	$-2.15E^{-01}$	*U(((*	$7.58E^{-01}$	$6.12E^{-01}$	$1.46E^{-01}$
03	*dP*	1.03	1.18	$-1.49E^{-01}$	*G...*	$4.57E^{-01}$	$2.05E^{-01}$	$2.52E^{-01}$
04	*zP*	$9.67E^{-01}$	1.03	$-6.33E^{-02}$	*A...*	$4.42E^{-01}$	$1.94E^{-01}$	$2.47E^{-01}$
05	*zQ*	$8.33E^{-01}$	$7.29E^{-01}$	$1.04E^{-01}$	*C...*	$4.31E^{-01}$	$1.84E^{-01}$	$2.47E^{-01}$
06	*dG*	$8.23E^{-01}$	$7.50E^{-01}$	$7.31E^{-02}$	*G.((*	$3.81E^{-01}$	$1.62E^{-01}$	$2.20E^{-01}$
07	*dQ*	$7.99E^{-01}$	$6.67E^{-01}$	$1.32E^{-01}$	*A(..*	$3.50E^{-01}$	$1.31E^{-01}$	$2.19E^{-01}$
08	*zD*	$7.92E^{-01}$	$6.70E^{-01}$	$1.23E^{-01}$	*A..(*	$3.28E^{-01}$	$1.17E^{-01}$	$2.11E^{-01}$
09	*dD*	$7.46E^{-01}$	$5.91E^{-01}$	$1.55E^{-01}$	*C((.*	$3.19E^{-01}$	$1.12E^{-01}$	$2.07E^{-01}$
10	$MFEI_2$	$4.41E^{-01}$	$1.53E^{-01}$	$2.88E^{-01}$	*G(..*	$3.07E^{-01}$	$9.75E^{-02}$	$2.10E^{-01}$
11	*%UA*	$3.87E^{-01}$	$1.56E^{-01}$	$2.31E^{-01}$	*U...*	$3.05E^{-01}$	$9.74E^{-02}$	$2.08E^{-01}$
12	*%G+C*	$3.06E^{-01}$	$1.04E^{-01}$	$2.02E^{-01}$	*C..(*	$2.97E^{-01}$	$9.54E^{-02}$	$2.02E^{-01}$
13	*zF*	$2.88E^{-01}$	$7.13E^{-02}$	$2.16E^{-01}$	*G(((*	$2.84E^{-01}$	$8.95E^{-02}$	$1.94E^{-01}$
14	*%UU*	$2.83E^{-01}$	$8.91E^{-02}$	$1.94E^{-01}$	*C(..*	$2.70E^{-01}$	$7.93E^{-02}$	$1.91E^{-01}$
15	*%GU*	$2.64E^{-01}$	$7.71E^{-02}$	$1.87E^{-01}$	*G((.*	$2.63E^{-01}$	$7.62E^{-02}$	$1.87E^{-01}$
16	*%GC*	$2.44E^{-01}$	$6.57E^{-02}$	$1.79E^{-01}$	*G..(*	$2.48E^{-01}$	$6.69E^{-02}$	$1.81E^{-01}$
17	*dF*	$2.42E^{-01}$	$5.16E^{-02}$	$1.90E^{-01}$	*U..(*	$2.19E^{-01}$	$5.20E^{-02}$	$1.67E^{-01}$
18	*%CC*	$2.04E^{-01}$	$4.59E^{-02}$	$1.58E^{-01}$	*C.((*	$1.89E^{-01}$	$3.92E^{-02}$	$1.50E^{-01}$
19	*%AA*	$1.83E^{-01}$	$3.73E^{-02}$	$1.46E^{-01}$	*C(((*	$1.87E^{-01}$	$3.88E^{-02}$	$1.48E^{-01}$
20	*%GG*	$1.82E^{-01}$	$3.68E^{-02}$	$1.45E^{-01}$	*G.(.*	$1.82E^{-01}$	$3.52E^{-02}$	$1.47E^{-01}$
21	*%CA*	$1.77E^{-01}$	$3.48E^{-02}$	$1.42E^{-01}$	*U.(.*	$1.71E^{-01}$	$2.88E^{-02}$	$1.42E^{-01}$
22	*%CG*	$1.73E^{-01}$	$3.30E^{-02}$	$1.40E^{-01}$	*U(..*	$1.56E^{-01}$	$2.69E^{-02}$	$1.30E^{-01}$
23	*%GA*	$1.41E^{-01}$	$2.13E^{-02}$	$1.19E^{-01}$	*U(.(*	$1.37E^{-01}$	$2.08E^{-02}$	$1.16E^{-01}$
24	*%AU*	$1.25E^{-01}$	$1.69E^{-02}$	$1.08E^{-01}$	*A.(.*	$1.22E^{-01}$	$1.52E^{-02}$	$1.07E^{-01}$
25	*%AG*	$1.08E^{-01}$	$1.28E^{-02}$	$9.54E^{-02}$	*C.(.*	$1.10E^{-01}$	$1.32E^{-02}$	$9.68E^{-02}$
26	*%UG*	$6.31E^{-02}$	$4.42E^{-03}$	$5.87E^{-02}$	*G(.(*	$1.02E^{-01}$	$1.13E^{-02}$	$9.05E^{-02}$
27	*%AC*	$3.71E^{-02}$	$1.53E^{-03}$	$3.55E^{-02}$	*C(.(*	$6.68E^{-02}$	$4.95E^{-03}$	$6.19E^{-02}$
28	*%CU*	$3.21E^{-02}$	$1.13E^{-03}$	$3.09E^{-02}$	*A(.(*	$6.06E^{-02}$	$4.06E^{-03}$	$5.65E^{-02}$
29	*%UC*	$2.18E^{-02}$	$5.21E^{-04}$	$2.13E^{-02}$	*A.((*	$5.90E^{-02}$	$3.87E^{-03}$	$5.52E^{-02}$
30	–	–	–	–	*A((.*	$3.21E^{-02}$	$1.14E^{-03}$	$3.10E^{-02}$
31	–	–	–	–	*U.((*	$3.28E^{-03}$	$1.20E^{-05}$	$3.26E^{-03}$
32	–	–	–	–	*U((.*	$6.80E^{-05}$	$0.00E^{+00}$	$6.80E^{-05}$
		0.429 ± 0.0711	0.332 ± 0.0872	–		0.252 ± 0.0336	0.103 ± 0.0277	–

†, *3SVM* model was trained on 200 human *pre-miRs* and 400 pseudo hairpins randomly selected using the latest libSVM 2.82 (the "-b 1" option was enabled) and the optimal hyperparameter pair (C, γ).

Table S7. Effects of feature selection on *miPred*'s accuracy.

Classifiers	*Human pre-miRs (TR-H and TE-H)*	*Non-human pre-miRs (IE-NH)*	*ncRNAs (IE-NC)*	*mRNAs (IE-M)*
miPred	93.60	95.64	68.68	87.10
$miPred_{3}$	94.12	95.69	68.31	87.10
$miPred_{3/5}$	92.67	95.36	71.20	100.00
$miPred_{3/10}$	93.40	95.64	69.82	83.87
$miPred_{3/15}$	93.40	95.79	60.93	80.65
$miPred_{3/20}$	92.67	94.68	72.18	100.00
$miPred_{3/21}$	92.67	95.29	72.01	100.00
$miPred_{3/22}$	92.57	95.15	71.26	100.00
$miPred_{3/23}$	92.67	95.22	70.15	100.00
$miPred_{3/24}$	92.98	95.39	64.56	100.00
$miPred_{3/25}$	91.64	93.52	63.16	96.77
$miPred_{I}$	77.30	76.35	67.53	90.32
$miPred_{II}$	93.81	95.83	61.38	54.84
$miPred_{III}$	93.60	95.69	66.13	70.97

$miPred_{3}$ contains a subset of 26 features from *miPred* that excludes *dQ*, *dD*, and *zD*. Derived from $miPred_{3}$, the remaining nine variants denoted as $miPred_{3/5}$, $miPred_{3/10}$, …, $miPred_{3/24}$, and $miPred_{3/25}$ only include the top ranking 21, 16, 11, 6, 5, 4, 3, 2, and 1 feature(s), respectively. $miPred_{I}$ (17 features: 16 dinucleotides frequencies and *%G+C*), $miPred_{II}$ (12 features; $MFEI_{1}$, $MFEI_{2}$, *dP*, *dG*, *dQ*, *dD*, *dF*, *zP*, *zG*, *zQ*, *zD*, and *zF*), and $miPred_{III}$ (9 features; a subset of $miPred_{II}$ that excludes *dQ*, *dD*, and *zD*).

Table S8. Putative viral-encoded *pre-miRs* having maximum length (≤ 95-nts), minimum size of terminal loop (≥ 3-nts), minimum free energy of folding (≤ -25 kcal/mol), and *miPred* scores ≥0.815 (except for [†]*ebv-mir-BHRF1-1* and [§]*mghv-mir-M1-8*). They are categorized according to *E. Barr virus*, *K. sarcoma-associated herpesvirus*, *M. γ-herpesvirus 68 strain WUMS*, and *H. cytomegalovirus strain AD169*; sorted in descending *miPred* scores.

S	*SP*	*L*	*E. Barr virus* (EBV; AJ507799.2); 5' → 3'
+	147303	92	CCAGAGGAGUGUCCCGGGGCCACCUCUUUGGUUCUGUACAUAU**uuu**GUUAUUGUACAUAACCAUGGAGUUGGCUGUGGUGCACUCCAUCUGG (*ebv-mir-BART10*)
+	7681	94	AUAUAGAUUAGGAUAGCAUAUGCUAUCCAGAUAUUUGGGUAGUA**uaugc**UACCCAGAUAUAAAUUAGGAUAGCAUAUACUACCCUAAUCUCUAU (*ebv-mir-p1*)
+	140016	92	UGACCUUGUUGGUACUUUAAGGUUGGUCCAAUCCAUAGGCUUUUU**uugu**GAAAACCCGGGGAUCGGACUAGCCUUAGAGUAACUCAAGGCCA (*ebv-mir-BART6*)
+	7693	95	AUAGCAUAUGCUAUCCAGAUAUUUGGGUAGUAUAUGCUACCCAGAU**auaa**AUUAGGAUAGCAUAUACUACCCUAAUCUCUAUUAGGAUAGCAUAU (*ebv-mir-p2*)
+	7932	94	GCAUAUGCUACCCAGAUAUAGAUUAGGAUAGCCUAUGCUACCCAG**aua**UAGAUUAGGAUAGCAUAUGCUAUCCAGAUAUUUGGGUAGUAUAUGC (*ebv-mir-p3*)
+	9007	91	UAGGACCCUUUUACUAACCCUAAUUCGAUAGCAUAUGCUUCCC**guu**GGGUAACAUAUGCUAUUGAAUUAGGGUUAGUCUGGAUAGUAUAUA (*ebv-mir-p4*)
+	7708	95	CAGAUAUUUGGGUAGUAUAUGCUACCCAGAUAUAAAUUAGGAUAG**cauaua**CUACCCUAAUCUCUAUUAGGAUAGCAUAUGCUACCCGGAUACAG (*ebv-mir-p5*)
+	146422	94	GGAUCCAGUGUCCUGAUCCUGGACCUUGACUAUGAAACAAUUCU**aaa**AAAAUGCAUCAUAGUCCAGUGUCCAGGGACAGUGCACUCGGAAGUCU (*ebv-mir-BART7*)
+	9031	95	UCGAUAGCAUAUGCUUCCCGUUGGGUAACAUAUGCUAUUGAAUUAGG**guuag**UCUGGAUAGUAUAUACUACUACCCGGGAAGCAUAUGCUACCCG (*ebv-mir-p6*)
+	152730	92	CUGGUGGACUUCCAGACUAUUUUCUGCAUUCGCCCUUGCGUGU**ccauu**GUUGCAAGGAGCGAUUUGGAGAAAAUAAACUGUGAGUUUCACAG (*ebv-mir-BART2*)
+	146753	95	GGUCGAUGGGUUCACUGAUUACGGUUUCCUAGAUUGUACAGAU**gaacuag**AACUGUCACAAUCUAUGGGGUCGUAGACAGUGUGCUUACCAGACU (*ebv-mir-BART8*)
+	42832	95	AAUGACCCGGCCCCCACUUUUAAAUUCUGUUGCAGCAGAUAGCUGAU**acccA**AUGUUAUCUUUUGCGGCAGAAAUUGAAAGUGCUGGCCAUAUCU (*ebv-mir-BHRF1-2*)
+	139064	93	AGGCAUUGUUAACCUUUGGUGGAACCUAGUGUUAGUGUUGUGCUGU**aaau**AAGUGUCCAGCGCACCACUAGUCACCAGGUGUCACCGGAGGCU (*ebv-mir-BART3*)
+	139898	95	AACAGGAUGUGGCACCCUAAGAGGACGCAGGCAUACAAGGUU**auuaccc**AGUCCUUGUAUGCCUGGUGUCCCCUUAGUGGGACGCAGGCCUAGGU (*ebv-mir-p7*)
+	12549	93	GGCAGAGGGUCGGCCUAGGCCCGGGGAAGUGGAGGGGGAUC**gcccg**GGUCUCUGUUGGCAGAGUCCGGGCGAUCCUCUGAGACCCUCCGGGCC (*ebv-mir-p8*)
+	139206	95	GGGGCUCUGUAACAUUUGGUGGGACCUGAUGCUGCUGGUGUGCUGU**aaau**AAGUGCCUAGCACAUCACGUAGGCACCAGGUGUCACCAGGGCUAC (*ebv-mir-BART4*)
+	42950	95	UAUACGCCUGUGGUGUUCUAACGGGAAGUGUGUAAGCACACACGUA**auu**UGCAAGCGGUGCUUCACGCUCUUCGUUAAAAUAACACAAGGACAAG (*ebv-mir-BHRF1-3*)
+	7754	94	AUAUACUACCCUAAUCUCUAUUAGGAUAGCAUAUGCUACCCGGAU**acag**AUUAGGAUAGCAUAUACUACCCAGAUAUAGAUUAGGAUAGCAUAU (*ebv-mir-p9*)
+	156856	94	UUUUGCGCCUGGAAGUUGUACUCCCGGAAGAUGCCCUCCAGGUCAA**agacg**UUGGAGGCACGCUGUUCGUCCCGUGAGUACAGCUCCAGGGAGG (*ebv-mir-p10*)
+	140356	94	CUGGAGACCUGCUAUGUGGCUAGACGUAUGGCCUACCCAAGACGU**ugg**GGGUCUCGGGUAGGCCAUGAUUCUUCCAGGCAUAGGUUACAACCAG (*ebv-mir-p11*)
+	146941	94	UGUGGCAGCUGUUGUUUGUACUGGACCCUGAAUUGGAAACAGUAACU**ugg**AUUCUGUAACACUUCAUGGGUCCCGUAGUGACAACUAUGCUGAA (*ebv-mir-BART9*)
+	139778	95	GCUUUCAGGUGUGGAAUUUAGAUAGAGUGGGUGUGUGCUCUUGUUU**aauu**ACACCAAGAUCACCACCCUCUAUCCAUAUCCCACAAUUGAUAAAC (*ebv-mir-p12*)
+	165115	94	UUCUUGGGUGAGCGAGUCACCCUGACCUCCUACUGGAGGAGG**gugag**CCUCGGUCCAGAGAUUGAGGUCAGCUGGUUUAAACUGGGCCCAGGAG (*ebv-mir-p13*)
+	139333	95	UAACAAACCCGUGGGGGUCUUAGUGGAAGUGACGUGCUGUGAAU**acag**GUCCAUAGCACCGCUAUCCACUAUGUCUCGCCCGGGCUAUAUGUCG (*ebv-mir-BART1*)
+	139658	93	GAUGCUCUGUGGCACCUCAAGGUGAAUAUAGCUGCCCAUCGACGUAUC**gcug**GAAACCGGUGGGCCGCUGUUCACCUAAAGUGACGCAAGGUC (*ebv-mir-BART5*)
+	140396	87	GACGUUGGGGGUCUCGGGUAGGCCAUGAUUCUUCCAGGCAUAGGU**uac**AACCAGUCACUGCUAUCAAGCCUACUCAGUUCCCAACGC (*ebv-mir-p14*)
+	7666	94	GCAUAUGCUACCCAGAUAUAGAUUAGGAUAGCAUAUGCUAUCCAG**auau**UUGGGUAGUAUAUGCUACCCAGAUAUAAAUUAGGAUAGCAUAUAC (*ebv-mir-p15*)
+	165097	95	GGCCAGGCUGUCACCGCUUUCUUGGGUGAGCGAGUCACCCUAGCCUCC**uacu**GGAGGAGGGUGAGCCUCGGUCCAGAGAUUGAGGUCAGCUGGUU (*ebv-mir-p16*)
+	48951	93	GACCGUGGCUCCCGCCUCUUGGAUGCCAUCAUCCCCUGCUUGG**gac**CCGACCGCACUUGCAUGCGGCCGGUGGUCCUGCGGGGGUGACGGUC (*ebv-mir-p17*)
+	103504	95	UGGCAGAGCUUUCACCGGUGGAACUCGUGACAGAUGUCUACGCCACC**cua**GGCAUCGUGGAGAUCAUCGACGAGCUCUACCGGAGCAGUCGCCUG (*ebv-mir-p18*)
+	68250	94	CUGAGUGUGGGGCCAUACGAGGCCUUCACUGGCCCUGUGGCC**aaggcuca**GGACGUGGGGGCCGUUGAGGCCCACGUUGUCUGCUCGGUAGCAG (*ebv-mir-p19*)
+	74472	94	UUGUGGCACAAACAAAACAGGCGGAAGCCCUCGUCAGGCCG**cgagagga**UGGCAUCGAGGAUGGCCUCCGCAAUGUCAGUGUUUGAGGCCACAA (*ebv-mir-p20*)
+	153549	95	UGUAGGCUGAGAGCUUGCGGCUGAGCUCCGUUGAAAAGCAGAGCUCCCC**cau**GGGGACCCUGCCUUCACGGAGGUCUGUGUAGGCCUGGUUUAGG (*ebv-mir-p21*)
+	145656	87	GAGUGGGGGAUGCUAGCCAAUUUAGCUUCCCCUCCCCUU**aac**AGGGGGUCUCGCGGGGUGCCAAUUGUCGCCUGCCUUCCCCCGCUU (*ebv-mir-p22*)
+	132971	95	CCGUGCUGGGCAGUCAGGGCCUGGAAGUCUUGGCGGCGUUGGUAUU**uaa**AAACCAGCGAUCCCUGAGAACGCUCCAGGUAGAGUUUCCAGCCCUG (*ebv-mir-p23*)
+	41458	94	AAGGACGGCUCCUUAUUAACCUGAUCAGCCCCGGAGUUGCCUGU**uuc**AUCACUAACCCCGGGCCUGAAGAGGUUGACAAGAAGGGUCAAGGUUU ([†]*ebv-mir-BHRF1-1*)

S	*SP*	*L*	*K. sarcoma-associated herpesvirus* (KSHV; U75698.1); 5' → 3'
−	119293	93	UCCAGUAGGUAUACCCAGCUGGGUCUACCCAGCUGCGUAAACCC**cgcu**GCGUAAACACAGCUGGGUAUACGCAGCUGCGUAAACCCGGCUGGG ([‡]*kshv-mir-K12-9*)
−	119273	94	UGGGUCUACCCAGCUGCGUAAACCCCGCUGCGUAAACACAGCUGGG**uaua**CGCAGCUGCGUAAACCCGGCUGGGUAAAUCCAGCUGUAAUUCUA ([Δ]*kshv-mir-K12-9*)
−	120744	94	GCGGGUUUAGAAAGACUUGUCCAGCAGCACCUAAUCCAUCGGC**ggucgg**GCUGAUGGUUUUCGGGCUGUUGAGCGAGUCUUUUUAUCUAGUCGC (*kshv-mir-K12-6*)
−	121535	89	GCCUGUGAUGGGCUAUCACAUUCUGAGGACGGCAGCGACGUG**uguc**UAACGUCAACGUCGCGGUCACAGAAUGUGACACCCCUCCAGGU (*kshv-mir-K12-3*)
−	23628	94	GUUUAAUUAUAGAAUUGCAGCUGGGUAUACCCAGCUGGGUUCACCC**accu**GGGUAUACCUACUGGAUUACCCAGCUGGGUAUACCUACUGGAAU (*kshv-mir-p1*)
−	119320	93	GCGCAGCUGAGUCAUCGCAGCCCCUAUUCCAGUAGGUAUACCC**agcu**GGGUCUACCCAGCUGCGUAAACCCCGCUGCGUAAACACAGCUGGGU (*kshv-mir-p2*)
−	121400	95	GAACCGGGCAGUAUAACUAGCUAAACCGCAGUACUCUAGGGCA**uucauu**UGUUACAUAGAAUACUGAGGCCUAGCUGAUUAUACUACCUCCGUCC (*kshv-mir-K12-4*)
−	81073	86	UACCCAGUUUGUCAUGACACCGACAGAAGCUGGUGCUGGCGA**cgucu**UCGCCGCGCCACUCGUCCGGUGGACAGGCUGAUUUGAAA (*kshv-mir-p3*)
−	133793	94	AUUAUGCAGGCUGUUAUGAUAUCCCUGGGGGCAGACCUGCUACCG**cugg**CGGUGCAGGCUUCAACCGGGGACAAUUAUAACGUGGCCAGGUACU (*kshv-mir-p4*)
−	79458	93	AGGACGCUGACGUUGGGGCCCCCGUAAGGACGUCGGCGAUCGU**ucg**GCGCUGUCGCCACUCGUACAAAAAAUAACCCUUACUGUCAGCGCCU (*kshv-mir-p5*)
−	9635	95	GCCGUGAUCUCGUUGGCCACAAAGUGGAAGCUGUCCUCGUGGUAGU**cugg**AUGGAGCGCGGGAAGGUUUUCCACAGUGCCAGCGGACACACGGC (*kshv-mir-p6*)
−	77286	95	GUAUAUCUUUGUUUUUUCAAGUUUGUGGACGAGGUGGUCCAUGCAUA**gac**UGGCAUGUGAUUACUCGCACAAGCGCUGACGAAAGCUAUGGUUUU (*kshv-mir-p7*)
−	121856	95	GCAGGGUGCGGUGCUGCCCAGGACGGCCGGAUGCGGGCGAUUACAG**gaaa**CUGGGUGUAAGCUGUACAUAAAUCCCCGGCAGCACCUGUUUCCUGC (*kshv-mir-p8*)
−	108693	94	UGGAGUGUGGAUGGUGAUAUGGUCUCCUGGGCCUGGCCGGCCACCG**ugu**CUAUGGUGGCCAACAGGAGGCCGGCCUGCUCCGCUUCCUGUACCA (*kshv-mir-p9*)
−	120342	95	GCGCAUAUUGGCGUUGAGCGCCACCGGACGGGGAUUUAUGCUGUA**ucu**UACUACCAUGAUCCCAUGUUGCUGGCGCUCACGGCCCGUGUGCCAGC (*kshv-mir-K12-7*)

S	*SP*	*L*	*M. γ-herpesvirus 68 strain WUMS* (MGHV68; U97553.2); 5' → 3'
+	1320	59	AACCACCUCCCACAAUUUCAGAGUCU**uagcc**AGAUUAUCUGAAACUGUGUGAGGUGGUU (*mghv-mir-p1*)
+	636	67	ACGAAGUAGCGAACCUCUGCUCACUGCCCGGG**ccc**UCCGGGAGGUGAGCAGGAGUUGCGCUUUUCUU (*mghv-mir-M1-3*)
+	739	93	CACGCUGCCAAUCUCACCCUGACAGCUGUCAGGGGUUACAUGAG**agaac**UUCAUGUAACCCCUGACAGCUGUCAACCUAAUCCUGACCGUGAG (*mghv-mir-p2*)
+	104268	94	CAGCUAACUGGUGUUGAGAGUACAUGUUUGCUUUGGAUACACUU**gug**AAGUUUAUUCAAAGUGUAGGGAUGUGUGCUACUAAACAUAACAGCUG (*mghv-mir-p3*)
+	548	91	CCCGAGCCCUGGUUGAGAGGGGGAGUGUGUGGUCUGUAGAGAGAC**auga**GUUGAUCGGCAGACCCCCUCUCCCCCUCUUUCCCUCUUUACG (*mghv-mir-M1-2*)
+	107005	94	UCUUUAGCAGACAGGUUAGAGCACUGUUGUGUGAUGUGAGAGGAGU**aagu**GUGUCUCCACCAUCGCAUAACAGUUGAUAGGUGGGCUUUAAAGA (*mghv-mir-p4*)
+	112453	95	GUUCUAUGGUACCAACAGACUCUUGUGUUUCUUGAAUGGUUCCAG**uuucauu**CUGGACAAGUAAGAAUACUUGAUCUGUUGUCACAUUAAGGAAU (*mghv-mir-p5*)
+	112662	95	GUGAGUAUUUCUUGGAUGGAGCACCUGACCUGUGGCAUCAUGGAC**cggau**GUAACAGGCGUGAGAAAGGUCUUUGGUCCAUUCUUGUAAUACUCU (*mghv-mir-p6*)
+	3794	94	UGUGAGCUCUUCUUUACCAGCACUCACUGGGGGUUUGGUCAGGAGAUC**aagua**GAUCUGACCAACCCUAAGUGAGUUUUUCUUCUUGCUUAACA (§*mghv-mir-M1-8*)

S	*SP*	*L*	*H. cytomegalovirus strain AD169* (HCMV; X17403.1); 5' → 3'
−	49486	94	GCAAGGUAAGCCCCACGUCGUUGAAGACACCUGGAAAGAGGACGUUC**gcuc**GGGCACGUUCUUUCCAGGUGUUUUCAACGUGCGUGGAUUUUUU (*hcmv-mir-UL36*)
+	49484	94	AGAAAAAAUCCACGCACGUUGAAAACACCUGGAAAGAACGUGCC**cga**GCGAACGUCCUCUUUCCAGGUGUCUUCAACGACGUGGGGCUUACCUU (*hcmv-mir-p1*)
−	174048	95	AUUGACGUCAAUGGGUGGAGUAUUUACGGUAAACUGCCCACUUGG**caguaca**UCAAGUGUAUCAUAUGCCAAGUACGCCCCCUAUUGACGUCAAU (*hcmv-mir-p2*)
+	203097	95	UCUUCGGAAACUGUGGACGCUGUUUCCGAAUACCGGGAGGAG**aucgugcuucc**CUCUUCCAAGGAUCGGAAAGUAGCGUCCGUCGUUUCCGCGGA (*hcmv-mir-p3*)
+	93409	94	GCCGCGGAAUGGACGGGACCCGGGGUCCGCGCCCUUCCCCUCC**ccccac**GGGGGGCUGGGUCGCGGACCCCGGUUCCUAGGCUCGUUCCGCGGU (*hcmv-mir-p4*)
−	155177	78	AAAGGACGACCCGUCUCCCCCCCGCACCCGGGUUUUUUC**ucuu**GGUCGAACCCGGCUUGCGACGACGGGUUGUUCCUUU (*hcmv-mir-p5*)
+	27628	95	GUUUCUUCCCAUAGCCUGUCUAACUAGCCUUCCCGUGAGAGUUUAU**gaac**AUGUAUCUCACCAGAAUGCUAGUUUGUAGAGGCUAUGCGGGAUGC (*hcmv-mir-UL22A*)
−	35809	94	CAGAAAUAGGGCGACGGUGUUUUUAUAACCGAAAGUAGCGUGUUU**gag**ACACGCGCUUCUGGUCGGUUUUUUCACCGUCGUCGCUCUAGGUUUG (*hcmv-mir-p6*)
−	147717	94	ACGUGCACGGUGAAAGUGGCGUCGUCGCUCGGCGGGUGCGCACC**gcc**GGUGCUGCUGCUGACUUCCACGACGUUGUUUUCACCGUCGCCGUCGU (*hcmv-mir-p7*)
+	38054	87	CUCGUCAGCUUCACGGAGCUGUUGUUACCGCCGCCGU**ccgucgcc**GCCGCUGCGGUGGCGGCGACAGCGACGAGCGAGGUGGGCGAG (*hcmv-mir-p8*)
−	65216	95	AGUACCUGCUCGACGACGCGUUCCAUCUGCUUCAGGUCCUCU**accggcaaa**AAGCCGUUAAGGAUGUGAUUGUGCACGCGCGUCAGCAGCUGCGU (*hcmv-mir-p9*)
+	116589	91	AUCACCGGCCUAUUACCGUCGGCGCGACUCUCCGGGCGGUAUGG**augaa**CCACCGUCCGGAUGGGAGCGUUACGACGGUGGUCACCGUGGU (*hcmv-mir-p10*)
−	7091	90	UUCGUCCCGUCUCCCUCUGUGGUCGUGGGUGGUGCGAGAGUACAC**gaugg**GUGGCUCUCGUCUCGGGGGACCACAGGGGGAGGGGGGUAA (*hcmv-mir-p11*)
−	25058	95	ACGCCGGUUUCAUCAUAAACCACCGUGAGAACCGGCGCGGGUUUC**aacac**GAAACCGCGUCACUCACGGACGUAGGUUAUUUCGAAAACCUACGU (*hcmv-mir-p12*)
+	174048	95	AUUGACGUCAAUAGGGGGCGUACUUGGCAUAUGAUACACUUGAUG**uac**UGCCAAGUGGGCAGUUUACCGUAAAUACUCCACCCAUUGACGUCAAU (*hcmv-mir-p13*)
−	194927	94	GUACGGUGUCGCCACCGUUGACGUGGGCGGCGAUGAGAACGUCA**gcgg**UGGCGAAACCGCCGUGCGGAAAGUCCCGGUGCCGAAAUCACCGUGU (*hcmv-mir-p14*)
−	49464	94	GAAGACACCUGGAAAGAGGACGUUCGCUCGGGCACGUU**cuuuccaggugguuuuc**AACGUGCGUGGAUUUUUUCUAUUCUCUACCAGGUGCUUAC (*hcmv-mir-p15*)
−	93410	92	CCGCGGAACGAGCCUAGGAACCGGGGUCCGCGACCCAGCCCCCCGU**ggg**GGGAGGGGAAGGGCGCGGACCCCGGGUCCCGUCCAUUCCGCGG (*hcmv-mir-p16*)
−	140853	95	GAAGUUUCGCGGCAGCGCAAGCCGUGGUAACCGUCGCCGCUGGCG**gcgc**UGCCGCGAGACGACGUGGACGGCACUACGGGCCGACGCAGGUUCUC (*hcmv-mir-p17*)
+	196047	90	UCUGAUCCAACACUGAACGCUUUCGUCGUGUUUUUCAUGCAGCUUU**uac**AGACCAUGACAAGCCUGACGAGAGCGUUCAUCGGGGCAUGA (*hcmv-mir-US5-1*)
+	174117	95	UAAAUACUCCACCCAUUGACGUCAAUGGAAAGUCCCUAUUGGCGU**uacu**AUGGGAACAUACGUCAUUAUUGACGUCAAUGGGCGGGGGUCGUUGG (*hcmv-mir-p18*)
−	27632	90	CAUCCCGCAUAGCCUCUACAAACUAGCAUUCUGGUGAGAUACAU**guuc**AUAAACUCUCACGGGAAGGCUAGUUAGACAGGCUAUGGGAAG (*hcmv-mir-p19*)
−	52491	89	CAUGUGCGCUCACCCGGCGUUCUGGCCACCGGUUACGCCG**ccaaca**UGGCGUAAUUGACGGUGAGAACUCGGAGACCGAGCGGUCCGUG (*hcmv-mir-p20*)
+	35813	92	CCUAGAGCGACGACGGUGAAAAAACCGACCAGAAGCGCGUGU**cuc**AAACACGCUACUUUCGGUUAUAAAAACACCGUCGCCCUAUUUCUGGG (*hcmv-mir-p21*)
−	90766	95	GGCGGUUCUUUGUGAUUUAAAAACACCGUGUUCGUGAAACGUGA**cuuu**UCACGGUUUGUUAGCUGAUGUGAUUUUGGAGGUCACAAACACCGUAC (*hcmv-mir-p22*)
+	25058	95	ACGUAGGUUUUCGAAAUAACCUACGUCCGUGAGUGACGCGGUUUCG**ugu**UGAAACCCGCGCCGGUUCUCACGGUGGUUUAUGAUGAAACCGGCGU (*hcmv-mir-p23*)
−	25024	94	CGCGGGUUUCAACACGAAACCGCGUCACUCACGGACGUAGGUUAUU**ucg**AAAACCUACGUUAAUCCUGAACGCGUUUUGUGUCACGCGUCCCCG (*hcmv-mir-p24*)
+	92228	94	GACGUAGCGAGCGUAGCGAGCUACGUCACGUAUGCGUGCGUCAUCU**ccggc**GGAAAUCAUCUCUGAUGACGUAGCGAGCGAAGCGAGCUACGUC (*hcmv-mir-p25*)
+	25023	94	GCGGGGACGCGUGACACAAAACGCGUUCAGGAUUAACGUAGGUUUU**cga**AAUAACCUACGUCCGUGAGUGACGCGGUUUCGUGUUGAAACCCGC (*hcmv-mir-p26*)
+	139178	89	GUUACCUGUUGUAUCGCAAGGCUCACGUGGAGCUGUCACUCU**ccagca**ACAAGGUGCAACACGUGGAAGCCGUGCUGCGACAGGUGUAC (*hcmv-mir-p27*)
+	146735	94	CGCGCCAGCUAGGGUGCGCUGGCCUCGGCCGUGACUACGGACGCC**gau**GAGCGUCGGCGCGGCCUAGAGCAGCGUAGCGCCGUGUUGGCGCGCG (*hcmv-mir-p28*)
+	37292	95	GCUUCGCUCUGGAUGGGCUCCGGGUCCGUCAACACGCGACUCGC**gcg**GCAAAAGGCACGCUGUUGACGGCGCGAGAGCCCGUCGUGAUAGUCCAU (*hcmv-mir-p29*)
+	173784	95	GCCCAUUUGCGUCAAUGGGGCGGAGUUGUUACGACAUUUUGGAAAGU**ccc**GUUGAUUUUGGUGCCAAAACAAACUCCCAUUGACGUCAAUGGGGU (*hcmv-mir-p30*)
+	25076	94	ACCUACGUCCGUGAGUGACGCGGUUUCGUGUUGAAACCCGCGCCGG**uucu**CACGGUGGUUUAUGAUGAAACCGGCGUUGGGGAUCUACGCGGGU (*hcmv-mir-p31*)
−	134672	93	AUCGUCAGCGAACCGCGCUUCAAACGCCAGAUCCGAAUACAGGUGCG**uuuc**CAUAUUCGGAACGCAUCUGUUUCAGAAGCGCGUCCUCGCGCU (*hcmv-mir-p32*)
+	32963	95	GGCCUUGCGGCGGCAGCGGUUGGCGUGGUUGCUCAGCUCGGCGUC**cga**GAGCGCCGAGCUGAACUGCGGCAGCCGCGUGCGAUCCUGCGGCGCGU (*hcmv-mir-p33*)
−	90873	95	GGCGGAGCGCAGCGAAAAUCGGUGGUGAUAGCGGCGAUUGAGGUUGCG**aga**CCAGAUUCAUCGCGCUUGUACCACCGUGGUGCGGUGUUUCUGCU (*hcmv-mir-p34*)
−	162576	95	CGGAGCAGAGGGUCGUUGUCCUCCUCGUCCUCGUGGCGGUUGUUUC**ucc**GUCGCGAUCUCCGAGAGGAGGAGGACGACGACGAUGCAGCCUGCCG (*hcmv-mir-p35*)
+	30965	94	CCAGAGCCGUUCGGGGCGUGCGGCCGCGCUAGCGCUUCAUUUUCU**cacguc**ACGAAAAGGAGUGACGGACGGCCAGUACGCCACGUCUCUGCGG (*hcmv-mir-p36*)
+	222717	95	UUACUCUCGAGUGCGGUCGGUGUCUCGUCGGUGAGACGAGGCCGCC**gccc**GACAAGUUCGAUCUCAUGUCGCUCUUGGAGCGCGAAGAGAGUUGG (*hcmv-mir-p37*)
−	210170	94	CGCUGCUUUCGCAUGCCCAAGUUCUUUCCGCCGCCCAUGUGCCGCGU**ucc**GUACAACGAAUGCGGCGUCGAAUUACCGGGCGGCGAUAGCAGCG (*hcmv-mir-p38*)
−	203097	95	UCCGCGGAAACGACGGACGCUACUUUCCGAUCCUUGGAAGAGG**gaagcacga**UCUCCUCCCGGUAUUCGGAAACAGCGUCCACAGUUUCCGAAGA (*hcmv-mir-p39*)
−	174118	92	AACGACCCCCGCCCAUUGACGUCAAUAAUGACGUAUGUUCCCAU**agua**ACGCCAAUAGGGACUUUCCAUUGACGUCAAUGGGUGGAGUAUUU (*hcmv-mir-p40*)
+	93225	94	CCCGCUCGACCCCCCCAUCCGACGGCCCGGCCGGGCUGGGACC**ccc**GCACCGGGGUCCCGGUUCCCGUCCGUGGCCCGGGGGGACCCGAGCGGG (*hcmv-mir-p41*)
+	20175	95	ACCCGCUGGGAGGAAAGCAACGUCGUGAGCCAGACGGCCACGCG**agua**CGUACGUGGUUCGUGGAAAGAACCACGUUUUGGCGUCGCACGUGGGU (*hcmv-mir-p42*)
+	163175	94	UCCCGGCGCUCUGACAGCCUCCGGAUCACAUGGUUACUCAGCGUCU**gcc**AGCCUAAGUGACGGUGAGAUCCAGGCUGUCCGUGCACCACGGUGA (*hcmv-mir-UL112*)
−	174625	92	GCCGAUGUGAGUUUCUGUGUAACUGAUAUCGCCAUUUUUCCAAAAG**uga**UUUUUGGGCAUACGCGAUAUCUGGCGAUAGCGCUUAUAUCGUU (*hcmv-mir-p43*)
+	124992	95	CGCGGUUUACGUAGGCUACGCAGGUAUUUGACGUGUAAC**ccagacccau**GUCUACGGUGUUAAUGUUCUGCGUGACGUGGUACGUAGUGCUGAUG (*hcmv-mir-p44*)

−	119625	94	AUCCUCGGCGACGGCGUGCACGUCGGGCGUUAUGACACGCGGCC**gccuuaa**GGCCGAGUCCACCGUCGCGCCCGAAGAGGACACCGACGAGGAU (*hcmv-mir-p45*)
−	36838	94	UCCUCUGCCUGGGCACGCGCGUCGGCCGCGUCGCAAACGCUGCUUGG**uac**CCGAGGUCUUUUGCACGCGCGACUUGGCCGACCUGUGCGUGCGA (*hcmv-mir-p46*)
−	197467	90	GUGGGUGCCCACGGACUUGGACCAUCUCACUCUGCAUUUGGUGC**cgu**GCACCAAAUGCAAACCCAUGUGGUGCCAGCCUCGGUACCAUAU (*hcmv-mir-p47*)
+	119625	94	AUCCUCGUCGGUGUCCUCUUCGGGCGCGACGGUGGACUCGGCCUU**aag**GCGGCCGCGUGUCAUAACGCCCGACGUGCACGCCGUCGCCGAGGAU (*hcmv-mir-p48*)
+	147719	93	GACGGCGACGGUGAAAACAACGUCGUGGAAGUCAGCAGCAGCACC**ggc**GGUGCGCACCCGCCGAGCGACGACGCCACUUUCACCGUGCACGUU (*hcmv-mir-p49*)
−	194965	93	UGACGUGACUCUUGACGUUUAUAAACCGCAUGGGAAAGUACGGUG**ucgc**CACCGUUGACGUGGGCGGCGAUGAGAACGUCAGCGGUGGCGAAA (*hcmv-mir-p50*)
+	128612	95	ACUGGGUCGUCUGUUACUGGGACCCGUGGCCGUACCCUGUUUUUG**cga**CGGUGAAGUGGAGGGCCACGGUGAACAUCUGGUACCUACGACGCAGU (*hcmv-mir-p51*)

S (+/− strand), *SP* (start position), and *L* (length of the putative *pre-miRs*). 25 true-positives and 1 false-negative match 25 published *pre-miR* sequences (red regions) and their mature miRNAs (underlined regions) as obtained from miRBase 8.2 (Griffiths-Jones *et al*., 2006); predicted terminal loop ≥3-nts (bold lowercase nucleotides). ‡/Δ*kshv-mir-K12-9* are the accepted and incorrect positives of *kshv-mir-K12-9*. †*ebv-mir-BHRF1-1* (0.437 *miPred* score) and §*mghv-mir-M1-8* (0.658).