COMPUTATIONAL IDENTIFICATION OF NOVEL MICRORNAS USING INTRINSIC RNA FOLDING MEASURES

COMPUTATIONAL IDENTIFICATION OF NOVEL MICRORNAS USING INTRINSIC RNA FOLDING MEASURES

NG KWANG LOONG STANLEY

NATIONAL UNIVERSITY OF SINGAPORE
2007/2008

NG KWANG LOONG STANLEY 2007/2008

COMPUTATIONAL IDENTIFICATION OF NOVEL MICRORNAS USING INTRINSIC RNA FOLDING MEASURES

NG KWANG LOONG STANLEY
(M.Eng. (Research), National University of Singapore)
(B.Eng. (Hons), National University of Singapore)

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE
SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE
2007/2008

Acknowledgments

My sincere gratitude to my two main supervisors Prof. Wong Lim Soon (2006−2008) and Dr. Santosh K. Mishra (2004−2006) for their overwhelming support and patience during my four graduate years at Bioinformatics Institute (BII). They provided constant academic guidance and inspired many of the ideas presented in my Ph.D project. Both supervisors are superb teachers, great communicators, and excellent manager of research projects. It was my fortune to be offered a chance to work closely with them. I look forward to develop our relationship further both as colleagues and as friends.

At BII, I have learned and acquired as much from the continuous interaction with other staffs and students as from my supervisors. I wish to acknowledge my colleagues Tan Yang Hwee, Stephen Wong, and Damien Leong from A*STAR Computational Resource Centre (ACRC) for their invaluable technical guidance and assistance concerning high-throughput grid computing. Prof. Gunaretnam Rajagopal, executive director of BII, motivated me with his enthusiastic encouragement and understanding, most critical to the development of my academic pursuit. In addition, I would like to extend special gratitude and heart-felt appreciation to two collaborators Beh Yee Ming Leslie and Leong Shiang Rong for sharing their knowledge of biology and genetics, and their understanding and advice on this academic project. I also acknowledge my thesis committee members Assist. Prof. Vinay Tergaonkar (2006−2008) and Prof. Barry Halliwell (2004−2006) for pointing me to the right direction during the long Ph.D journey. Special appreciation to the Reproductive Genomics Group members Kwan Hsiao Yuen, Wang Xin Gang, Ng Say Aik, Liew Woei Chang, Alex Chang, Rajini Sreenivasan, and Assoc. Prof. Laszlo Orban from Temasek Life Sciences Laboratory (TLL), for their warm support and expertise in zebrafish. They provided significant collaboration on the construction of small RNA library, real-time RT-PCR, and *in situ* hybridization.

I wish to dedicate this thesis to my mother, for without her love, self-sacrifice, constant guidance, and encouragement throughout my life, I would not have this great opportunity to pursue and fulfill my academic ambition, and being provided the best possible education. I also would like to thank my wife for her support and for having absolute confidence in me.

Assoc. Prof. Christian Schoenbach from School of Biological Sciences, Nanyang Technological University (NTU), and Assoc. Prof. Lee Mong Li Janice from School of Computing (SOC), National University of Singapore (NUS) were specially invited to review the final pre-submission draft of this thesis. I am especially indebted to the first reviewer and his coworker Ng Sze Wei for performing the Northern Blotting validation of novel miRNAs expressed in zebrafish.

Finally, I am grateful to my three examiners Prof. Peter Clote (Biology Department of Boston College), Prof. Vladimir B Bajic (Deputy Director of South African National Biodiversity Institute), and Prof. Peter Stadler (University of Leipzig), whom have provided invaluable comments for improving greatly the quality of this dissertation.

This work is supported by the Agency for Science, Technology and Research (A*STAR).

Table of Contents

Page

Abstract

MicroRNAs (miRNAs) are small endogenous ncRNAs participating in diverse cellular and physiological processes by post-transcriptionally suppressing the target genes. Critically associated with the early stages of the mature miRNA biogenesis, the hairpin motif is a crucial structural prerequisite for the prediction of authentic and novel precursor miRNAs (pre-miRs). Majority of the abundant genomic inverted repeats (pseudo hairpins) are dysfunctional pre-miRs and can be filtered by comparative genomic-driven approaches, but genuine specie-specific pre-miRs are likely to remain elusive.

Motivated by the incomplete knowledge on the number of miRNAs present in the genomes of vertebrates, worms, plants, and even viruses, an in-depth statistical study (Ng and Mishra 2007b) was conducted to elucidate the unique hairpin folding of an entire pre-miR. The comprehensive and heterogeneous datasets comprised of a collection of 2,241 published (non-redundant) pre-miRs across 41 species, 8,494 pseudo hairpins, 12,387 (non-redundant) ncRNAs spanning 457 types, 31 full-length mRNAs, and 4 sets of synthetically generated genomic background corresponding to each of the native RNA sequence. The global and intrinsic hairpin folding features include the *%G+C* content, normalized base pairing propensity *dP*, normalized Minimum Free Energy of folding *dG*, normalized Shannon Entropy *dQ*, normalized base pair distance *dD*, and degree of compactness *dF*, as well as their normalized *Z*-scores. These features distinguish unambiguously pre-miRs from other types of ncRNAs, pseudo hairpins, mRNAs, and genomic background.

A new *de novo* Support Vector Machine classifier *miPred* (Ng and Mishra 2007a) was developed for identifying pre-miRs without relying on phylogenetic conservation information, while able to handle arbitrary secondary structures. It achieved significantly higher sensitivity and specificity than existing (quasi) *de novo* predictors, by incorporating a Gaussian Radial Basis Function kernel as a similarity measure for the 29 combinatoric attributes. They characterized a pre-miR with the sequence motifs at the dinucleotide sequence level, hairpin structural characteristics, and topological descriptors. The predictor *miPred* achieved 93.50% (five-fold cross-validation accuracy) and 0.9833 (AUC or ROC score) on the human training

dataset; 84.55% (sensitivity), 97.97% (specificity), and 93.50% (accuracy) for the remaining human testing dataset; 87.65% (sensitivity), 97.75% (specificity), and 94.38% (accuracy) for 1,918 pre-miRs in 40 non-human species.

Two novel miRNAs *dre-miR-N1* and *dre-miR-N2* identified by *miPred* in the brain and gonads of juvenile and adult zebrafish, were validated experimentally as *bona fide* through Northern Blot, and were found to be localized in the adult ovary and testis via frozen section *in situ* hybridization (Beh and Ng *et. al.* 2007; *in preparation*).

Keywords: classification, intrinsic RNA folding measures, microRNAs, precursor microRNAs, pseudo hairpins, secondary structure, support vector machine

List of Tables

Table *Page*

List of Figures

Figure *Page*

List of Abbreviations

ACC	**ACCURACY**
DNA	**DEOXYRIBONUCLEIC ACID**
DS	**DINUCLEOTIDE SHUFFLING**
EGFP	**ENHANCED GREEN FLUORESCENT PROTEIN**
ESTS	**EXPRESSED SEQUENCE TAGS**
FM	**FIRST-ORDER MARKOV MODEL**
MFE	**MINIMUM FREE ENERGY OF FOLDING**
MIRNA	**MICRORNA**
MRNA	**MESSENGER RNA**
MS	**MONONUCLEOTIDE SHUFFLING**
NCRNA	**NON-CODING RNA**
PCR	**POLYMERASE CHAIN REACTION**
POL-II	**RNA POLYMERASE TYPE II**
PRE-MIR	**PRECURSOR MICRORNA**
PRI-MIR	**PRIMARY MICRORNA**
RBF	**GAUSSIAN RADIAL BASIS FUNCTION**
RISC	**RNA-INDUCED SILENCING COMPLEX**
RNA	**RIBONUCLEIC ACID**
ROC	**RECEIVER OPERATING CHARACTERISTIC CURVE**
RRNA	**RIBOSOMAL RNA**
RT-PCR	**REVERSE TRANSCRIPTION POLYMERASE CHAIN REACTION**
SE	**SENSITIVITY**
SIRNA	**SMALL-INTERFERING RNA**
SNORNA	**SMALL NUCLEOLAR RNA**
SP	**SPECIFICITY**
SVM	**SUPPORT VECTOR MACHINE**
TF	**TRANSCRIPTION FACTOR**
TRNA	**TRANSFER RNA**
TU	**TRANSCRIPTION UNIT**
ZM	**ZERO-ORDER MARKOV MODEL**

List of Mathematical Symbols and Notations

$\%G+C$	AGGREGATE DINUCLEOTIDE FREQUENCY %G+C RATIO
dD	ADJUSTED BASE PAIR DISTANCE
dF	SECOND (OR THE FIEDLER) EIGENVALUE
dG	ADJUSTED MINIMUM FREE ENERGY OF FOLDING
dP	ADJUSTED BASE PAIRING PROPENSITY
dQ	ADJUSTED SHANNON ENTROPY
FN	FALSE NEGATIVES
FP	FALSE POSITIVES
F_X	MONONUCLEOTIDE FREQUENCIES
F_{XY}	DINUCLEOTIDE FREQUENCIES
$MFEI_1$	MFE INDEX 1
$MFEI_2$	MFE INDEX 2
TN	TRUE NEGATIVES
TP	TRUE POSITIVES
zD	Z-SCORE OF ADJUSTED BASE PAIR DISTANCE
zF	Z-SCORE OF SECOND (OR THE FIEDLER) EIGENVALUE
zG	Z-SCORE OF ADJUSTED MINIMUM FREE ENERGY OF FOLDING
zP	Z-SCORE OF ADJUSTED BASE PAIRING PROPENSITY
zQ	Z-SCORE OF ADJUSTED SHANNON ENTROPY

Chapter 1.
Introduction

Precise genetic control is an essential survival feature of cellular systems, as they must respond to a multitude of metabolic requirements and developmental programs by varying spatial and temporal genetic expression patterns. Since the early 1960s, the concept of operon (Beckwith 1996) was postulated that all protein-coding transcriptional units are controlled by means of operons subject to mechanisms of genetic control. Presumably, such mechanisms always involve protein factors that can sense biochemical signals and environmental cues, and then modulate the expression of corresponding genes by selectively interacting with the relevant Deoxyribonucleic acid (DNA) or Ribonucleic acid (RNA) sequences.

Although proteins fulfill most requirements that biology has for enzyme, receptor, and structural functions, it is rediscovered lately that a plethora of functional non-coding RNA molecules can also serve in these capacities. Unlike mRNA, non-coding RNAs (ncRNAs) are characterized uniquely as functional RNAs that are not translated into proteins after being transcribed from genomic DNA. Inadvertently, ncRNA was widely perceived as "junk" RNA functionally unimportant in the cell, and merely performed as "accessory components to aid protein functioning" (Huttenhofer *et al*., 2005). These functional ncRNAs are emerging gradually as the central player participating in multiple regulatory layers and influencing a wide range of vital cellular processes including chromatin modification, mRNA stability and localization, transcription initiation, RNA processing, mRNA and protein synthesis, as well as post-translational RNA modification (Mattick and Makunin 2005; Storz 2002; Eddy 2001; Gray and Wickens 1998).

Functional ncRNAs that have been discovered to date, namely, the ribozymes (Puerta-Fernandez *et al*., 2003), small nuclear RNA (snRNA) (Storz *et al*., 2005), transfer RNAs (tRNAs) (Sprinzl and Vassilenko 2005), ribosomal RNAs (rRNAs), endogenous small-interfering RNAs (siRNAs) (Huttenhofer *et al*., 2005), and most recently the riboswitches (Soukup and Soukup 2004; Mandal and Breaker 2004; Nudler and Mironov 2004; Vitreschak *et*

al., 2004; Winkler and Breaker 2003; Stormo 2003; Lai 2003; Hesselberth and Ellington 2002) are relatively short in length compared to protein-coding mRNAs. Others ncRNAs are long ranging from hundreds of base pairs to more than 10 kilobases and resemble mRNAs in that they are spliced, polyadenylated, and possibly 5' capped (Erdmann *et al*., 2000), but may only contain short ORFs. These mRNA-like ncRNAs include the mouse air RNA required for gene imprinting (Sleutels *et al*., 2002), the yeast meiRNA involved in meiosis control (Yamashita *et al*., 1998), and the mammalian XIST RNAs required for X chromosome inactivation (Xiao *et al*., 2007).

This series of unexpected and exciting discoveries have led to a new paradigm of RNA-directed gene expression regulation, defying the central dogma that DNA acts purely as a storage of information, RNA is solely the intermediate, and protein performs as the vehicle for catalytic reactions. Multiple challenges laid ahead as exact mechanism of action for some ncRNAs especially microRNAs in relation to their structures (Ahmed and Duncan 2004) and how the underlying sequence relates to and their biological functions (Vogel *et al*., 2003; Kitagawa *et al*., 2003) are still largely unclear. Notably, two international scientific consortiums, namely, the ENCyclopedia Of DNA Elements (ENCODE) Project (The ENCODE Project Consortium 2004) and the Functional Annotation of Mouse (FANTOM) (Maeda *et al*., 2006) are making significant progress in applying high-throughput computational and laboratory-based approaches for detecting all sequence elements, especially those that undergo non-coding transcription, that confer biological function.

1.1. Background of MicroRNAs

Several large families of functional RNAs associated with essential protein synthesis are ubiquitous among all three kingdoms of life i.e., eukaryota, bacteria, and archaea (Griffiths-Jones *et al*., 2005) − rRNA (decodes mRNA into amino acid) and tRNA (delivers amino acid to growing polypeptide chain), along with RNase P (tRNA maturation) and SRP RNA (protein export). In contrast, microRNAs (miRNAs) constitute an abundant class of small ~21–23 nucleotides in length evolutionary conserved ncRNA molecules (Figure 1.1; colored in red) found exclusively in eukaryotes. They play important roles in gene regulation by mediating post-transcriptionally the production of intra-cellular proteins in most eukaryotes via sequence-specific target mechanisms (Bartel 2004; Mallory and Vaucheret 2004; Ambros 2001). The founding members of the miRNA gene family *lin-4* (Lee *et al*., 1993) and *let-7* (Reinhart *et al*.,

2000) unraveled respectively in 1993 and 2000, are essential heterochronic regulators directing temporal aspects of development timing in the early larval nematode *Caenorhabditis elegans* by repressing target genes *lin-14*, *lin-28*, and *lin-41* (Banerjee and Slack 2002). Since the inception of this epic regulatory RNA phenomenon, thousands of novel miRNA genes have been discovered across plants, worms, flies, vertebrates, and even viruses (Griffiths-Jones *et al*., 2006). (Figure 1.2) Among them, 474 and 373 mouse miRNAs were found in human and mouse genomes, respectively.

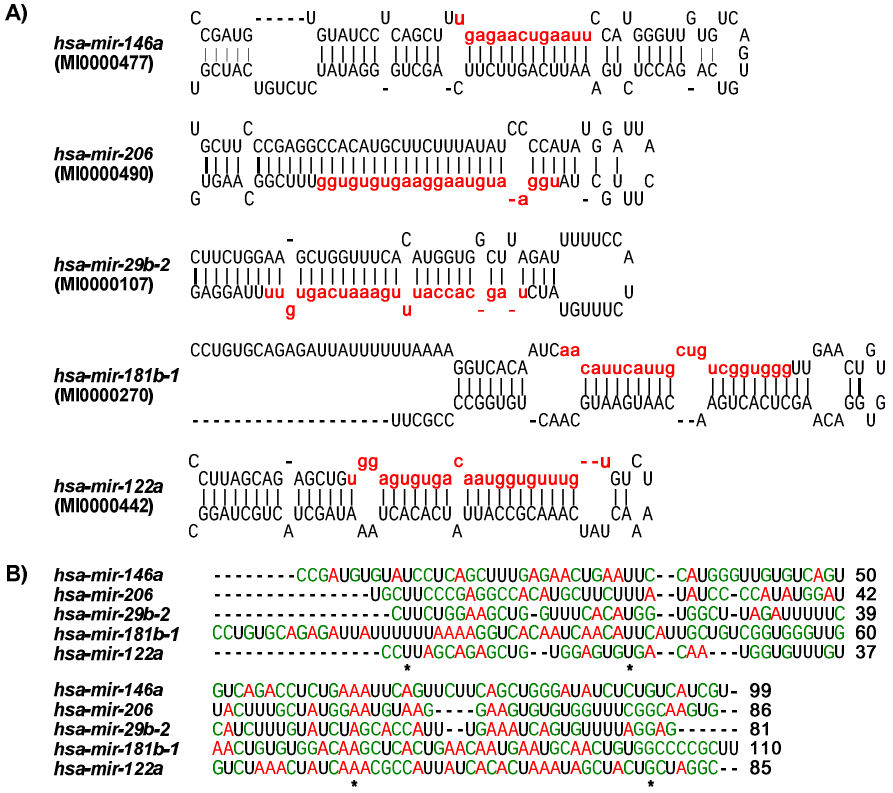

Figure 1.1: A) Secondary structures of sample human miRNA precursors. Red regions denote mature miRNAs. B) Multiple alignments of sample human miRNA precursors.

Majority of the endogenous miRNA genes originate from the polycistronic genes residing in the intergenic regions overlapping with the introns of protein-coding genes (Lee *et al*., 2002), or in the exons of the pseudo-ncRNA genes (Rodriguez *et al*., 2004). Lately, miRNAs have also been discovered in the introns (Ying and Lin 2005) of *Caenorhabditis elegans* (Ohler *et al*., 2004). These intronic miRNAs differ uniquely from intergenic miRNAs in the requirement of RNA polymerases type II (Pol-II) and spliceosomal components for its biogenesis. (Figure 1.3)

MiRNA genes originate primarily from intronic and independent genomic regions of protein-coding and mRNA-like ncRNA transcription units, but fewer from exons and untranslated regions (Rodriguez *et al*., 2004). Details of miRNA biogenesis are described in section 2.1.

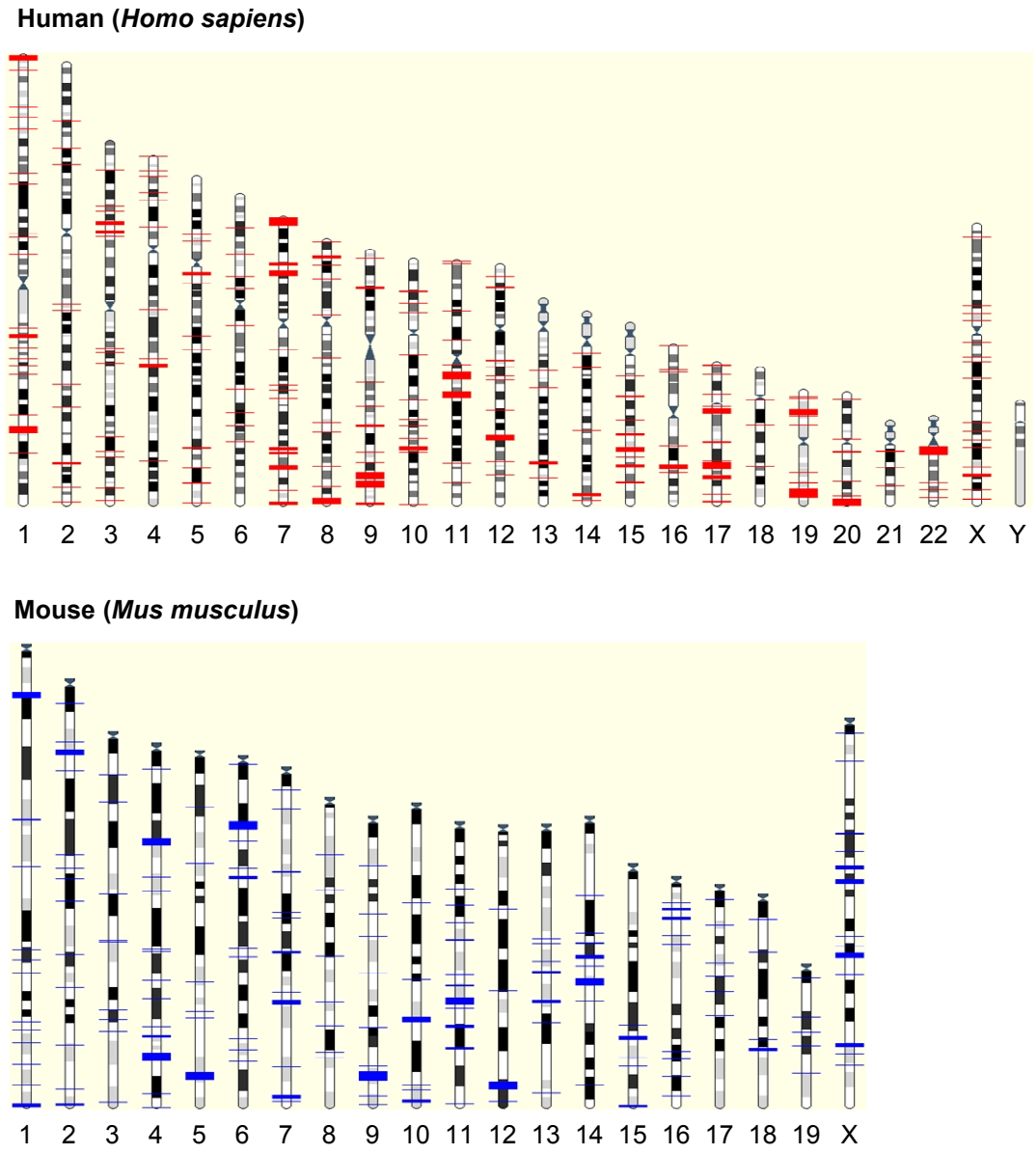

Figure 1.2: Distribution of known 474 human and 373 mouse miRNAs with respect to the chromosome loci.

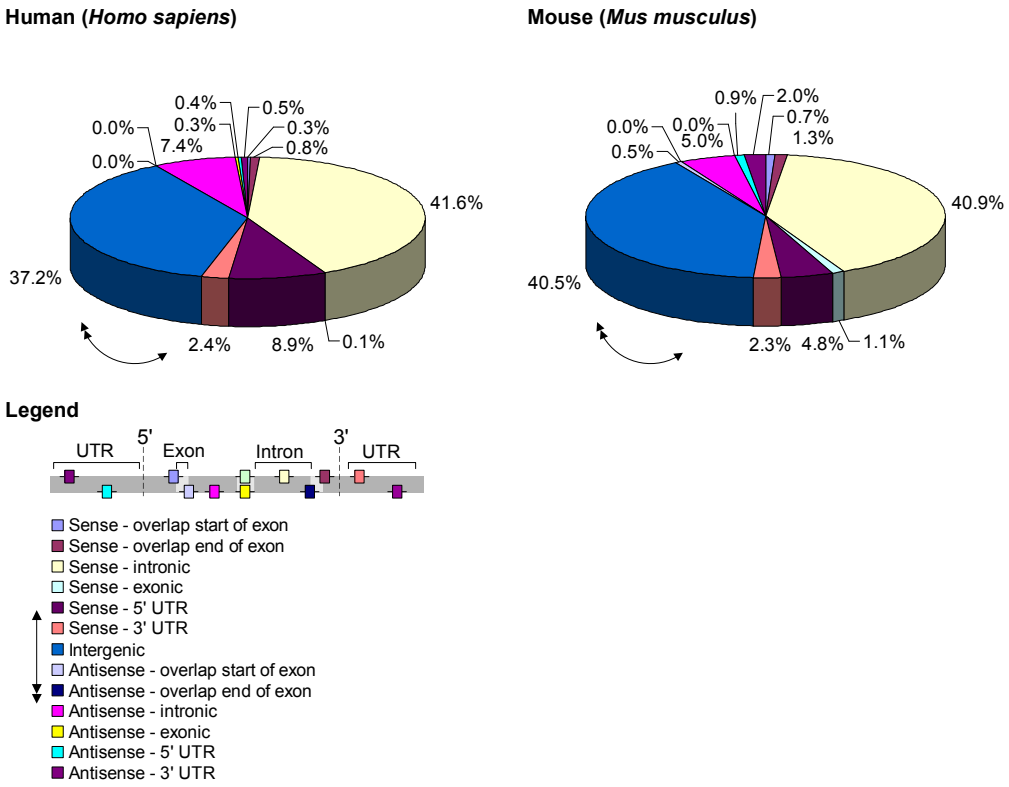

Figure 1.3: Distribution of known 474 human and 373 mouse miRNAs with respect to the nearest transcription unit.

Biologically pivotal and more prevalent genomically than presumed, emerging body of experimental evidence from those (relatively few) miRNAs whose biological function have been characterized, substantiates that miRNAs perform key regulatory roles for diverse developmental and physiological processes. For example, the *Caenorhabditis elegans lsy-6* determines the left-right asymmetry of chemo-receptor expression (Johnston and Hobert 2003); *Caenorhabditis elegans lin-57*/*hbl-1* ensures post-embryonic developmental events are appropriately timed (Abrahante *et al*., 2003); *Caenorhabditis elegans let-7* negatively regulates *let-60*/RAS associated with lung tumors (Johnson *et al*., 2005); *Drosophila melanogaster miR-14* miRNA is involved in apoptosis, stress resistance, and fat metabolism (Xu *et al*., 2003); *D melanogaster bantam* represses the gene *hid* associated with apoptosis and proliferation (Brennecke *et al*., 2003); *Mus musculus miR-181a* modulates hematopoietic differentiation (Chen *et al*., 2004); *Mus musculus miR-196* induces directed-cleaving of *Hox-B8* transcripts (Yekta *et al*., 2004); *Arabidopsis thaliana* miRNAs regulate the expression of transcription factor genes (Li and Zhang 2005); viral-encoded miRNAs hijack the host immune defense to

sustain their viral replication and pathogenesis (Stern-Ginossar *et al*., 2007; Pfeffer *et al*., 2005; Samols *et al*., 2005; Grey *et al*., 2005; Pfeffer *et al*., 2004). This dynamic range of biological findings underscores the functional importance of miRNAs, and the need for expanding our limited knowledge concerning them.

1.2. Contributions of this Thesis

MicroRNAs (miRNAs) are small ncRNAs participating in diverse cellular and physiological processes through the post-transcriptional gene regulatory pathway. Critically associated with the early stages of the mature miRNA biogenesis, the hairpin motif is a crucial structural prerequisite for the computational prediction of authentic and novel precursor miRNAs (pre-miRs). Though many of the abundant genomic inverted repeats (pseudo hairpins) can be filtered computationally by comparative genomic-driven approaches, genuine specie-specific pre-miRs are likely to remain elusive. A definitive criterion for identifying and classifying accurately promising precursor transcripts as *bona fide* pre-miRs within a single genome has not yet been discovered. Moreover, discriminative features used in existing (quasi) *de novo* classifiers have achieved far from satisfactory predictive performances.

Motivated by the incomplete knowledge on the number of miRNAs present in the genomes of vertebrates, nematodes, plants, and even viruses, an in-depth statistical study (Ng and Mishra 2007b) was conducted to elucidate the unique hairpin folding of an entire pre-miR based on their sequence motifs, hairpin structural characteristics, and topological descriptors. The comprehensive and heterogeneous datasets comprised of a collection of 2,241 published (non-redundant) pre-miRs across 41 species (*Sanger miRBase* 8.2), 8,494 pseudo hairpins extracted from the human RefSeq genes, 12,387 (non-redundant) ncRNAs spanning 457 types (*Sanger Rfam* 7.0), 31 full-length mRNAs randomly selected from *GenBank*, and four sets of synthetically generated genomic background corresponding to each of the native RNA sequence. The combinatoric (intrinsic and global) features include the *%G+C* content, normalized base pairing propensity *dP*, normalized Minimum Free Energy of folding *dG*, normalized Shannon Entropy *dQ*, normalized base pair distance *dD*, and degree of compactness *dF*, as well as their corresponding *Z*-scores *zP*, *zG*, *zQ*, *zD*, and *zF*. The large-scale characterization analysis revealed that these features distinguish distinctively pre-miRs from other types of ncRNAs, pseudo hairpins, mRNAs, and genomic background according to the non-parametric Kruskal-Wallis ANOVA ($p < 0.001$).

Based on the earlier findings (Ng and Mishra 2007b), a new *de novo* Support Vector Machine classifier *miPred* (Ng and Mishra 2007a) was developed for identifying pre-miRs without relying on phylogenetic conservation information, while able to handle arbitrary secondary structures. It achieved significantly higher sensitivity and specificity than existing (quasi) *de novo* predictors, by incorporating a Gaussian Radial Basis Function kernel as a similarity measure for the 29 global and intrinsic hairpin folding attributes. They characterized 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* achieved 93.50% (five-fold cross-validation accuracy) and 0.9833 (AUC or ROC score). Tested on the remaining 123 human pre-miRs and 246 pseudo hairpins, it reported 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 yielded 87.65% (92.08%), 97.75% (97.42%), and 94.38% (95.64%) for the mean (overall) sensitivity, specificity, and accuracy. Notably, *Apis mellifera*, *Ateles geoffroyi*, *Canis familiaris*, *Epstein barr virus*, *Herpes simplex virus*, *Human cytomegalovirus*, *Ovis aries*, *Physcomitrella patens*, *Rhesus lymphocryptovirus*, *Simian virus*, and *Zea mays* were unambiguously classified with 100.00% (sensitivity) and more than 93.75% (specificity).

Given the promising performances of the proposed *de novo* SVM classifier *miPred*, it was incorporated into a computational pipeline for the screening of novel miRNAs expressed in the brain and gonads of juvenile and adult zebrafish. Two novel miRNAs *dre-miR-N1* and *dre-miR-N2* found to be expressed in the adult testis and juvenile female brain small RNA libraries, possessed Minimum Free Energy of -45.90 kcal/mol and -56.30 kcal/mol, as well as *miPred* scores of 0.999978 and 0.999681 as predicted by a SVM-based classifier *miPred* (Ng and Mishra 2007a), respectively. They were validated experimentally as *bona fide* miRNAs through Northern Blotting (Beh and Ng *et. al.* 2007; *in preparation*). Further characterization via frozen section *in situ* hybridization revealed their differential expression in the stage I/II oocytes (but not in stage III oocytes) of adult ovary and primary spermatocytes (but not secondary spermatocytes) of adult testis, and they exhibited sexual dimorphism in non-canonical sex-related organs including the brain, gill and muscle/connective tissue between both sexes.

1.3. Publications

A series of peer-reviewed publications, international conferences, and working papers were

authored during the course of this thesis. Arranged chronologically; bold and underlined name(s) denote corresponding and first author(s), respectively.

Beh,E.M., Ng,K.L.S., Schoenbach,C., Ng,S.W., **Wong,L.S.**, and **Orban,L.** (2008) Small RNA Profiling in Zebrafish Gonads and Brain: Novel miRNAs with Sexually Dimorphic Expression (*in preparation*). Both first authors contributed equally.

Ng,K.L.S. and Mishra,S.K. (2007a) De Novo SVM Classification of Precursor MicroRNAs from Genomic Pseudo Hairpins Using Global and Intrinsic Folding Measures. *Bioinformatics*, **23**, 1321-1330.

Ng,K.L.S. and Mishra,S.K. (2007b) Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification. *RNA*, **13**, 170-187.

Ng,K.L.S. and Mishra,S.K. (2006a) Spectral Graph Partitioning Analysis of In Vitro Synthesized RNA Structural Folding, in Proceedings of the *International Workshop on Pattern Recognition in Bioinformatics (PRIB 2006)*, Hong Kong, China, August 20, 2006. Also published in *Lecture Notes in Computer Science* (Springer), **4146**, 81-92.

Ng,K.L.S. and Mishra,S.K. (2006b) Virus on the Grid: Grid-enabling Viral-encoded MicroRNAs Identification, in Proceedings of the *Third International Life Science Grid Workshop (LSGRID 2006)*, Yokohama Kanagawa, Japan, October 13-14, 2006.

1.4. Thesis Organization

The thesis is organized into six chapters:

Chapter 2 introduces the biogenesis model of mature miRNA. Notably, the hairpin motif is a crucial structural prerequisite for the computational prediction of authentic and novel precursor miRNAs (pre-miRs). State-of-the-art approaches for identifying *bona fide* miRNAs (namely, experiment-based, comparative-genomics driven, and prediction-based) are then discussed.

Chapter 3 summarizes the material and methods described in both works (Ng and Mishra 2007a; Ng and Mishra 2007b). They are the biologically relevant datasets, intrinsic RNA folding measures, implementation of *de novo* classifier *miPred*, and statistical analysis metrics.

Chapter 4 and **Chapter 5** cover the results and discussion presented in both works (Ng and Mishra 2007b) and (Ng and Mishra 2007a), respectively. An in-depth statistical study (Ng and Mishra 2007b) was conducted to elucidate the unique hairpin folding of an entire pre-miR based on their sequence motifs, hairpin structural characteristics, and topological descriptors.

Follow up from the new findings, a *de novo* Support Vector Machine classifier *miPred* (Ng and Mishra 2007a) based on intrinsic folding measures was developed for identifying novel pre-miRs without relying on phylogenetic conservation information.

Chapter 6 describes the application of *miPred* as part of a computational pipeline for the identification of novel miRNAs expressed in the brain and gonads of juvenile and adult zebrafish (Beh and Ng *et. al.* 2007; *in preparation*). Two selected putative miRNAs were validated by northern blot and subjected to characterization by *in situ* hybridization.

Chapter 7 concludes this dissertation and outlines the future directions including the ESTs analysis of miRNAs; research on miRNA target prediction algorithms to improve accuracy of miRNA target binding sites associated with human diseases; research on the mechanisms for transcriptional regulation of miRNAs given that most of their expression are highly cell/tissue specific.

Chapter 2.
Background of MicroRNA Identifications

2.1. Biogenesis of MicroRNAs and Small-Interfering RNAs

(Figure 2.1) The prevailing biogenesis model of miRNA maturation points to five or six compartmentalized stepwise processing within the nucleus/cytoplasm in plants and vertebrates, respectively (Kim 2005; Anthony and Peter 2005). Briefly, (**1**) majority of the primary miRNAs (pri-miRs) are transcribed by the RNA polymerase II (Pol-II) into long primary transcripts. (**2**) These capped and polyadenylated pri-miRs of varying length (more than 1,000 nucleotides) tend to fold with specific "hairpin-shaped" secondary structure, serve as substrates for recognition by the nuclear endonuclease RNase III Drosha/Pasha complex (Lee *et al*., 2003; Zeng and Cullen 2003). Cleaving asymmetrically at sites near the bases of their primary stems release approximately 60−120 nucleotides intermediate precursor transcripts (pre-miRs). (**3**) Those pre-miRs possessing characteristic imperfect and extended hairpin structures with a 5' phosphate and a 2 nucleotides 3' overhang, are exported into the cytoplasm by the cargo transporter protein Exportin-5 in a Ran-GTP dependent manner or by HASTY, the orthologue of Exportin-5 (Zhang *et al*., 2006b). (**4**) Cytoplasmic RNase III-type endonuclease Dicer excises the pre-miRs, about 2 helical turns away from the termini of the stem-loop of pre-miRs, into 22–23 nucleotides asymmetric mature miRNA duplexes miRNA:miRNA*. On the contrary, Dicer-like 1 enzyme DCL1, a plant orthologue of Drosha, performs both cleavage steps in the nucleus i.e., pri-miRs $\rightarrow$ 80−200 nucleotides pre-miRs $\rightarrow$ miRNA:miRNA* (Anthony and Peter 2005). Plant mature miRNA duplexes miRNA:miRNA* exhibit greater frequency of base pairings and have tighter length distribution centering on 21 nucleotides (Anthony and Peter 2005). (**5**) The strand miRNA with the less thermo-stable 5' termini is preferentially incorporated into a ribonucleoprotein to form a RNA-induced silencing complex (RISC) (Rivas *et al*., 2005; Maniataki and Mourelatos 2005; Tang 2005; Gregory *et al*., 2005; Tijsterman and Plasterk 2004; Cullen 2004a). Every RISC contains a member of the Argonaute protein family

that tightly binds the single-strand RNA in the complex. (**6**) The bound strand guides the RISC to the target mRNAs, for which the mechanistic modes of miRNA-directed post-transcriptional silencing of target genes differ between vertebrates and plants (Anthony and Peter 2005).

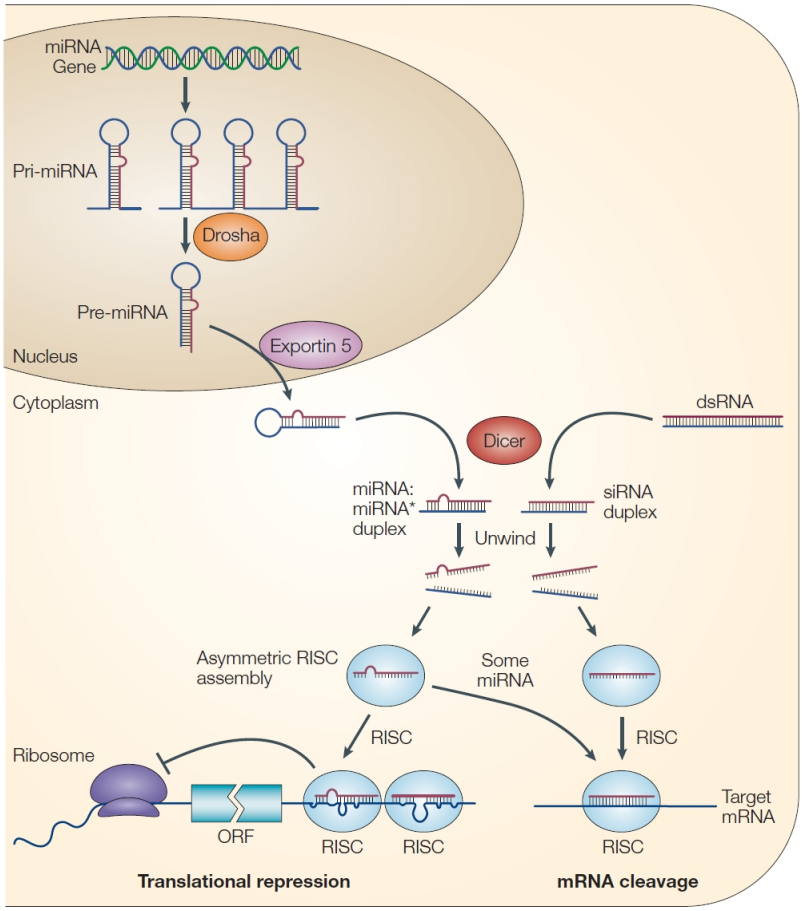

Figure 2.1: Simplified model of miRNA and siRNA biogenesis and regulation of target gene expression (He and Hannon 2004).

Primarily in vertebrates, through imperfect complementary base pairing to the 3' untranslated regions of specific mRNA transcripts, the RISC represses post-transcriptionally the target gene expression via translational arrest of protein synthesis (Doench and Sharp 2004; Reinhart *et al*., 2000; Olsen and Ambros 1999; Moss *et al*., 1997) and occasionally deadenylation (Wu *et al*., 2006). Exceptions include the miRNA-guided cleaving of *Mus musculus Hox-B8* transcripts (Yekta *et al*., 2004) and of *Epstein barr virus BALF5* (virus DNA polymerase) transcripts (Pfeffer *et al*., 2004) by *miR-196* and *miR-BART2*, respectively. For plants, mRNA cleavage-degradation occurs with exact (or quasi) complementarity of not more

than 4 mismatches at the protein-coding regions of mRNAs (Brennecke *et al*., 2005; Yekta *et al*., 2004). Nevertheless, *Arabidopsis thaliana* non-protein coding gene IPS1 (Induced by Phosphate Starvation1) contains a motif with sequence complementarity to the phosphate (Pi) starvation-induced miRNA *miR-399*, but the pairing was found to be interrupted by a mismatched loop at the expected miRNA cleavage site. The IPS1 RNA is not cleaved, instead sequesters *miR-399* (Franco-Zorrilla *et al*., 2007).

In comparison, small-interfering RNAs (siRNAs) are another family of short 21−22 nucleotides ncRNAs, functionally equivalent to miRNAs. Like the mature miRNA, the mature siRNA possesses a 5' phosphate and a two nucleotides 3' overhang, and is incorporated as a single-stranded RNA into the RISC. The RISC binds with exact (or quasi) anti-sense complementarity to the mRNA of the target genes. It cleaves between the 10^{th} and 11^{th} nucleotides (Elbashir *et al*., 2001a; Elbashir *et al*., 2001b), resulting in the post-transcriptional silencing of the target gene. At least demonstrated in mammalian tissue cells culture (Zeng *et al*., 2003; Doench *et al*., 2003), exogenously supplied siRNA can repress expression of a target mRNAs with partial complementarity to the 3' untranslated regions without inducing detectable RNA cleavage, while endogenously encoded human miRNA can direct cleaving of an mRNA bearing fully complementary target sites. Experimental evidence points to partial overlap in the protein composition of RISCs used by siRNAs and miRNAs (Filipowicz *et al*., 2005), explaining why both species of small ncRNAs are able to utilize largely similar or entirely identical post-transcriptional regulatory machinery (Cullen 2004b).

Both miRNA and siRNA differ mainly in their biogenesis and evolutionary conservation (Murchison and Hannon 2004; Bartel 2004; Ambros *et al*., 2003b). For biogenesis, identical copies of mature miRNAs originate from one arm of each precursor hairpin, which is the stem region of shorter hairpins of endogenously encoded transcripts. In contrast, numerous different mature siRNAs are derived from each exogenously long double-stranded RNA precursor via the RNA interference pathway (Hannon 2002). The mature miRNAs and their precursor hairpins are often evolutionarily conserved. These hairpins are also transcribed from the miRNA genomic loci that are distinct from and usually distant from other gene types. In contrast, siRNAs generally display less sequence conservation, and they often correspond perfectly to the sequences of known or predicted mRNAs, transposons, or regions of heterochromatic DNA.

2.2. State-of-the-arts for MicroRNA Identification

The strategies for identifying systematically novel miRNAs can be broadly categorized into *in vivo* and *in silico* (Berezikov *et al*., 2006; Ambros *et al*., 2003a). The latter can be subclassified into approaches based on comparative-genomics, machine learning, machine learning coupled with comparative-genomics, and others.

2.2.1. Experimental Approaches

To date, a handful of miRNA genes, namely, *Caenorhabditis elegans lin-4*, *let-7*, and *lsy-6*; *Drosophila melanogaster bantam*, *miR-14*, and *miR-278* were yielded by forward genetic screening coupled with standard positional cloning of genetic loci. In particular, forward genetic screening methods require no prior knowledge of the sequence function. The standard methodology is to apply a chosen mutagen to organisms with a phenotype that was selected to facilitate the screening for the desired type of mutation. For example, when screening for lethal mutations in a specific chromosomal region, an appropriate marker gene should be used. The absence of progeny with the marker phenotype indicates a linked lethal mutation. Forward genetic screens can be used to select for mutations in the entire genome or in localized regions.

Earlier experiment-driven discoveries were low-throughput and technically challenging, since many miRNA mutants might be unrecognized in a phenotype-driven screen due to pathway redundancy. Currently, novel miRNAs were discovered almost exclusively through intensive direct cloning and sequencing of cDNA libraries derived from the size-fractionated RNA transcripts. Breakdown products of mRNA transcripts in the background, endogenous ncRNAs contaminants (e.g., rRNAs, tRNAs, and snRNAs) as well as exogenous siRNAs are dominant players coexisting in the small RNA samples isolated from the cytoplasmic total RNA extracts. To thwart designating them erroneously as putative miRNAs, isolated approximately 22 nucleotides small RNAs are assessed computationally against annotated mRNA and ncRNA databases (Lagos-Quintana *et al*., 2002; Lagos-Quintana *et al*., 2001; Lee and Ambros 2001; Lau *et al*., 2001). Directional cloning routes are neither exhaustive nor straightforward in discovering all the known miRNAs for two reasons. They are highly biased towards abundantly and/or ubiquitously expressed miRNAs that usually dominate the cloned products, rendering the isolation of novel miRNAs difficult (Lagos-Quintana *et al*., 2003). Moreover, miRNAs expressed constitutively at low abundance or have preferentially restrictive/specific temporal

(cell-phase) and spatial (tissue-/cell-type) expression patterns, are intricate to detect experimentally (Lagos-Quintana *et al*., 2001). To express them sufficiently for cloning efforts under controlled cellular conditions and non-abundant cell types is technically involving. In principle, this issue can be overcome by high-throughput deep sequencing of small RNA libraries using Massively Parallel Signature Sequencing (MPSS) (Brenner *et al*., 2000) on an appropriately pooled biological samples (Lu *et al*., 2006).

To be characterized as *bona fide* mature miRNAs, selected small RNAs must be assessed whether they conform according to a combination of criteria for both their expression and biogenesis (Ambros *et al*., 2003a). (**1**) The 22 nucleotides RNA sequence should originate from the genomic regions of the organism from which they were cloned. (**2**) The genomic sequence encoding the novel mature miRNAs should potentially display characteristic hairpin-shaped secondary structures that fold in the absence of large internal loops or bulges especially large asymmetric ones with the lowest Minimum Free Energy of folding (MFE). (**3**) The putative miRNA should occupy entirely one arm of the hairpin, or at least 16 base pairs involving the first 22 nucleotides of the novel mature miRNA embedded within one arm of the fold-back precursor. (**4**) The distinct short RNA transcript should then be validated by experimental means, for example Northern blotting. (**5**) Accumulation of the fold-back precursor should be detected when Dicer is down-regulated.

The short sequence length of small RNAs, however, confers relatively low specificity whereby matching regions are readily encoded in overwhelming number of unwanted genomic segments that can potentially fold into hairpin-shaped structures. To eliminate the over-represented false-positives or simply pseudo hairpins, earlier computation-driven approaches relied on identifying close homologs of these putative pre-miRs as used for *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 putativeorthologues of evolutionary conserved miRNAs genes should conform to the expression and biogenesis criteria (Ambros *et al*., 2003a). Apparently, mere application of simple alignment queries and positive-selection rules is likely to overlook novel families lacking clear homologues to published mature miRNAs.

2.2.2. Comparative-genomics Approaches

Advanced comparative-based identification techniques like *MiRscan* (Lim *et al*., 2003a; Lim *et al*., 2003b), *MIRcheck* (Jones-Rhoades and Bartel 2004), *miRFinder* (Bonnet *et al*., 2004a), *miRseeker* (Lai *et al*., 2003), *findMiRNA* (Adai *et al*., 2005), and *MiRAlign* (Wang *et al*., 2005) were developed to systematically exploited the greater availability of genomic sequences in nematodes, human, insects, and plants. Similar to the computational identification of ncRNA genes, they were largely based on cross-species sequence and structural conservations to identify evolutionarily conserved regions in the genome for miRNA candidates, and to distinguish phylogentically well-conserved pre-miR candidates from irrelevant (often over-represented) genomic dysfunctional hairpins. For example, *MiRscan* (Lim *et al*., 2003a; Lim *et al*., 2003b) relies on the observation that the known miRNAs are derived from phylogenetically conserved stem-loop precursor RNAs with characteristic features. It successfully predicted hundreds of miRNAs in nematodes and human with a high sensitivity. *MiRAlign* (Wang *et al*., 2005) aligns the secondary structure of pre-miRs to detect miRNAs. Typically, conserved regions are first identified by aligning the entire genome of phylogentically related species and masking out those regions most unlikely to be occupied by miRNAs (e.g., tRNAs and rRNAs). Sliding windows of the unmasked regions are folded at both strands by *Mfold* (Zuker 2003) or *RNAfold* (Hofacker 2003), two commonly used RNA secondary structure predictors. The secondary folds are scored according to a set of several characteristic features like MFE, length of the symmetric/asymmetric regions, and size of the terminal loop. The composite scores are thresholded, those high-ranking ones deem similar to pre-miRs published in *Sanger miRBase* (Griffiths-Jones *et al*., 2006) are then reserved for further experimental validation.

Alternatively, an extensive set of novel miRNAs based on genome-wide human-mouse-rat comparisons was identified from a characteristic conservation profile of ten primate species using a technique known as Phylogenetic shadowing (Berezikov *et al*., 2005). Phylogenetic shadowing is a variant of phylogenetic footprinting, which examines genomic sequences of closely related species and takes into consideration the phylogenetic relationship of the set of species analyzed (Boffelli *et al*., 2003). Out of the 69 representative human candidates, 16 were validated with Northern blotting. From which, it was observed that there was a striking drop in conservation for sequences immediately flanking the miRNA hairpins. A similar comparative analysis of the human, mouse, rat, and dog genomes revealed that a proportion of the common regulatory motifs in the promoters and 3' untranslated regions are likely to be associated with

miRNAs (Xie *et al*., 2005).

Evidently, these comparative approaches seem to be utmost promising for genome-wide screening for closely related species, but they are unable to predict non-conserved genes in divergent evolutionary distance with sufficient high sensitivity (Berezikov *et al*., 2005; Boffelli *et al*., 2003). As extensive genomics datasets for computationally intensive multiple genome alignments are involved, this renders identification of miRNAs impossible especially for organisms whose closest relatives have partial or yet-to-start sequenced genomes. Another significant drawback is that non-conserved pre-miRs with genus-specific patterns are likely to evade detection. Thus, identification of pre-miRs that differ significantly or evolve rapidly at the sequence level while retaining their characteristic evolutionary conserved hairpin-shaped structures poses an issue. Pathogenic viral-encoded pre-miRs have been uncovered in *Epstein barr virus*, *Kaposi sarcoma-associated herpesvirus*, *Mouse γ-herpesvirus 68*, *Human cytomegalovirus*, and *Simian virus 40* that share little or no sequence homologies among themselves or with those of hosts (Pfeffer *et al*., 2005; Samols *et al*., 2005; Grey *et al*., 2005; Pfeffer *et al*., 2004), are likely to remain elusive to comparative-based detection.

2.2.3. Machine Learning Approaches

To surmount the technical shortfalls of comparative approaches for distinguishing species-specific and non-conserved pre-miRs, predictors based on *ab initio* or *de novo* methodologies have been extensively developed. A critical and necessary feature for the mature miRNAs biogenesis is that they reside primarily on one arm of the pre-miRs that form characteristic imperfect hairpin-shaped structures. This criterion points to only those small RNA sequences occupying the 20 nucleotides matched regions on one arm of the hairpin-shaped precursors should be curated as novel miRNAs after experimentally validating them. Genome-wide screening for novel pre-miRs is technically complicated considering that the hairpin-shaped structures are rampant in the eukaryotic genomes and are not unique to miRNAs exclusively. These dysfunctional inverted repeats (termed as pseudo hairpins) are genomically prevalent in the *Homo sapiens* (1.1×10^7) (Bentwich *et al*., 2005) and *Caenorhabditis elegans* (4.4×10^4) (Pervouchine *et al*., 2003) genomes. 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 nucleotides in animal and 100−400 nucleotides in plants) and have highly diverse base compositions (Zhang *et al*., 2006b).

De novo or *ab initio* predictors characterize the variable-length sequence of pre-miRs as a fixed-length vector containing exclusively intrinsic descriptors, analogous to the face- or handwriting-pattern recognition techniques. Unlike protein-coding genes possessing statistically significant primary-sequence signals such as the open reading frames (ORFs), promoter motifs, and codon signatures, pre-miRs display defined "hairpin-shaped" secondary structure that have been readily exploited by existing *de novo* methods for reliable and high-throughput detection.

Typically, they first decompose the individual pre-miR into a modularized RNA substructures comprising of dangling termini, (a)symmetric stem, and terminal loop. Derived from these specific regions are a complex array of sequence (e.g., nucleotide composition) and structural characteristics (e.g., thermodynamic stability). This is fashioned analogously to the protein-coding gene identification techniques that scan the genomic regions for signature signals of protein-coding genes without relying on external transcripts or genomic sequences. A supervised machine learning classification algorithm e.g., Support Vector Machine (SVM) is trained on a binary-labeled positive set of genuine pre-miRs and a negative set of pseudo hairpins. Through this inductive machine learning on their feature vectors, a classifier model and a set of decision rules are devised to discriminate between them. With the classification model, any unlabelled non- or well-conserved hairpins can be designated simply as a putative pre-miR or a dysfunctional inverted repeat with higher sensitivity/specificity and significantly efficient than previous comparative methods. (Table 2.1) Generally, better recognition accuracy are obtained according to a combination of structural features like Minimum Free Energy of folding or MFE by *miR-abela* (Sewer *et al*., 2005; Pfeffer *et al*., 2005), normalized MFE (z-score) by *RNAmicro* (Hertel and Stadler 2006); local continuous substructure-sequence attributes by *Triplet-SVM* (Xue *et al*., 2005).

An inaugural and definitive work, *miR-abela* (Sewer *et al*., 2005; Pfeffer *et al*., 2005) compiled 40 distinctive sequence and structural features gathered from the experimental domain knowledge of pre-miRs that obviates the use of comparative genomics information − stem length, length of the longest symmetrical region, number of complementary base pairs in the "relaxed symmetry" region, MFE, number of nucleotides in symmetrical and asymmetrical loops in the "relaxed symmetry" region, and the average size of the asymmetrical loops. The SVM classifier-based method named *miR-abela*, was trained with the binary-labeled feature vectors extracted from human pre-miRs (as positive examples) and random sequences like tRNAs, rRNAs and mRNA genes (as negative examples). It recovered 71.00% of the positive pre-miRs with a remarkably low false-positive rate of ~3.00%. It also predicted ~50 to 100

novel clustered pre-miRs for several species of human, mouse and rat by applying to their genomic regions around already known miRNAs; ~30.00% of these were previously experimentally validated. The validation rate among the predicted cases that were conserved in at least one other species was higher at ~60.00%; many had not been detected by comparative genomics approaches. The significance of *miR-abela* is its ability to detect non-conserved miRNA candidates that did not have any sequence homology to the existing known miRNA genes at the time, demonstrating the power of machine learning in overcoming the limitations of comparative approaches relying on phylogenetic conservation.

The accuracy of predicting novel miRNAs was improved to ~90.00% in human and up to 90.00% for other species, by another *de novo* classifier *Triplet-SVM* (Xue *et al*., 2005). This approach proposed a set of novel encoding features that combines the local continuous structure and sequence information of known pre-miRs' stem-loop structures and represented them as a set of 32 triplet elements − a nucleotide type and three continuous sub-structures e.g., "A(((" and "G(..". Albeit its methodological simplicity, promising performances, and independence of comparative genomics information, *Triplet-SVM* was largely limited to classifying RNA sequences that fold stringently into hairpin secondary structures without containing multiple loops.

Alternatively, *ProMiR* (Nam *et al*., 2005) exploited a probabilistic co-learning technique Hidden Markov Model (HMM) that has a topology of hidden states to discriminate miRNA genes according to their pairwise aligned sequences. Notably, HMM is a statistical model in which the system being modeled is assumed a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. Applying HMM to the identification of miRNAs, *ProMiR* was trained and validated through 5- fold cross validation with a positive dataset comprising of 136 human mature miRNAs and a negative dataset comprising of 1000 extended stem-loop structures randomly extracted from the human genome. It achieved a promisingly low false-positive rate of 4.00%, but compromised for a less performing sensitivity of only 73.00%; out of 23 novel candidates detected, nine were further validated.

Table 2.1: Existing (quasi) *de novo* classifiers for distinguishing novel pre-miRs from genomic pseudo hairpins.

Works	Classifiers	Num	Description of Features	Datasets	Real pre-miRs	Pseudo Hairpins	Sensitivity	Specificity
miR-abela (Sewer *et al*., 2005; Pfeffer *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.	Human	178	5,395	71.00	97.00
ProMiR (Nam *et al*., 2005)	HMM + CI	−	A hairpin structure is represented as a pairwise sequence. Each position of the pairwise sequence has two states, structural and hidden.	Human	136	1,000	73.00	96.00
Triplet-SVM (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(..".	Human Human	30 39	1,000 2,444	93.30 92.30	88.10 89.00
BayesMIRfinder (Yousef *et al*., 2006)	NBI + CI	84	62 secondary structural features derived from the foot, mature, and head of a hairpin-loop; 12 sequence features extracted from the candidate sequence.	Worm Mouse	11 22	150 150	83.00 97.00	96.00 91.00
RNAmicro (Hertel and Stadler 2006)	SVM + CI	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); CI (Comparative genomics information). (Num) Number of features.

2.2.4. Machine Learning with Comparative-genomics Approaches

A relatively recent work *BayesMIRfinder* (Yousef *et al*., 2006) adopted an alternative discriminative machine learning algorithm Naïve Bayesian Induction (NBI) as its underlying classifier algorithm in combination with multi-species genomic data a conservation filter to reduce the number of false positives. NBI is based on "Bayes theorem" and strong independence assumption. Similar to SVM, with the supply of a set of structural and sequence features,

BayesMIRfinder was trained using a variety of miRNAs from multiple organisms to predict novel and nonconserved miRNAs. Notwithstanding its technical novelty, *BayesMIRfinder* relied on the comparative analysis of conserved genomics regions for post-processing of candidates to yield a considerably higher sensitivity of 97.00% and comparable specificity of 91.00% in mouse to existing algorithms.

Another SVM-based work *RNAmicro* (Hertel and Stadler 2006) incorporating 12 sequence and structural descriptors as part of its feature vector, reported an incredibly promising efficiency of 91.16% (sensitivity) and 99.47% (specificity). Two key characteristics of its classification pipeline were: (**1**) computationally expensive multiple sequence alignments were required for its inputs. (**2**) It implemented a structural filter that identified conserved 'almost-hairpins' in a multiple sequence alignment. The filter excluded assessment of alignment windows whose consensus structure contained a stem with less than 10 base pairs or at least 2 hairpins with at least 5 base pairs each, and classified them instantly as non pre-miRs. *RNAmicro* was applied to three independent and genome-wide comparative genomics surveys for candidate functional ncRNAs possessing evolutionary conserved sequence and RNA secondary structures − vertebrate (Washietl *et al*., 2005a), nematode (Missal *et al*., 2006), and urochordate (Missal *et al*., 2005). These datasets were generated from *RNAz* (Washietl *et al*., 2005b) screening methodology (a machine learning technique relying on distinctive features of thermodynamic stability and conservation of secondary structure of functional ncRNAs) that neither incorporate nor provide membership information of disparate classes of ncRNAs; alternatively, *Evofold* (Pedersen *et al*., 2006) could also be used. Annotating the extensive collection of newly identified ncRNAs into specific classes is a resource-intensive and error-prone task, which was first undertaken in an automated manner using *RNAmicro* from the perspective of miRNA. A strong association between the identified miRNAs and those published in previous reports was observed.

2.2.5. Hybrid Approaches

The following works do not belong to any of the three categories: comparative, machine learning, and machine learning coupled with comparative genomics.

89 novel human miRNAs, nearly doubling the number of known human miRNAs, were previously reported using an integrative approach that combined computational identification of hairpin-shaped secondary structures, expression analysis based on microarray profiles, and

sequence directed cloning results (Bentwich *et al*., 2005). A novel 'target-driven' approach was developed for identifying miRNAs (Chan *et al*., 2005) that relied on comparative genomic studies between closely related flies and worms to first screen for miRNA binding sites in the 3' untranslated regions of target mRNAs. Since the miRNA sequences are complementary to some degree to their binding targets, putative mature miRNAs that potentially hybridize to the predicted targets were then identified.

Two independent groups had developed algorithms specifically for viral-encoded miRNAs in small genomes of less than 500 kilobases. *VirMir* (Sullivan *et al*., 2005; Sullivan and Ganem 2005) scanned the viral genome in both orientations with a window of 100 nucleotides in step of 10 nucleotides. The secondary structure of each window was scored and the MFE was computed. The high-scoring candidates would then be validated experimentally by Northern blotting. A refined version of *VirMir*, *Vmir* (Grundhoff *et al*., 2006) had two improvements. First, the hairpin structures were directed to a structural analysis, and a scoring algorithm based on the statistical comparison of a positive and negative training sets were used for classification. Second, microarray analysis was employed to scan the high-scoring candidates. Another research group computationally screened the genome of *Herpes simplex viruses 1* for hairpin-like structures (Cui *et al*., 2006) and obtained a set of pre-miR candidates via several filters, namely, the *%G+C* content, repeats, protein-coding sequence, and MFEs.

2.3. Summary

MicroRNAs (miRNAs) perform critical roles in the gene regulation network by targeting mRNAs for cleavage or translational repression. The ~22 nucleotides mature miRNAs originate from the transcription of long primary miRNAs, which are then processed into precursor miRNAs (pre-miRs) by nuclear RNase III Drosha. Validated miRNAs are involved in the developmental timing and left/right asymmetry of chemoreceptor expression in nematodes, programmed cell death in Drosophila, hematopoietic differentiation in mammals, apoptosis, and metabolism in insects, cellular proliferation, and immune response inhibition in viruses. Since past several years, studies on the biological roles of miRNAs in cancers have been emerging, pointing to miRNA as an invaluable and potential therapeutic target in human diseases.

Detecting systematically miRNAs from a genome using current experimental techniques is labor-intensive and technically difficult, two main challenges gradually being resolved by computational approaches. Comparative genomics methods were first adopted to identify novel

miRNAs in specific animals and plants, according to reports that miRNA genes are conserved in the primary sequences and secondary structures. Obviously, novel miRNAs that have no known close homologies due to the limitation of the data for specie that does not have a closely related one sequenced, or due to the possible evolution of miRNAs, are unable to be identified. *Ab initio* prediction methods were recently developed that rely mainly on the characteristic of hairpin-shaped structures of pre-miRs for identifying novel miRNAs. Major limitations include using phylogenetic information to improve prediction accuracy, restricted to only strict hairpin-shaped structures, and using extrinsic parameters of pre-miRs. Given that a large population of pre-miR-like hairpins can be screened from many genomes, it remains a challenge to distinguish the *bona fide* pre-miRs from pseudo ones.

Chapter 3.
Materials and Methods

3.1. Biologically Relevant Datasets

3.1.1. Precursor MicroRNA Sequences

4,028 curated pre-miRs spanning across 45 species were retrieved from *Sanger miRBase* 8.2 available at *http://microrna.sanger.ac.uk/sequences* as of July 2006 (Griffiths-Jones *et al*., 2006). As strong sequence homologies existed among pre-miRs for both within a single and between different specie(s), the original dataset was filtered to 90% identity using a greedy incremental clustering algorithm (Li and Godzik 2006). Briefly, all the sequences were first sorted in order of decreasing length and the longest one became the representative of the first cluster. Each remaining sequence was compared with the existing representatives and grouped into their cluster if the similarity with any representative was above a given threshold (default value is 0.9), else that sequence became the representative of a new cluster. Consequently, 2,241 non-redundant pre-miRs spanning 41 species categorized into arthropoda, nematoda, verterbrata, viridiplantae, and viruses, were used for analysis. None of the sequences from *G. gorilla*, *M. nemestrina*, *P. paniscus*, and *P. pygmaeus* was retained. See details in Table 3.1.

3.1.2. Functional Non-coding RNA Sequences

All available curated seed ncRNA sequences were retrieved from *Sanger Rfam* 7.0 available at *http://www.sanger.ac.uk/Software/Rfam* as of 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 families. These functional ncRNAs have similar length distribution to the known pre-miRs, and can fold with hairpin(s) or stem-loop(s) (Svoboda and Cara 2006; Storz 2002; Eddy 2001). See Table 3.1 for details.

Briefly, *cis*-regulatory elements are well-conserved untranslated mRNA leader region

capable of adopting alternate structural conformations that result in transcription termination or transcription elongation into the downstream region. For example, the T-box leader regulates transcription of many bacterial aminoacyl-tRNA synthetases, amino acid biosynthesis, and amino acid transport genes using uncharged tRNA as the effector (Winkler *et al*., 2001).

Internal ribosome entry site (IRES) is a nucleotide sequence that allows for translation initiation in the middle of an mRNA. It mimics the 5' cap structure, critical for the assembly of the initiation complex.

Riboswitches are highly conserved RNA regulatory elements, embedded within the 5' untranslated region of biosynthesis genes or operons, and *cis*-modulate their expressions upon binding to metabolite (e.g., guanine and thiamine pyrophosphate), without involving protein cofactors (Soukup and Soukup 2004; Mandal and Breaker 2004; Nudler and Mironov 2004; Vitreschak *et al*., 2004; Winkler and Breaker 2003; Stormo 2003; Lai 2003; Hesselberth and Ellington 2002).

Thermoregulators are *cis*-regulatory elements commonly found in the 5' untranslated regions of mRNAs, whose secondary structure is regulated by temperature. For example, the structural motif of *PrfA* thermoregulator represses translation at 30°C by masking the Shine-Dalgarno sequence, but conformational change frees it for ribosome binding to allow maximal translation when the temperature rises to 37°C (Johansson *et al*., 2002).

Antisenses are characterized by a long hairpin-shaped structure interrupted by several unpaired residues or bulged loops, involved in negative regulation. For instance, the *micF* gene is a *E.coli* stress response gene encoding an untranslated 93 nucleotides antisense that binds to its target *ompF* mRNA of the outer membrane porin gene (Delihas and Forst 2001). It regulates *ompF* expression post-transcriptionally by causing translational repression.

Ribozymes e.g., the Hepatitis δ-virus ribozyme and Hammerhead ribozyme, possess endonuclease function and catalyze a range of reactions such as self-cleavage of hepatitis δ-virus transcript (Puerta-Fernandez *et al*., 2003).

Small nucleolar RNAs (snoRNAs) can be functionally divided into C/D snoRNAs or H/ACA snoRNAs acting as guides for site-specific 2'-O-ribose methylation or as guides for pseudouridylation in the post-transcriptional processing of rRNAs (Weinstein and Steitz 1999).

Spliceosomal RNAs or splicing RNAs e.g., U1-2 and U4-6 (Storz *et al*., 2005), are small nuclear RNAs constituting the spliceosome that process pre-mRNA into mRNA by excising the intronic regions.

Transfer RNAs (tRNAs) exist as approximately 54−93 nucleotides hydrogen-bonded

cloverleaf structures, involved in transporting amino acids to the site of protein synthesis during translation (Sprinzl and Vassilenko 2005).

Group I/II intron RNAs are large self-splicing ribozymes catalyzing their own excision from mRNA, tRNA, and rRNA precursors (Bonen and Vogel 2001; Cech 1990).

3.1.3. mRNA Sequences

31 mRNA sequences with the *GenBank* accession numbers shown in Table 3.1 were randomly selected from *NCBI GenBank* available at *http://www.ncbi.nlm.nih.gov/GenBank* (Benson *et al*., 2005). They tend to fold into complex RNA structures with extremely negative MFEs as previously reported (Freyhult *et al*., 2005).

3.1.4. Pseudo Hairpin Sequences

8,494 pseudo hairpins were extracted from the protein-coding regions (CDSs) according to the UCSC refGene annotation tables (Karolchik *et al*., 2003) and human RefSeq genes from *NCBI GenBank* available at *http://www.ncbi.nlm.nih.gov/RefSeq* (Pruitt and Maglott 2001). As wrongly assumed 'negative samples' can distort the decision boundary of classifier in an unpredictable and/or significant manner, special requirements were imposed on the selection of genomic inverted repeats. First, they must 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 about 90 nucleotides, hairpin-shaped structures with stem at least 8 base pairs including the GU wobble pairs, and Minimum Free Energy of Folding (MFE) of at most -15 kcal/mol. In addition, they fold without multiple loops in their RNA structures as verified by the *RNAfold* program in Vienna RNA package (Hofacker 2003). The *RNAfold* program is utilized with default parameter values (T = 37°C) to predict the secondary structures, based on Zuker's minimum free energy algorithm (Zuker and Stiegler 1981). The current study only utilized optimal folding results.

3.1.5. Random Sequences

In practice, randomization methods are often used to generate random sequences for extracting statistical significance for properties from biological sequences. The random sequences mimic

the "background noise" from which it is possible to differentiate the real biological information. However, a simple randomization method of RNA sequence obscures the frequencies of the mononucleotides and dinucleotides, which are often biased and are crucial for the physical stability of the secondary structure (Clote *et al*., 2005; Bonnet *et al*., 2004b; Katz and Burge 2003; Rivas and Eddy 2000). It is consequently essential to rule out the bias of base composition in the robustness analysis.

In this work, four sets of $R = 10^4$ shuffled or randomized RNAs $\mathbf{r}_n = r_1 r_2 \ldots r_L$ serving as the genomic background are synthesized from each n^{th} native RNA sequence $\mathbf{s}_n = s_1 s_2 \ldots s_L$, using four sequence randomization algorithms, namely the Mononucleotide Shuffling (MS), Dinucleotide Shuffling (DS), Zero-order Markov Model (ZM), and First-order Markov Model (FM) that preserved the exact or nearly exact mononucleotide and dinucleotide base composition as the native sequence, correspondingly. These randomization methods as Adapted from Clote et al. (2005)have been widely used in the thermodynamic stability study of RNA secondary structure (Clote *et al*., 2005; Bonnet *et al*., 2004b; Katz and Burge 2003; Rivas and Eddy 2000). Here, L denotes the length of sequence in nucleotides; biochemical nucleotide at the i^{th} position $r_i \in \sum$ and $s_i \in \sum$, where $\sum$ = [A, C, G, U] is the nucleotide alphabet.

(Figure 3.1) Mononucleotide Shuffling (MS) algorithm implements the "Fisher-Yates shuffle algorithm" that sequentially swaps the mononucleotides at all positions of $\mathbf{s}_n$ with another at a randomly selected position. It consumes $\Theta(L \log L)$ bits and runs in linear time. The order of the shuffled nucleotides is truly random, preserving the mono- but not the di-nucleotide frequencies.

(Figure 3.2) In previous Dinucleotide Shuffling (DS) algorithms (Bonnet *et al*., 2004b; Workman and Krogh 1999), a trinucleotide is randomly selected (e.g., **ATT**), then all the non-overlapping trinucleotides that start and end with the same bases (e.g., **AAT**, **ACT**, **AGT**, and **ATT**) are shuffled at random. This procedure is repeated 10 times the length of the native sequence. Consequently, the shuffled RNA sequences are heuristically-based and may not guarantee to preserve correctly the exact mono and dinucleotide frequencies as the native RNA. In this work, the exact "Altschul-Erikson algorithm" (Altschul and Erickson 1985) was implemented such that it shuffles $\mathbf{s}_n$ while preserving exactly both the mono- and di-nucleotide frequencies. The native and shuffled sequences always share the same first and last nucleotides (Coward 1999). The order of the shuffled nucleotides is 'less random' due to fewer possible dinucleotide preserving permutations.

1. Let random sequence $\mathbf{r}_n$ be a copy of native sequence $\mathbf{s}_n$.
2. **For** nucleotide position i from L to 1, **do**
3. nucleotide position j is sampled from *uniform*(1, i).
4. **If** $i \neq j$, **then** *swap*(r_i, r_j).

Figure 3.1: Pseudo codes of Mononucleotide Shuffling (Fisher-Yates shuffle) algorithm.

1. Let random sequence $\mathbf{r}_n$ be a copy of native sequence $\mathbf{s}_n$.
2. **Foreach** nucleotide r of $\mathbf{r}_n$, **do**
3. create an edge-list L_r of edge-pairs (r, x) with nucleotides r and x occurring as a dinucleotide rx in $\mathbf{s}_n$.
4. Last nucleotide r_L is s_L.
5. **Foreach** nucleotide $r \neq r_L$ of $\mathbf{r}_n$, **do**
6. append an edge-pair randomly selected from L_r to $E(\mathbf{s}_n)$. $E(\mathbf{s}_n)$ contains at most three edge-pairs.
7. Let $G = (V, E)$ be the last-edge graph such that $(r, x) \in V$ and $(r, x) \in E(\mathbf{s}_n)$. **If** any vertex in G is not connected to r_L, **then** goto (4). **Else**, goto (7) as all vertices are connected in graph G to last nucleotide r_L.
8. **Foreach** nucleotide r of $\mathbf{r}_n$, **do**
9. permute the remaining edge-pairs in $L_r - E(\mathbf{s}_n)$. Append to each L_r any edges from $E(\mathbf{s}_n)$ that had been removed.
10. First nucleotide r_1 is s_1.
11. **For** nucleotide position i from 1 to $L - 1$, **do**
12. generate nucleotide r_{i+1} such that $(r_i, r_{i+1}) \in L_r$.

Figure 3.2: Pseudo codes of Dinucleotide Shuffling (Altschul-Erikson) algorithm. Adapted from Clote *et al.* (2005).

(Figure 3.3) For Zero-order Markov Model (ZM) algorithm, a new random sequence $\mathbf{r}_n$ is formed by iteratively adding nucleotide r_i sampled with expected mononucleotide frequencies $F(\sum, \mathbf{s}_n)$. The process is stopped when the random sequence $\mathbf{r}_n$ has exactly the same length as the original $\mathbf{s}_n$. The sequence $\mathbf{r}_n$ is 'truly' random and its mononucleotide frequencies fluctuate about the native ones. The dinucleotide frequencies are completely distorted using this method.

(Figure 3.4) For First-order Markov Model (FM) algorithm, a new random sequence $\mathbf{r}_n$ is formed by first choosing a nucleotide r_1 sampled with expected mononucleotide frequencies $F(\sum, \mathbf{s}_n)$. Iteratively add the next nucleotide r_i+1 sampled with conditional probabilities $P(r_i+1|r_i)$ i.e., the probability of occurrence of a nucleotide at a particular position depends only on the previous nucleotide. The process is stopped when the random sequence $\mathbf{r}_n$ has exactly the same length as the native sequence $\mathbf{s}_n$. The shuffled sequence $\mathbf{r}_n$ is 'truly' random such that its dinucleotide frequencies fluctuate around the native ones but that do not have exactly the same values. Mononucleotide frequencies are not preserved.

1. Compute mononucleotide frequencies $F(\sum, \mathbf{s}_n)$ from native sequence $\mathbf{s}_n$.
2. **For** nucleotide position i from 1 to L, **do**
3. generate nucleotide r_i by sampling with $F(\sum, \mathbf{s}_n)$.

Figure 3.3: Pseudo codes of Zero-order Markov Model algorithm.

1. Compute mononucleotide frequencies $F(\sum, \mathbf{s}_n)$ and conditional probabilities $P(r_i+1|r_i)$ from native sequence $\mathbf{s}_n$.
2. Generate first nucleotide r_1 by sampling with $F(\sum, \mathbf{s}_n)$.
3. **For** nucleotide position i from 2 to L, **do**
4. generate nucleotide r_i by sampling with $P(r_i+1|r_i)$.

Figure 3.4: Pseudo codes of First-order Markov Model algorithm.

Table 3.1: Annotation information of biologically relevant datasets.

Datasets	Counts	Annotation Information
Precursor miRNAs[†]	2,241	Arthropoda (**4**/**171**): *Anopheles gambiae*, *Apis mellifera*, *Drosophila melanogaster*, *Drosophila pseudoobscura* Nematoda (**2**/**189**): *Caenorhabditis briggsae*, *Caenorhabditis elegans* Vertebrata (**19**/**1203**): *Xenopus laevis*, *Xenopus tropicalis*, *Gallus gallus*, *Canis familiaris*, *Ateles geoffroyi*, *Lagothrix lagotricha*, *Saguinus labiatus*, *Macaca mulatta*, *Homo sapiens*, *Pan troglodytes*, *Lemur catta*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Ovis aries*, *Sus scrofa*, *Danio rerio*, *Fugu rubripes*, *Tetraodon nigroviridis* Viridiplantae (**9**/**606**): *Arabidopsis thaliana*, *Glycine max*, *Medicago truncatula*, *Oryza sativa*, *Physcomitrella patens*, *Populus trichocarpa*, *Saccharum officinarum*, *Sorghum bicolor*, *Zea mays* Viruses (**7**/**72**): *Epstein barr virus*, *Herpes simplex virus*, *Human cytomegalovirus*, *Kaposi sarcoma-associated herpesvirus*, *Mouse γ-herpesvirus*, *Rhesus lymphocryptovirus*, *Simian virus*
Non-coding RNAs[‡]	12,387	*Cis*-reg (**77**/**4002**): **X**031, **X**032, **X**036, **X**037, **X**040, **X**041, **X**048, **X**109, **X**114, **X**140, **X**161, **X**164, **X**165, **X**171, **X**172, **X**175, **X**176, **X**179, **X**180, **X**182, **X**183, **X**184, **X**185, **X**192, **X**193, **X**194, **X**196, **X**197, **X**207, **X**214, **X**215, **X**220, **X**227, **X**230, **X**232, **X**233, **X**243, **X**250, **X**252, **X**259, **X**260, **X**290, **X**362, **X**374, **X**375, **X**376, **X**384, **X**385, **X**386, **X**389, **X**390, **X**391, **X**434, **X**436, **X**437, **X**453, **X**454, **X**459, **X**460, **X**463, **X**465, **X**467, **X**468, **X**469, **X**470, **X**481, **X**485, **X**490, **X**491, **X**496, **X**497, **X**498, **X**499, **X**500, **X**501, **X**502, **X**506 *Cis*-reg|frameshift (**5**/**808**): **X**381, **X**382, **X**383, **X**480, **X**507 *Cis*-reg|IRES (**24**/**1201**): **X**061, **X**209, **X**210, **X**216, **X**222, **X**223, **X**224, **X**225, **X**226, **X**228, **X**229, **X**261, **X**387, **X**447, **X**448, **X**449, **X**457, **X**458, **X**461, **X**462, **X**483, **X**484, **X**487, **X**495 *Cis*-reg|riboswitch (**12**/**917**): **X**050, **X**059, **X**080, **X**162, **X**167, **X**168, **X**174, **X**234, **X**379, **X**380, **X**442, **X**504 *Cis*-reg|thermoregulator (**4**/**21**): **X**038, **X**433, **X**435, **X**466 Gene (**24**/**480**): **X**006, **X**013, **X**017, **X**019, **X**023, **X**024, **X**025, **X**044, **X**058, **X**060, **X**062, **X**063, **X**064, **X**100, **X**102, **X**107, **X**169, **X**170, **X**198, **X**199, **X**235, **X**240, **X**262, **X**503 Gene|antisense (**10**/**147**): **X**033, **X**039, **X**042, **X**043, **X**106, **X**236, **X**238, **X**242, **X**388, **X**489 Gene|ribozyme (**9**/**561**): **X**008, **X**009, **X**010, **X**011, **X**030, **X**094, **X**163, **X**173, **X**373 Gene|rRNA (**3**/**1010**): **X**001, **X**002, **X**177 Gene|snRNA (**1**/**28**): **X**066 Gene|snRNA|guide|C/D-box (**165**/**1050**): **X**012, **X**016, **X**046, **X**049, **X**054, **X**055, **X**065, **X**067, **X**068, **X**069, **X**070, **X**071, **X**085, **X**086, **X**087, **X**088, **X**089, **X**093, **X**095, **X**096, **X**097, **X**099, **X**105, **X**108, **X**132, **X**133, **X**134, **X**135, **X**136, **X**137, **X**138, **X**141, **X**142, **X**145, **X**146, **X**147, **X**149, **X**150, **X**151, **X**152, **X**153, **X**154, **X**157, **X**158, **X**159, **X**160, **X**181, **X**186, **X**187, **X**188, **X**189, **X**200, **X**201, **X**202, **X**203, **X**204, **X**205, **X**206, **X**208, **X**211, **X**212, **X**213, **X**217, **X**218, **X**219, **X**221, **X**266, **X**267, **X**268, **X**270, **X**271, **X**273, **X**274, **X**275, **X**276, **X**277, **X**278, **X**279, **X**280, **X**281, **X**282, **X**283, **X**284, **X**285, **X**287, **X**288, **X**289, **X**292, **X**294, **X**295, **X**296, **X**297, **X**299, **X**300, **X**301, **X**304, **X**305, **X**306, **X**308, **X**309, **X**310, **X**311, **X**312, **X**313, **X**314, **X**315, **X**316, **X**317, **X**318, **X**320, **X**321, **X**323, **X**324, **X**325, **X**326, **X**327, **X**328, **X**329, **X**330, **X**331, **X**332, **X**333, **X**335, **X**336, **X**337, **X**338, **X**339, **X**341, **X**342, **X**343, **X**344, **X**345, **X**346, **X**347, **X**348, **X**349, **X**350, **X**351, **X**352, **X**353, **X**355, **X**356, **X**357, **X**358, **X**359, **X**360, **X**361, **X**377, **X**439,

Datasets	Counts	Annotation Information
		X440, **X**441, **X**450, **X**471, **X**472, **X**473, **X**474, **X**475, **X**476, **X**477, **X**478, **X**479, **X**492, **X**493, **X**494, **X**509 Gene|snRNA|guide|H/ACA-box (**71**/**419**): **X**045, **X**056, **X**072, **X**090, **X**091, **X**092, **X**098, **X**139, **X**155, **X**156, **X**190, **X**191, **X**231, **X**263, **X**264, **X**265, **X**272, **X**286, **X**291, **X**293, **X**302, **X**303, **X**307, **X**319, **X**322, **X**334, **X**340, **X**392, **X**393, **X**394, **X**395, **X**396, **X**397, **X**398, **X**399, **X**400, **X**401, **X**402, **X**403, **X**404, **X**405, **X**406, **X**407, **X**408, **X**409, **X**410, **X**411, **X**412, **X**413, **X**414, **X**415, **X**416, **X**417, **X**418, **X**419, **X**420, **X**421, **X**422, **X**423, **X**424, **X**425, **X**426, **X**427, **X**428, **X**429, **X**430, **X**431, **X**432, **X**438, **X**443, **X**482 Gene|snRNA|splicing (**7**/**250**): **X**003, **X**004, **X**007, **X**015, **X**020, **X**026, **X**488 Gene|sRNA (**42**/**233**): **X**014, **X**018, **X**021, **X**022, **X**034, **X**035, **X**057, **X**077, **X**078, **X**079, **X**081, **X**082, **X**083, **X**084, **X**101, **X**110, **X**111, **X**112, **X**113, **X**115, **X**116, **X**117, **X**118, **X**119, **X**120, **X**121, **X**122, **X**124, **X**125, **X**126, **X**127, **X**128, **X**166, **X**195, **X**368, **X**369, **X**370, **X**371, **X**372, **X**378, **X**444, **X**505 Gene|tRNA (**1**/**1114**): **X**005 Intron (**2**/**146**): **X**028, **X**029
mRNAs[§]	31	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

†, e.g., phylum Arthropoda (4/171) has four species of pre-miRs containing 171 sequences. ‡, e.g., family *Cis*-reg (77/4002) has 77 types of ncRNAs containing 4,002 sequences; *miRBase* accession X005 abbreviates RF00005. §, *GenBank* accession numbers.

3.1.6. Four Complete Viral Genomes

They were downloaded from *NCBI GenBank* (Benson *et al*., 2005), namely the *Epstein barr virus* (EBV; 171,823 base pairs; DNA circular; AJ507799.2), *Kaposi sarcoma-associated herpesvirus* (KSHV; 137,508 base pairs; DNA linear; U75698.1), *Mouse γ-herpesvirus 68 strain WUMS* (MGHV68; 119,451 base pairs; DNA linear; U97553.2), and *Human cytomegalovirus strain AD169* (HCMV; 229,354 base pairs; DNA linear; X17403.1).

3.2. Intrinsic RNA Folding Measures (Feature Vector)

Adjusted base pairing propensity, *dP* measures the total number of base pairs present in the RNA secondary structure *S* normalized to the sequence length *L* in nucleotides (Schultes *et al*., 1999). 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 from a vast

population of predicted RNA secondary structures, normalized to the sequence 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). The computation of MFE structures uses a specialized dynamic programming algorithm (Zuker and Stiegler 1981).

MFE Index 1, $MFEI_1$ in Eq. (3.1) is the ratio of *dG* and *%G+C* content (Zhang *et al*., 2006a).

$$MFEI_1 = \frac{dG}{\%G + C}. \tag{3.1}$$

Here, *%G+C* ratio = 100 × (f_G + f_C). f_G and f_C represent the occurring frequencies of nucleotides G and C in a given RNA sequence, respectively.

Adjusted shannon entropy, *dQ* in Eq. (3.2), characterizes the base pairing probability distribution per base (BPPD) in a RNA structure *S* as a chaotic dynamical system (Freyhult *et al*., 2005; Schultes *et al*., 1999; Huynen *et al*., 1997). The local dominance of a single structure within the Boltzmann distribution of alternative secondary structures is strongly correlated with the reliability of the MFE structure. 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.

$$dQ = -\frac{1}{L}\sum_{i<j} p_{ij}\log_2(p_{ij}). \tag{3.2}$$

Here, the McCaskill base pair probability p_{ij} in Eq. (3.3) denotes the probability of base pairing between bases i and j (McCaskill 1990); $\delta_{ij}^{\alpha} = 1$ 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). The implementation of McCaskill's algorithm in *RNAfold* program (Hofacker 2003) was used to compute base-pair probabilities.

$$\begin{aligned} p_{ij} &= \sum_{S_\alpha \in S(\mathbf{x})} P(S_\alpha)\delta_{ij}^{\alpha}, \\ \text{where } P(S_\alpha) &= \frac{e^{-E_\alpha/RT}}{\aleph}, \\ \aleph &= \sum_{S_\alpha \in S(\mathbf{s})} \frac{e^{-E_\alpha}}{RT}. \end{aligned} \tag{3.3}$$

Here, E_α is the free energy of S_α, R is the molar gas constant given by 8.31451 $Jmol^{-1}K^{-1}$, and T is the absolute temperature taken 310.15 K or 37°C.

Adjusted base pair distance, *dD* in Eq. (3.4), 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).

$$dD = \frac{1}{2L} \sum_{S_\alpha, S_\beta \in \mathbf{S(s)}} P(S_\alpha) P(S_\beta) d_{BP}(S_\alpha, S_\beta) \,. \tag{3.4}$$

The base-pair distance $d_{BP}(S_\alpha, S_\beta)$ in Eq. (3.5) between two structures S_α and S_β on **s** is defined as the number of base-pairs not shared by the structures S_α and S_β. Here, the number of base pairs in S_α is $|S_\alpha| = \sum_{i<j} \delta_{ij}^{\alpha}$; $\delta_{ij}^{\alpha} = 1$ if i and j pair, 0 otherwise.

$$\begin{aligned} d_{BP}(S_\alpha, S_\beta) &= \left|S_\alpha \cup S_\beta\right| - \left|S_\alpha \cap S_\beta\right| \,, \\ &= \left|S_\alpha\right| + \left|S_\beta\right| - 2\left|S_\alpha \cap S_\beta\right| \,, \\ &= \sum_{i<j} \left(\delta_{ij}^{\alpha} + \delta_{ij}^{\beta} - 2\delta_{ij}^{\alpha}\delta_{ij}^{\beta}\right) . \end{aligned} \tag{3.5}$$

The computable form of *dD* in Eq. (3.6) can be obtained by substituting the Eq. (3.5) into Eq. (3.4), expanding the terms, and simplifying with known definition of p_{ij} in Eq. (3.3).

$$\begin{aligned} dD &= \frac{1}{2L} \sum_{S_\alpha, S_\beta \in \mathbf{S(s)}} \left[P(S_\alpha) P(S_\beta) \sum_{i<j} \left(\delta_{ij}^{\alpha} + \delta_{ij}^{\beta} - 2\delta_{ij}^{\alpha}\delta_{ij}^{\beta}\right) \right] , \\ &= \frac{1}{2L} \sum_{i<j} \left[\begin{array}{c} \sum_{S_\alpha \in \mathbf{S(s)}} P(S_\alpha)\delta_{ij}^{\alpha} \sum_{S_\beta \in \mathbf{S(s)}} P(S_\beta) \\ + \sum_{S_\alpha \in \mathbf{S(s)}} P(S_\alpha) \sum_{S_\beta \in \mathbf{S(s)}} P(S_\beta)\delta_{ij}^{\beta} \\ -2 \sum_{S_\alpha \in \mathbf{S(s)}} P(S_\alpha)\delta_{ij}^{\alpha} \sum_{S_\beta \in \mathbf{S(s)}} P(S_\beta)\delta_{ij}^{\beta} \end{array} \right] , \\ &= \frac{1}{2L} \sum_{i<j} \left[p_{ij} + p_{ij} - 2p_{ij} \bullet p_{ij} \right] , \\ &= \frac{1}{L} \sum_{i<j} p_{ij}(1 - p_{ij}) \,. \end{aligned} \tag{3.6}$$

Second (or the Fiedler) eigenvalue, *dF* in Eq. (3.7), measures the compactness of a tree-graph $G = (V, E)$ (Gan *et al*., 2004; Fera *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. See Appendix A for details.

$$dF = FidlerEigen[\mathbf{L}(G)] \Leftrightarrow \mathbf{L}(G)\mathbf{X}=\lambda\mathbf{X}\,. \tag{3.7}$$

MFE Index 2, $MFEI_2$ in Eq. (3.8) is the ratio of *dG* and the number of stems *m*, which are structural motifs containing more than three contiguous base pairs.

$$MFEI_2 = \frac{dG}{m}\,. \tag{3.8}$$

***Z*-scores of RNA folding measure** or **normalized feature vectors** i.e., the *Z*-score $Z(\mathbf{s}_n)$ in Eq. (3.9) normalizes the feature $S(\mathbf{s}_n) \in [dG, dP, dQ, dD, dF]$ of n^{th} native RNA sequence $\mathbf{s}_n$. $Z(\mathbf{s}_n)$ is defined as the number of standard deviations by which $S(\mathbf{s}_n)$ differs from the mean of inferred $R = 10^4$ randomized RNA sequences $\mathbf{r}_n$. The corresponding *Z*-scores of $S(\mathbf{s}_n) \in [dG, dP, dQ, dD, dF]$ are denoted as *zG*, *zP*, *zQ*, *zD*, and *zF* using the four sequence randomization algorithms.

$$\begin{aligned} Z(\mathbf{s}_n) &= \frac{S(\mathbf{s}_n) - \mu_n}{\sigma_n}, \\ \text{where } \sigma_n^2 &= \frac{1}{R-1}\sum_{i=1}^{R}\left[S_i(\mathbf{r}_n) - \mu_n\right]^2. \end{aligned} \tag{3.9}$$

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 $S(\mathbf{s}_n)$ for *R* random RNA sequences $\mathbf{r}_n$. The entire set of *R* random sequences $\mathbf{r}_n$ is synthesized via a Monte Carlo randomization approach (Clote *et al*., 2005; Bonnet *et al*., 2004b; Workman and Krogh 1999) e.g., by the "Altschul-Erikson algorithm" (Altschul and Erickson 1985), an exact form of dinucleotide shuffling algorithm. 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.

3.3. Statistical Analysis

To measure the statistical differences inherent within pre-miRs' global structural and intrinsic stability features as well as to compute the probability that the samples are drawn from the same distribution, the non-parametric Kruskal-Wallis one-way Analysis of Variance (ANOVA) or non-parametric Mann-Whitney-Wilcoxon (Wilcoxon rank-sum) were conducted. The former tests for statistically significant difference in the median values ($p < 0.001$) among the experimental groups against the control are greater than would be expected by chance. To isolate the groups that differ from the control, Dunn's method of multiple comparisons test is conducted at $p < 0.01$. It does not include an adjustment for ties but allows the sample sizes of the experimental groups to be different. The latter tests for statistically significant difference in the median values between two experimental groups ($p < 0.001$). Unlike parametric statistical test like student t-test, both ANOVA and Wilcoxon rank-sum compare the ranks of the data values instead of the actual data values. Thus, they are robust to samples drawn from populations with non-normal distribution or have unequal variances. (Systat® SigmaPlot™ 9.0 and SigmaStat™ 3.11).

To quantify the correlation between measures for native pre-miRs, the Pearson correlation coefficients $C_p(f, g)$ in Eq. (3.10) is computed; statistically significant at $p < 0.001$. Knowing that C_p is not robust to outliers and to non-Gaussian distributions, as it assumes a pseudo-Gaussian distribution of the dataset. Thus, the results of C_p were also validated against those of non-parametric Spearman-rank C_s (ranks-based) and Kendall's C_k (relative ranks-based) correlation metrics. Both C_s and C_k are robust to samples containing outliers, or drawn from population with unequal variances, non-normality distribution, and non-linearity. (Mathworks® Matlab™ 7.1).

$$C_p(f,g) = \frac{(f-\overline{f})\cdot(g-\overline{g})}{\left\|f-\overline{f}\right\|\left\|g-\overline{g}\right\|}. \tag{3.10}$$

Here, f and g denote the vector of values for measure f and g, respectively.

3.4. De Novo Classifier miPred

3.4.1. Background on Support Vector Machine

Derived from the structural risk minimization principle of statistical learning theory, "Support Vector Machine" (SVM) is a supervised-learning technique that generates a classifier by simultaneously minimizing the empirical classification error and maximizing the geometric margin (Burges 1998; Vapnik 1998). Classifiers based on the special property of SVMs, are also known as maximum margin classifiers. Briefly, given a set of P and N binary-labeled samples ($\mathbf{x}_i$, y_i) as training vectors, the primary objective of SVM is to explicitly construct an optimal hyperplane i.e., a multi-dimensional orthogonal plane that divides the feature vectors $\mathbf{x}_i$ into binary-labeled classes y_i with a maximum margin of separation while maintaining reasonable computing efficiency. Finding this hyperplane translates effectively to solving a convex quadratic programming optimization problem given in Eq. (3.11). This new formulation trades off the two goals of finding a hyperplane with large margin (i.e., minimizing $\|\mathbf{w}\|$), and finding a hyperplane that separates the data well (i.e., minimizing the ξ_i).

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i, \quad C > 0, \\ \text{s.t} \quad & y_i(\mathbf{w}^T\varphi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i. \end{aligned} \tag{3.11}$$

Here, $y_i \in \pm 1$ represents the positive (+1) and negative (−1) labeled classes to which the i^{th} vector $\mathbf{x}_i \in R^M$ (i = 1, 2,…, $P + N$) having M attributes; b is a measure of the perpendicular distance from the hyperplane in the direction of $\mathbf{w}$ to the origin; soft-margin slack variable ξ_i measures the degree of misclassification for $\mathbf{x}_i$; C is the penalty parameter of the training error.

Typically, the training vectors of input variables $\mathbf{x}_i$ are not linearly separable and must be transformed uniquely to high-dimensional feature space by the function φ. SVM handles this non-linearity by simply incorporating a kernel transformation in order that only the function $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j)$ is required for training. A commonly used kernel is the Gaussian Radial Basis Function (RBF) kernel in Eq. (3.12), which maps the data to the Hilbert space of infinite dimensions. The parameter γ (the radius) controls the degree of smoothing of the decision surface in input space. Small values lead to an extremely flat and smooth decision surface, while large values tend to give a very convoluted decision surface that fits tightly around the training points. (Figure 3.5) General application of SVM is conducted using three straightforward steps,

namely the 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).

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}, \quad \text{where } \gamma = \frac{1}{2\sigma^2}. \tag{3.12}$$

After obtaining the classifier model, any unlabeled testing instance $\mathbf{x}$ can then be classified according to the decision function in Eq. (3.13).

$$f(\mathbf{x}) = \operatorname{sgn}\left[\mathbf{w}^T \varphi(\mathbf{x}) + b \right]. \tag{3.13}$$

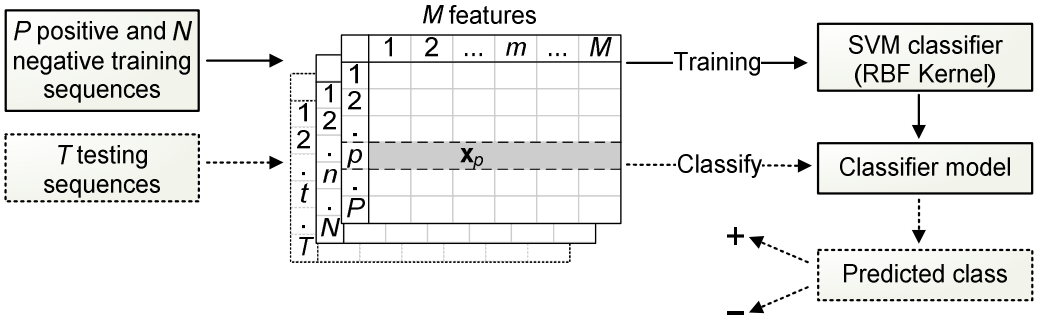

Figure 3.5: Computational pipeline of vectorization and SVM classification.

3.4.2. Grid-search Strategy for Parameter Estimation

All classifier models were generated with the optimal values of hyperparameter pair (C, γ) given in Eq. (3.11) and (3.12), which were obtained from the following model selection procedure. Briefly, at each hyperparameter pair (C, γ) where $C \in [C_1, C_2, \ldots, C_n]$ and $\gamma \in [\gamma_1, \gamma_2, \ldots, \gamma_m]$, the training dataset was randomly partitioned into approximately k distinct equal-sized subsets. Repeating the validation process k times for each subset i.e., retaining a subsets for testing and the remaining $k-1$ subsets for training, the average accuracy of the k models gave the k -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 k-fold LOOCV accuracy rate served as the default setting for training *miPred*. In this work, k = 5 for five-fold cross validation, search space $\log_2 C \in$ [-10, -9, …, 15] and $\log_2 \gamma \in$ [-15, -14, …, 10].

The search was terminated when the mean of the k prediction accuracies $acc(C, \gamma)$ was maximized. The corresponding pair (C, γ) was selected to train the entire training set and to generate the final classifier model. Finally, the classification was conducted on the testing and independent evaluation datasets with "*svm-predict* -b 1".

3.4.3. Training, Testing, and Independent Datasets

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) were 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) were used for testing. They, denoted as TR-H and TE-H, take into account that the training and testing human datasets should be uncorrelated, potentially to avoid overly optimistic classification performances. 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* was 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 *Sanger Rfam* 7.0 (Griffiths-Jones *et al*., 2005); and 31 mRNAs (negatives) from *NCBI GenBank* (Benson *et al*., 2005), respectively. Details of all five datasets can be found at section 3.1 (Ng and Mishra 2007b). To avoid having paralogous miRNAs in the training and testing datasets, the original miRNA dataset download from *Sanger Rfam* 7.0 (Griffiths-Jones *et al*., 2005) was filtered to 90% identity using a greedy incremental clustering algorithm (Li and Godzik 2006).

3.4.4. Implementation of miPred

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 discriminative machine learning tool to address diverse bioinformatics problems (Liu *et al*., 2006; Dror *et al*., 2005; Han *et al*., 2004). Considering that a single criterion to filter pseudo hairpins has not yet been identified, *miPred* undertakes a novel approach that posits the

entire hairpin-shaped structure of each pre-miR can be characterized solely into a feature vector x_i containing 29 RNA global and intrinsic folding attributes, without relying on phylogenetic conservation information (Ng and Mishra 2007b).

17 base composition variables: 16 dinucleotide frequencies $\%\boldsymbol{XY} = f_{XY}/(f_X \times f_Y)$ such that $\boldsymbol{X}$, $\boldsymbol{Y} \in \sum = [A, C, G, U]$, and 1 aggregate dinucleotide frequency $\%G+C$ ratio = $100 \times (f_G + f_C)$. Here, f_X and f_{XY} represent the mononucleotide and dinucleotide frequencies, respectively. RNA intrinsic structural constraints may affect the dinucleotide base compositions and may deviate from approximately the $\%A = \%T$ and $\%C = \%G$ (Xia *et al*., 1998). Previous studies have also suggested that the base composition features $\%G+C$ ratio and dinucleotide frequencies may serve as indicators of ncRNAs (Schattner 2002; Klein *et al*., 2002). Thus, 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 (MFE) denoted as *dG* (Freyhult *et al*., 2005; Seffens and Digby 1999), MFE 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 MFE index 2 $MFEI_2$. 1 topological descriptor: degree of compactness *dF* (Gan *et al*., 2004; Fera *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. The 17 sequence composition variables as well as the non-linear statistical thermodynamics measures *dQ* and *dD* were computed by a custom-made *Perl* program *genRNAStats.pl* 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 structural folds of single sequences and their corresponding MFEs; the topological descriptors *S* and *dF* by a custom-made program *RNAspectral* (see Appendix A for details). After synthesizing the set of random RNA sequences via a custom-made *Perl* program *genRandomRNA.pl*, the normalized variants *zP*, *zG*, *zQ*, *zD*, and *zF* were computed in a similar manner using *genRNARandomStats.pl*. All intensive computations were performed on three clusters comprising of 192 dual-cores computational nodes.

The proposed *miPred*'s binary classifier was developed using libSVM version 2.82 (Chang and Lin 2001), a free implementation of SVM. Samples were randomly selected without replacement via a custom-made python script. Foremost, the 29 attributes of *miPred* were rescaled linearly by the *svm-scale* program to the interval [-1.0, 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 were 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 were optimally calibrated by an exhaustive grid-search strategy using *k*-fold cross-validation as described earlier.

3.4.5. Classification Performance Metrics

Sensitivity or recall (SE), Specificity (SP), Accuracy (ACC), F-measure (Fm) (Liu *et al*., 2006), and Matthew's Correlation Coefficient (MCC) (Bhasin *et al*., 2006) are defined in Eq. (3.14). 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. Here True Positives (*TP*), False Negatives *(FN)*, False Positives (*FP*), and True Negatives (*TN*) denote the number of true/false samples (which are pre-miRs in this work) that are detected/missed by the classifier, correspondingly. Figure 3.6 shows a two by two confusion matrix containing the information about the actual and predicted outcomes evaluated by a binary classification.

$$
\begin{aligned}
\mathrm{SE} &= \frac{TP}{TP+FN}, \\
\mathrm{SP} &= \frac{TN}{TN+FP}, \\
\mathrm{ACC} &= \frac{TP+TN}{TP+TN+FP+FN}, \\
\mathrm{Fm} &= \frac{2(SP \bullet PPV)}{SP+PPV}, \\
\text{where } PPV &= \frac{TP}{TP+FP}, \\
\mathrm{MCC} &= \frac{TP \bullet TN - FN \bullet FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}.
\end{aligned}
\tag{3.14}
$$

Briefly, SE (or SP) measures the proportion of actual positives (or negatives) which are correctly identified; a test with a high SE (or SP) has fewer Type II errors (or Type I error rate),

and *vice versa*. ACC is the proportion of true results (both TP and TN) in the experiment, and measures how well a binary classification test correctly identifies or excludes a sample. When a binary-labeled dataset is unbalanced i.e., the number of positive and negative samples differ greatly like ratio of 1:5 or 1:10, the ACC of a classifier is not representative of the true performance of the classifier. Unlike ACC, Fm and MCC account for unbalanced datasets and are regarded as balanced measures. Fm is the harmonic mean of SE and positive predictive value (PPV).

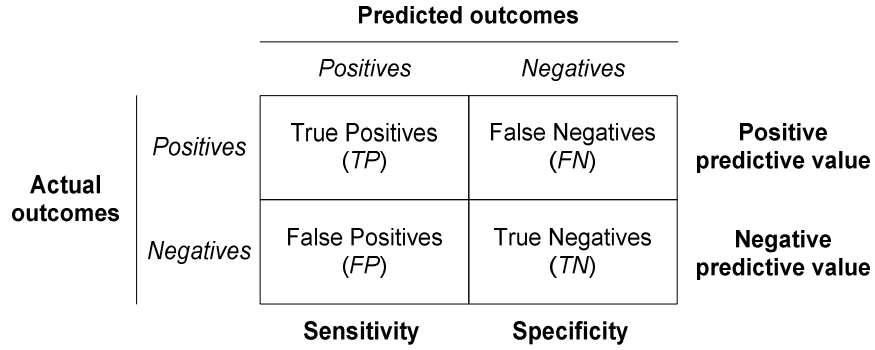

Figure 3.6: Confusion matrix for a binary-class classifier.

The "quality" of a binary classification is commonly shown by the Receiver Operating Characteristic curve (ROC) that plots the trade-off between the SE and the false-positive rate (FPR = 1 − SP) across all possible classification thresholds (Hou *et al*., 2003). The normalized area under the ROC curve, denoted simply as the AUC or ROC score, is a measure of the discriminative power of the classes using the given features and classifier. AUC has the advantage over the ROC of quantifying the performance over the full range of classification costs. The AUC can be interpreted as the probability that two random samples selected from two classes will be ranked correctly, and is invariant to changes in class proportions (unlike ACC). It ranges [0.5, 1.0]; closer to 0.5 (about the upward diagonal) for a totally random classifier for non-distinguishable classes; near to 1.0 (along the left-top boundary) signify a perfect classifier for separable classes (Lasko *et al*., 2005). An efficient algorithm for computing the AUC or ROC score (Hou *et al*., 2003) is shown in Figure 3.7.

1. **Output variable**: $AUC = 0$.
2. **Local variables**: $TP = 0$; $FP = 0$.
3. Sort the SVM scores of the positve and negative test samples. This gives a single column of sorted class labels [1, -1], which is denoted as *sortedlabels*.
4. **Foreach** *label* of *sortedlabels*, **do**
5. **If** *label* == 1, **then** $TP = TP + 1$.
6. **Else** $FP = FP + 1$; $AUC = AUC + TP$.
7. **If** $TP == 0$, **then** $AUC = 0$.
8. **Else if** $FP == 0$, **then** $AUC = 1$.
9. **Else** $AUC = AUC/(TP * FP)$.

Figure 3.7: Pseudo codes for computing efficiently AUC or ROC score. Adapted from Hou *et al.*, (2003).

3.4.6. 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. (3.15). The larger their values for the i^{th} feature, the more likely this feature possesses discriminative importance/power.

$$F1 = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\left|\sigma_i^+ + \sigma_i^-\right|},$$
$$F2 = \frac{\left(\mu_i^+ - \bar{\mu}_i\right)^2 + \left(\mu_i^- - \bar{\mu}_i\right)^2}{\left(\sigma_i^+\right)^2 + \left(\sigma_i^-\right)^2}. \quad (3.15)$$

Here μ_i^+ / μ_i^- and σ_i^+ / σ_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.

3.4.7. Benchmarking miPred

Both *Triplet-SVM* (Xue *et al*., 2005) and Naïve Bayesian Classifier (*NBC*) served as

independent baseline models to benchmark the performance improvements or deterioration (if any) of *miPred*. The original *Triplet-SVM* was previously trained on 163 human pre-miRs and 168 pseudo hairpins using the older libSVM version 2.36 with the "-b 1" option disabled. Here, *Triplet-SVM* was trained on randomly selected 200 human pre-miRs and 400 pseudo hairpins using the latest libSVM version 2.82 (the "-b 1" option is enabled) and the optimal hyperparameter pair (C, γ). *Triplet-SVM* was 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* available at *http://fuzzy.cs.unimagdeburg.de/~borgelt/bayes.html*, was used for training and testing with the exact samples and attributes employed by *Triplet-SVM* and *miPred*; denoted as *Triplet-SVM-NBC* and *miPred-NBC*. For training, "*bci* -L1" yielded better classification results than the default "-L0". In theory, *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 $C = (T, F)$.

The detailed prediction performances for miRNAs are found in Table C.1 (*miPred*), Table C.2 (*miPred-NBC*), Table C.3 (*Triplet-SVM*), and Table C.4 (*Triplet-SVM-NBC*); their mean sensitivities and specificities are summarized in Table C.5. The detailed prediction performances for non-miRNAs ncRNAs comparing the four classifiers are found in Table C.6; their mean specificities are summarized in Table C.7.

3.5. Availability of Datasets and Software

Supplemental materials including the entire datasets (RNA sequences in FASTA format), source codes (implementation of RNAspectral in *ANSI C*, shuffling/randomizing algorithms in *Perl*, and *miPred* in *Perl*), raw results (feature extraction of RNA sequences), and auxillary (*Bash* and *Perl*) scripts are available publicly at *http://web.bii.a-star.edu.sg/~stanley/Publications*.

Chapter 4.
Unique Folding of Precursor MicroRNAs: Quantitative Evidence and Implications for De Novo Identification

4.1. Comparison between Vertebrate and Plant Precursor MicroRNAs

Among the arthropoda, nematoda, verterbrata, viridiplantae, and viruses available from *Sanger miRBase* 8.2 (Griffiths-Jones *et al*., 2006), no orthologous miRNA gene shared by vertebrates and plants has ever been reported (Anthony and Peter 2005). Pathogenic viral-encoded pre-miRs present in *Kaposi sarcoma-associated herpesvirus*, *Mouse γ-herpesvirus 68*, and *Human cytomegalovirus* should be treated as exceptions, though they have also been demonstrated to neither share significant sequence homology with known host pre-miRs nor among themselves (Pfeffer *et al*., 2005; Samols *et al*., 2005; Grey *et al*., 2005). Viral-encoded pre-miRs do not possess genes homologous to host miRNA processing proteins e.g., Drosha, Dicer, and RISC, but are likely to hijack these proteins to facilitate their viral replication after infecting the host cells (Sarnow *et al*., 2006). Despite the apparent similarities of miRNAs biogenesis between vertebrates and plants, their evolutionarily ancient processing pathways (beyond 400 million years ago) were not operating in a common ancestor and could have evolved independently from a more ancient system (Anthony and Peter 2005). This suggests that both vertebrate and plant pre-miRs are likely to exhibit distinct folding features that warrant careful structural analysis.

Vertebrate and plant pre-miRs have significantly distinct $MFEI_2$, $MFEI_1$, *%G+C*, *dP*, *dG*, *dQ*, *dD*, and *dF* from ncRNAs and mRNAs ($p < 0.001$). (Figure 4.1 and Figure 4.3) Foremost, the sequence length (in nucleotides) differs considerably between and among pre-miRs

(vertebrate; 90.4522 ± 0.4164 and plants; 137.9175 ± 2.0309), ncRNAs (frameshift; 53.2599 ± 0.2543 to IRES; 276.0841 ± 2.4342), and mRNA (332.3226 ± 16.3064). The sequence lengths of ncRNAs and mRNAs are strongly and positively correlated with their Minimum Free Energy of Folding (MFE), as previously demonstrated (Zhang *et al*., 2006a; Bonnet *et al*., 2004b; Seffens and Digby 1999). Longer sequence length tends to results in a greater degree of freedom such that the native RNA sequences can fold into complex secondary structures with corresponding higher thermo-stability or lower MFEs. By normalizing the MFE with the sequence length, the normalized MFE *dG* ensures that it serves as a comparable measure without unduly penalizing the shorter pre-miRs or favoring the longer mRNAs (Zhang *et al*., 2006a; Freyhult *et al*., 2005; Seffens and Digby 1999). In agreement with earlier findings (Zhang *et al*., 2006a; Freyhult *et al*., 2005), vertebrate and plant pre-miRs possess statistically distinct *dG* of -0.4308 ± 0.0025 and -0.4456 ± 0.0038 and are the lowest except frameshift (-0.4814 ± 0.0023). Interestingly, a single criterion based on a variant of *dG* greater than a threshold value $\varepsilon = 0.68$ has been applied to genome-wide detection of *Caenorhabditis elegans* pre-miRs (Pervouchine *et al*., 2003). This yielded ~4.4×10^4 stable hairpins localized to ~4.00% of the genome, covering 64.29% (36/56) of the published ones (Lau *et al*., 2001).

Vertebrate and plant pre-miRs possess significantly highest normalized base-pairing propensity *dP* of 0.3518 ± 0.0009 and 0.3545 ± 0.0013, accounting for ~70.36 − 70.9% of their nucleotides forming complementary base pairings within their highly thermo-stable hairpin-shaped structures. Similar >72.00% for *dP* has also been reported corroborating our findings, albeit a smaller dataset of 513 plants pre-miRs across seven species was analyzed (Zhang *et al*., 2006a). The presence of more hydrogen bonds and base pairings in the plant pre-miRs might benefit their recognition, processing, and nucleus-cytoplasm transport (Zhang *et al*., 2006a). Emerging experimental evidence also points to the hairpin motif of vertebrate pre-miRs as a critical feature for the miRNAs maturation (Zeng and Cullen 2004). Human *pre-miR-30* binding by Exportin-5 involved recognition of almost the entire hairpin, except the terminal loop (Zeng and Cullen 2004). A hairpin-shaped structure >16 base pairs was required for detectable binding and >18 base pairs for high-affinity binding such that the stacking of contiguous paired nucleotides tended to reduce the MFE of the overall folded structure for greater thermo-stability. Contrary to the common belief that the unpaired regions tended to disrupt the RNA structure with greater MFE, deleting the 2 nucleotides bulge of *pre-miR-30* left the binding unaffected or reduced binding modestly, unless the stem length was suboptimal. There was negligible or no significant effect on the correct recognition for varying size of the terminal loop, until it was

shortened from the normal 15 to 4 nucleotides. Besides nuclear export of pre-miR, the binding of Exportin-5 served to stabilize the pre-miR in the nucleus and during export by inhibiting the *in vitro* exonucleolytic cleavage (Zeng and Cullen 2004).

Vertebrate and plant pre-miRs encode higher *%A+U* content than *%G+C* content of 48.3079 ± 0.2504 and 46.6719 ± 0.3513; similarly observed by Zhang *et al.,* (2006a). The higher *%A+U* content in the plant pre-miRs (likewise for vertebrate pre-miRs) might possibly serve as a biochemical signal for miRNA biogenesis by the RISC (Zhang *et al*., 2006a). The *%G+C* contents for vertebrate and plant pre-miRs were also found to be not considerably different from mRNAs (50.4626 ± 1.4654) and common families of ncRNAs like *cis*-regulator (48.9672 ± 0.1188), frameshift (46.4785 ± 0.1477), riboswitch (50.5054 ± 0.3381), thermoregulator (42.6490 ± 3.2009), HACA-box snoRNA (46.3048 ± 0.3160), splicing RNA (47.6933 ± 0.3731), sRNA (46.3963 ± 0.3513), tRNA (48.2725 ± 0.3541), and intron (44.7871 ± 0.8350). Unlike the *%G+C* content, the $MFEI_1$ (divides *dG* by *%G+C* content, a newly proposed folding energy score to analyze plant pre-miRs (Zhang *et al*., 2006a) for vertebrate and plant pre-miRs of -0.0091 ± 0.0001 and -0.0096 ± 0.0001 are statistically highest except antisense (-0.0083 ± 0.0001) and frameshift (-0.0104 ± 0.0000). Our finding and another (Zhang *et al*., 2006a) point to the $MFEI_1$ as a potential discriminative criterion to distinguish pre-miRs from mRNAs and ncRNAs, which a recent comparative classifier *RNAmicro* has included into its feature set (Hertel and Stadler 2006).

Notably, vertebrate pre-miRs possess statistically higher normalized Shannon Entropy *dQ* and normalized base pair distance *dD* of 0.1161 ± 0.0025 and 0.0431 ± 0.0009 than plant pre-miRs of 0.1424 ± 0.0036 and 0.0502 ± 0.0011. Generally, RNA sequences having relatively high values of both advanced folding measures are either unstructured, or long in length that fold with the assistance of accessory proteins, or have repertoire of alternative (pseudoknot) structures (Freyhult *et al*., 2005). This suggests that vertebrate pre-miRs will likely to fold into well-defined hairpins restricted to relatively fewer alternative conformations, possibly due to shorter sequence length (90.4522 ± 0.4164 nucleotides) compared to plants (137.9175 ± 2.0309 nucleotides). The different "structureness" of vertebrate and plant pre-miRs causes the former to display significantly lowest and distinct *dQ* and *dD* except antisense (0.1336 ± 0.0061 and 0.0468 ± 0.0020). The latter is not significantly unique from *cis*-regulator (0.2124 ± 0.0021 and 0.0689 ± 0.0006), frameshift (0.1396 ± 0.0024 and 0.0552 ± 0.0009), antisense (0.1336 ± 0.0061 and 0.0468 ± 0.0020), snRNA (0.2305 ± 0.0260 and 0.0741 ± 0.0074), and intron (0.1802 ± 0.0089 and 0.0620 ± 0.0026). Maturation of plant miRNA:miRNA* duplex is

performed exclusively by Dicer-like 1 enzyme (DCL1) via two cleavage steps pri-miR → pre-miR → miRNA:miRNA* within the nucleus. In contrast to vertebrates (Zhang *et al*., 2006b; Anthony and Peter 2005), the two reactions are compartmentalized and directed separately by the nuclear Drosha (pri-miR → pre-miR) and cytoplasmic Dicer (pre-miR → miRNA:miRNA*). Moreover, plant pre-miRs are less conserved (conservation of plants mature miRNAs is well preserved) than those in vertebrates (Zhang *et al*., 2006b; Anthony and Peter 2005). Our structural analysis substantiates both experimental findings, pointing to the plant pre-miRs as very transient molecules (Zhang *et al*., 2006b) that possess less "structureness" indicative of lower *dQ* and *dD* compared to their vertebrate counterparts.

Lastly, two newly proposed topological measures were analyzed i.e., degree of compactness *dF* and $MFEI_2$ (divides *dG* by number of stems *m*). Vertebrate pre-miRs have significantly higher *dF* of 0.2197 ± 0.0042 than plant pre-miRs of 0.1251 ± 0.0033. Generally, RNAs possessing lower *dF* have less structured folds (Barash 2004b; Barash 2003) like mRNAs (0.0391 ± 0.0059). Both vertebrate and plant pre-miRs fold into topologically distinct structures with *dF* being statistically different but is not the extreme among mRNAs (0.0391 ± 0.0059) and common families of ncRNAs like frameshift (0.8865 ± 0.0079), IRES (0.0442 ± 0.0013), antisense (0.3734 ± 0.0133), rRNA (0.0933 ± 0.0020), snRNA (0.5372 ± 0.0415), and tRNA (0.5333 ± 0.0093). The other folding measure $MFEI_2$ was inspired by the formation of the critical hairpin-shaped structure in the early stages of miRNA maturation. Reasonably, MFE should be largely localized to the stem(s) within the hairpin such that the higher $MFEI_2$ corresponds to greater thermo-stability per stem. The $MFEI_2$ for vertebrate and plant pre-miRs of -0.0761 ± 0.0013 and -0.0539 ± 0.0010 are significantly different except antisense (-0.0811 ± 0.0030), snRNA (-0.0764 ± 0.0088), and tRNA (-0.0676 ± 0.0007); *cis*-regulator (-0.0793 ± 0.0017), snRNA (-0.0764 ± 0.0088), and intron (-0.0604 ± 0.0029).

In summary, the 1,203 vertebrate and 606 plant pre-miRs are statistically distinct from 12,387 ncRNAs and 31 mRNAs according to the measures $MFEI_2$, $MFEI_1$, *%G+C*, *dP*, *dG*, *dQ*, *dD*, and *dF*. Except for two recent published works investigating 513 plant pre-miRs (Zhang *et al*., 2006a) and 135 pre-miRs from different species (Freyhult *et al*., 2005), no larger-scale and in-depth statistical analysis highlighting these results on the folding characteristics of published pre-miRs have ever been reported.

Vertebrate and plant pre-miRs have significantly distinct *Z*-scores of *dG*, *dQ*, *dD*, *dP*, and *dF* compared to the ncRNAs and mRNAs. (Figure 4.2 and Figure 4.3) Evolutionarily conserved vertebrate and plant pre-miRs possess considerably lowest *zG* except frameshift and antisense,

regardless of the sequence randomization algorithms. Our finding and another (Freyhult *et al*., 2005) affirm the hypothesis that pre-miRs fold into highly thermo-stable secondary structures with significantly lower MFEs relative to their synthetically generated sequence randomized controls (Bonnet *et al*., 2004b; Workman and Krogh 1999). Therefore this unique structural characteristic of vertebrate and plant pre-miRs is not expected to occur by chance, it is indispensable for correct recognition and processing by Dicer-like enzymes (Bonnet *et al*., 2004b). Earlier works (Bonnet *et al*., 2004b; Workman and Krogh 1999) were inconclusive as their dinucleotide shuffling algorithms were heuristically-based and the resulting shuffled RNAs might not guarantee to preserve the exact dinucleotide frequencies as the native RNAs (Clote *et al*., 2005). Instead, considerably larger dataset of pre-miRs and ncRNAs were investigated as well as the exact "Altschul-Erikson algorithm" (Altschul and Erickson 1985) for synthesizing 10^4 dinucleotide shuffled RNAs. Two computational studies (Clote *et al*., 2005; Washietl and Hofacker 2004) also demonstrated that structural ncRNAs displayed lower MFEs than dinucleotide shuffled RNAs, but pre-miRs were not analyzed.

Both *zQ* and *zD* of vertebrate and plant pre-miRs are statistically different and are the lowest except antisense, irrespective of the sequence randomization algorithms. Recent computational study reported that pre-miRs and ncRNAs (like hammerhead ribozymes type III, and tRNAs) possessed significantly fewer *k*-locally optimal structures (potential kinetic traps) than their dinucleotide shuffled RNAs (Clote 2005). Both findings suggest pre-miRs are probable to undergo evolutionary pressure in adopting relatively fewer alternative folds of significantly lower MFEs than the random background, in order to function properly in the post-transcriptional gene regulatory pathway.

Vertebrate and plant pre-miRs report significantly highest *zP* i.e., more complementary base pairings are present in their RNA secondary structures than the genomic background, irrespective of the sequence randomization methods. They also have statistically distinct *zF* except common families of ncRNAs like *cis*-regulator, IRES, thermoregulator, CD-box snoRNA, and HACA-box snoRNA, as well as mRNAs.

In summary, the 1,203 vertebrate and 606 plant pre-miRs are significantly different from the 12,387 ncRNAs and 31 mRNAs, after examining their *zG*, *zQ*, *zD*, *zP*, and *zF* based on 4 sequence randomization algorithms and 10^4 random sequences corresponding to each native RNA. This statistical finding confirms that to reliably identify pre-miRs from the genomic background requires them more than possessing characteristic and well-defined secondary structures of statistically significant MFEs (Washietl and Hofacker 2004; Rivas and Eddy 2000).

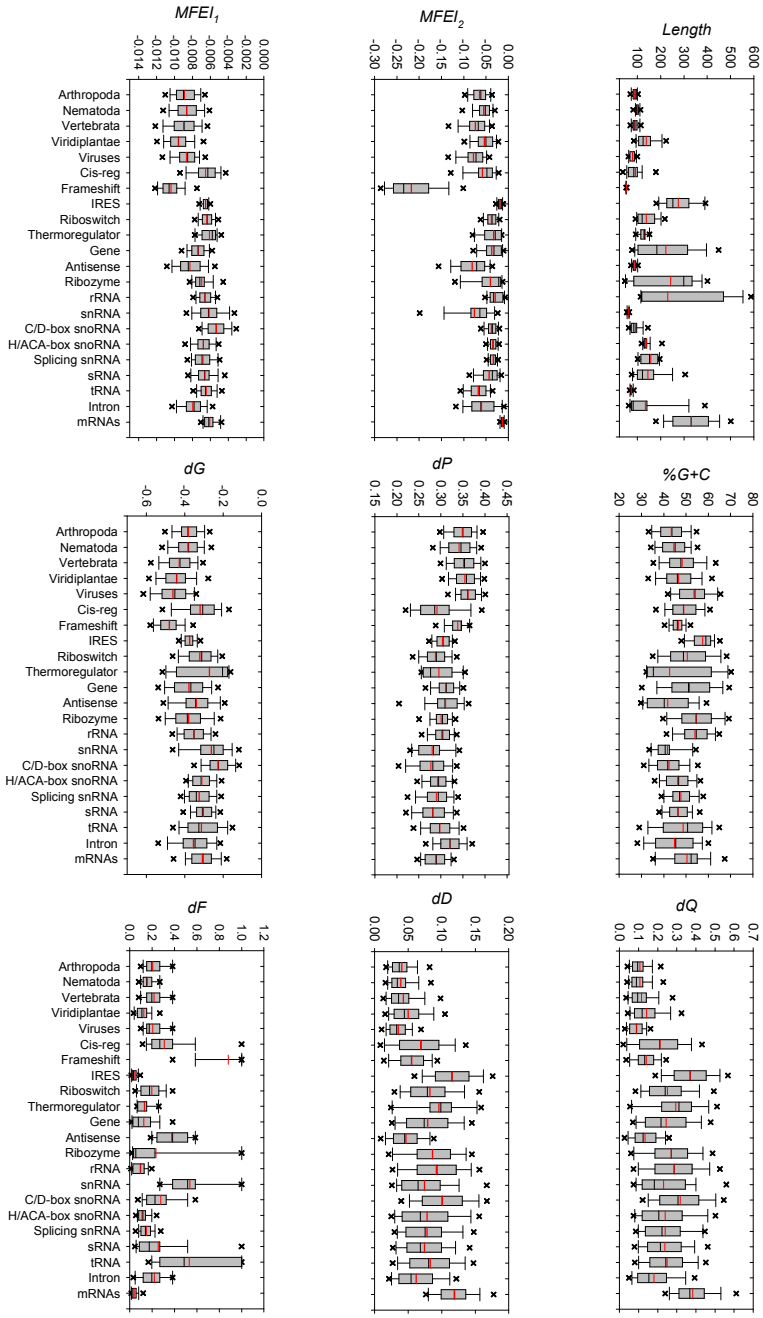

Figure 4.1: Distribution profiles of pre-miRs, ncRNAs, and mRNAs for *Length*, $MFEI_2$, $MFEI_1$, *%G+C*, *dP*, *dG*, *dQ*, *dD*, and *dF*. Box lines indicate the lower quartile, median, mean, and upper quartile; whisker lines extend to the most extreme data value or at most 1.5 times the box height; outliers beyond 5th and 95th percentile are not shown. See Table B.1 for details.

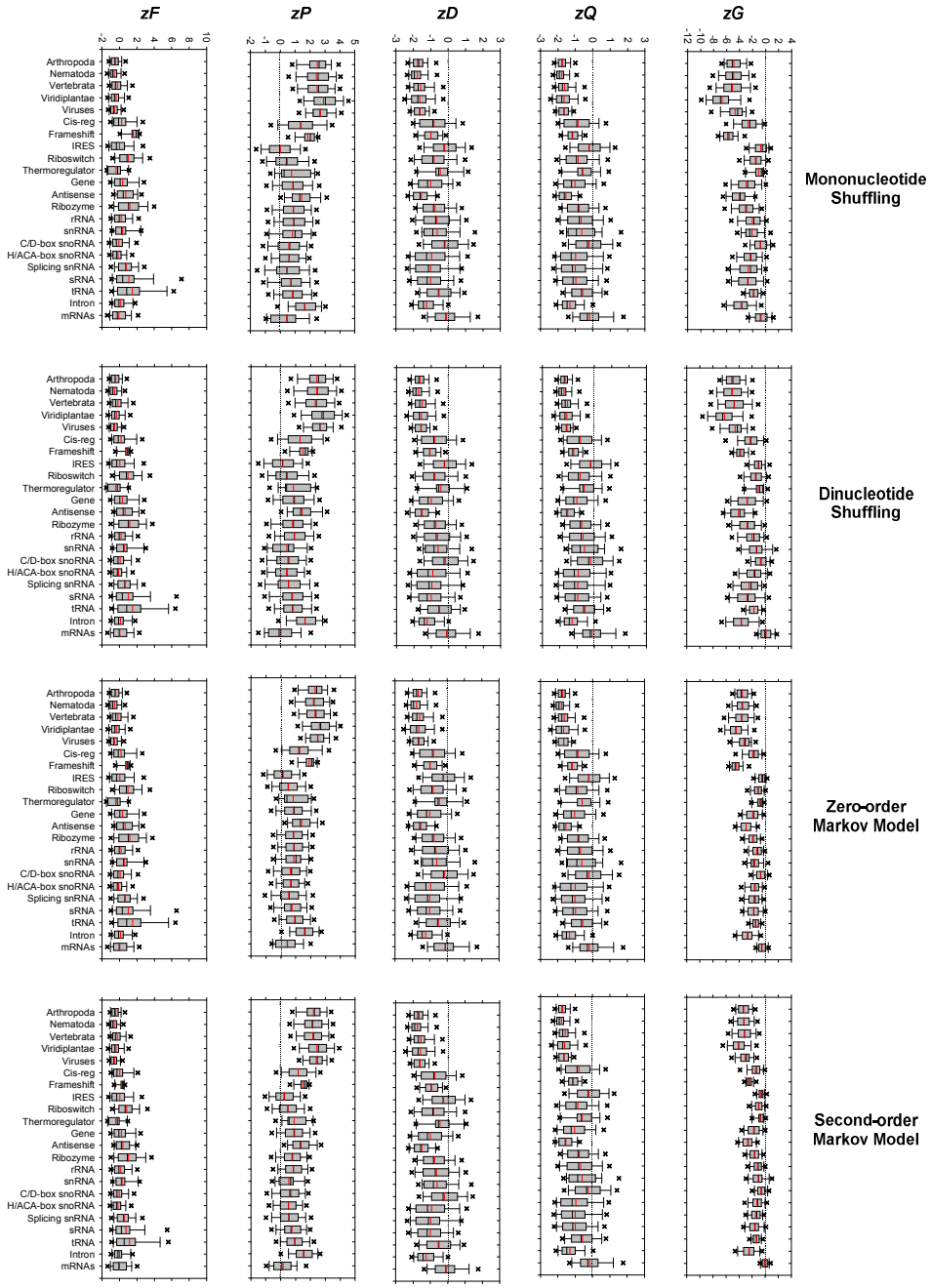

Figure 4.2: Distribution profiles of pre-miRs, ncRNAs, and mRNAs for *zG*, *zQ*, *zD*, *zP*, and *zF*. The horizontal dashed line indicates *Z*-score at zero. Box lines indicate the lower quartile, median, mean, and upper quartile; whisker lines extend to the most extreme data value or at most 1.5 times the box height; outliers beyond 5th and 95th percentile are not shown. See Table B.2 for details.

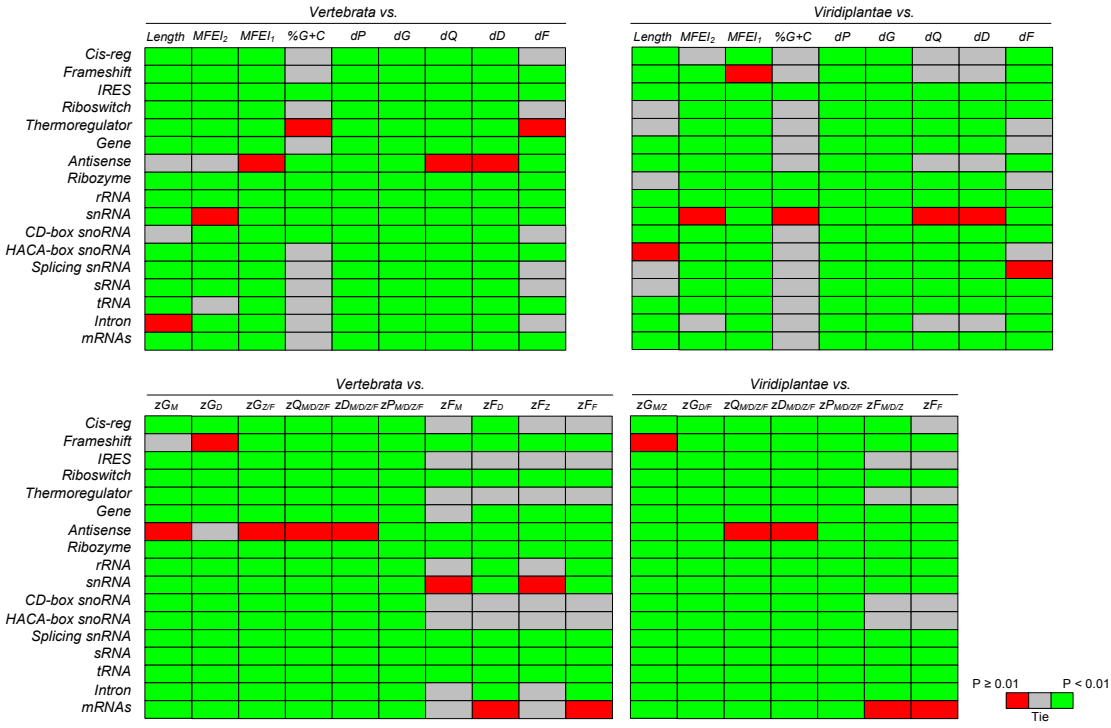

Figure 4.3: Heatmap of vertebrate and plants pre-miRs *vs.* ncRNAs, and mRNAs. $zG_{M/D/Z/F}$ denotes *zG* with respect to Mono- and Di-nucleotide shuffling, Zero- and First-Order Markov Model; green represents statistically different median; red for no statistical difference; grey for ties according to the ANOVA ($p < 0.001$) and Dunn's Method of multiple comparisons tests ($p < 0.01$). See Table B.3 for details.

4.2. Comparison with Previous Studies on Structural Folding Analysis of ncRNAs and mRNAs

(Figure 4.2 and Figure 4.3) For completeness of this large-scale study, three notable points were outlined to revisit previous works investigating whether ncRNAs and mRNAs fold into statistically significant and thermodynamically stable secondary structures. First, 51 mRNAs had significantly lower MFEs than their corresponding sets of 10 mononucleotide shuffled RNAs (Seffens and Digby 1999) and a subset of 46 mRNAs did not display any statistically lower MFEs than their corresponding sets of 10 dinucleotide shuffled RNAs (Workman and Krogh 1999). Our study (mononucleotide shuffling; -0.7223 ± 0.2089 and dinucleotide shuffling; 0.1021 ± 0.1625) and another using dinucleotide shuffling (Freyhult *et al.*, 2005)

support both previous conclusions (Workman and Krogh 1999; Seffens and Digby 1999). Unique to this work, the mRNAs were observed to have considerably lower MFEs than the genomic background for Zero-order markov model (-0.4770 ± 0.1098), but not for First-order markov model (-0.0830 ± 0.0845).

Second, our investigated 1114 tRNAs possess significantly lower MFEs than the genomic background for the four sequence randomization methods. This finding agrees with earlier results (Clote *et al*., 2005; Freyhult *et al*., 2005; Washietl and Hofacker 2004) that relied on dinucleotide shuffled RNAs, but differs from another work (Workman and Krogh 1999) in which the dinucleotide shuffling algorithm was heuristically-based as previously explained (Clote *et al*., 2005). Similar findings were reported for the hammerhead ribozymes type III (Clote *et al*., 2005; Washietl and Hofacker 2004), spliceosomal RNAs (Clote *et al*., 2005; Washietl and Hofacker 2004), riboswitches (Clote *et al*., 2005), and introns (Washietl and Hofacker 2004) that have considerably lower MFEs than corresponding sets of dinucleotide shuffled RNA sequences.

Third, previously discussed (Clote *et al*., 2005; Bonnet *et al*., 2004b; Workman and Krogh 1999), the controls serving as the genomic background would give erroneous conclusions if they destroyed certain non-random composition of the native sequence. Our results highlight that detectable systematic bias of *zG* distribution profiles exist among the four sequence randomization algorithms. Generally, the mean *zG* for pre-miRs, ncRNAs, and mRNAs are ordered from the lowest mononucleotide shuffling, marginally below those of dinucleotide shuffling, followed by Zero- and First-Order Markov Model. This result agrees with earlier works (Clote *et al*., 2005; Bonnet *et al*., 2004b; Workman and Krogh 1999) that disrupting the naturally occurring biases in the inherent dinucleotide frequencies of the sequences base composition should be avoided for determining the significance of secondary structure. Preserving the dinucleotide frequencies of the native sequences is critical so as not to affect the critical energy contributions of stacked base pairs and the corresponding accuracy of the RNA structural predictions (Clote *et al*., 2005; Bonnet *et al*., 2004b; Workman and Krogh 1999).

4.3. Vertebrate and Plant Precursor MicroRNAs are Uniquely Different from Pseudo Hairpins

To elucidate the unique folding of pre-miRs present in vertebrates and plants, the preceding two

statistical experiments were repeated by evaluating them against 8,494 pseudo hairpins instead of ncRNAs and mRNAs. Pseudo hairpins are genomic inverted repeats extracted from the protein coding regions of human RefSeq genes with no known alternative splicing (AS) events. They were first introduced as negative samples in *Triplet-SVM* (Xue *et al*., 2005), a *de novo* classifier based on triplet-encoding features e.g., "A(((" and "G(..". However, no structural analysis or comparison to published pre-miRs has been reported about them.

(Figure 4.4 and Figure 4.6) Generally, the vertebrate and plant pre-miRs have significantly higher *dP* and *dF* as well as lower $MFEI_2$, $MFEI_1$, *%G+C*, *dG*, *dQ*, and *dD* than pseudo hairpins ($p < 0.001$). (Figure 4.5 and Figure 4.6) The distribution profiles of vertebrate and plant pre-miRs for *zG*, *zQ*, *zD*, and *zP* differ distinctively from pseudo hairpins ($p < 0.001$), irrespective of the sequence randomization algorithms. Unlike pseudo hairpins, pre-miRs tend to fold into secondary structures with significantly higher thermodynamic structural stability (lower *zG*), fewer alternative folds (lower *zQ* and *zD*), and more base pairings (higher *zP*). Except plants, vertebrate pre-miRs clearly have significantly higher *zF* (more compactness) than pseudo hairpins ($p < 0.001$).

In summary, both findings invalidate conclusively the hypothesis that pseudo hairpins share comparable degree of structural folding characteristics with known vertebrate and plant pre-miRs. Our statistical results clearly points to the $MFEI_2$, $MFEI_1$, *%G+C*, *dP*, *dG*, *dQ*, *dD*, and *dF* as well as *zG*, *zQ*, *zD*, *zP*, and *zF* as potential discriminative descriptors. They effectively expand the triplet-encoding features in *Triplet-SVM* (Xue *et al*., 2005) to classify more accurately the genuine pre-miRs from pseudo hairpins in genome-wide screens.

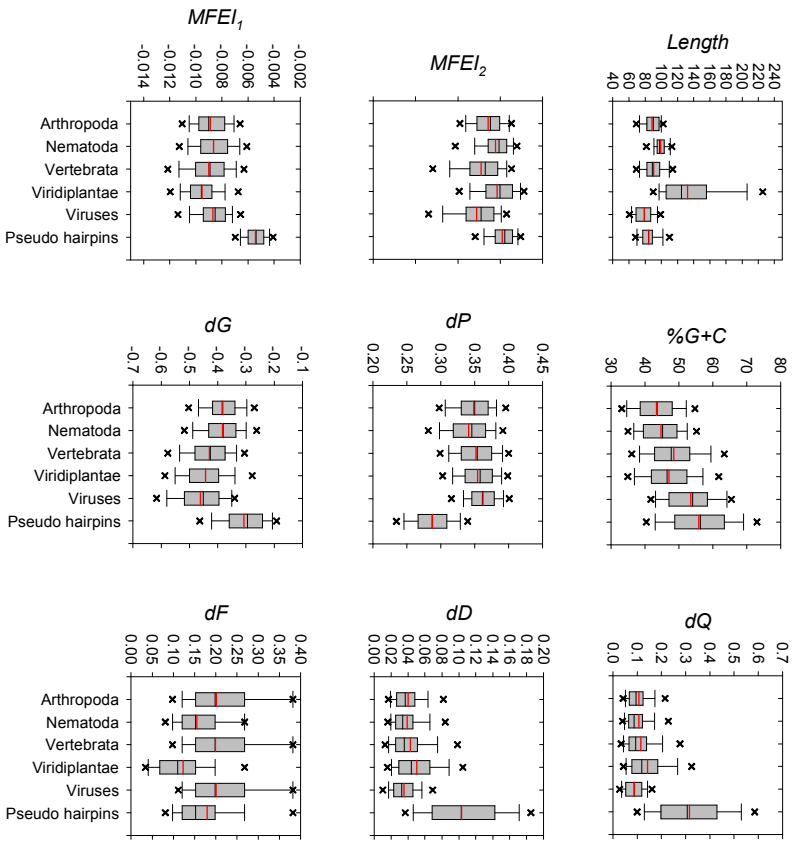

Figure 4.4: Distribution profiles of the pre-miRs for *Length*, $MFEI_2$, $MFEI_1$, *%G+C*, *dP*, *dG*, *dQ*, *dD*. Box lines indicate the lower quartile, median, mean, and upper quartile; whisker lines extend to the most extreme data value or at most 1.5 times the box height; outliers beyond 5th and 95th percentile are not shown. See Table B.1 for details.

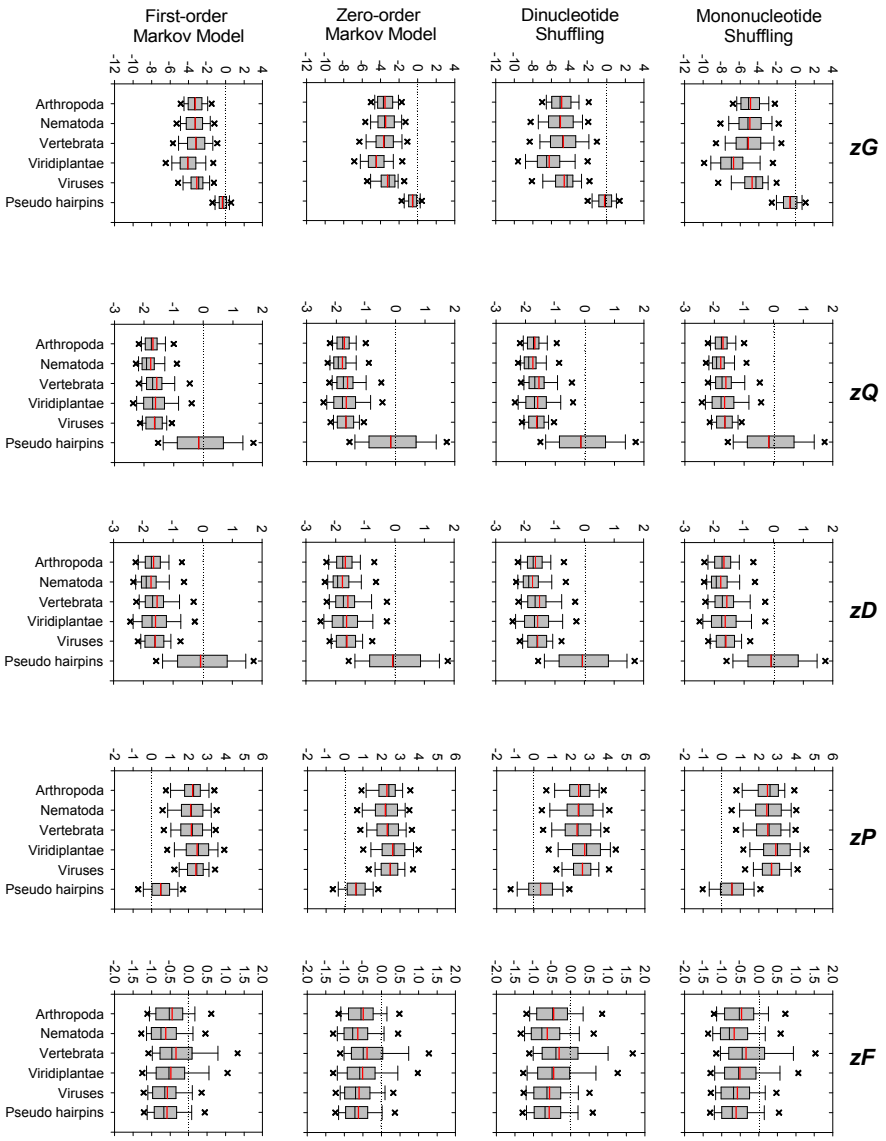

Figure 4.5: Distribution profiles of the pre-miRs for *zG*, *zQ*, *zD*, *zP*, and *zF*. The horizontal dashed line indicates Z-score at zero. Box lines indicate the lower quartile, median, mean, and upper quartile; whisker lines extend to the most extreme data value or at most 1.5 times the box height; outliers beyond 5th and 95th percentile are not shown. See Table B.2 for details.

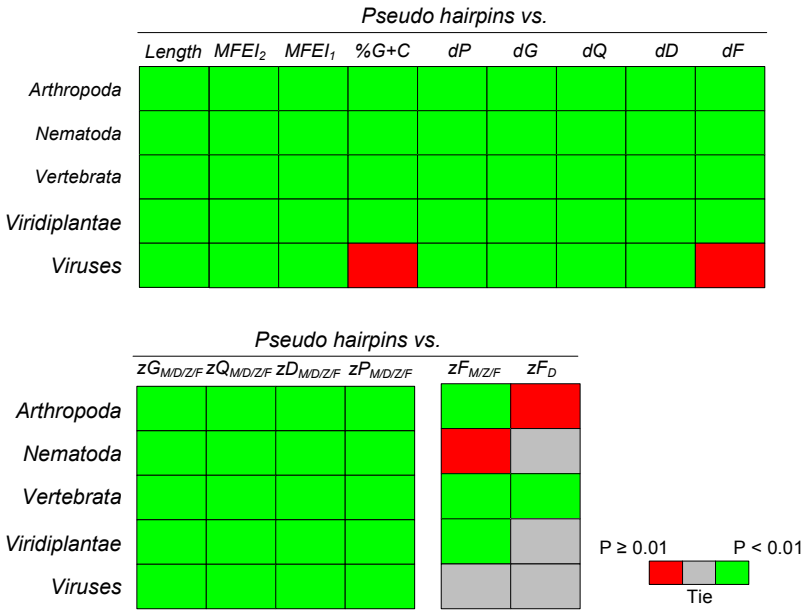

Figure 4.6: Heatmap of pre-miRs *vs.* pseudo hairpins. $zG_{M/D/Z/F}$ denotes *zG* with respect to Mono- and Di-nucleotide shuffling, Zero- and First-Order Markov Model; green represents statistically different median; red for no statistical difference; grey for ties according to the ANOVA ($p < 0.001$) and Dunn's Method of multiple comparisons tests ($p < 0.01$). See Table B.3 for details.

4.4. Correlation between Intrinsic RNA Folding Measures

(Figure 4.7) Correlation tests were conducted on 2,241 non-redundant known pre-miRs according to the following metrics: *Length*, $MFEI_2$, $MFEI_1$, *%G+C*, *dP*, *dG*, *dQ*, *dD*, and *dF* as well as the *zG*, *zQ*, *zD*, *zP*, and *zF* (normalized forms of *dG*, *dQ*, *dD*, *dP*, and *dF* using the four sequence randomization algorithms). The Pearson correlation coefficients C_p are also validated against Spearman-rank C_s (ranks-based) and Kendall's C_k (relative ranks-based) correlation coefficients, as C_s and C_k are extremely robust to non-normal distribution.

Generally, all of the metrics are weakly ($|C_p| < 0.4$) and moderately ($0.4 \leq |C_p| < 0.9$) correlated except *dQ*, *dD*, *zQ*, and *zD*, regardless of the sequence randomization algorithms. Both *dQ* and *dD* are computed from the McCaskill base pair probability p_{ij} (Freyhult *et al.*, 2005), explaining the strong quasi-linear relationship ($C_p \geq 0.9$) for the two pairs *dQ* and *dD* as well as their corresponding normalized form *zQ* and *zD*. There exist moderate pearson correlations within the three pairs *dG* and *zG*, *dP* and *zP*, as well as *dF* and *zF* for the four

sequence randomization algorithms. Initially, two pairs of features *dQ* and *zQ* as well as *dD* and *zD* were expected to behave similarly. Interestingly and currently unclear is why a strong association is observed within themselves. As a guide for future studies especially where computational resources is limited, only *dQ* instead of *dD* should be included (Freyhult *et al*., 2005), while *zQ* and *zD* are extremely time-consuming to compute beyond 10^3 random RNA sequences.

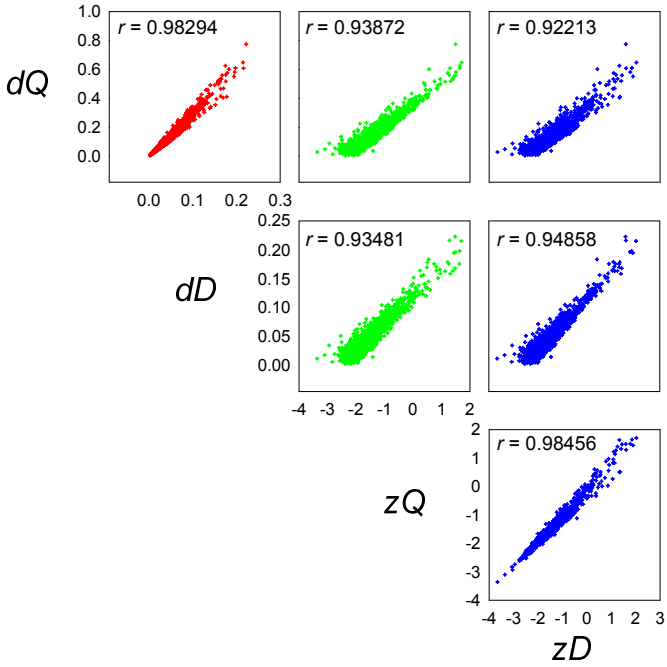

Figure 4.7: Correlation between *dQ*, *dD*, *zQ*, and *zD* for pre-miRs; *zQ*, and *zD* correspond to dinucleotide shuffling; *r* indicates Pearson correlation coefficients *Cp*. $p < 10^{-30}$ for all correlation. The pearson C_p, Spearman-rank C_s (ranks-based), and Kendall's C_k (relative ranks-based) correlation coefficients for all the metrics and sequence randomization methods studied in this work are provided in Table B.4−7.

4.5. Summary

In this large-scale investigation characterizing the entire hairpin-shaped structure of known precursor miRNAs (pre-miRs), notably vertebrate and plant pre-miRs were found to possess a set of 13 statistically significant global features. This *in silico* findings has greatly advanced our understanding of miRNA functions and biogenesis in relation to their structural features and

distinct folding patterns. A definitive criterion for identifying and classifying accurately promising precursor transcripts as *bona fide* pre-miRs, while discriminating against abundant pseudo hairpins within a single genome has not yet been discovered. Moreover, discriminative features used in existing (quasi) *de novo* classifiers have achieved far from satisfactory specificity and sensitivity, especially when cross-specie conservation is unavailable. Our investigated features relating to the intrinsic folding and topological characteristics of pre-miRs, can potentially serve as discriminative measures in improving the designs and performances of current *de novo* predictors. The 13 features have been incorporated into the development of a new and better performing *de novo* classifier for identifying specie-specific and non-conserved pre-miRs, wholly independent of phylogenetic conservation information.

Chapter 5.
De Novo Classification of Precursor MicroRNAs from Genomic Pseudo Hairpins Using Global and Intrinsic Folding Measures

5.1. Training and Classifying Human Precursor MicroRNAs

The optimal hyperparameter pair (C, γ) of the SVM classifier *miPred* was calibrated using TR-H (see section 3.4.3 for details), giving (C, γ) = (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, designating it as a putative 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) (Liu *et al*., 2006; Dror *et al*., 2005). The ROC analysis of *miPred*'s classification model reported that the AUC or ROC score is approximately unity i.e., 0.9833.

(Figure 5.1-A) With the default *miPred*'s threshold predefined at 0.5, the Sensitivity (SE), Specificity (SP), and Accuracy (ACC) reported for TR-H are 88.00%, 97.50%, and 94.33%, respectively. Here, SP is greater than 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 *Caenorhabditis 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.

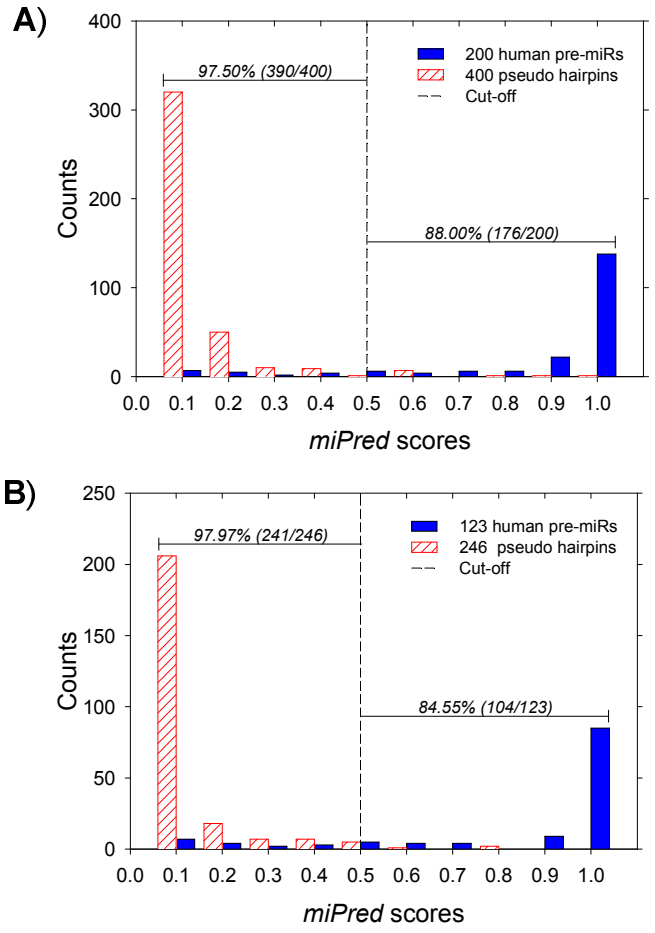

Figure 5.1: A−B) Distribution of TR-H (200 human pre-miRs and 400 pseudo hairpins) and TE-H (remaining 123 human pre-miRs and 246 pseudo hairpins) by *miPred* scores. Default *miPred* decision boundary (vertical dash line at 0.5). See Table C.1 for details.

(Figure 5.1-B) Next, conducting *miPred* onto TE-H (see section 3.4.3 for details) 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 more than 1 clone or differentially expressed in a Dicer-deficient cell-line (Cummins *et al*., 2006). Nevertheless, 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).

In contrast, *Triplet-SVM* based on triplet-encoding scheme (Xue *et al*., 2005) yields slightly poorer results: 86.00% (SE), 97.00% (SP), and 93.33% (ACC) for TR-H; 73.15%, 95.37%, and 87.96% for TE-H; or overall 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.

5.2. Improved Classification of Non-human Precursor MicroRNAs

(Figure 5.2) The validation of *miPred* is extended to IE-NH (see section 3.4.3 for details) and its mean (overall) SE, SP, and ACC were quantified. 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 C.1) In contrast, *Triplet-SVM* 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 for distinguishing reliably specie-specific and evolutionary well-conserved pre-miRs across plants, worms, flies, vertebrates, and viruses (Griffiths-Jones *et al*., 2006).

Notably, those pre-miRs present in the genomes of *Physcomitrella patens*, *Apis mellifera*, *Ateles geoffroyi*, *Canis familiaris*, *Ovis aries*, *Epstein barr virus*, *Herpes simplex virus*, *Human cytomegalovirus*, *Rhesus lymphocryptovirus*, *Simian virus*, and *Zea mays* are unambiguously identified by *miPred* with 100.00% (SE) and >93.75% (SP). Moreover, pre-miRs encoded in *Caenorhabditis briggsae* and *Caenorhabditis 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 *Mouse γ-herpesvirus* and *Kaposi 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.

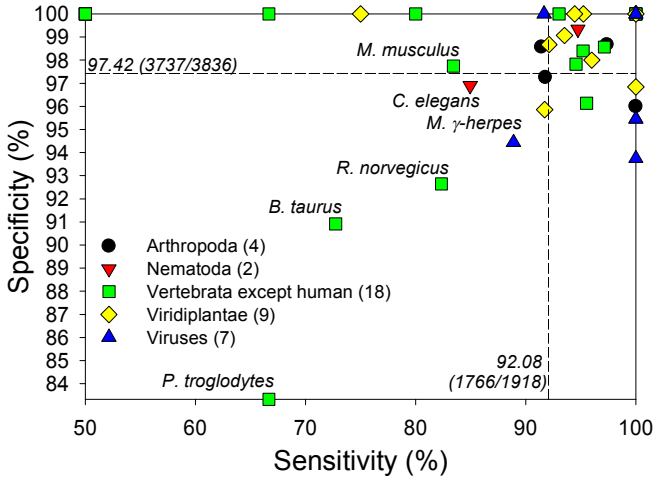

Figure 5.2: Distribution of IE-NH (1,918 pre-miRs across 40 non-human species and 3,836 pseudo hairpins) by specificity and sensitivity. Dash lines denote overall performances. For clarity, only specie names are assigned in left-bottom quarter. See Table C.1 for details.

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 (Sarnow *et al*., 2006; Cullen 2006). As there are few experimental studies elucidating their biological activities and biogenesis (Sullivan *et al*., 2005), it is reasonable to infer that pathogenic viruses do not possess homologous genes, which 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 (Sarnow *et al*., 2006; Cullen 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.

5.3. Performance Comparison with Existing Predictors

(Figure 5.3) By evaluating the published results of existing (quasi) *de novo* classifiers (Table 2.1) 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 F-measure and Matthew's Correlation Coefficient of 98.90% and 92.97% (pre-miRs from various animals) *vs. miPred* of 95.29% and 85.47% (human pre-miRs), or 95.34% and 90.14% (non-human pre-miRs). In contrast, *Triplet-SVM* (Xue *et al*., 2005) is the worst performer among the remaining classifiers that report 20.85−91.87% and 30.80−79.51%, respectively.

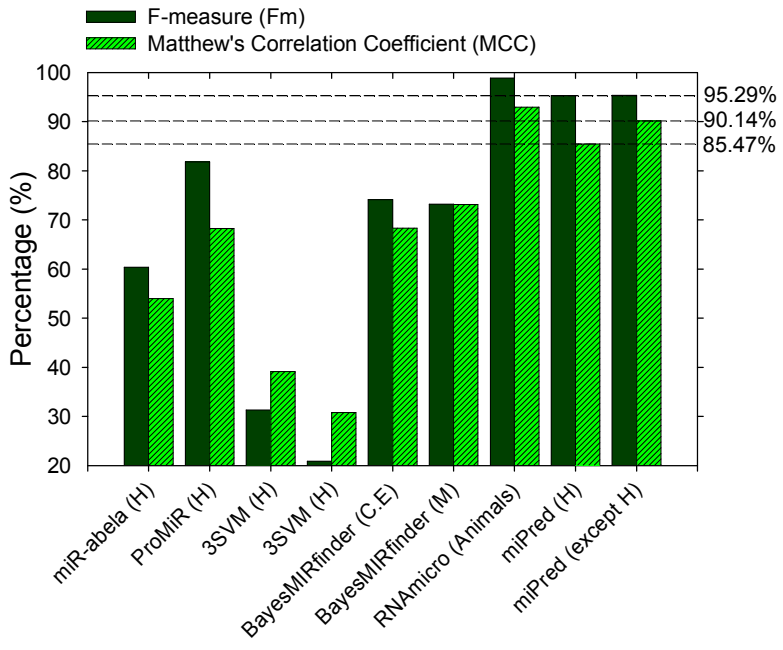

Figure 5.3: Performance comparison with existing (quasi) *de novo* classifiers listed in Table 2.1. H (*Homo sapiens*), C.E (*Caenorhabditis elegans*), and M (*Mus musculus*).

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., vertebrates *vs*. nematode as well as urochordate) is too exceptionally diverged in rendering reliable multi-genomes alignment technically difficult or impossible. Second, distinct from classifiers by *miR-abela* (Sewer *et al*., 2005; Pfeffer *et al*., 2005) and *Triplet-SVM* (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-shaped structures as well as RNA sequences that fold with multiple loops.

5.4. 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 2.1). (Figure 5.4) Evaluating *miPred* and *Triplet-SVM* (Xue *et al*., 2005) onto IE-NC and IE-M, the former reports mean (overall) SP of 76.15% (68.68%; 8507/12,387 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, *Triplet-SVM* 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 2.1) that extract modularized features from predefined RNA sub-structures. The comparison with *Triplet-*

SVM 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.

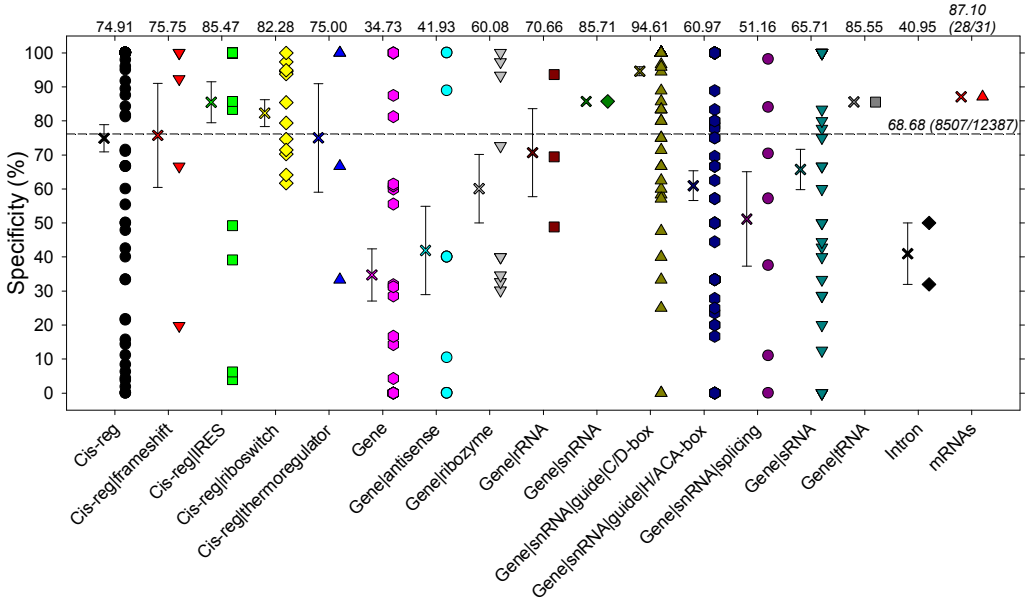

Figure 5.4: Distribution of IE-NC (12,387 ncRNAs) and IE-M (31 mRNAs) by specificity. Dash line denotes overall specificity. See Table C.6 and Table C.7 for details.

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 5.4) 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 pre-processing 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 5.4) *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 5.4) Several classes of ncRNA are poorly classified by *miPred* as potential pre-miRs with SP not more than 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.

5.5. Discriminative Power Contributed by Individual Feature

The essential attributes of *miPred* were investigated on how they contribute substantially to the class distinctions between pre-miRs and pseudo hairpins, or whether exclusion of selected feature(s) can further enhance/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 *Triplet-SVM* (Xue *et al*., 2005) among the existing (quasi) *de novo* classifiers (Table 2.1) has conducted an analysis (less detailed than ours) on its feature selection.

(Figure 5.5) The F-scores F1 and F2 (defined in section 3.4.6) were evaluated 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 $p = 1.272 \times 10^{-19}$. 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. Thus, they are likely to be indispensable for rendering more robustness to the multi-feature capability of *miPred* against erroneous classifications of novel pre-miRs.

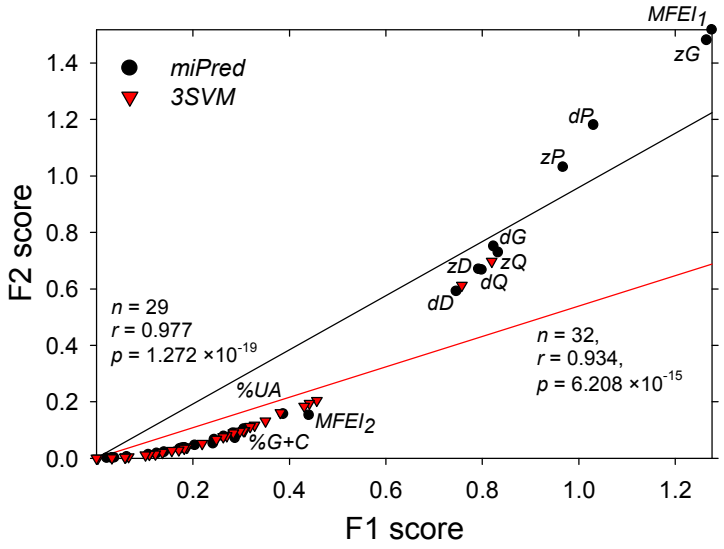

Figure 5.5: F1 and F2 scores for features of *miPred* and *Triplet-SVM*. For clarity, only the names for the top 12 ranking attributes of *miPred* are shown. See Table C.8 for details.

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). The previous experiments were repeated using 10 variants of *miPred* i.e., they have a smaller collection of features and are 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 (Ng and Mishra 2007b) 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 5.6). Derived from $miPred_3$, the remaining nine variants $miPred_{3/5}$, $miPred_{3/10}$, …, $miPred_{3/24}$, and $miPred_{3/25}$ include only the top ranking 21, 16, 11, 6, 5, 4, 3, 2, and 1 feature(s), respectively.

(Figure 5.6) 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 5.6) 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, another three variants of *miPred* were evaluated: $miPred_I$ (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 does 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.

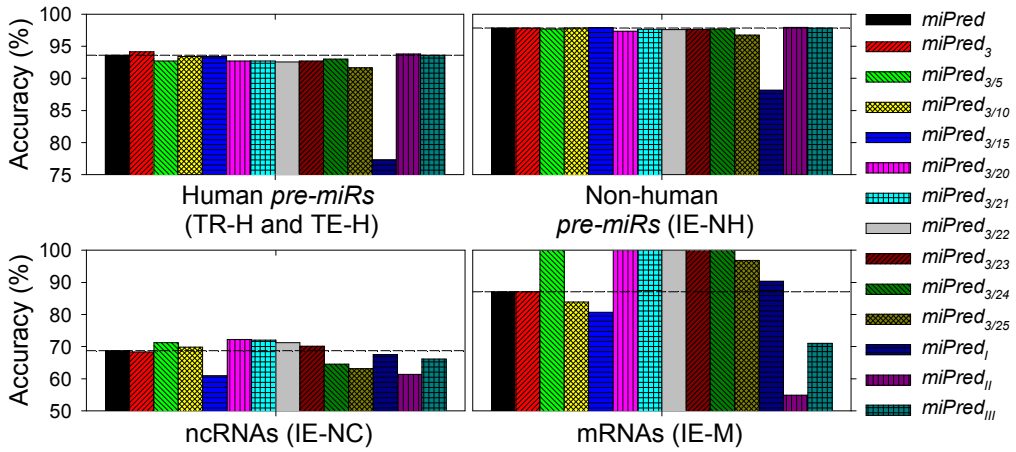

Figure 5.6: Effects of feature selection on *miPred*'s accuracy. Dash lines denote accuracies of original *miPred*. See Table C.9 for details.

5.6. Screening Viral-encoded MicroRNA Genes

rna22 is a pattern-based method for the identification of microRNA-target sites and their corresponding RNA/RNA complexes. 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): *Caenorhabditis elegans* (359), *Drosophila melanogaster* (654), *Mus musculus* (>25,000) and *Homo sapiens* (>25,000). As an illustrative application of *miPred*, four complete viral genomes were randomly selected for screening novel pre-miRs via a similar methodology (Miranda *et al*., 2006): *Epstein barr virus* (EBV), *Kaposi sarcoma-associated herpesvirus* (KSHV), *Mouse γ-herpesvirus 68 strain WUMS* (MGHV68), and *Human cytomegalovirus strain AD169* (HCMV). To date, *Sanger 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 nucleotides in 1 nucleotide steps) for potential viral-encoded hairpins. Those genomic regions satisfying the maximum length (≤ 95 nucleotides), minimum size of terminal loop (≥ 3 nucleotides), and MFEs (≤ -25 kcal/mol) were 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 *srnaloop* was described previously by Grad *et al*. (2003) with differences in the parameter settings as mentioned earlier. Briefly, *srnaloop* uses a BLAST-like algorithm to search for short complementary words (stem-shaped structure) within a specified distance and dynamic programming to determine the complete alignment. In searching a sequence for hairpins of a certain length, *srnaloop* might find two or more hairpins on the same strand that overlap for a considerable percentage of their lengths, a phenomenon called "stuttering". Stutter filtering was applied to cycle iteratively through predicted hairpins on a strand-by-strand basis, to detect overlaps whose length exceeded a threshold fraction of the smaller of the two overlapping hairpin lengths, and to eliminate the hairpins with the smaller srnaloop score. Finally, MFEs were predicted by *RNAfold* program (Hofacker 2003) with default parameters.

(Figure 5.7) In total 1,081 genomic hairpins were screened from the four viruses via *srnaloop*. 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 were classified as putative pre-miRs (positives) at the default *miPred* score cut-off ≥ 0.5; remaining ones were regarded as negatives. (Table C.10) The viral-encoded hairpins were manually mapped to the published pre-miRs, 25 true positives (and 1 false negative) matched 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 *kshv-miR-K12-9*, the remaining true-positive predictions had one or two mature miRNAs embedded exclusively in either arms of their (a)symmetric stem. *kshv-miR-K12-9* was subsequently eliminated as it was a duplicate copy containing the exact sequence of *kshv-miR-K12-9*, and the encoded mature miRNAs overlap the most with its predicted 4 nucleotides (**UAUA**) terminal loop. Together, 44.64% (25/56) of the known pre-miRs for the four viruses were identified as hairpins, and 96.00% (24/25) of them were recovered as true positives.

The 25 identified positives reported 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 had not yet been discovered (more will arise due to innate evolutionary mutations), suggesting that previous estimates of viral-encoded pre-miRs and miRNAs especially in EBV and HCMV might be grossly understated. (Figure 5.8) 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), p1 was observed to overlap exactly with but was shorter than m6 by 3 nucleotides (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 was reassigned as a false-positive. p2 resided immediate downstream of m3 and within a known miRNA cluster ~1.5 kb consisting of m1−7 that were transcribed by RNA Polymerase III (Pol-III) (Pfeffer *et al*., 2005), which indicated p2 was likely to be regulated by similar Pol-III promoter. Known host miRNA transcripts were 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 (Lee *et al*., 2004; Cai *et al*., 2004), however, there were emerging evidence of them being transcribed from the exons of protein-coding genes like in *Oryza sativa* (Sunkar *et al*., 2005). Thus, p3, p4, and p5−6 located in the exons of three proteins might also undergo distinct processing and nuclear export mechanism from the host cell's miRNA maturation machinery.

5.7. Summary

In this work, a *de novo* SVM classifier model *miPred* was proposed to address specifically the challenges in improving the classification accuracy of existing (quasi) *de novo* approaches. Without relying on phylogenetic conservation information, *miPred* achieved significantly higher sensitivity and specificity by incorporating a Gaussian Radial Basis Function kernel as a similarity measure for the 29 global and intrinsic hairpin folding attributes. The comprehensive analysis reported that it yielded comparable or significantly better predictive 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. Applying *miPred* to the screening of four viral genomes, numerous numbers of sequence segments have the potential to fold into pre-miR like hairpins. The successful *ab initio* classification of real pre-miRs from pseudo ones opens a new approach for identifying novel miRNAs.

Deployment of *miPred* will likely to translate into considerable saving on precious and scarce experimental resources devoted to validating significantly fewer false-positives, since it is highly assured that 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, this research prototype was revamped as part of a computational pipeline for the identification of novel miRNAs expressed in the gonads and brain of zebrafish.

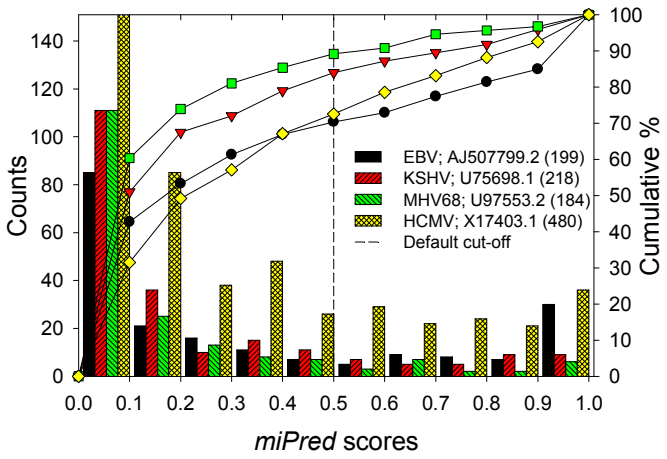

Figure 5.7: Distribution of viral-encoded hairpins according to *miPred* scores. See Table C.10 for details.

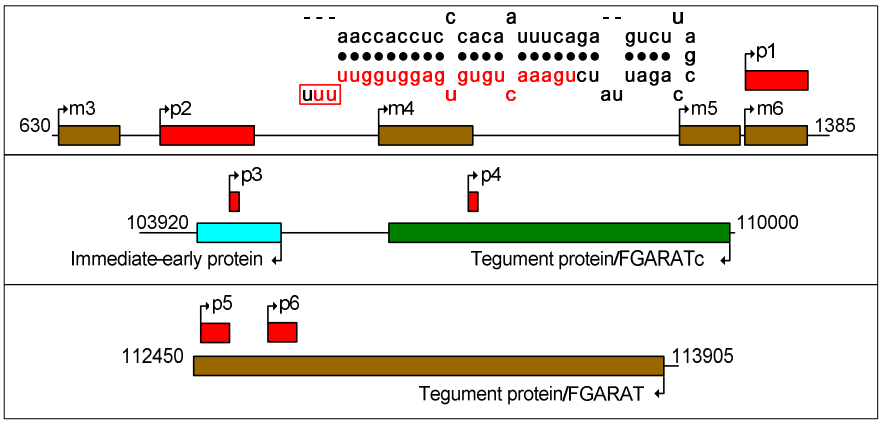

Figure 5.8: Genomic map of predicted (pX denotes *mghv-miR-pX*) and published (mX denotes *mghv-miR-M1-X*) MGHV68-encoded pre-miRs, drawn not to scale by Genepalette 1.2 (Rebeiz and Posakony 2004); RNA structure of m6 (inset; *mghv-miR-M1-6*) was obtained from *Sanger miRBase* 8.2 (Griffiths-Jones *et al.*, 2006); red region denotes mature miRNA. See Table C.10 for details.

Chapter 6.
Small RNA Profiling in Zebrafish Gonads and Brain: Novel MicroRNAs with Sexually Dimorphic Expression

6.1. Introduction

The zebrafish *Danio rerio* has become an invaluable vertebrate model system for development and functional genetics, and is arguably the most widely-studied teleost to date. Sex determination in teleosts is a fundamental but poorly understood process crucial to continuation of the germ line. The genetic mechanisms controlling the sex determination and differentiation of zebrafish remain largely unknown, not well-understood, or at best contradicting (Uchida *et al*., 2002). The established model proposed that zebrafish is sexually mature after approximately three months, and distinct sexes can be detected after 21–23 days post fertilization (dpf) (Uchida *et al*., 2002). Prior to sex differentiation, all zebrafish develop ovary-like gonads by default, a process that is initiated after 10 dpf and progresses till 20 dpf. Between 21 dpf and 30 dpf, this gonad development is initiated simultaneously in males alongside the ovarian apoptosis. Synaptonemal complex karyotype revealed that the diploid genome of zebrafish consists of 50 chromosomes, but no specific sex chromosomes (Wallace and Wallace 2003) nor sex linked genes have been identified to date (von Hofsten and Olsson 2005). *FTZ-F1* genes have been suggested recently to be involved in the sex determination process, however, many key questions remain unresolved (von Hofsten and Olsson 2005). Furthermore, teleosts display an enormous diversity of sex determination systems, which can also be influenced by environmental factors (Devlin and Nagahama 2002).

Recent functional studies indicated that microRNAs (miRNAs) play essential roles in zebrafish development. *Dicer1* mutants that were defective in miRNA processing, experienced arrest in overall growth and development, possibly caused by the depletion of miRNAs

ubiquitously required for cell proliferation or specific ones required in various tissues and organs (Wienholds *et al*., 2003). Injection of synthetic double-stranded *dre-let-7* miRNAs caused specific phenotypic defects in the zebrafish embryo, as demonstrated that two *dre-let-7* target sites from the zebrafish *lin-41* gene were mediated during post-transcriptional silencing (Kloosterman *et al*., 2004). Through maternal-zygotic dicer (*MZdicer*) mutants that disrupted the Dicer ribonuclease III and double-stranded RNA-binding domains, miRNAs expressed in zebrafish were experimentally shown to be indispensable for cell fate determination, axis formation, and cell differentiation (Giraldez *et al*., 2005). Moreover, *MZdicer* mutants displayed abnormal morphogenesis during gastrulation, brain formation, somitogenesis, and heart development.

Known miRNAs were essentially absent from the early zygote stage at 0 hours post fertilization (hpf), given that a mere 3% miRNA content was derived from part of the small RNA library, and cloned miRNAs could not be detected during that period (Chen *et al*., 2005). The miRNAs expression commenced during the blastula stage (4 hpf) with a zebrafish-specific family of miRNAs encoded by closely spaced multi-copy genes (Chen *et al*., 2005). Most of the known miRNAs were expressed preferentially in the later stages of development and approximately one-third of them were expressed at the onset stage of the embryonic brain (Kloosterman *et al*., 2006). Majority of the known miRNAs could not be detected up to the segmentation stage, but became visible between the pharyngula stage (24 hpf) and hatching stage (48 hpf). They showed strong expression when organogenesis was largely completed at 96 hpf (Wienholds *et al*., 2005). Generally, known miRNAs were expressed in a highly tissue-specific manner during segmentation (12 hpf) and later stages, but not in the early development, suggesting their role in differentiation or maintenance of tissue identity and not in tissue fate establishment (Wienholds *et al*., 2005). In another study, miRNA expressions were found to be highly differential across ten adult tissues (in the order listed in text) i.e., the total, brain, eye, muscle, gills, fins, skin, liver, gut, and heart (Kloosterman *et al*., 2006). Through recent work, miRNAs possessed a diverse expression profiles in neural cells when detected by *in situ* hybridizations (Kapsimali *et al*., 2007). Interestingly, miRNA profiles of two fibroblast cell lines derived from both caudal fin and liver epithelium closely resembled each other, despite the cell lines were established independently from various tissue sources including the liver and caudal fin (Chen *et al*., 2005).

In earlier expression profiling studies conducted using small RNA cloning (Chen *et al*., 2005), microarray analysis (Kloosterman *et al*., 2006; Wienholds *et al*., 2005), and *in situ*

hybridizations experiments using locked-nucleic acid (LNA) modified oligonucleotide probes (Kapsimali *et al*., 2007; Kloosterman *et al*., 2006; Wienholds *et al*., 2005), organ and tissue-specific miRNA expression profiles were observed spatially and temporally at different developmental stages in zebrafish. Considering that pooled RNA samples from both male and female zebrafish were used, as well as gonads from juveniles and adults were excluded, none of these studies has attempted to analyze the sex determination associated to and sexually dimorphic expression of miRNAs in zebrafish gonads and brain – a gap that this present study seeks to fill.

6.2. Results and Discussion

6.2.1. Cloning of Known and Novel MicroRNAs from Zebrafish Gonads and Brain

In order to obtain a comprehensive miRNA expression profiles of zebrafish sex-related organs and brains, a large-scale sequencing experiment of six small RNA libraries was conducted, namely, ovary and testis of 35 days post fertilization (dpf) juveniles (*5WO* and *5WT*), ovary and testis of adults (*AOV* and *ATE*), and the brain of 35 dpf female and male juveniles (*5WFB* and *5WMB*). The inclusion of *5WFB* and *5WMB* was motivated by a recent investigation that majority of the miRNAs were expressed in the onset stage of the embryonic brain (Kloosterman *et al*., 2006) and that sexually dimorphic cell proliferation was also observed in the teleost brains (Ampatzis and Dermon 2007; Zikopoulos *et al*., 2001).

Roughly 1,500 clones was randomly picked from each library for sequencing (see Figure 6.10 (*Left*) for details), except for *5WFB* requiring twice as many due to lower cloning efficiency of small RNAs. Through an in-house computational pipeline consisting of four stages as shown in Figure 6.10 (*Right*), (Figure 6.1) 19,016 small RNAs (of which 11,791 were unique) were extracted from the 10,456 concatamers. The obtained sequences were then functionally annotated against 32,540 ncRNA sequences from *Sanger Rfam 8.0* (Griffiths-Jones *et al*., 2005) and 60,067 others from another published dataset (Chen *et al*., 2005), 4,584 pre-miR and 4,430 mature miRNA sequences (377 pre-miRs and 219 mature miRNAs in zebrafish) from *Sanger miRBase 9.2*, as well as 218,100 published piRNA sequences that were identified from zebrafish and mouse (Houwing *et al*., 2007; Grivna *et al*., 2006); see section 6.3.3 for details.

Majority of the small RNAs corresponded to fragments of published ncRNAs (1,469), known miRNAs (780), and piRNAs (7,415). Interestingly, a small subset of 133 small RNAs mapping to known pre-miRs but not the mature region could likely be miRNA*, 720 matched to both piRNAs and miRNAs, and 31 others could not be functionally annotated due to higher occurrences of sequencing errors. Expression profiles of known miRNAs are described in section 6.2.2. The average percentage of small RNAs cloned that were identified as known and putative miRNAs roughly matched that of previous miRNA profiling experiments conducted in mouse testis (Ro *et al*., 2007) and a recent large-scale mammalian miRNA expression atlas based on small RNA library sequencing (Landgraf *et al*., 2007).

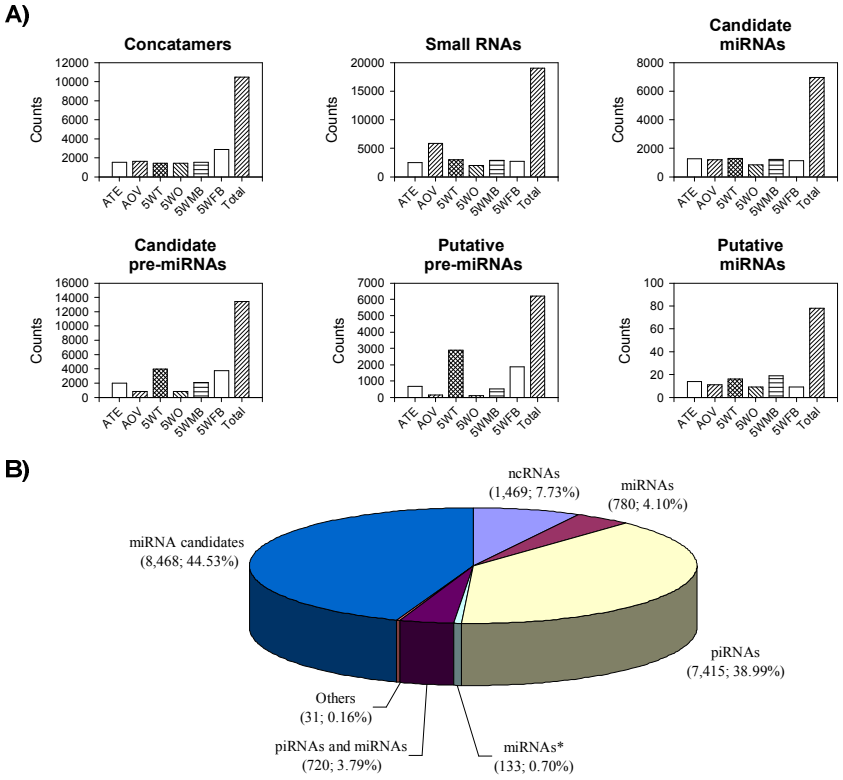

Figure 6.1: A) Distribution of 10,456 concatamers, 19,016 small RNAs, 8,468 non-annotated small RNAs (candidate miRNAs), 13,448 candidate pre-miRs, 6,202 putative pre-miRs, and 78 putative miRNAs across six libraries. Adult Testis and Ovary (*ATE* and *AOV*); Juvenile Testis and Ovary (*5WT* and *5WO*); Juvenile Male and Female Brain (*5WMB* and *5WFB*). See Table D.1 for details. B) Functional annotation of 19,016 small RNAs extracted from 10,456 concatamers.

8,468 non-annotated small RNAs (among them 6,964 were unique) that were clearly not belonging to any of the annotated ncRNAs, piRNAs, pre-miRs, and mature miRNAs, were considered as miRNA candidates. As the next stage was a computationally expensive exercise, only the 6,964 unique miRNA candidates were mapped to the sixth assembly of the zebrafish genome and folded structurally to screen for 13,448 candidate pre-miRs. They were then subjected to classification by *miPred* (Ng and Mishra 2007a; Ng and Mishra 2007b) into 6,202 putative pre-miRs corresponding to 78 putative miRNAs, and the remaining 7,246 as pseudo hairpins. Majority of the putative miRNAs with flanking regions had been observed to have significantly higher occurrences of folding into a putative miRNA hairpin (Chen *et al*., 2005). This phenomena is not unusual as known mature miRNA *dre-miR-430* family composing of five members has ~100 gene copies distributed over two large clusters of 30 and 17 kilobases within unassembled genome sequence, and a very small (500 base pairs) cluster of three miRNAs positioned on chromosome 13 (Chen *et al*., 2005).

6.2.2. Expression Profile Analysis of Known and Novel MicroRNAs based on Small RNA Libraries

(Figure 6.2) In order to analyze the temporal and organ-specific miRNA expression in zebrafish sex-related organs and brains, the expression profile of known miRNAs was generated from the normalized cloning frequency of 780 small RNAs that were homologous to 88 zebrafish mature miRNAs. This took into consideration the relative fraction of known miRNAs identified within the total pool of cloned small RNAs of a given RNA sample across the six libraries. The values in each row corresponding to each known miRNA were linearly rescaled to the interval [-1.0, 1.0]; -1.0 and 1.0 indicate weakly expressed and strongly expressed, respectively. The latest *Sanger miRBase* 9.2 (Griffiths-Jones *et al*., 2006) reported 219 mature miRNAs in the zebrafish genome. Among them, 88 (~40.00%) matched to 780 small RNAs expressed across the six miRNA libraries i.e., *ATE*, *AOV*, *5WT*, 5WO, *5WMB*, and *5WFB*.

The gonadal miRNA expression pattern obtained from this study revealed many highly correlated instances with those of previously published microarray analysis on zebrafish (Wienholds *et al*., 2005). Notably, *dre-let-7a*, *dre-let-7c*, *dre-miR*-7 family (*dre-miR-[7a, 7b]*), and *dre-miR-143* were preferentially enriched in the *ATE* than *AOV*, while *dre-let-7i*, *dre-miR-92* family (*dre-miR-[92a, 92b]*), and *dre-miR-132* were strongly expressed in *AOV* than in *ATE*. Interestingly, *dre-let-7b* expression was formerly reported to be restricted to the proliferative

ciliary marginal zone of the retina and absent from all mature retinal neurons, while *dre-let-7a* and *dre-let-7c* (differing from *dre-let-7b* by two and one nucleotide, respectively) lacked this retinal expression (Kapsimali *et al*., 2007). Notwithstanding the similarities, several differences between both miRNA expression profiles could also be observed. The *dre-miR-125b* was reported previously in both libraries *ATE* and *AOV*, and at a higher level in the latter (Wienholds *et al*., 2005). In this present study, it was identified in *ATE* but absent in *AOV*. *In situ* hybridization data reported its detection in the brain and spinal cord (Ason *et al*., 2006). Interestingly, abundant *dre-miR-214* expression was detected in the adult gonad libraries *AOV* and *ATE*, but significantly stronger expression in *AOV* than in *ATE* was reported by the previous study (Wienholds *et al*., 2005). This miRNA was known to be expressed during the early segmentation stages in somites and varying its expression altered the genes expression regulated by the Hedgehog signaling (Flynt *et al*., 2007). The *su*(*fu*) mRNA encoding a negative regulator of Hedgehog signaling was targeted by *dre-miR-214* for post-transcriptional suppression, and inhibiting the miRNA resulted in a reduction or loss of slow-muscle cell types, suggesting its involvement in the specification of muscle cell fate during somitogenesis (Flynt *et al*., 2007). Moreover, differential regulation of germline-specific gene expressions in the primordial germ cells (PGCs) and somatic cells involves *dre-miR-214* targeting the 3' untranslated region of germline-specific genes *nanos1* and *TDRD7* (Mishima *et al*., 2006).

A specific set of miRNAs was observed to be differentially expressed in both adult gonad libraries *ATE* and *AOV* such as the *dre-let-7e* and *dre-miR-101* family (*dre-miR-[101a*, *101b]*). The embryonic specific miRNA *dre-miR-430* was previously known to be strongly detected at the onset of zygotic transcription, with functions in promoting the deadenylation and clearance of maternal mRNAs, while rescuing the brain morphogenesis phenotype (Giraldez *et al*., 2006). Interestingly, *dre-miR-430c*, a member of the *dre-miR-430* family was detected exclusively in *AOV*, suggesting that this miRNA family could be expressed at the later stages in zebrafish development with additional and uncharacterized biological functions.

Another small subset of miRNAs was preferentially expressed in an organ-specific and/or time-specific manner such that *dre-miR-122* was found exclusively in both ovarian libraries of adult and juvenile zebrafish *AOV* and *5WO*. Its homologue was reported to be specifically expressed in mouse liver (Pfeffer *et al*., 2005) and antagonism of *mmu-miR-122* by systemically administered LNA-antimiR triggered up-regulation of a large set of predicted target mRNAs in the liver (Elmen *et al*., 2007). The *dre-miR-29a* displayed "male specific" expression in *ATE*, *5WT*, and *5WMB*. Additionally, *dre-miR-145* was detected in the juvenile but not in the adult

gonads, suggesting an early role in gonadal development, which was also observed in zebrafish pharyngeal arches and fins with weaker expression observed in the gut and gall bladder (Ason *et al*., 2006). Conversely, the *dre-miR-19* family (*dre-miR-[19a, 19b, 19c, 19d]*) was expressed strongly in the adult rather than the juvenile ovary, pointing to a later role in ovarian development of zebrafish.

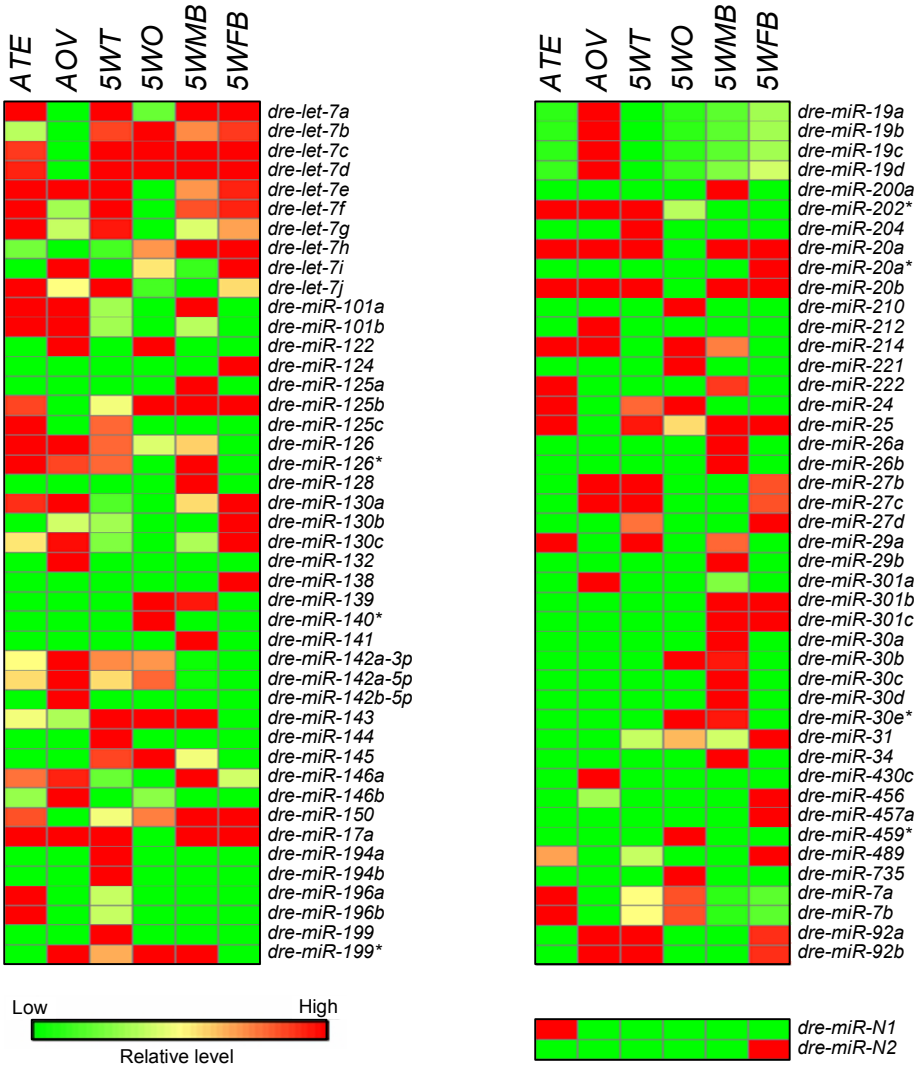

Figure 6.2: Expression profiles of 88 known miRNAs and 2 novel miRNAs expressed across six miRNA Libraries. Adult Testis and Ovary (*ATE* and *AOV*); Juvenile Testis and Ovary (*5WT* and *5WO*); Juvenile Male and Female Brain (*5WMB* and *5WFB*). See Table D.2 for details.

Sexual dimorphism of miRNA expression was distinctively observed at the juvenile stage of development such that the *dre-miR-140**, *dre-miR-199*, *dre-miR-[34, 138, 141]*, and *dre-miR-[124, 200a]* were expressed exclusively in one of the juvenile organ libraries *5WO*, *5WT*, *5WFB*, and *5WMB*, respectively. Interestingly, the presence of *dre-miR-140* was specifically restricted to the cartilage of pharyngeal arches, head skeleton, and fins at 72 hpf, (Wienholds *et al*., 2005), suggesting that *dre-miR-140** could also be expressed in some or all these tissues. Our results on *dre-miR-34* and *dre-miR-124* were consistent with a recent *in situ* hybridization study of miRNA expression in neuronal system such that the constitutive expression of *dre-miR-124* was detected in the mature neurons, as well as its expression associated with transition from proliferation to differentiation; predicted targets for *dre-miR-124* included diverse 'early' neural genes *zic2a*, *pou5f1*, *otx2*, and *slit2*; *dre-miR-200a* was expressed in the peripheral sensory neural cells; *dre-miR-34* was expressed in neural cells in restricted subdivisions along the rostro-caudal axis of the larval brain (Kapsimali *et al*., 2007). The homologue of *dre-miR-124* was also reported to be specifically expressed in the mouse brain as determined by Northern blotting (Lagos-Quintana *et al*., 2002). These result suggested that miRNAs could also serve as a regulatory vehicle contributing towards the differential brain phenotypes observed in both sexes.

miRNA:miRNA* pair originate from a common pre-miR hairpin, where the less stable strand miRNA* tends to be degraded and is not incorporated into RISC for post-transcriptional silencing of the target genes (Ambros *et al*., 2003a; Ambros 2001). Interestingly, both strands of two studied miRNA:miRNA* pairs could perform the role of mature miRNAs such that *dre-miR-199* was expressed only in *5WT* while its counterpart strand *dre-miR-199** was detected in the *AOV*, *5WT*, *5WO*, and *5WMB*; *dre-miR-20a* was cloned in all libraries except *5WO*, while *dre-miR-20a** was expressed solely in *5WFB*. It remains unknown what underlying mechanisms were involved or how this notable contrasting expression pattern could be achieved by both miRNA:miRNA* pairs, since a miRNA:miRNA* pair is initially generated from a common pre-miR hairpin. A possibility could be that either of the two mature miRNAs from a hairpin was selectively degraded in an organ or time-specific context.

Other detected known miRNAs that could not be classified temporally and/or spatially include the *dre-miR-7* family (*dre-miR-[7a, 7b]*), which was strongly expressed in *ATE* and was previously detected in the endocrine pancreas of Langerhans islets (Wienholds *et al*., 2005). The *dre-miR-92* family (*dre-miR-[92a, 92b]*) was strongly expressed in diverse libraries *AOV*, *5WT*, and *5WFB*, in addition to the *dre-miR-92b* expression in neuronal precursors and stem cells (Kapsimali *et al*., 2007) and brain (Kloosterman *et al*., 2006). Besides being strongly expressed

in the two male-specific libraries *ATE* and *5WMB*, *dre-miR-222* was previously detected in a specific groups of differentiating cells of the forebrain and midbrain (Kapsimali *et al*., 2007).

Finally, the two novel miRNAs *dre-miR-N1* and *dre-miR-N2* were strongly detected in the *ATE* and *5WFB*, respectively. Though their sequences showed little or no similarities to existing miRNAs, the closest expression patterns to *dre-miR-N1* belonged to the *dre-miR-196* family (*dre-miR-[196a, 196b]*), while *dre-miR-N2* had identical profile as that of two known miRNAs *dre-miR-124* and *dre-miR-138*. Besides *pre-miR-N1* and *pre-miR-N2*, 8 and 16 known miRNAs were located on the chromosome 1 and 23, respectively, for *Sanger miRBase* 9.2 (Griffiths-Jones *et al*., 2006). Interestingly, none of those on chromosome 1 (*dre-miR-15a*, *dre-miR-16b*, *dre-miR-155*, *dre-miR-218a*, *dre-miR-220*, *dre-miR-722*, dre-*miR-734*, and *dre-miR-740*) were cloned in the six small RNA libraries; more than half of those on chromosome 23 (as indicated in bold; ***dre-let-7g***, ***dre-let-7h***, *dre-miR-1*, *dre-miR-10b*, ***dre-miR-26b***, ***dre-miR-29a***, ***dre-miR-29b***, ***dre-miR-34***, ***dre-miR-124***, *dre-miR-133a*, *dre-miR-135c*, ***dre-miR-196a***, ***dre-miR-200a***, *dre-miR-200b*, and *dre-miR-429*) were detected.

6.2.3. Real-time RT-PCR Analysis of Known MicroRNAs Shows Sexually Dimorphic Expression

The miRNA expression profile shown in Figure 6.2 was obtained from a miRNA library construction consisting of multi-steps experimental procedures that were likely to be prone to cloning fluctuations caused by sequencing aberrations, cloning techniques, and adaptors ligation efficiencies. In consideration of its inherent limitation, the miRNA expression profile portrayed a semi-quantitative measure on the abundance of miRNAs based on their clone numbers, which was an adequate but not fully accurate method of quantifying miRNA expression levels. Thus, it was evaluated with alternative miRNA quantitative methods, in particular real-time RT-PCR, which is more specific and sensitive. (Figure 6.3) The real-time RT-PCR analysis was performed on a selected set of five known zebrafish miRNAs, namely, the *dre-let-7g*, *dre-let-7j*, *dre-miR-125b*, *dre-miR-130a*, and *dre-miR-143* on the six existing RNA samples *ATE*, *AOV*, *5WT*, *5WO*, *5WMB*, and *5WFB*.

In comparison with the miRNA library expression profiles of two miRNAs *dre-let-7g* and *dre-let-7j* belonging to the abundant *dre-let-7* family, their real-time RT-PCR results reported that the expression patterns generally correlated with each other except that they had stronger expressions in *5WMB*. The third miRNA *dre-miR-125b* was expressed at much lower levels in

both the adult and juvenile gonads (*AOV*, *ATE*, *5WO*, and *5WT*), than in the adult and juvenile brains of both sexes (*AMB*, *AFB*, *5WMB*, and *5WFB*) indicating a level of consistency with the miRNA library cloning data for which *dre-miR-125b* was weakly expressed in *AOV*, *ATE*, *5WT* and *5WO* than in *5WMB* and *5WFB*. Notably, the remaining two miRNAs *dre-miR-130a* and *dre-miR-143* reported significantly different expression levels between miRNA cloning data and real-time RT-PCR data. For example, *dre-miR-143* was moderately expressed in *5WMB* and strongly expressed in *5WO* and *5WMB* according to miRNA cloning data. Instead, real-time RT-PCR results reported stronger expression in *5WMB* than in *5WO* and *5WMB*. The discrepancies between both technologies could be due to (but not limited to) an inherent bias in miRNA cloning with respect to these two miRNAs across the six libraries, causing a small fluctuation in the population of cloned miRNAs to significantly influence the resultant expression profile.

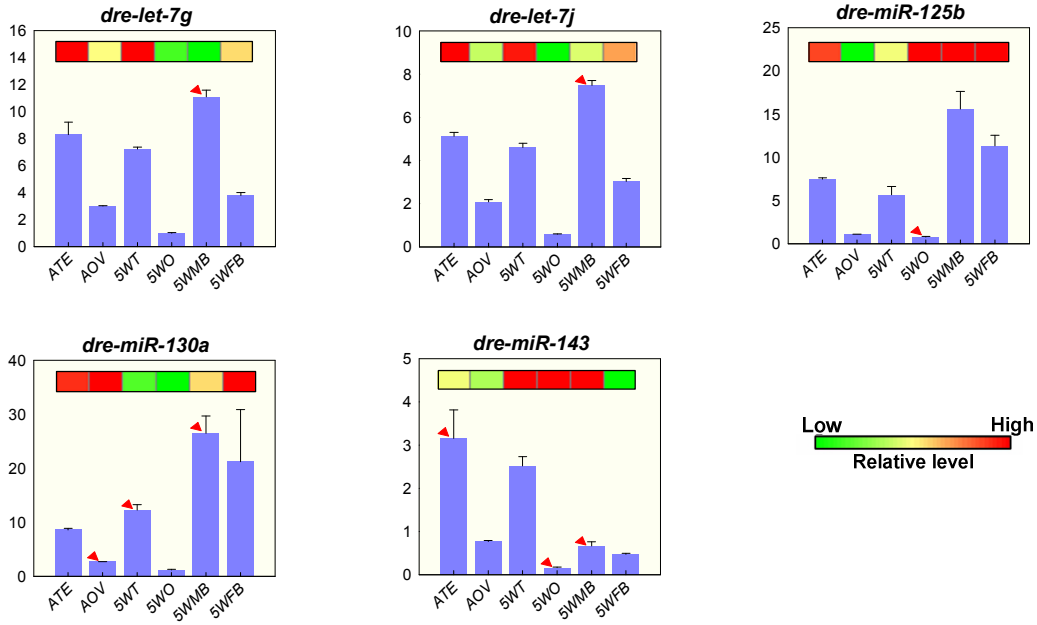

Figure 6.3: Real-time RT-PCR results of five selected known miRNAs expressed in gonads and brains of juvenile and adult zebrafish. Mean and standard deviations were derived from triplicates.

Interestingly, characteristic sexual dimorphism of miRNA expression was observed in the male and female gonads of zebrafish by real-time RT-PCR data. The miRNA expression pattern for the five tested miRNAs reported significantly higher expression of approximately 2-3 folds in the testis than ovary for both juvenile and adult zebrafish i.e., *ATE vs. AOV* and *5WT vs.*

5WO. Similar sexual dimorphism of miRNA expression was demonstrated for *dre-let-7g* and *dre-let-7j* (but less distinctive for *dre-miR-125b*, *dre-miR-130a*, and *dre-miR-143*) in the male and female brain of juvenile zebrafish i.e., *5WMB vs. 5WFB*.

6.2.4. Computational Identification of Novel MicroRNAs

Two novel miRNAs *dre-miR-N1* and *dre-miR-N2* were observed to be expressed exclusively in the adult testis (*ATE*) and juvenile female brain (*5WFM*) small RNA libraries (see section 6.2.2 for details), (Figure 6.4) as well as originated from one arm of promising precursor transcripts that tend to fold into energetically stable and high-scoring hairpin-shaped secondary structures.

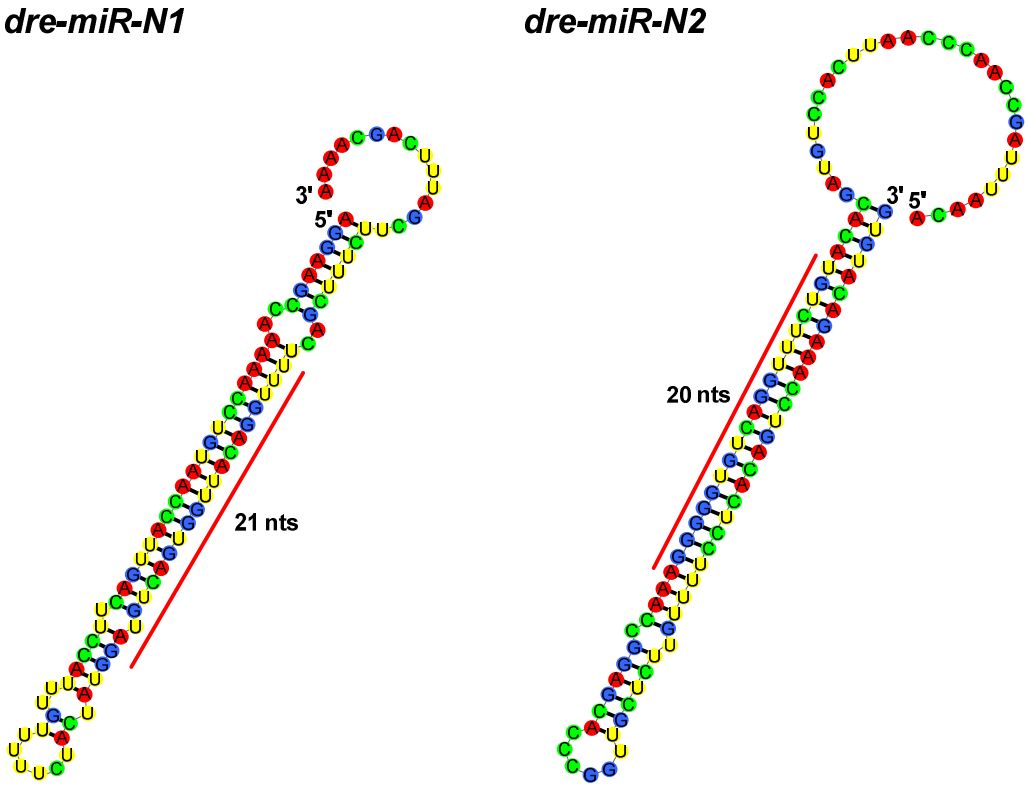

Figure 6.4: Secondary structures of two selected novel miRNAs *dre-miR-N1* and *dre-miR-N2*. Sequence region underlined in red indicates the novel mature miRNA. Size in nucleotides (nt) indicates length of novel miRNA.

(Figure 6.5) Their corresponding putative pre-miRs possessed minimum free energy of folding (MFE) of -45.90 kcal/mol and -56.30 kcal/mol as predicted by *RNAfold* program (Hofacker 2003) with default parameters, as well as *miPred* scores of 0.999978 and 0.999681 as

predicted by the SVM-based classifier *miPred* using intrinsic RNA folding measures (Ng and Mishra 2007a; Ng and Mishra 2007b), respectively.

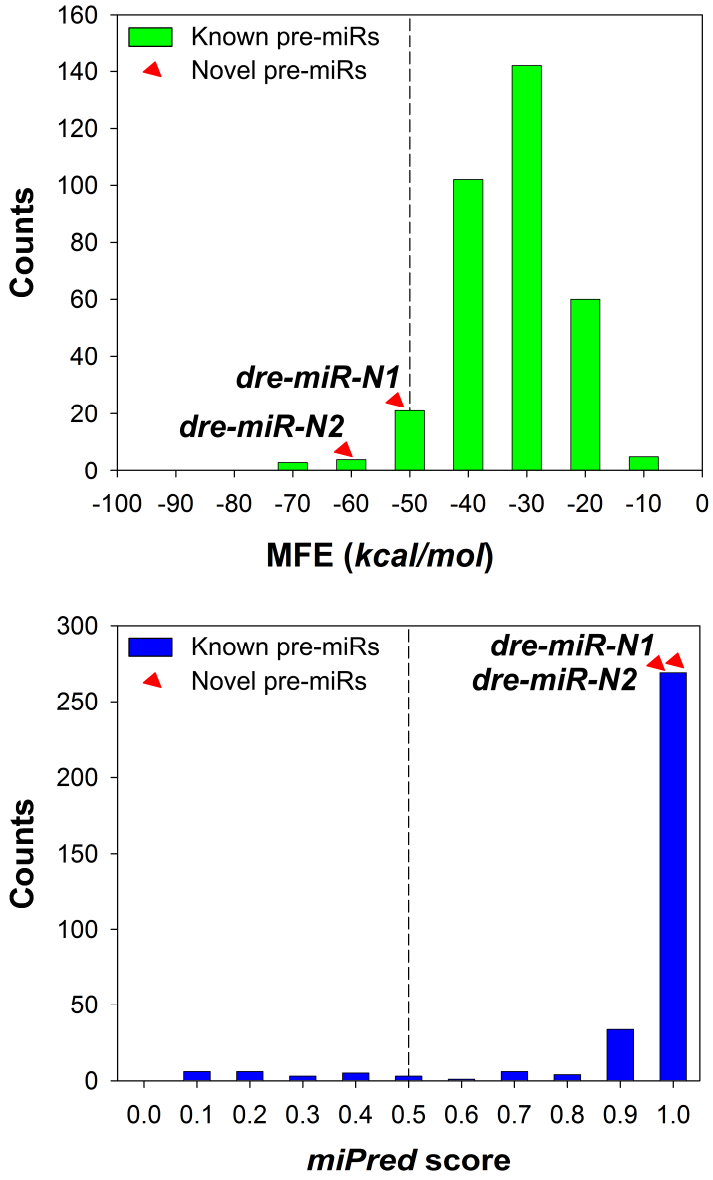

Figure 6.5: Distribution of 377 known pre-miRs and 2 novel miRNAs *dre-miR-N1* and *dre-miR-N2* with respect to their MFE (kcal/mol) and *miPred* score.

(Table 6.1) These transcripts were derived from the anti-sense strand of chromosome 23 and sense strand of chromosome 1 of the zebrafish genomic loci, respectively. Additional evidence for substantiating computationally both novel pre-miRs as likely genuine ones, were

provided by two recently published pre-miR classifiers *miPred-J* (Jiang *et al*., 2007) and *ProMiR II* (Nam *et al*., 2006) using their default parameters: *miPred-J* score of 0.828 and 0.680 (above default cut-off of 0.5 indicates likely real pre-miR); *ProMiR II* score of 23.6223 and 233.662 (above most stringent cut-off of 3.3 indicates likely genuine pre-miR). The former was coincidently named *miPred*, but independently developed using random forest prediction model with combined features, namely, local contiguous triplet structure composition, MFE, and *p*-value of dinucleotide shuffling. The latter used a probabilistic co-learning model with additional features like G/C ratio, MFE, and entropy of candidate sequences for identifying putative ones.

Table 6.1: Sequence and structural statistics of two selected novel miRNAs *dre-miR-N1* and *dre-miR-N2*.

	dre-miR-N1	***dre-miR-N2***
Sample	Adult Testis (*ATE*)	35 dpf Female Brain (*5WFB*)
Mature miRNA		
Sequence	GAUGUCAGUGGUUACAGGUUU	UGUCUUUGGACUGUGGGGGA
Length (nts)	21	20
Chromosomal coordinates	Chr23|36750719|36750739	Chr1|55786236|55786255
Precursor miRNA		
Sequence	AGGAAGCCAAAAACCUGUAACC AUUGACUUCCAUUUGUUUUUCU ACUAUGGAUGUCAGUGGUUACA GGUUUUCAGCUUUCUUCGAUUU CAGCAAAA	ACAAUUUAGCCAACCCAAUUCA CCUGUAGCACAUGUCUUUGGAC UGUGGGGGAAACCGGAGCACCC GGUUGCUCUUGUUUUCCUCACA GUCCAAAGACAUGUG
Length (nts)	96	103
Chromosomal coordinates	Chr23|36750669|36750764	Chr1|55786203|55786306
Direction	Antisense	Sense
dG (kcal/mol)	-45.90	-56.30
miPred score	0.999978	0.999681
miPred-J score†	0.828	0.680
ProMiR II score‡	23.6223	233.662

†, above 0.5 indicates real pre-miR by *miPred-J* (Jiang *et al*., 2007); ‡, above 3.3 (most stringent) indicates real pre-miR by *ProMiR II* (Nam *et al*., 2006).

In order to validate the computational pipeline and the miRNA cloning construction, both novel miRNAs were reserved for confirmation and characterization based on Northern Blot analysis (see section 6.2.5 for details) and *in situ* hybridization (see section 6.2.6 for details), respectively.

6.2.5. Northern Blot Validation of Novel MicroRNAs

(Figure 6.6) To provide experimental evidence for the existence of *bona fide* miRNAs, Northern Blot is the preferred method for validating novel miRNAs, as it is extremely sensitive for detecting miRNAs and it also determines the sequence length of the RNA species (Kloosterman *et al*., 2006). Total RNA samples were derived from three adult zebrafish organs of both sexes, namely, the adult ovary and testis (*AOV* and *ATE*), adult male and female brain (*AMB* and *AFB*), as well as adult male and female gill (*AMG* and *AFG*). Juvenile zebrafish was excluded, as the total RNA samples yielded from the corresponding organs were insufficient for Northern Blot analysis.

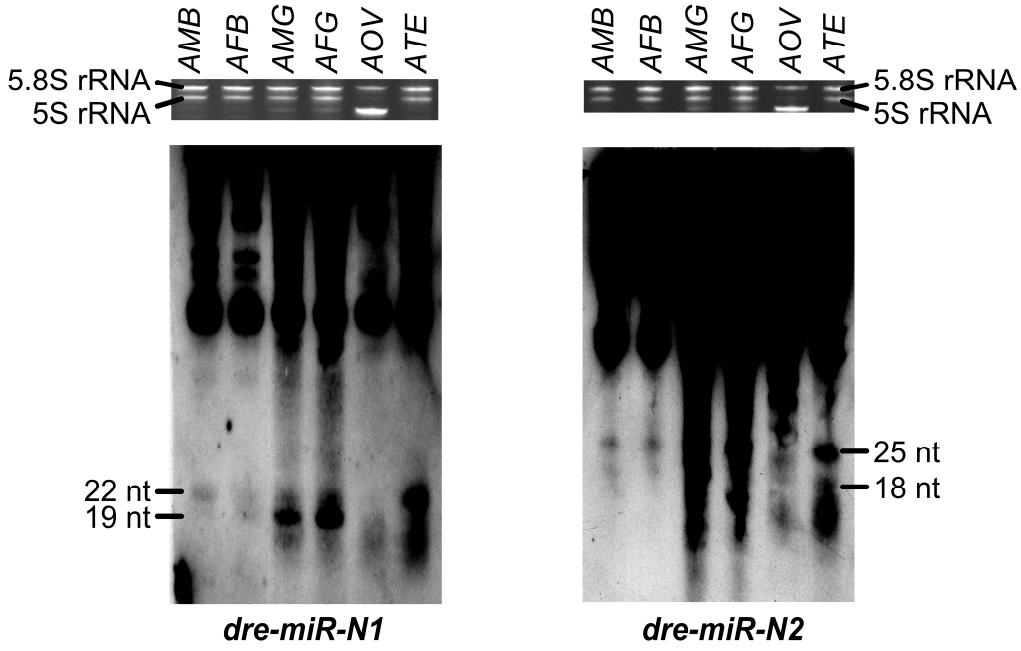

Figure 6.6: Northern Blot validation of two selected novel miRNAs *dre-miR-N1* and *dre-miR-N2*. Adult Male and Female Brain (*AMB* and *AFB*); Adult Male and Female Gill (*AMG* and *AFG*); Adult Ovary and Testis (*AOV* and *ATE*). Size in nucleotides (nt) indicates RNA length.

(Figure 6.6, *Left*) The first novel miRNA *dre-miR-N1* was moderately expressed in the *AMG* and *AFG* as well as the *ATE*, and weakly expressed in the *AMB* in the range of ~19-22 nt in length, but deficient in the remaining two samples *AFB* and *AOV*. (Figure 6.6, *Right*) Similarly, the second novel miRNA *dre-miR-N2* was strongly expressed in the *AMG* and *AFG*

as well as *ATE*, and weakly expressed in the *AMB* and *AFB* as well as *AOV* at about 25nt in length, which is comparable (within the limit of Northern Blot resolution) to the canonical miRNA of ~22 nt. These Northern Blot results provided the experimental evidence to authenticate that both novel miRNAs *dre-miR-N1* and *dre-miR-N2* are *bona fide* novel miRNAs expressed in a sexually dimorphic manner across the ovary, gill, and testis in adult zebrafish of both sexes. It is reasonable to postulate that similar visible Northern Blot bands and positive observations could be derived from juvenile zebrafish.

6.2.6. Characterization of Novel MicroRNAs using In Situ Hybridization

(Figure 6.7) To examine the macroscopic distribution and cellular localization of novel miRNAs in a heterogeneous cell population, expression pattern characterization by frozen section *in situ* hybridization was performed on juvenile and adult gonads. Generally, both novel miRNAs *dre-miR-N1* and *dre-miR-N2* were preferentially expressed in a germ-cell specific manner. Cross-hybridization with other members of the respective miRNA families was negligible or at least minimally controlled (technically infeasible to be eliminated entirely), given that a single mismatch in the locked-nucleic acid (LNA) modified oligonucleotide probe significantly reduced the hybridization signal (Kapsimali *et al*., 2007).

(Figure 6.7 A/B) In the adult ovary, their LNA modified probes were expressed exclusively in stage I and II oocytes but not in stage III oocytes, (Figure 6.7 C/D) while in the adult testis they were expressed mainly in primary spermatocytes and absent in secondary spermatocytes. Their distinctive expression patterns characterized by *in situ* hybridization were shared by two selected known miRNAs *dre-miR-19a* and *dre-miR-25* for (Figure 6.8 A/B) adult ovary and (Figure 6.8 C/D) testis, respectively. Interestingly, little or no detectable miRNAs were expressed in the newly fertilized stage V oocytes (0 hpf embryos) as reported by a recent miRNA microarray study (Wienholds *et al*., 2005). Together with our *in situ* hybridization data on stage III oocytes were likely to be devoid of miRNA expression, they point to mature, unfertilized oocytes (stage III – V) in the adult ovary are likely to share this trait.

(Figure 6.7 E/F) In the juvenile ovary, little or no *dre-miR-N1* was detected in the stage I oocytes as compared to the surrounding tissue, while *dre-miR-N2* was visible at a significant level indicating that the former was likely to be expressed at a temporally later stage than *dre-miR-N2* in the female germ cells of the juvenile ovary. (Figure 6.8 E/F) Similar observation could be made for *dre-miR-19a* and *dre-miR-25*, except that the former was expressed at a

temporally earlier stage than *dre-mir-25* in the female germ cells of the juvenile ovary, for which further functional experiments will likely to elucidate the mechanism and biological significance of this phenomenon. Juvenile testis was inadvertently excluded from the entire *in situ* hybridization experiments as repeated attempts with both tested pairs of known and novel miRNAs was technically unsuccessful, possibly due to the minuscule size of the juvenile testis in comparison to the other evaluated organs and tissues.

(Figure 6.9) Given that *dre-miR-N2* was well expressed across the six samples based on the northern blot analysis shown in Figure 6.6, it was selected for follow-up frozen section *in situ* hybridization experiments with various non-sex tissues in juvenile and adult zebrafish. Interestingly, *dre-miR-N2* was differentially expressed at a much higher level in a variety of female tissues than that of the male including (Figure 6.9 A/B) the epithelium of gills in 35 dpf individuals, as well as (Figure 6.9 C/D) the muscle and connective tissue in the trunk of juveniles. (Figure 6.9 E/F) Furthermore, *dre-miR-N2* was strongly expressed in the periphery of the corpus cerebelli in the adult female brain, but absent in the adult male brain. This finding is in corroboration with a previous study demonstrating that sexual differences occurred in the zebrafish brain with respect to cell differentiation (Ampatzis and Dermon 2007), which together raised the possibility that miRNAs could generally contribute to the "higher order" differences in the brains. These results also demonstrated that sexually dimorphic expression of *dre-miR-N2* was not limited to "canonical" sex-related organs such as the gonads, and that seemingly "sexes-unrelated" tissues associated with the brain, gill, and muscle/connective tissues possessed the capacity to exhibit this sexually dimorphic expression.

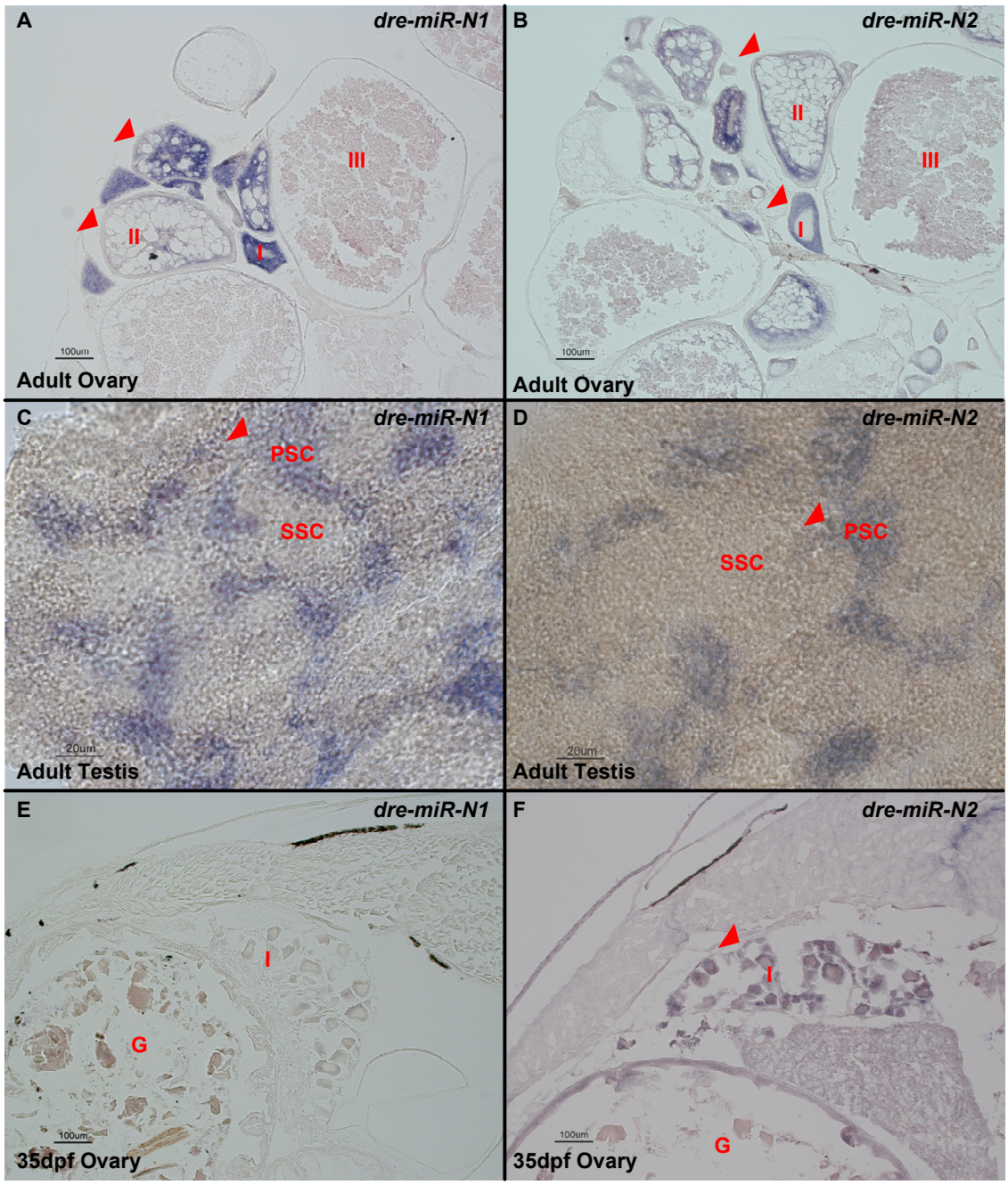

Figure 6.7: *In situ* hybridization of novel miRNAs *dre-miR-N1* and *dre-miR-N2* showing expression patterns in zebrafish gonads. Stage I/II oocytes (I/II); Primary spermatocytes (psc); Secondary spermatocyte (ssc); Gut (G).

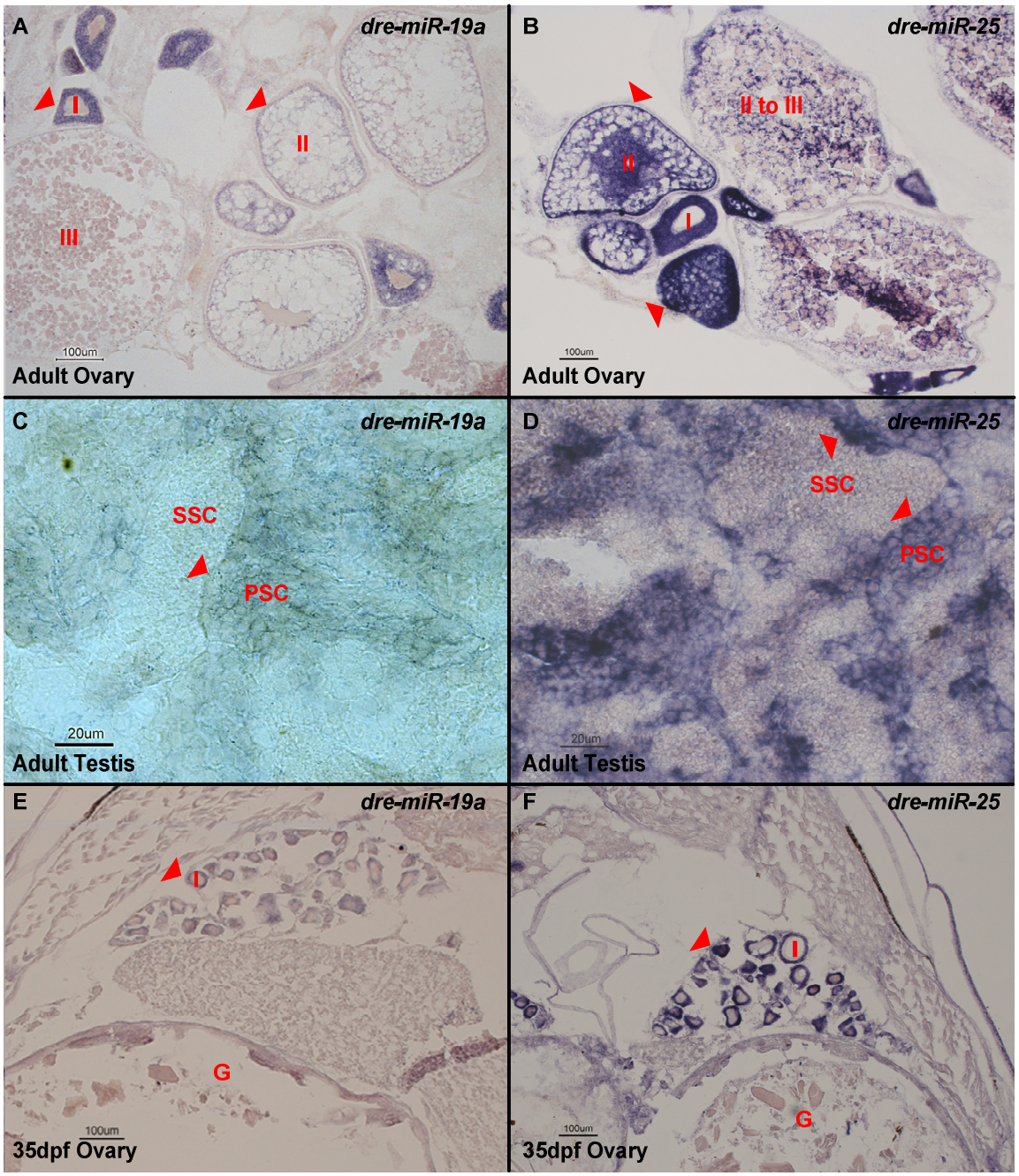

Figure 6.8: *In situ* hybridization of two known miRNAs *dre-miR-19a* and *dre-miR-25* showing expression patterns in zebrafish gonads. Stage I/II oocytes (I/II); Primary spermatocytes (psc); Secondary spermatocyte (ssc); Gut (G).

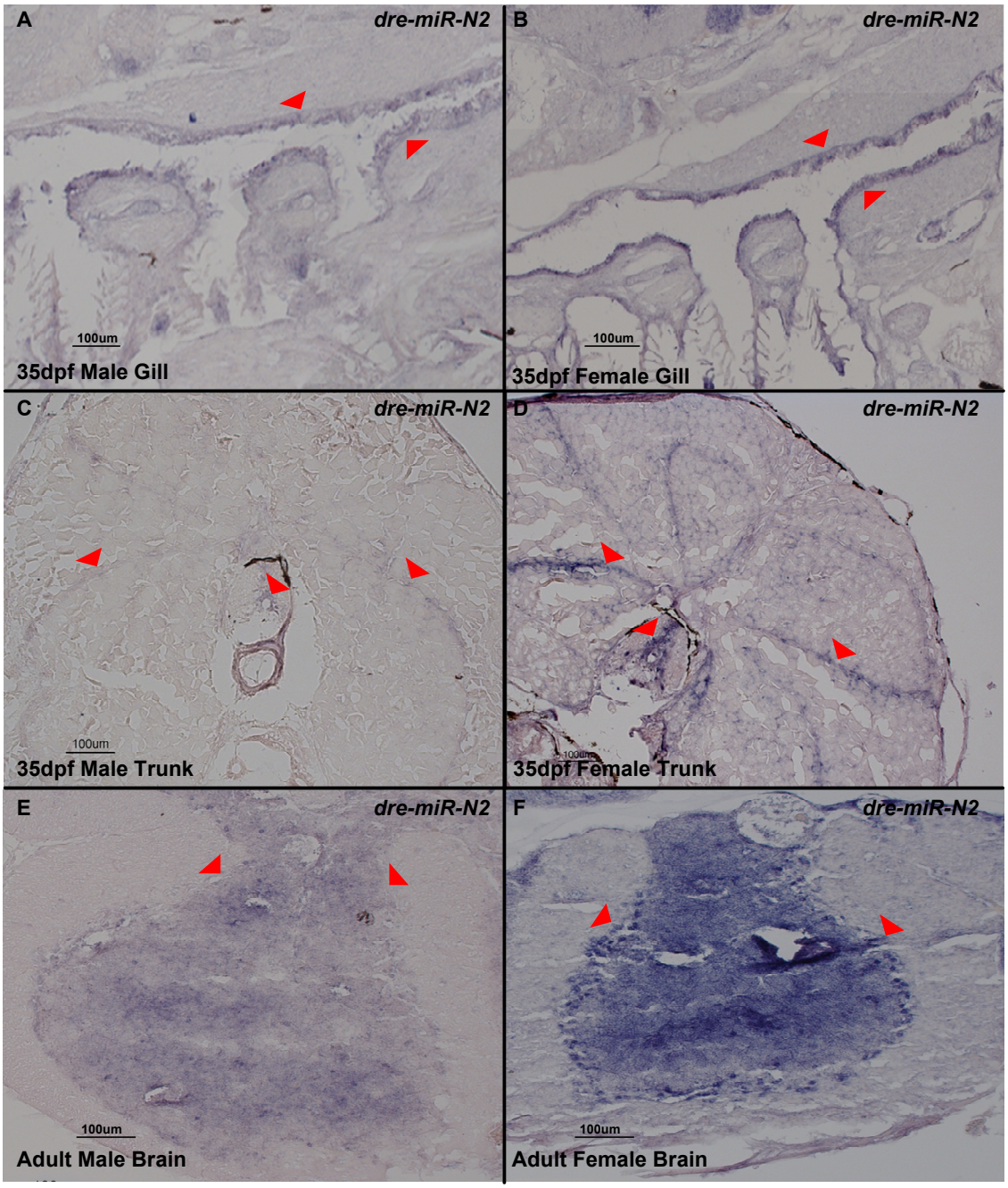

Figure 6.9: *In situ* hybridization of novel miRNA *dre-miR-N2* showing sexually dimorphic expression across juvenile gill, muscle tissue, and adult brain.

6.3. Methods and Materials

6.3.1. RNA Isolation

All organs were dissected from a homozygous *vas::egfp* transgenic line (Wang and Orban 2007). 35 days post fertilization (dpf) zebrafish were sexed according to the expression of enhanced green fluorescent protein (EGFP). Individuals with strong EGFP-derived fluorescence in their gonads were classified as females, while those with no or weak detectable EGFP-derived fluorescence in their gonads were classified as males. Small RNAs with length less than 200nt were isolated using the mirVana miRNA Isolation Kit (Ambion) according to the manufacturer's instructions.

For Northern Blot analysis, total RNA from zebrafish organs was isolated using Trizol reagent (Invitrogen), but with the following modifications: after addition of 100% isopropanol for RNA precipitation, the samples were incubated overnight at -20^{o}C; one RNA pellet wash was performed using 0.5ml 80% ice cold ethanol; the RNA pellet was dissolved in RNase-free H_2O.

For real-time RT-PCR analysis, total RNA from zebrafish organs was isolated using mirVana miRNA Isolation Kit (Ambion) and DNase-treated using DNA-free (Ambion). RNA isolations were separate from those used for small RNA library construction. 50ηg of total RNA was used for each real-time RT-PCR analysis.

6.3.2. Small RNA Library Construction

The miRNA Amplification Profiling (mRAP) protocol (Takada *et al*., 2006) was used for library construction, along with several modifications:

(Figure 6.10, *Left*) For the adult ovary and testis libraries (*AOV* and *ATE*), 55μg of small RNA was electrophoresed on a 15% denaturing polyacrylamide gel. The eluted RNA from 19-24nt was dephosphorylated, ligated to 3' adaptor (5' TGTAAGCTTTAACCGCGAATTCG 3'), subjected to RT-PCR, ligated with 5' adaptor (5' GCACCACGTATGCTATCGATCGTGAGA-TGGG 3'), and filling in as previously described (Takada *et al*., 2006). The RT-PCR amplicons were fractionated on a 10% nondenaturing polyacrylamide gel and the 85-90bp fragments were eluted, precipitated in 100% EtOH overnight at -20°C and resuspended in distilled water. 20ηg of DNA was used for re-amplification by PCR using an exponential number of cycles. The re-amplification cycles were 94°C for 15sec, 55°C for 20 sec, and 72°C for 2min.

PCR re-amplification primers were identical to those used in the first round of amplification. Re-PCR products were then purified with QIAquick PCR Purification Kit (Qiagen), digested with *BanI* endonuclease (New England Biolabs), and subjected to Chroma-Spin 30 (Clontech) purification, before concatamerization using T4 DNA Ligase (New England Biolabs) by incubation at 4°C overnight. DNA concatamers were further purified with Chroma-Spin 100 (Clontech), 3'-A tailed using Taq DNA polymerase (New England Biolabs) at 72°C for 30min, and ligated into the pGEM-Teasy vector (Promega).

For the 35 dpf ovary, testis, and brain libraries (*5WO*, *5WTE*, *5WMB*, and *5WFB*), 200ηg of small RNA was subjected to mRAP (Takada *et al*., 2006) with the abovementioned modifications, except that prior electrophoresis on a 15% denaturing gel was omitted. The isolated small RNA fraction was directly subjected to dephosphorylation and subsequent processing as mentioned previously (Takada *et al*., 2006) without initial size fractionation.

6.3.3. Computational Pipeline for Identification of Novel MicroRNAs

(Figure 6.10, *Right*) The computation pipeline for sequence analysis of small RNAs consisted of four stages. In the first, cloning vector, as well as 5' adapter sequence (5' GCACCACGTATG-CTATCGATCGTGAGATGGG 3') and 3' adapter sequences (5' TGTAAGCTTTAACCGCGA-ATTCG 3') were masked and redundancy removed from each concatamer using a custom-made *Perl* program *extractsmallRNAs.pl*. The implemented algorithm took into consideration that not all cloned inserts matched perfectly to the zebrafish genome. Manual curation of non-matching sequences pointed to three sources of errors: mutations occurring in the 5' and/or 3' adapter sequences; deletion of entire 5' and/or 3' adaptors; duplications of entire 5' and/or 3' adaptors. These anomalies might be artifacts of the cloning, PCR, and sequencing procedures, or a consequence of non-templated modification of small RNAs. The insets of length ranging 18-30bp inclusively, were then mapped to five independent databases (denoted as *rfam_ncRNAs*, *literature_ncRNAs*, *miRBase_mature*, *miRBase_hairpins*, and *literature_piRNAs*) using the *NCBI Blastn* 2.2.12 (McGinnis and Madden 2004); parameters optimized for short sequences were word-size = 7, as well as no masking of low compositional complexity and lower case.

For the functional annotation, *rfam_ncRNAs* database used the entire data set of 32,540 full-length ncRNA sequences excluding miRNAs from *Sanger Rfam* 8.0 (Griffiths-Jones *et al*., 2005), which is publicly available at the *http://www.sanger.ac.uk/Software/Rfam*. The 60,067 ncRNA sequences constituting the *literature_ncRNAs* database were obtained from a previous

publication related to the small RNA cloning on zebrafish (Chen *et al*., 2005). It was assembled from rRNA, tRNA, snRNA, snoRNA, and mRNA sequences by querying the *NCBI GenBank* online database from *http://www.ncbi.nih.gov/Genbank/index.html* for numerous species including the *Danio rerio*, *Mus musculus*, *Homo sapiens*, *Barbus barbus*, *Carassius carassius*, *Cynoscion nebulosus*, *Cyprinus carpio*, *Gobio gobio*, *Notropis hudsonius*, *Pimephales promelas*, *Rutilus rutilus*, *Oncorhynchus mykiss*, *Salvelinus alpinus*, and *Salmo trutta*. The entire 4,584 pre-miR and 4,430 mature miRNA sequences were directly downloaded from *Sanger miRBase* 9.2 (Griffiths-Jones *et al*., 2006), which is publicly available at the *http://microrna.sanger.ac.uk/sequences* for both *miRBase_hairpins* and *miRBase_mature* databases, respectively. Lastly, a total of 218,100 published piRNA sequences for *literature_piRNAs* were identified from two sources for zebrafish and mouse (Houwing *et al*., 2007; Grivna *et al*., 2006).

For the third stage, miRNA candidates (small RNAs that were clearly not belonging to any of the annotated ncRNAs, piRNAs, pre-miRs, and mature miRNAs) were mapped to the close to completion zebrafish Zv6 genome, *ftp://ftp.ensembl.org/pub/assembly/zebrafish/Zv6release/* using the same *NCBI Blastn* 2.2.12 program (McGinnis and Madden 2004) and parameters. The zebrafish genome assembly (Zv6 March 2006) containing 1,626,077,335bp, is approximately half the size of the available human genome sequence. Hits found on the zebrafish genome were then extended with 50 bases flanking sequences from both ends, and their most stable secondary structures predicted by the *RNAfold* program (Hofacker 2003). Only selected regions that folded into hairpins satisfying three criteria were reserved for subsequent analysis − had length of at least 70 bases; possessed termini loops of at least three nucleotides; were embedded with an inset in one of their hairpin arms with at least 75.00% overlap. These filtering steps were implemented in a custom-made *Perl* program *extracthairpins.pl*.

In the final stage, candidate pre-miRs were classified by a custom-made *Perl* program *mipred.pl* into putative pre-miRs or pseudo hairpins, based on the implementation of a SVM-based classifier *miPred* using intrinsic RNA folding measures from earlier works (Ng and Mishra 2007a; Ng and Mishra 2007b). Hairpin candidates that passed the classification with *miPred* score (0.5, 1.0] were assigned as putative pre-miRs; those with [0.0, 0.5] were considered as pseudo hairpins. Selected putative miRNAs embedded in one arm of the putative pre-miRs were reserved for further validation using Northern Blotting and *in situ* hybridization.

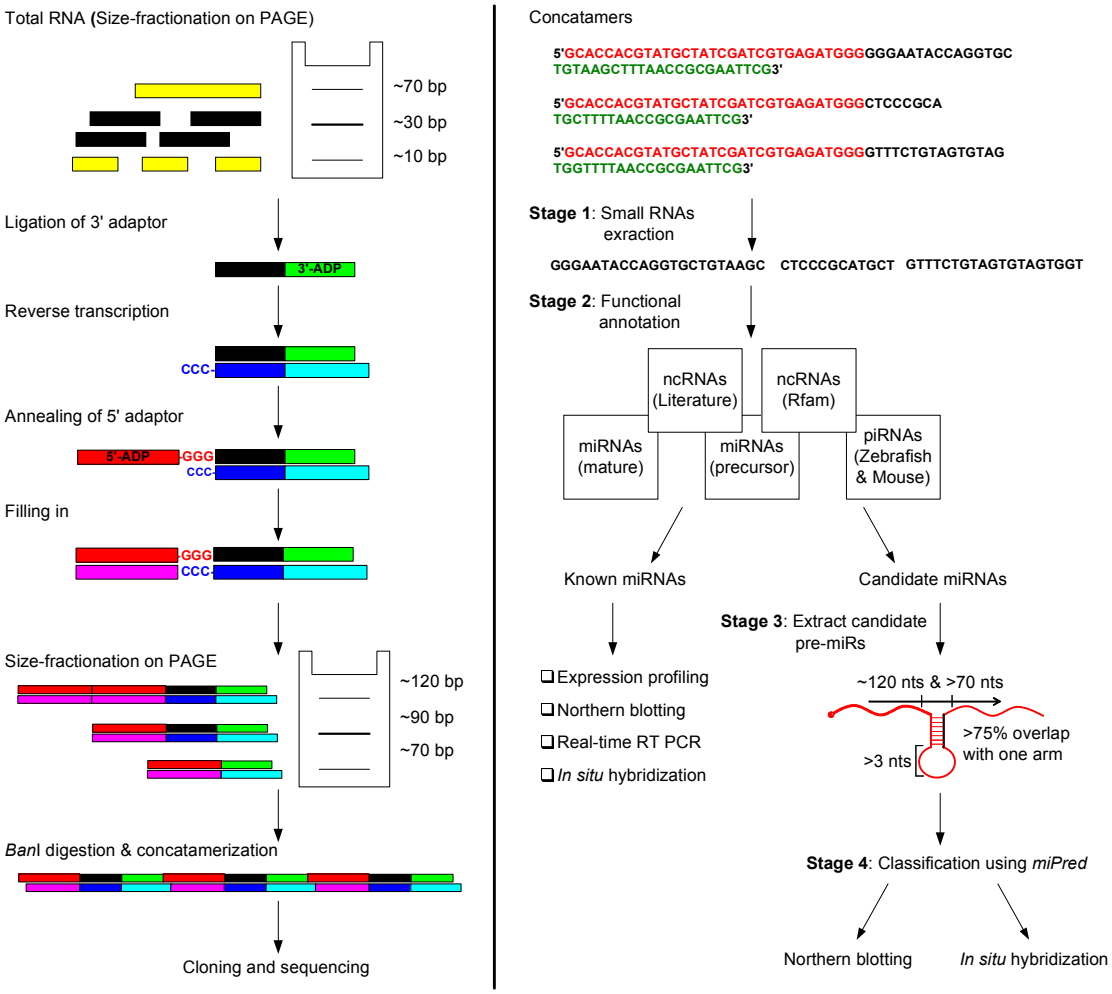

Figure 6.10: Experimental and computational pipeline for small RNAs cloning and sequencing, as well as candidate precursor miRNAs screening and classification.

6.3.4. Real-time RT-PCR

Real-time RT-PCR was performed using mirVana qRT-PCR miRNA Detection Kit (Ambion) and Hairpin-it miRNA Real-Time PCR Quantitation Kit (Genepharma) for *dre-let-7g*, *dre-let-7j*, dre-*miR-125b*, *dre-miR-130a*, and *dre-miR-143*. Reactions were carried out according to the manufacturer's instructions, and using the MyiQ Single-Color Real-Time PCR Detection System (Bio-Rad).

β-actin was used with iQ Supermix (Bio-Rad) for normalization. The β-actin forward and reverse primer sequences were 5' CCATCCTTCTTGGGTATGGAATC 3' and 5' GGTGGGGC-AATGATCTTGATC 3', respectively. The forward and reverse primers used for the known

miRNAs tested were proprietary Intellectual Property to Ambion and Genepharma.

6.3.5. Northern Blotting

Northern blot analysis of both novel miRNAs *dre-miR-N1* and *dre-miR-N2* was based on the protocol described in previous work (Kloosterman *et al*., 2006), except with the following modifications: each RNA sample of 1µg was ran on a 15% denaturing polyacrylamide gel (PAGE) until the bromophenol blue dye front of the loading buffer migrated to the bottom of the gel. The gel was then stained in 50ml of EtBr/0.5x TBE for 30min, followed by destaining and equilibration in 50ml of 0.5x TBE for 15min before imaging. Further destaining and equilibration was performed as described, after imaging. Hybond N^+ nylon membrane (Amersham) pre-equilibrated in 0.5x TBE for 15min was used for the individual Northern Blot experiments. Transfer was performed at 120mA for 30min. Crosslinking was subsequently performed at 1200µJ for 25-50sec using Stratalinker (Stratagene). Membranes were pre-dried for 15min before further use.

After crosslinking, membranes were prehybridized at 65^{o}C with DIG Easy Hyb (Roche). Pre-heated probe solution (100ηg/µl, pre-heated to 65^{o}C) was incubated with blots overnight at 28^{o}C. Stringency washes were then performed: 2x 5min in 2x SSC, 0.1% SDS at RT; and 2x 15min in 2x SSC, 0.1% SDS at 42^{o}C with manual agitation. Following this, immunological detection was performed as according to manufacturer's instructions (Roche).

Size determination was performed using 2µg of 10bp DNA ladder (Invitrogen) denatured at 70^{o}C for 5min and taking into account the fact that DNA migrates approximately 10% faster than RNA in denaturing polyacrylamide gel (Grivna *et al*., 2006).

The probe sequences were *dre-miR-N1* 5' AAACCUGUAACCACUGACAUC 3', and *dre-miR-N2* 5' UCCCCCACAGUCCAAAGACA 3'.

6.3.6. Frozen Sections In situ Hybridization

In situ hybridization was performed exactly as previously described in Exiqon's manual "MicroRNA Protocol for In-situ Hybridization on Frozen Sections", publicly available at the *http://www.exiqon.com/uploads/Frozen_sections_in_situ_hybridization(5).pdf*. 3'-DIG labeled locked-nucleic acid (LNA) modified oligonucleotide probes were purchased from Integrated DNA Technologies.

Specimens were fixed in 4% paraformaldehyde in PBS at 4°C overnight, embedded in 2%

agar, and soaked in 30% sucrose overnight at 4°C overnight. Embedded sections were mounted in Jung freezing medium (Leica) and sectioned at 12μm using a cryomicrotome (Leica). Sections were fixed in 4% paraformaldehyde in PBS for 4°C for 10 min, acetylated, and hybridized with 25-40ηM LNA modified probes overnight at a temperature 20-22°C below probe melting temperature. After the hybridization, slides were washed and blocked at 4°C overnight. Anti-digoxigenin-AP, Fab fragments (Roche) at 1:1500 were incubated with sections at 24°C for 4h before NBT/BCIP (Sigma) development at 24°C overnight. Frozen sections from *in situ* hybridizations were observed and analyzed using Axioplan inverted microscope. Images were captured with Metamorph ACT-1 software and processed using Adobe Photoshop 7.0.

The probe sequences were *dre-miR-19a* 5' TCAGTTTTGCATAGATTTGCACA 3', *dre-miR-25* 5' TCAGACCGAGACAAGTGCAATG 3', *dre-miR-N1* 5' AAACCTGTAACCACTG-ACATC 3', and *dre-miR-N2* 5' TCCCCCACAGTCCAAAGACA 3'.

6.4. Summary

Expression profiles of 88 (out of 219 existing ones) known miRNAs based on six small RNA libraries, revealed that the cloned miRNAs were generally expressed in the gonads and brains of juvenile and adult zebrafish during specific stages of development (e.g., *dre-let-7e*, *dre-miR-101a*, and *dre-miR-101b* in adult gonads; *dre-miR-140**, *dre-miR-199*, and *dre-miR-141* in juvenile gonads), some were expressed ubiquitously (e.g., *dre-let*-7 families), but many were expressed in a tissue specific manner (e.g., *dre-miR-122* in ovarian libraries and *dre-miR-29a* in "male only" organs). Interestingly, two pairs of miRNA:miRNA* (i.e., *dre-miR-199* and *dre-miR-199**; *dre-miR-20a* and *dre-miR-20a**) had dissimilar expression pattern, suggesting the hypothesis that selective degradation of miRNA and/or miRNA* by RISC occurs in an organ and/or tissue-specific fashion.

Remaining small RNAs that did not match annotated databases containing published sequences of ncRNAs, piRNAs, and known miRNAs, were subjected to *miPred* classification and experimental validation. Two novel miRNAs *dre-miR-N1* and *dre-miR-N2* predicted by *miPred* were confirmed by Northern Blotting as *bona fide* miRNAs. Furthermore, they were detected exclusively via *in situ* hybridization in stage I and II oocytes but not in stage III oocytes of adult ovary; they were expressed preferentially in the primary spermatocytes but absent in the secondary spermatocytes of the adult testis.

Chapter 7.
Conclusion and Future Directions

7.1. Conclusion

In this thesis, an extensive literature survey was conducted to investigate comprehensively existing works on the identification of novel precursor miRNAs (pre-miRs), which faced technical limitations in distinguishing them from dysfunctional pseudo hairpins that are pervasive in many genomes. Because experimental techniques are labor-intensive and highly biased towards abundant miRNAs, comparative-based approaches seek to detect miRNA genes that are conserved in the primary sequences and secondary structures similar to known ones. Improvements were proposed to resolve identification of novel miRNAs that have no known close homologies due to the lack of genomic data for species that do not have any closely related ones, or due to the possible evolution of miRNAs. However, recent *ab initio* prediction methods relying exclusively on the characteristic of hairpin-shaped structures of pre-miRs for identifying novel miRNAs, also suffer major limitations from the use of phylogenetic information to improve prediction accuracy, are restricted to only strict hairpin-shaped structures, and the use of extrinsic parameters that define physical parts of a pre-miR.

A definitive criterion for identifying and classifying accurately promising precursor transcripts as *bona fide* pre-miRs, while discriminating against abundant pseudo hairpins within a single genome has not yet been discovered. Moreover, discriminative features incorporated in existing (quasi) *de novo* predictors have reported far from satisfactory performances, especially when cross-species conservation information is unavailable. Through a comprehensive large-scale characterization study on the entire hairpin-shaped structure of known pre-miRs from diverse species including that of vertebrate, plant, nematodes, and viruses, pre-miR was found to possess a set of thirteen statistically significant global and intrinsic features (Ng and Mishra 2007b). This *in silico* findings has greatly advanced our understanding of miRNA functions and biogenesis in relation to their structural features and distinct folding patterns.

The investigated features relating to the intrinsic folding and topological characteristics of pre-miRs were integrated into the development of an improved *de novo* SVM classifier *miPred* for identifying specie-specific and non-conserved pre-miRs, wholly independent of phylogenetic conservation information (Ng and Mishra 2007a). It yielded comparable or significantly better predictive 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.

Application of *miPred* for the identification of novel miRNAs expressed in the gonads and brain of zebrafish yielded two novel ones, which were validated by Northern Blotting as *bona fide* miRNAs (Beh and Ng *et. al.* 2007; *in preparation*). Both of them were detected exclusively via *in situ* hybridization in stage I and II oocytes but not in stage III oocytes of adult ovary; expressed preferentially in the primary spermatocytes and not in secondary spermatocytes in the adult testis. These results clearly showed that deployment of *miPred* in future screening projects would likely to yield considerable saving on precious and scarce experimental resources devoted to validating significantly fewer false-positives, since it is highly assured that those precursor transcripts predicted would be experimentally confirmed as functional pre-miRs.

7.2. Expressed Sequence Tags Analysis of MicroRNAs

In practice, designating putative pre-miRs as authentic members is conditional upon them conforming to a set of strict (but constantly refined) empirical criteria (Ambros *et al*., 2003a). First, these putatives must fold into hairpin-shaped structures with sufficient base-pairings more than 16 nucleotides (nt) in the stem to facilitate the maturation of miRNAs, which effectively referred to those consensus sets of pre-miRs. Second, the expression of pre-miRs or mature forms must be quantifiable by wet-lab experimental means (Berezikov *et al*., 2006). Alternatively, Expressed Sequence Tags (ESTs) could be utilized as an economically feasible and high-throughput vehicle for the transcriptome analysis and authentication of pre-miRs.

Briefly, ESTs are partial and single-pass sequence reads (~200−500 nt) generated from either ends of randomly sampled cDNA libraries of expressed genes (Adams *et al*., 1991). EST sequencing strategy generally favors long stretch of stable and polyadenylated mRNAs. Together with direct experimental evidence consistently pointing to RNA Polymerase II (Pol-II) as the transcription machinery of miRNAs (Lee *et al*., 2004; Bracht *et al*., 2004; Cai *et al*.,

2004), pre-miRs encoded within the primary transcripts should be present in ESTs. Circumstantial and direct experimental works have demonstrated the effectiveness of EST analysis in providing expression evidence for the existence of known pre-miRs in plants (Zhang *et al*., 2005; Bonnet *et al*., 2004a; Palatnik *et al*., 2003) and vertebrates (Jin *et al*., 2006; Li *et al*., 2006; Smalheiser 2003). It is tantalizing to extend the scope of EST analysis to minimize the false-positive rate especially of candidate pre-miRs from human and mouse, as both species have captured the largest repository of EST entries in the latest *dbEST* (Boguski *et al*., 1993).

Several outcomes concerning the preliminary EST analysis of latest published human pre-miRs are highlighted. Pairwise sequence similarities were assessed via *NCBI Blastn* 2.2.12 (McGinnis and Madden 2004); parameters optimized for short sequences were word-size = 7, as well as no masking of low compositional complexity and lower case. First, pre-miRs located and sense oriented within introns of pre-mRNAs share the same promoter with their encoded genes but undergo spliceosomal excision from the Pol-II driven transcript when the mRNA serves as a template for protein synthesis (Lin *et al*., 2006). In principle, EST analysis should be ineffective towards sense intronic pre-miRs. However, 19.01% (27/142) of these pre-miRs were readily detected; none matched *hsa-miR-[28*, *101b*, *103*, *107*, *140*, *152*, *153-1*, *153-2*, *218-1*, *218-2]* (Lin *et al*., 2006). Second, EST analysis was sufficiently sensitive to 25.36% (70/276) of intergenic or antisense oriented pre-miRs to neighboring genes, which are suspected to be transcribed as independent units possessing their own (not necessarily Pol-II) promoters. Two highly intra-related intergenic clusters specific to human embryonic stem cells: *hsa-miR-[367*, ***302a***, ***302b***, ***302c***, *302d]* (antisense oriented; ~700bp; chromosome 4) and *hsa-miR-[**371**, **372]* (sense oriented; 500bp; chromosome 19), are expressed as polycistronic and polyadeylated primary transcripts (Suh *et al*., 2004). Several of them (bold) were detected by EST analysis. Surprisingly, a well-characterized intergenic cluster *hsa-miR-[24-2*, *27a*, *23a]* (antisense oriented; ~2.2 kbp; chromosome 19) had no EST matches. The last two pre-miRs are 5' capped and polyadenylated 1.8 kb downstream of the 3' termini of *hsa-miR-24-2* that has a minimal ~600 bp Pol-II dependent promoter (Lee *et al*., 2004). Third, *hsa-miR-[515-1*, ***517a***, *517c*, ***519a-1]* residing in the chromosome 19 miRNA cluster (C19MC), of which two (bold) were present in the human EST population. These pre-miRs, unlike conventional ones expressed exclusively by Pol-II, undergo RNA Pol-III mediated synthesis (Borchert *et al*., 2006). Additional EST searches for 18 predicted C19MC miRNAs and human miRNAs with upstream Alu-, tRNA- or mammalian-wide interspersed repeat (MWIR) dependent promoter elements that are strong candidates for Pol-III mediated transcription (Borchert *et al*., 2006), reported

merely three matching hits *hsa-miR-[34a*, *517b*, *594]*.

Considering that the annotated 116 human and 82 mouse pre-miRs from *Sanger miRBase* 9.0 had EST matches, transcriptome analysis at the finishing stage of the computational pipeline would provide more reliable expression evidence concerning the existence of novel pre-miRs that had not yet been experimentally characterized.

7.3. Prediction of MicroRNA Target Sites Associated with Human Diseases

Latest statistics suggests approximately 800 human miRNA genes (Bentwich *et al*., 2005; John *et al*., 2004) constituting ~1-2% of the known ~22,000 protein-coding genes, may actually regulate as many as one-third of the human genes (Du and Zamore 2005). However, the majority of their detailed regulatory functions remain largely unknown. A major obstacle that is stalling progress towards elucidating the exact causation of cellular processes linked to miRNAs is, our knowledge is greatly limited to only a few experimentally characterized miRNAs and their corresponding known regulated targets (Jiang *et al*., 2005; Ambros 2004). Certainly, prediction of human miRNA targets for post-transcriptional regulation would provide invaluable insights in at least two aspects. Firstly, knowing the miRNA targets provides an alternative approach to assign biological functions to the many miRNAs. Secondly, we would have a deeper understanding as to how dysfunctional miRNA might be associated with cancers, or might contribute to human diseases. However, this endeavor is a major challenge because high-throughput experimental methods for identifying human miRNA targets are not yet available.

Experimentally, miRNAs have shown to display differential expression levels in embryonic stem cells (Suh *et al*., 2004; Houbaviy *et al*., 2003), temporal and spatial expression patterns in normal tissues (Lagos-Quintana *et al*., 2002), and mammalian organs (Eder and Scherr 2005; Sempere *et al*., 2004; Krichevsky *et al*., 2003). Recently, comprehensive phenotypic and expression analysis even suggests an intrinsic association between oncogensis and human miRNAs (Lu *et al*., 2005) in tumor tissues (Iorio *et al*., 2005; Jiang *et al*., 2005; Michael *et al*., 2003; Calin *et al*., 2002). miRNAs may function as oncogenes or tumour suppressors, as they are frequently located at genomic regions involved in cancers (Gregory and Shiekhattar 2005) or are mediating antiviral defense in human cells (Lecellier *et al*., 2005). Expression profiles for 217 miRNAs distributed across 334 samples including cancers such as

leukemia and lymphomas, reflect informatively the developmental lineage and differentiation state of tumorgenesis (Lu *et al*., 2005). Experimental evidence (Bottoni *et al*., 2005; Calin *et al*., 2002; Stilgenbauer *et al*., 1998) also suggest that the human *miR-15a* and *miR-16* miRNAs located within 0.5 kb on chromosome 13q14, may be related to B cell Chronic Lymphocytic Leukemia (CLL), mantle cell lymphoma, multiple myeloma, and prostate cancer cases. This region has been known to be excised in these cancer types, and both genes are deleted or co-repressed in more than two third of CLL cases, strongly suggesting their active involvement in tumorigenesis.

In relation to human disease and potential therapies, islet-specific miRNA *miR-375* has been identified to be a regulator of insulin secretion and may constitute a novel pharmacological target for the treatment of diabetes (Poy *et al*., 2004). Overexpression and inhibition of endogenous *miR-375*'s function suppressed and enhanced glucose-induced insulin secretion, respectively. The mechanism by which insulin secretion is modified by *miR-375* is independent of changes in glucose metabolism or intracellular Ca^{2+}-signalling but correlated with a direct effect on insulin exocytosis. Myotrophin (*Mtpn*) was predicted to be and validated as a target of *miR-375*. Inhibition of *Mtpn* by siRNA mimicked the effects of *miR-375* on glucose-stimulated insulin secretion and exocytosis. Similarly, in zebrafish, *miR-375* is essential for formation of the insulin-secreting pancreatic islet. Loss of *miR-375* function interfered by morpholino oligonucleotides, causes defects in the morphology of the pancreatic islet. Although the islet is still intact at 24 hours post fertilization (hpf), the islet cells become scattered by 36 hpf (Kloosterman *et al*., 2007). Another miRNA *miR-133* has been reported to have a critical role in determining cardiomyocyte hypertrophy, suggesting its therapeutic application in heart disease. *In vitro* overexpression of *miR-133* inhibited cardiac hypertrophy (Care *et al*., 2007). In contrast, suppression of *miR-133* by 'decoy' sequences induced hypertrophy, which was more pronounced than that after stimulation with conventional inducers of hypertrophy. *In vivo* inhibition of *miR-133* by a single infusion of an antagomir caused marked and sustained cardiac hypertrophy. Specific targets of *miR-133* were identified and validated, namely, *RhoA*, a GDP-GTP exchange protein regulating cardiac hypertrophy; *Cdc42*, a signal transduction kinase implicated in hypertrophy; and *Nelf-A/WHSC2*, a nuclear factor involved in cardiogenesis. A cluster of cellular miRNAs (i.e., *miR-[*28, *125b*, *150*, *223*, *382]*) are pivotal in the latency of Human Immunodeficiency Virus type 1 (HIV-1) in resting primary $CD4^+$ T lymphocytes (Huang *et al*., 2007), which is the major barrier for the eradication of the virus in patients on suppressive highly active anti-retroviral therapy (HAART). Even with optimal HAART treatment,

replication-competent HIV-1 exists in resting primary $CD4^+$ T cells. A breakthrough discovery was that the 3' ends of HIV-1 mRNAs RNAs are targeted by the five miRNAs, which are enriched in resting $CD4^+$ T cells but not in activated $CD4^+$ T cells. Specific inhibitors of these miRNAs substantially counteracted their effects on the target mRNAs, measured either as HIV-1 protein translation in resting $CD4^+$ T cells transfected with HIV-1 infectious clones, or as HIV-1 virus production from resting $CD4^+$ T cells isolated from HIV-1–infected individuals on suppressive HAART. These results suggested suggest that manipulation of cellular miRNAs could be a novel approach for purging the HIV-1 reservoir.

Only recently has the development of computational approaches gain prominence and acceptance as a research-advancement tools (Brown and Sanseau 2005), namely the RNAhybrid, miRanda, PicTar, MovingTargets, TargetScanS, and miRanda. In part, they have gradually overcome the folding complexity caused by the imperfect and interrupted hybridization (Lewis *et al*., 2003) between the relatively short miRNAs and the target mRNAs, and the rules governing miRNA-mRNA target interactions are gradually defined from those already deduced from *lin-4*, *let-7*, and *bantam*.

To date, the identification for human miRNA targets to inform cancer diagnosis has not been systematically explored. Future contribution includes development of a computational approach for identifying all potential mRNA target sites for each miRNA sequence by considering the combinatoric folding of miRNA-mRNA duplex in the tripartite regions of mRNA and without relying on homology to other organisms. Several open questions will then be investigated − which miRNAs are the key regulators in the cellular system and which miRNA targets are highly and co-regulated by this set of miRNAs? What proportion of all genes is regulated by miRNAs in tumors? How many of those genes are regulated by each miRNA in tumors? Are specific cellular processes targeted by specific miRNAs or by miRNAs in general? What is the extent of cooperativity and multiplicity in miRNA-mRNA binding? Together, they will provide quantitative correlation between the degree of regulation and quality of the individual interactions (or their sum). This future work will serve as a small, but an important contribution towards identifying novel human miRNA targets and elucidating the full details of miRNA regulatory functions in tumorgenesis.

7.4. Transcriptional Regulation of MicroRNAs

Besides the identification of all potential mRNA target sites, another important (but lesser

researched) problem is elucidating the mechanisms for transcriptional regulation of miRNAs themselves given that most of their expression are highly cell/tissue specific − which promoter(s) and transcription factor(s) regulate the miRNA expression.

In a recent study, through a combination of mouse and human cells, *miR-10b* was reported to be highly expressed in metastatic breast cancer cells and positively regulated cell migration and invasion (Ma *et al*., 2007). Overexpression of *miR-10b* in non-metastatic breast tumours initiated robust invasion and metastasis. Expression of *miR-10b* is induced by the pleiotropic transcription factor TWIST (part of an undescribed regulatory pathway) that binds directly to the putative promoter of *miR-10b*. The *miR-10b* inhibits translation of the target mRNA encoding homeobox *HoxD10*, resulting in increased expression of a well-characterized pro-metastatic gene *RhoC*.

As part of the ongoing research towards systematic identification of miRNA promoters, novel and clustered pre-miRs from various species (human, mouse, and viruses) are being actively identified in combination with direct whole-genome measurement of *cis*-regulatory promoter activities by technologies including the paired-end ditag (Ng *et al*., 2005) and single-end Cap Analysis Gene Expression (CAGE) (Shiraki *et al*., 2003). An expanded repertoire of miRNA genes and regulatory mechanisms will definitely signify both a huge opportunity and technical challenge for mRNA target identification and comprehensive genome annotation, as we delve into the functional roles of miRNAs interplay with other genetic regulatory networks, biological pathways, and signaling cascades.

Appendix A.
RNAspectral

A.1. Representing RNA Secondary Structure as Planar Tree-graph

The primary structure of a linear RNA chain molecule is the nucleotide sequence $\mathbf{s} = s_1 s_2 \ldots s_i \ldots s_L$, and runs in the direction 5' $\rightarrow$ 3' terminus. L defines the number of nucleotides and $s_i \in \sum = (A, C, G, U)$ is the biochemical nucleotide at the i^{th} position. The RNA molecule **s** folds upon itself relatively rapid into a two-dimensional RNA secondary structure S (Tinoco and Bustamante 1999). The structure S is stabilized by the canonical Watson-Crick $G \equiv C$ and $A=U$, and wobble $G=U$ base pairings.

(Figure A.1) A planar RNA secondary structure S is mathematically described by a set of base pairings $(i, j) \in S$ connecting bases s_i and s_j, where $i < j$ (Moulton *et al*., 2000). Given (i, j) and $(k, l) \in S$, a nucleotide can base pair to at most one other nucleotide i.e., $i = k \Leftrightarrow j = l$. A set of $\Delta \in \mathbf{Z}^+$ consecutive base pairs defines a stem for stabilizing the structure against thermal fluctuations. The number of unpaired nucleotides between paired s_i and s_j should at most be $\theta \in \mathbf{Z}^+$ i.e., $i < j + \theta$; otherwise, the structural motif is considered an unpaired-loop of multi-branch, bulge, hairpin, or internal. Hairpin loop, folds upon itself; Internal loop, an unpaired region between two stems due to mismatched (e.g., AG and CU) or unpaired bases; Bulge loop, an asymmetrical internal loop formed from one strand; Multi-branch loop or junction, more than two stems coincide with some unpaired bases; Stem, a base paired region. Short and long dashed lines indicate unpaired nucleotides and paired bases. "•" and "—" represent vertex and edge.

(Figure A.1) The RNA structure S has two hairpin loops, an internal loop, a bulge loop, a multi-branch loop, and five stems. It is represented as a RNA planar tree-graph $G = (V, E)$ consisting of six vertices "•" and five edges "—" according to the following pair of vertex-edge rules (Gan *et al*., 2004; Fera *et al*., 2004). (**1**) Vertex V i.e., "•" denotes a set of $\theta \geq 1$ mismatched nucleotides or unmatched pairs of bases for hairpin loop, bulge loop, internal loop, the 5' and 3'

unpaired termini, and the multi-branch loop. In general, the vertices are arbitrarily labeled in the direction 5' $\rightarrow$ 3' terminus. (**2**) Edge E i.e., "—" denotes a RNA stem having $\Delta \geq 2$ consecutive complementary pairs stabilized by the canonical Watson-Crick $G \equiv C$ and $A=U$, and wobble $G=U$ base pairings.

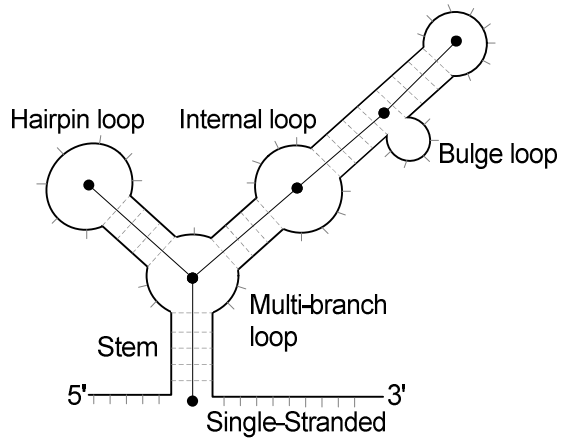

Figure A.1: Planar schematic of RNA secondary structure and its embedded motifs.

A.2. Converting RNA Planar Tree-graph to Laplacian Matrix

A RNA planar tree-graph $G = (V, E)$ is a mathematical formalism composed of n vertices $v_i \in V$, $i = (1, 2,\ldots,|V|)$ connected by m incident undirected edges $(v_i, v_j) \in E$, each of which is assigned an edge weight E_{ij}. Without loss of generality, edges are unweighted i.e., $E_{ij} = 1$ (Barash 2004b; Barash 2003). The tree-graph G in Eq. (A.1) is uniquely represented by the Laplacian matrix $\mathbf{L}(G)_{n\times n}$.

$$G = (V, E) \leftrightarrow \mathbf{L}(G) = \mathbf{D}(G) - \mathbf{A}(G). \tag{A.1}$$

Here $\mathbf{D}(G)_{n\times n}$ and $\mathbf{A}(G)_{n\times n}$ are known as the degree and adjacency matrices of the tree-graph G, respectively. The diagonal elements d_{ij} of $\mathbf{D}(G)_{n\times n}$ specify the degree or the minimum number of incident edges that each vertex v_i connects with the other vertices $v_j \neq v_i$, denoted by $\deg(v_i)$. d_{ij} takes on values of $\deg(v_i) = 1$ for hairpin loop, as well as 5' and 3' unpaired termini; $\deg(v_i) = 2$ for internal and bulge loops; and $\deg(v_i) > 2$ for multi-branch loop. The off-diagonal elements a_{ij} of $\mathbf{A}(G)_{n\times n}$ specify whether there exists an incident edge connecting the vertices v_i and v_j. If v_i and v_j are adjacent $a_{ij} = 1$, otherwise $a_{ij} = 0$.

$\mathbf{L}(G)_{n\times n}$ is a symmetric matrix having each of its rows and columns indexed by V, and individually total to zero. The value of element l_{ij} in Eq. (A.2) is given by the difference between d_{ij} and a_{ij}. It specifies the degree of connectivity between the vertices v_i and v_j of the tree-graph G.

$$l_{ij} = \begin{cases} d_{ij} = \deg(v_i), \text{ if } i = j, \\ -a_{ij} = -1, \text{ if edge } (v_i, v_j) \in E \ \wedge \ i \neq j, \\ 0, \text{ if edge } (v_i, v_j) \notin E. \end{cases} \tag{A.2}$$

Applying the "Eigen-decomposition theorem" onto $\mathbf{L}(G)_{n\times n}$, as shown in Eq. (A.3),

$$\mathbf{L}(G)\mathbf{X}=\lambda\mathbf{X} \Leftrightarrow \left[\mathbf{L}(G) - \lambda\mathbf{I}\right]\mathbf{X} = \mathbf{O}. \tag{A.3}$$

Here, the eigenvalue λ is taken as some scalar of $\mathbf{L}(G)_{n\times n}$ along with its corresponding eigenvector $\mathbf{X} \in \Re^n \neq 0$. $\mathbf{I}$ and $\mathbf{O}$ are the identity and null matrices. Equation (A.3) has non-trivial solutions if and only if the condition in Eq. (A.4) is satisfied,

$$\det\left|\mathbf{L}(G) - \lambda\mathbf{I}\right| = 0. \tag{A.4}$$

Solving the n^{th}-degree characteristic polynomial in Eq. (A.4) generates the entire set of ordered eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. This set is the matrix's eigenvalue spectrum quantifying the connectivity as well as characterizing the graph similarity. Generally, $\mathbf{L}(G)$ is always positive semi-definite such that the first eigenvalue $\lambda_1 = 0$ and those of higher orders $\lambda_{k>1} \in \Re^+$ (Barash 2004b; Barash 2003). According to the concept of "Spectral Graph Partitioning" that originates from the field of domain decomposition in parallel computing (Alex *et al*., 1990), the second (also known as the Fiedler) eigenvalue λ_2 represents mathematically the algebraic connectivity of the tree-graph G (Barash 2004b; Barash 2003). In relation to the RNA secondary structure, λ_2 measures the degree of compactness of the RNA topology at the coarsest scale (Barash 2004b; Barash 2003). RNA structures having similar values of λ_2 tend to be similar in topologies. Typically, the value of λ_2 increases monotonically with greater compactness in the RNA structure. Large values correspond to vertices of high degree that are in close proximity, while small values for more equally dispersed edge set. Maximum value of λ_2 is either 1 or 2 for an $n > 2$ perfectly connected star-shaped tree-graph or for $n = 2$ linear tree-graph, respectively (Barash 2004b; Barash 2003).

A.3. Pseudo Codes of RNAspectral Algorithm

"Spectral Graph Partitioning" has been extensively applied to a variety of bioinformatics problems: the prediction of multiple mutation to disrupt motifs in riboswitches (Barash 2003), the prediction of RNA conformational switch by mutation (Barash 2004a), the search and analysis of RNA secondary structures (Barash 2004b), the classification of RNA coarse-grained tree-graph structures (Gan *et al*., 2004; Fera *et al*., 2004), and lastly for systematically partitioning complex RNA structures into simpler fragments with maximal decoupling between them (Gan *et al*., 2003). These applications underscore the potential of "Spectral Graph Partitioning" as an invaluable computational tool to elucidate the topological patterns hidden in the post-genomic sequences and to offer a tremendous opportunity for an enhanced understanding of both functional and structural genomics.

"RNA Matrix Computer Program" (Gan *et al*., 2004; Fera *et al*., 2004) was the pioneering and only implementation of "Spectral Graph Partitioning" analysis on RNA structural folding. It is available online and provides a user-friendly web-interface for uploading a "ct file" produced by Zuker's *Mfold* prediction server (Zuker 2003; Zuker and Stiegler 1981) or equivalent. As an attempt to address the high-throughput demands of our in-house projects, *RNAspectral* was designed from scratch based on the mathematical formalisms gathered from literature, and iteratively validated against the "reference" results of "RNA Matrix Computer Program" (Gan *et al*., 2004; Fera *et al*., 2004).

The algorithm ***RNAspectral***(*S*) presents our strategy geared towards two tasks. Given a RNA secondary structure *S* described in a Vienna dot-bracket notation containing ".", "(", and ")" (Hofacker 2003), it first abstract *S* at the coarsest-scale into a planar tree-graph representation. This transforms uniquely the RNA structural motifs (hairpin loops, internal loops, bulge loops, and multi-branch loops, as well as stems) into a network of vertices connected by incident edges. Next, it computes the Fidler eigenvalue λ_2 from the Laplacian matrix corresponding to the tree-graph.

RNAspectral(*S*) uses two primary functions in *Line* 2−3, whose pseudo-codes are described in both functions ***optimizeStruct***(*S*) and ***parseStruct***(*S*), respectively. The former returns RNA structure *S′* and the latter returns the values for five global variables *totalpath, path*, *stems*, *ld*, *ls*, and *hs*. *Line* 4−5, sets the value of adjacency matrix **A** at row *path*[*i*] and column *path*[*i* + 1] to 1; 6−7, sets the value of degree matrix **D** at row *i* and column *i* to *ld*[*i*]; 8, computes the Laplacian matrix **L**; 9, the auxiliary function ***computeEigVals***(**L**) computes the

eigenvalue spectrum using the well-established "Eigen-decomposition theorem" and $\det|\mathbf{L}-\lambda\mathbf{I}|=0$.

In function ***optimizeStruct***(*S*), it implements the pair of vertex-edge rules described in subsection A.1. *Line* 1, vector *pt* contains the values returned by the auxiliary function ***makePBTable***(*S*), such that the *pt*[*i*] of nucleotide at position *i* has value of UNPAIRED when that nucleotide is unpaired or denotes the position of the base to which it is paired; 2−8, internal loops with only one pair of mismatches are identified and then paired; 9−12, stems with only one complementary pair are identified and then unpaired; 13−17, bulges having unpaired mono-nucleotide are deleted; 18, the resulting RNA structure *S′* is returned after applying the pair of vertex-edge rules.

In function ***parseStruct***(*S*), it implements the Eq. (A.1) and (A.2) described in subsection A.2. *Line* 1, *S′* is a RNA secondary structure specified in an extended dot-bracket format with additional symbols "[", and "]", returned by the auxiliary function ***auxStruct***(*S*), to track the onset of a helical stem-loop; 2−14 computes the Euclidean *path* transverse from the first to the final (*stems* + 1)th vertex, in the direction of 5' → 3' terminus of the given RNA sequence; the size of vector *path* is stored in the variable *totalpath*. The size of each vertex and stem measured by the number of unpaired bases and number of pairs, respectively, are tracked by two variables *ls* and *hs*; the degree of each vertex is stored in the variable *ld*.

1. **Global variables**: *totalpath* = 0; *path* = ϕ; *stems* = 0; *ld* = ϕ; *ls* = ϕ; *hs* = ϕ.

2. **Local variables**: adjacency matrix **A** = ϕ; degree matrix **D** = ϕ; Laplacian matrix **L** = ϕ.

3. Generate RNA structure *S*′ from ***optimizeStruct***(*S*),

4. Generate a vector consisting of 6 elements (*totalpath, path*, *stems*, *ld*, *ls*, *hs*) from ***parseStruct***(*S*′).

4. **For** *i* from 1 to *totalpath*, **do**

5. **A**[*path*[*i*]][*path*[*i* + 1]] = 1.

6. **For** *i* = 1 to *stems* + 1, **do**

7. **D**[*i*][*i*] = *ld*[*i*].

8. **L** = **D** – **A**.

9. ***computeEigVals***(**L**).

Figure A.2: Pseudo codes of algorithm RNAspectral(*S*). See section A.3 for details.

```
1. Variables: L = len(S), pt = makePBTable(S), S′ = S, j = 1.
2. For i = 1 to L − 1, do
3.     If pt[i] == UNPAIRED, then
4.         If min(pt[i − 1], pt[i + 1]) == UNPAIRED, then continue.
5.         If abs(pt[i − 1] − pt[i + 1]) == 2, then
6.             pt[i] = max(pt[i − 1], pt[i + 1]) − 1.
7.         pt[pt[i]] = i.
8.         S′[i] = "(", S′[pt[i]] = ")".
9.     If pt[i] ≠ UNPAIRED, then
10.        If pt[i − 1] == pt[i + 1], then
11.            S′[i] = S′[pt[i]] = ".".
12.            pt[pt[i]] = pt[i] = UNPAIRED.
13. For i = 1 to L − 2, do
14.    If pt[i] == UNPAIRED, then
15.        If abs(pt[i − 1] − pt[i + 1]) == 1, then continue.
16.        S′[j] = S′[i], j = j + 1.
17. S′[j] = S′[L − 1], j = j + 1, S′[j] = ϕ.
18. return S′.
```

Figure A.3: Pseudo codes of function *optimizeStruct*(*S*). See section A.3 for details.

```
1. Variables: L = len(S), pt = ϕ, stack = ϕ, j = 0.
2. Foreach S[i] such that i = 1 to L − 1, do
3.     If S[i] == ".", then pt[i] = UNPAIRED.
4.     If S[i] == "(", then stack[j] = i,  j = j + 1.
5.     If S[i] == ")", then
6.         i = i − 1, pt[i] = stack[j],
7.         pt[pt[i]] = i.
8. return pt.
```

Figure A.4: Pseudo codes of function *makePBTable*(*S*). See section A.3 for details.

```
1. Variables: L = len(S), S′ = auxStruct(S), loop = ϕ, lp = 0, j = 0.
2. Foreach S′[i] such that i = 1 to L − 1, do
3.     If S′[i] == ".", then ls[loop[lp]] = ls[loop[lp]] + 1.
4.     If S′[i] == "[", then
5.         path[totalpath] = loop[lp], totalpath = totalpath + 1, lp = lp + 1,
6.         stems = stems + 1, ld[stems] = 1,
7.         loop[lp] = stems.
8.     If S′[i] == ")", then j = j + 1.
9.     If S′[i] == "]", then
10.        hs[loop[lp]] = j + 1,
11.        j = 0,
12.        path[totalpath] = loop[lp] , totalpath = totalpath + 1,
13.        lp = lp − 1, ld[loop[lp]] = ld[loop[lp]] + 1.
14.  path[totalpath] = 0.
```

Figure A.5: Pseudo codes of function *parseStruct*(*S*). See section A.3 for details.

1. **Variables**: $L = \textbf{\textit{len}}(S)$, $mp = \phi$, $S' = S$, $o = 0$, $j = 0$.
2. **Foreach** $S'[i]$ such that $i = 1$ to $L - 1$, **do**
3. **If** $S'[i] =$ "(", **then** $o = o + 1$, $mp[o] = i$.
4. **If** $S'[i] =$ ")", **then**
5. $j = i$.
6. **While** $S'[j + 1] =$ ")" $\wedge$ $mp[o - 1] = mp[o] - 1$, **do**
7. $j = j + 1$, $o = o - 1$.
8. $S'[j] =$ "]", $i = j$, $S'[mp[o]] =$ "[", $o = o - 1$.
9. **return** S'.

Figure A.6: Pseudo codes of function *auxStruct*(*S*). See section A.3 for details.

A.4. ANSI C Source Codes of RNAspectral Algorithm

RNAspectral is an efficient and rapid algorithm, implemented in portable ANSI C programming language using the development platform Intel Pentium M 2.0 GHz, and 1.0 GB RAM; Cygwin 1.5.19-Windows XP. It provides a user-friendly command-line interface and four user-adjustable parameters: *-v1*, to enable the level of verbosity for obtaining output identical to that of "RNA Matrix Computer Program" (Gan *et al*., 2004; Fera *et al*., 2004); *-v2*, to enable detailed debugging and further analysis into *RNAspectral* internalities; *--noopt*, to disable the pair of vertex-edge rules; *--monitor*, to monitor the execution time. Together, these options and functionalities allow the inexperienced user to integrate the information from "Spectral Graph Partitioning" analysis such as the second eigenvalue λ_2 and the number of vertices as part of their experimental methodologies, in an intuitive manner.

```c
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <time.h>
#include <gsl/gsl_matrix.h>
#include <gsl/gsl_math.h>
#include <gsl/gsl_eigen.h>
#include <gsl/gsl_statistics_int.h>
#include "gopt.h"

/* Maximum length of RNA sequence */
#define MAXLEN 10000
/* Maximum number of loops at min stack length 2 */
#define STRUC MAXLEN/5
/* Definition of UNPAIRED */
#define UNPAIRED -1
#define VERBOSE_DEBUG 2
#define VERBOSE_RNARAG 1
#define VERBOSE_DEFAULT 0
#define TURN_OFF 0
#define TURN_ON 1
#define OPTIMIZE_DEFAULT TURN_ON
#define MONITOR_DEFAULT TURN_OFF

#define PUBLIC
#define PRIVATE static

/* Functions */
/* Process commandline parameters */
PRIVATE void processParams(int argc, char *argv[]);
/* Print usage and available commandline parameters */
PRIVATE void usage(void);
/* Process input file containing RNA structure of form ".()" */
PRIVATE void processInput(FILE * tgFile, FILE * input);
/* Create new pointer */
PRIVATE void *space(const unsigned size);
/* Write output to file */
PRIVATE void writeFile(FILE * tgFile, const char *sequence,
                       const char *structure, const char *header,
                       const float energy);
/* Checks whether RNA structure in format ".()" is well-structured */
PRIVATE int checkStruct(const char *structure);
/* Computes the Laplacian matrix */
PRIVATE gsl_matrix *computeL(const char *structure, FILE * tgFile);
/* Implements the pair of vertex-edge rules described in subsection A.1 */
PRIVATE char *optimizeStruct(const char *structure);
/* Implements the Eq. (A.1) and (A.2) described in subsection A.2 */
PRIVATE void parseStruct(const char *structure);
/* Track the onset of a helical stem-loop */
PRIVATE char *auxStruct(const char *structure);
/* Initialize five global variables totalpath, path, stems, ld, ls, and hs */
PRIVATE void zeroVars(void);
/* Returns array representation of RNA structure */
PRIVATE int *makePBTable(const char *structure);
/* Get graph statistics */
PRIVATE void makeTopo(void);
/*
 * Computes the eigenvalue spectrum using the well-established
 * "Eigen-decomposition theorem" and  det|L - lambda*I|=0
 */
PRIVATE gsl_vector *computeEigVals(const gsl_matrix * L, int vertices);
/* Output message upon error */
PRIVATE void nrerror(const char *file, const char *function, int line);

/* Global Variables */
PRIVATE char line[] =
    "-------------------------------------------------------------\n";
/*
 * ls[0] ..... number of external digits.
```

```
 * ls[1 <= i <= loops] ..... size of i-th vertex measured by
 * the number of unpaired bases
 */
PRIVATE int ls[STRUC];
/* hs[1 <= i <= loops] .... size of i-th stem measured by the number of pairs*/
PRIVATE int hs[STRUC];
/* ld[1 <= i <= loops] .. degree (branches) of i-th vertex (loop) */
PRIVATE int ld[STRUC];
/* num of stems in a structure */
PRIVATE int stems;
/*
 * The Euclidean path transverse from the first to the final (stems + 1)th
 * vertex, in the direction of 5' to 3' terminus of the given RNA sequence
 */
PRIVATE int path[2 * STRUC];
/* stores size of vector path */
PRIVATE int totalpath;

/* Global Variables set by command line */
PRIVATE int verbose = VERBOSE_DEFAULT;
PRIVATE int optimize = OPTIMIZE_DEFAULT;
PRIVATE int monitor = MONITOR_DEFAULT;

/* Global Time Variables to monitor execution time */
clock_t start, end;
double elapsed;

int main(int argc, char *argv[]) {
    char *fname = "-";
    FILE *tgFile = stdout;

    processParams(argc, argv);

    if (strcmp(fname, "-") != 0)
        if ((tgFile = fopen(fname, "a")) == NULL)
            nrerror(__FILE__, "main", __LINE__);

    if (monitor == TURN_ON)
        start = clock();

    processInput(tgFile, stdin);

    if (monitor == TURN_ON) {
        end = clock();
        elapsed = ((double) (end - start)) / CLOCKS_PER_SEC;
        fprintf(tgFile, "Time Taken (seconds): %.4f\n", elapsed);
    }

    if (strcmp(fname, "-") == 0)
        fclose(tgFile);
    else
        fflush(tgFile);

    exit(0);
}

/* Process commandline parameters */
void processParams(int argc, char *argv[]) {
    void *options;
    const char *params;

    if ((options = gopts("hvc", &argc, &argv)) == NULL)
        nrerror(__FILE__, "processParams", __LINE__);

    if (gopt(options, 'h', "help", NULL))
        usage();

    if (gopt(options, 'v', NULL, &params))
        verbose = atoi(params);
```

```
139
140     if (gopt(options, 0, "noopt", NULL))
141         optimize = TURN OFF;
142
143     if (gopt(options, 0, "monitor", NULL))
144         monitor = TURN ON;
145
146     free(options);
147 }
148
149 /* Print usage and available commandline parameters */
150 void usage(void) {
151     fprintf(stderr, "RNAspectral -h v12 --noopt < \"RNAfold File\"\n");
152     fprintf(stderr,
153             "Example usage 1: RNAspectral < \"RNAfold File\" > \"Output\"\n");
154     fprintf(stderr,
155             "Example usage 2: RNAfold < \"Fasta File\" | RNAspectral >
156             \"Output\"\n");
157     exit(0);
158 }
159
160 /* Process input file containing RNA structure of form ".()" */
161 void processInput(FILE * tgFile, FILE * input) {
162     char line[MAXLEN + 1];
163     char *sequence = NULL, *structure = NULL, *header = NULL;
164     float energy;
165     int n = 1;
166
167     if (verbose == VERBOSE_DEFAULT)
168         fprintf(tgFile,
169                 "ID\tMFE\tLen\tVer\tStems\tJunct\tEndpts\tMidpts\tSecEigen\n");
170
171     while (fgets(line, MAXLEN, stdin) != NULL) {
172         if (strcmp(line, "@") == 0)
173             break;
174
175         switch (line[0]) {
176         case '>':
177             header = (char *) space(strlen(line));
178             sscanf(line, "> %s", header);
179             break;
180         case '.':
181         case '(':
182         case ')':
183             structure = (char *) space(strlen(line));
184             if (sscanf(line, "%s (%f)", structure, &energy) != 2
185                     && sscanf(line, "%s", structure) != 1) {
186                 free(structure);
187                 structure = NULL;
188                 break;
189             }
190             if (structure != NULL) {
191                 if (verbose == VERBOSE_DEFAULT || header == NULL) {
192                     header = (char *) space(10);
193                     sprintf(header, "%d", n);
194                 }
195
196                 writeFile(tgFile, sequence, structure, header, energy);
197
198                 free(sequence);
199                 free(structure);
200                 free(header);
201                 sequence = structure = header = NULL;
202             }
203             n++;
204             break;
205         default:
206             sequence = (char *) space(strlen(line) + 1);
207             sscanf(line, "%s", sequence);
```

```
208             break;
209         }
210     }
211 }
212
213 /* Create new pointer */
214 void *space(const unsigned size) {
215     void *pointer = (void *) calloc(1, size);
216
217     if (pointer == NULL)
218         nrerror(__FILE__, "Space", __LINE__);
219
220     return pointer;
221 }
222
223 /* Write output to file */
224 void writeFile(FILE * tgFile, const char *sequence, const char *structure,
225                const char *header, const float energy) {
226     int length = strlen(structure);
227
228     if (checkStruct(structure) == 0) {
229         fprintf(tgFile, "%s\t%.1f\t%d\tStructure incorrect\n", header,
230                 energy, length);
231         return;
232     }
233
234     gsl_matrix *L = computeL(structure, tgFile);
235
236     if (L == NULL) {
237         fprintf(tgFile, "%s\t%.1f\t%d\tLaplacian matrix incorrect\n",
238                 header, energy, length);
239         return;
240     }
241
242     int junctions = 0, endpoints = 0, midpoints = 0, vertices = L->size1;
243     int i = vertices, j;
244     gsl_vector_view diagonal = gsl_matrix_diagonal(L);
245
246     while (i--) {
247         int element = gsl_vector_get(&diagonal.vector, i);
248         switch (element) {
249         case 0:
250             break;
251         case 1:
252             endpoints++;
253             break;
254         case 2:
255             midpoints++;
256             break;
257         default:
258             junctions++;
259             break;
260         }
261     }
262
263     gsl_vector *eval = computeEigVals(L, L->size1);
264
265     if (verbose == VERBOSE_DEFAULT) {
266         double eval_i = gsl_vector_get(eval, 1);
267         fprintf(tgFile, "%s\t%.1f\t%d\t%3d\t%3d\t%3d\t%3d\t%3d\t%6f\n",
268                 header, energy, length, vertices, vertices - 1, junctions,
269                 endpoints, midpoints, eval_i);
270     }
271
272     if (verbose >= VERBOSE_RNARAG) {
273         fprintf(tgFile, "ENERGY =  %7.1f    %s\n", energy, header);
274         fprintf(tgFile, "%s\n", sequence);
275         fprintf(tgFile, "%s\n", structure);
276         fprintf(tgFile,
```

```
277                 "There are %d nucleotides for the following sequence.\n",
278                 length);
279
280         fprintf(tgFile,
281                 "There are %d junction(s), %d endpoint(s), %d midpoint(s),
282                 and %d stem(s).\n",
283                 junctions, endpoints, midpoints, vertices - 1);
284         fprintf(tgFile, "Vertex#= %d\n", vertices);
285         fprintf(tgFile, "iteration=1\n\n");
286         fprintf(tgFile, "LAPLACIAN MATRIX!\n");
287
288         for (i = 0; i < vertices; i++) {
289             for (j = 0; j < vertices; j++)
290                 fprintf(tgFile, "%5g", gsl_matrix_get(L, i, j));
291             fprintf(tgFile, "\n");
292         }
293         fprintf(tgFile, "\n");
294
295         double eigen_mul = 1;
296         for (i = 0; i < vertices; i++) {
297             double eval_i = gsl_vector_get(eval, i);
298             if (i > 0)
299                 eigen_mul *= eval_i;
300             fprintf(tgFile, "eigenvalue %3d=     %6f\n", vertices - i, eval_i);
301         }
302         fprintf(tgFile, "\n");
303
304         fprintf(tgFile,
305                 "Vertices (determined by multipying eigenvalues):     %6f\n",
306                 eigen_mul);
307         fprintf(tgFile, line);
308     }
309
310     gsl_matrix_free(L);
311     gsl_vector_free(eval);
312 }
313
314 /*
315  * Checks whether RNA structure in format ".()" is well-structured
316  * return 0: brackets do not match
317  * return 1: brackets match
318  */
319 int checkStruct(const char *structure) {
320     int i, o = 0, length = strlen(structure);
321
322     for (i = 0; i < length; i++) {
323         switch (structure[i]) {
324         case '(':
325             o++;
326             break;
327         case '.':
328             break;
329         case ')':
330             o--;
331             break;
332         default:
333             return 0;
334         }
335     }
336
337     if (o != 0)
338         return 0;
339     return 1;
340 }
341
342 /* Computes the Laplacian matrix */
343 gsl_matrix *computeL(const char *structure, FILE * tgFile) {
344     int i, vertices;
345     int length = strlen(structure);
```

```
346     char *new_structure = (char *) space(sizeof(char) * (length + 1));
347
348     strcpy(new_structure, structure);
349
350     if (verbose >= VERBOSE_DEBUG)
351         fprintf(tgFile, "old structure %s\n", structure);
352
353     if (optimize == TURN_ON)
354         new_structure = optimizeStruct(structure);
355
356     if (verbose >= VERBOSE_DEBUG)
357         fprintf(tgFile, "new structure %s\n", new_structure);
358
359     parseStruct(new_structure);
360     free(new_structure);
361
362     makeTopo();
363
364     vertices = stems + 1;
365
366     int actual vertices = 0;
367     i = vertices;
368     while (i--) {
369         if (ld[i] > 0)
370             actual vertices++;
371     }
372
373     if (verbose >= VERBOSE_DEBUG) {
374         fprintf(tgFile, "vertex\tdegree\tsize\tstem\tsize\n");
375         for (i = 0; i < vertices; i++)
376             fprintf(tgFile, "%d\t%d\t%d\t%d\t%d\n", i, ld[i], ls[i], i, hs[i]);
377
378         fprintf(tgFile,
379                 "number of vertices = %d, total unpaired bases = %d\n",
380                 actual vertices, topo.sumunpaired);
381         fprintf(tgFile,
382                 "number of stems = %d, total paired bases = %d\n",
383                 stems, topo.sumpaired * 2);
384     }
385
386     if (actual_vertices == 0)
387         return NULL;
388
389     gsl_matrix *A = gsl_matrix_calloc(vertices, vertices);
390     gsl_matrix *D = gsl_matrix_calloc(vertices, vertices);
391
392     /*
393      * Sets the value of adjacency matrix A at row path[i]
394      * and column path[i + 1] to 1
395      */
396     for (i = 0; i < totalpath; i++)
397         gsl_matrix_set(A, path[i], path[i + 1], 1);
398
399     /* Sets the value of degree matrix D at row i and column i to ld[i] */
400     for (i = 0; i < vertices; i++)
401         gsl_matrix_set(D, i, i, ld[i]);
402
403     /* Computes the Laplacian matrix L = D - A by using D as L */
404     gsl_matrix_sub(D, A);
405     gsl_matrix_free(A);
406
407     /* Using D as L = D - A */
408     return (D);
409 }
410
411 /* Implements the pair of vertex-edge rules described in subsection A.1 */
412 char *optimizeStruct(const char *structure) {
413     int i, j, length = strlen(structure);
414     char *new_structure = (char *) space(sizeof(char) * (length + 1));
```

```
415
416     strcpy(new_structure, structure);
417
418     /* vector pt contains the values returned by makePBTable */
419     int *pt = makePBTable(structure);
420     if (pt == NULL)
421         nrerror(__FILE__, "optimizeStruct", __LINE__);
422
423     for (i = 1; i < length - 1; i++) {
424         /*
425          * Internal loops with only one pair of mismatches are identified
426          * and then paired; >0 -1 >0
427          */
428         if (pt[i] == UNPAIRED) {
429             if (pt[i - 1] == UNPAIRED)
430                 continue;
431             if (pt[i + 1] == UNPAIRED)
432                 continue;
433
434             if ((pt[i - 1] - pt[i + 1]) == 2)
435                 pt[i] = pt[i - 1] - 1;
436             else if ((pt[i + 1] - pt[i - 1]) == 2)
437                 pt[i] = pt[i - 1] + 1;
438             else
439                 continue;
440
441             pt[pt[i]] = i;
442             new_structure[i] = '(';
443             new_structure[pt[i]] = ')';
444             continue;
445         }
446
447         /*
448          * Stems with only one complementary pair are identified
449          * and then unpaired; -1 >0 -1
450          */
451         if (pt[i] > 0) {
452             /* Both pt[i - 1] and pt[i + 1] are UNPAIRED */
453             if (pt[i - 1] == pt[i + 1]) {
454                 new_structure[i] = new_structure[pt[i]] = '.';
455                 pt[pt[i]] = pt[i] = UNPAIRED;
456             }
457         }
458     }
459
460     for (j = 1, i = 1; i < length - 1; i++) {
461         /* Bulges having unpaired mono-nucleotide are deleted */
462         if (pt[i] == UNPAIRED) {
463             if ((pt[i - 1] - pt[i + 1]) == 1)
464                 continue;
465             if ((pt[i + 1] - pt[i - 1]) == 1)
466                 continue;
467         }
468         new_structure[j++] = new_structure[i];
469     }
470
471     new_structure[j++] = new_structure[length - 1];
472     new_structure[j] = '\0';
473     free(pt);
474
475     /*
476      * Resulting RNA structure S' is returned after applying
477      * the pair of vertex-edge rules
478      */
479     return new_structure;
480 }
481
482 /*
483  * Returns array representation of RNA structure.
```

```
 * pt[i] of nucleotide at position i has value of UNPAIRED
 * when that nucleotide is unpaired or denotes the position
 * of the base to which it is paired
 */
int *makePBTable(const char *structure) {
    int i, j = 0, length = strlen(structure);
    int *stack = (int *) space(sizeof(int) * (length + 1));
    int *pt = (int *) space(sizeof(int) * (length + 1));

    for (i = 0; i < length; i++) {
        switch (structure[i]) {
        case '.':
            pt[i] = UNPAIRED;
            break;
        case '(':
            stack[j++] = i;
            break;
        case ')':
            pt[i] = stack[--j];
            pt[pt[i]] = i;
            break;
        default:
            break;
        }
    }

    free(stack);
    return (pt);
}

/* Implements the Eq. (A.1) and (A.2) described in subsection A.2 */
void parseStruct(const char *structure) {
    int i, lp = 0, p = 0;
    int length = strlen(structure);
    int *loop = (int *) space(sizeof(int) * (length / 3 + 1));
    char *string = auxStruct(structure);

    if (string == NULL)
        nrerror(__FILE__, "parseStruct", __LINE__);

    zeroVars();

    /*
     * Computes the Euclidean path transverse from the first
     * to the final (stems + 1)th vertex, in the direction
     * of 5' to 3' terminus of the given RNA sequence
     */
    for (i = 0; i < length; i++) {
        switch (string[i]) {
        case '.':
            ls[loop[lp]]++;
            break;
        case '[':
            path[totalpath++] = loop[lp++];
            ld[++stems] = 1;
            loop[lp] = stems;
            break;
        case ')':
            p++;
            break;
        case ']':
            hs[loop[lp]] = p + 1;
            p = 0;
            path[totalpath++] = loop[lp];
            ld[loop[--lp]]++;
            break;
        default:
            break;
        }
```

```
553     }
554
555     path[totalpath] = 0;
556     free(string);
557     free(loop);
558 }
559
560 /*
561  * Track the onset of a helical stem-loop
562  * Returns a RNA secondary structure specified in an extended dot-bracket
563  * format with additional symbols "[", and "]"
564  */
565 char *auxStruct(const char *structure) {
566     int length = strlen(structure);
567     int i, o = 0, p = 0;
568     int *mp = (int *) space(sizeof(int) * (length / 2 + 1));
569     char *auxStruct = (char *) space(sizeof(char) * (length + 1));
570
571     strcpy(auxStruct, structure);
572
573     for (i = 0; i < length; i++) {
574         switch (auxStruct[i]) {
575         case '.':
576             break;
577         case '(':
578             mp[++o] = i;
579             break;
580         case ')':
581             p = i;
582             while ((auxStruct[p + 1] == ')') && (mp[o - 1] == mp[o] - 1)) {
583                 p++;
584                 o--;
585             }
586             auxStruct[p] = ']';
587             i = p;
588             auxStruct[mp[o--]] = '[';
589             break;
590         default:
591             nrerror(__FILE__, "auxStruct", __LINE__);
592         }
593     }
594
595     free(mp);
596     return (auxStruct);
597 }
598
599 /* Initialize five global variables totalpath, path, stems, ld, ls, and hs */
600 void zeroVars(void) {
601     int i = STRUC;
602     while (i--)
603         ls[i] = hs[i] = ld[i] = 0;
604     totalpath = stems = 0;
605 }
606
607 /* Get graph statistics */
608 void makeTopo(void) {
609     topo.vertices = stems + 1;
610     topo.stems = stems;
611
612     topo.sumpaired = topo.sumunpaired = 0;
613     int i = topo.vertices;
614     while (i--) {
615         topo.sumunpaired += ls[i];
616         topo.sumpaired += hs[i];
617     }
618
619     topo.meanpaired = gsl_stats_int_mean(hs, 1, topo.vertices);
620     topo.sdpaired = gsl_stats_int_sd(hs, 1, topo.vertices);
621     topo.minpaired = gsl_stats_int_min(hs, 1, topo.vertices);
```

```
622     topo.maxpaired = gsl_stats_int_max(hs, 1, topo.vertices);
623     topo.meanunpaired = gsl_stats_int_mean(ls, 1, topo.stems);
624     topo.sdunpaired = gsl_stats_int_sd(ls, 1, topo.stems);
625     topo.minunpaired = gsl_stats_int_min(hs, 1, topo.stems);
626     topo.maxunpaired = gsl_stats_int_max(hs, 1, topo.stems);
627 }
628
629 /*
630  * Computes the eigenvalue spectrum using the well-established
631  * "Eigen-decomposition theorem" and  det|L - lambda*I|=0
632  */
633 gsl_vector *computeEigVals(const gsl_matrix * L, int vertices) {
634     gsl_matrix *tempL = gsl_matrix_alloc(vertices, vertices);
635     gsl_vector *eval = gsl_vector_alloc(vertices);
636     gsl_matrix *evec = gsl_matrix_alloc(vertices, vertices);
637     gsl_eigen_symmv_workspace *w = gsl_eigen_symmv_alloc(vertices);
638
639     gsl_matrix_memcpy(tempL, L);
640     gsl_eigen_symmv(tempL, eval, evec, w);
641     gsl_eigen_symmv_sort(eval, evec, GSL_EIGEN_SORT_VAL_ASC);
642
643     gsl_eigen_symmv_free(w);
644     gsl_matrix_free(evec);
645     gsl_matrix_free(tempL);
646
647     return (eval);
648 }
649
650 /* Output message upon error */
651 void nrerror(const char *file, const char *function, int line) {
652     fprintf(stderr, "Error: %s %s %d\n", file, function, line);
653     exit(0);
654 }
```

A.5. Experimental Methodology

A typical experimental setup using *RNAfold* and *RNAspectral* programs in an automated manner is outlined in Figure A.7. In this example, sequence of THI element (AC084406.7; thiamine pyrophosphate riboswitch) (Sudarsan *et al*., 2003) was extracted from *Sanger Rfam* 7.0 (Griffiths-Jones *et al*., 2005). Given a primary RNA sequence described in FASTA format, (Step **A**) its optimal secondary structure is predicted using *RNAfold* program (Hofacker 2003). The output of *RNAfold* is a FASTA-like format appended with the optimal structure in Vienna dot-bracket notation with the base pairs and unpaired bases represented by brackets "( )" and dots "." (Hofacker 2003), respectively and the Minimum Free Energy of folding (MFE). In this example, the RNA secondary structure predicted by *RNAfold* has two hairpin loops, 5' and 3' termini, two internal loops, one bulge loop, and one multi-branch loop - all of these stabilized by six stems. (Step **B**) This is read by *RNAspectral* that converts the structure in bracket notation into a planar tree-graph consisting of seven arbitrarily labeled vertices "•" connected by six unweighted edges "—". (Step **C**) *RNAspectral* computes the seven by seven Laplacian matrix and the eigenvalue spectrum. (Step **D**) The output of *RNAspectral* is described in a tab-delimited ASCII flat format for convenient import into numerical processing applications such as Mathworks® Matlab™ and Microsoft® Excel™. The labeled header shows the following rows of columnated values corresponding to the identifier (**ID** starts at 1 and increases monotically), Minimum Free Energy of folding (**MFE** in kcal/mol), length of sequence (**Len** in nucleotides), number of vertices (**Ver**), number of stems (**Stems**), number of junctions (**Junct**, more than 2 stems), number of endpoints (**Endpts**, 1 stem), number of midpoints (**Midpts**, 2 stems), and the second eigenvalue λ_2 (**SecEigen**).

(Figure A.8) The benchmarking platform was an AMD Opteron Processor 850 2.4 GHz and 1.5 GB RAM; GNU compiler v3.4.5 on Linux 2.6.9-5. The average speed of *RNAspectral* was computed by running it five times on 6,656 sets of 10^4 random RNA sequences. The random sequences were synthesized from each of the 6,656 sequences (average length of 113.451 ± 0.803 nucleotides) gathered from *Sanger miRBase* 7.1 (Griffiths-Jones 2004) and *Sanger Rfam* 7.0 (Griffiths-Jones *et al*., 2005). *RNAspectral* required at most ~7.0 seconds or mean 427.8 milliseconds for processing the entire dataset.

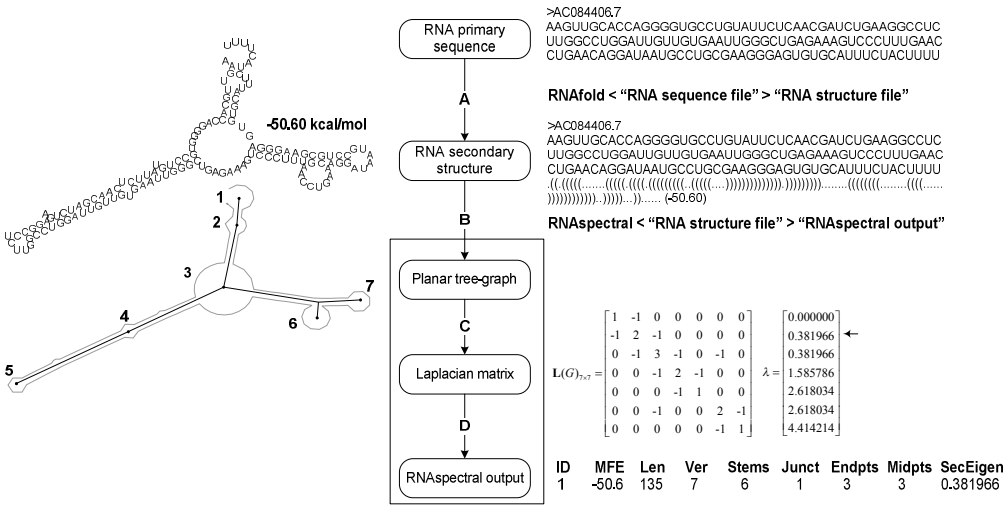

Figure A.7: Typical workflow using *RNAspectral* for "Spectral Graph Partitioning" analysis on RNA structures. ←, second eigenvalue λ_2 shows the same results as "RNA Matrix Computer Program" (Gan *et al*., 2004; Fera *et al*., 2004); bold, Unix commands.

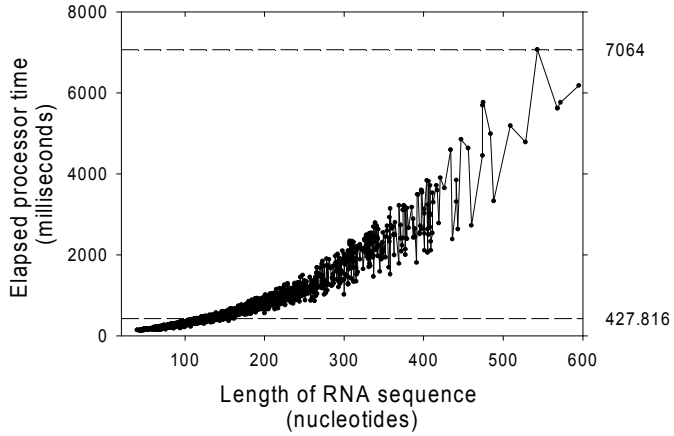

Figure A.8: Average speed performance of *RNAspectral*. Unlike the actual wall-clock time, elapsed processor time excludes time spent queuing for free I/O or waiting for other processes to complete execution.

Appendix B.

Supplemental for Chapter 4

Continue on next page.

Table B.1: Statistical comparison between pre-miRs, ncRNAs, mRNA, and pseudo hairpins based on *Length*, *MFEI$_2$*, *MFEI$_1$*, *%G+C*, *dP*, *dG*, *dQ*, *dD*, and *dF*.

Datasets	*Counts*	*Length*	*MFEI$_2$*	*MFEI$_1$*	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*
Arthropoda	171	88.6901 ± 0.8213	-0.0645 ± 0.0016	-0.0089 ± 0.0001	43.3811 ± 0.4752	0.3488 ± 0.0023	-0.3824 ± 0.0050	0.1067 ± 0.0047	0.0403 ± 0.0016	0.2059 ± 0.0067
Nematoda	189	99.0212 ± 0.6723	-0.0556 ± 0.0015	-0.0086 ± 0.0001	44.5725 ± 0.4641	0.3411 ± 0.0025	-0.3831 ± 0.0056	0.1075 ± 0.0059	0.0398 ± 0.0019	0.1577 ± 0.0050
Vertebrata	1203	90.4522 ± 0.4164	-0.0761 ± 0.0013	-0.0091 ± 0.0001	48.3079 ± 0.2504	0.3518 ± 0.0009	-0.4308 ± 0.0025	0.1161 ± 0.0025	0.0431 ± 0.0009	0.2197 ± 0.0042
Viridiplantae	606	137.9175 ± 2.0309	-0.0539 ± 0.0010	-0.0096 ± 0.0001	46.6719 ± 0.3513	0.3545 ± 0.0013	-0.4456 ± 0.0038	0.1424 ± 0.0036	0.0502 ± 0.0011	0.1251 ± 0.0033
Viruses	72	78.8750 ± 1.4665	-0.0780 ± 0.0032	-0.0087 ± 0.0002	53.5111 ± 0.9219	0.3619 ± 0.0029	-0.4615 ± 0.0097	0.0893 ± 0.0051	0.0352 ± 0.0020	0.2059 ± 0.0114
Cis-reg	4002	90.7511 ± 0.8069	-0.0793 ± 0.0017	-0.0065 ± 0.0000	48.9672 ± 0.1188	0.2905 ± 0.0008	-0.3233 ± 0.0017	0.2124 ± 0.0021	0.0689 ± 0.0006	0.3871 ± 0.0064
Cis-reg|frameshift	808	53.2599 ± 0.2543	-0.2210 ± 0.0021	-0.0104 ± 0.0000	46.4785 ± 0.1477	0.3382 ± 0.0010	-0.4814 ± 0.0023	0.1396 ± 0.0024	0.0552 ± 0.0009	0.8865 ± 0.0079
Cis-reg|IRES	1201	276.0841 ± 2.4342	-0.0192 ± 0.0002	-0.0065 ± 0.0000	57.5340 ± 0.1745	0.3039 ± 0.0006	-0.3757 ± 0.0013	0.3702 ± 0.0034	0.1156 ± 0.0010	0.0442 ± 0.0013
Cis-reg|riboswitch	917	138.6358 ± 1.4673	-0.0381 ± 0.0005	-0.0064 ± 0.0000	50.5054 ± 0.3381	0.2877 ± 0.0010	-0.3223 ± 0.0026	0.2515 ± 0.0041	0.0826 ± 0.0012	0.1960 ± 0.0042
Cis-reg|thermoregulator	21	127.0476 ± 4.0447	-0.0330 ± 0.0047	-0.0061 ± 0.0002	42.6490 ± 3.2009	0.2955 ± 0.0075	-0.2713 ± 0.0301	0.2935 ± 0.0269	0.0956 ± 0.0080	0.1312 ± 0.0138
Gene	480	222.2708 ± 5.8445	-0.0372 ± 0.0012	-0.0074 ± 0.0000	51.6146 ± 0.5262	0.3109 ± 0.0012	-0.3808 ± 0.0046	0.2435 ± 0.0060	0.0794 ± 0.0018	0.1258 ± 0.0058
Gene|antisense	147	86.0476 ± 0.8681	-0.0811 ± 0.0030	-0.0083 ± 0.0001	41.7778 ± 0.8673	0.3106 ± 0.0034	-0.3414 ± 0.0076	0.1336 ± 0.0061	0.0468 ± 0.0020	0.3734 ± 0.0133
Gene|ribozyme	561	242.0428 ± 5.4441	-0.0406 ± 0.0017	-0.0070 ± 0.0000	54.4837 ± 0.3930	0.3000 ± 0.0011	-0.3811 ± 0.0040	0.2704 ± 0.0053	0.0863 ± 0.0016	0.2335 ± 0.0145
Gene|rRNA	1010	244.3208 ± 5.8418	-0.0295 ± 0.0005	-0.0066 ± 0.0000	53.8479 ± 0.2508	0.3022 ± 0.0008	-0.3545 ± 0.0023	0.2870 ± 0.0043	0.0921 ± 0.0012	0.0933 ± 0.0020
Gene|snRNA	28	62.0357 ± 0.7024	-0.0764 ± 0.0088	-0.0061 ± 0.0003	41.6782 ± 1.2105	0.2803 ± 0.0064	-0.2631 ± 0.0187	0.2305 ± 0.0260	0.0741 ± 0.0074	0.5372 ± 0.0415
Gene|snRNA|guide|CD-box	1050	91.5867 ± 1.0464	-0.0379 ± 0.0004	-0.0053 ± 0.0000	42.3681 ± 0.2301	0.2764 ± 0.0013	-0.2265 ± 0.0022	0.3174 ± 0.0041	0.1012 ± 0.0012	0.2772 ± 0.0058
Gene|snRNA|guide|HACA-box	419	139.3675 ± 1.2446	-0.0348 ± 0.0005	-0.0068 ± 0.0001	46.3048 ± 0.3160	0.2929 ± 0.0013	-0.3125 ± 0.0029	0.2383 ± 0.0068	0.0783 ± 0.0021	0.1194 ± 0.0028
Gene|snRNA|splicing	250	157.1200 ± 4.4708	-0.0341 ± 0.0006	-0.0068 ± 0.0001	47.6933 ± 0.3731	0.2898 ± 0.0021	-0.3251 ± 0.0042	0.2399 ± 0.0076	0.0781 ± 0.0023	0.1470 ± 0.0043
Gene|sRNA	233	145.6524 ± 4.5117	-0.0432 ± 0.0016	-0.0066 ± 0.0001	46.3963 ± 0.3513	0.2815 ± 0.0024	-0.3036 ± 0.0041	0.2371 ± 0.0077	0.0745 ± 0.0023	0.2531 ± 0.0170
Gene|tRNA	1114	73.4354 ± 0.1529	-0.0676 ± 0.0007	-0.0064 ± 0.0000	48.2725 ± 0.3541	0.2975 ± 0.0010	-0.3138 ± 0.0029	0.2488 ± 0.0035	0.0831 ± 0.0011	0.5333 ± 0.0093
Intron	146	134.4384 ± 8.6225	-0.0604 ± 0.0029	-0.0080 ± 0.0001	44.7871 ± 0.8350	0.3204 ± 0.0024	-0.3551 ± 0.0081	0.1802 ± 0.0089	0.0620 ± 0.0026	0.2200 ± 0.0107
mRNAs	31	332.3226 ± 16.3064	-0.0132 ± 0.0006	-0.0061 ± 0.0001	50.4626 ± 1.4654	0.2881 ± 0.0045	-0.3087 ± 0.0131	0.3828 ± 0.0175	0.1192 ± 0.0049	0.0391 ± 0.0059
Pseudo hairpins	8494	84.7020 ± 0.1268	-0.0476 ± 0.0002	-0.0054 ± 0.0000	56.1466 ± 0.1108	0.2874 ± 0.0003	-0.3070 ± 0.0009	0.3185 ± 0.0016	0.1048 ± 0.0005	0.1818 ± 0.0008

(*Counts*) Number of sequences being investigated. Values are stated as mean ± standard error.

Table B.2: Statistical comparison between pre-miRs, ncRNAs, mRNA, and pseudo hairpins based on *zG*, *zQ*, and *zD* using the four sequence randomization algorithms.

		zG				*zQ*				*zD*			
Datasets	*Counts*	*MS*	*DS*	*ZM*	*FM*	*MS*	*DS*	*ZM*	*FM*	*MS*	*DS*	*ZM*	*FM*
Arthropoda	171	-4.8894 ± 0.1127	-4.8985 ± 0.1177	-3.5250 ± 0.0839	-3.3032 ± 0.0818	-1.7166 ± 0.0321	-1.6873 ± 0.0316	-1.7259 ± 0.0320	-1.7067 ± 0.0311	-1.6782 ± 0.0377	-1.6526 ± 0.0365	-1.6706 ± 0.0379	-1.6544 ± 0.0364
Nematoda	189	-4.9930 ± 0.1457	-5.0228 ± 0.1443	-3.4481 ± 0.1003	-3.2885 ± 0.0927	-1.7836 ± 0.0399	-1.7548 ± 0.0394	-1.7936 ± 0.0401	-1.7797 ± 0.0390	-1.7598 ± 0.0440	-1.7358 ± 0.0432	-1.7523 ± 0.0443	-1.7433 ± 0.0432
Vertebrata	1203	-5.2058 ± 0.0645	-4.7608 ± 0.0642	-3.6575 ± 0.0463	-3.2317 ± 0.0426	-1.6090 ± 0.0170	-1.5465 ± 0.0164	-1.6252 ± 0.0170	-1.5779 ± 0.0164	-1.5755 ± 0.0196	-1.5209 ± 0.0187	-1.5762 ± 0.0198	-1.5368 ± 0.0188
Viridiplantae	606	-6.9286 ± 0.1033	-6.4395 ± 0.1037	-4.5333 ± 0.0718	-4.1132 ± 0.0693	-1.6602 ± 0.0248	-1.5957 ± 0.0243	-1.6725 ± 0.0248	-1.6211 ± 0.0242	-1.6440 ± 0.0276	-1.5879 ± 0.0267	-1.6422 ± 0.0277	-1.5982 ± 0.0267
Viruses	72	-4.7038 ± 0.1952	-4.5972 ± 0.1908	-3.2593 ± 0.1325	-3.0913 ± 0.1280	-1.6475 ± 0.0414	-1.6214 ± 0.0403	-1.6722 ± 0.0416	-1.6524 ± 0.0405	-1.6088 ± 0.0495	-1.5848 ± 0.0481	-1.6191 ± 0.0498	-1.6016 ± 0.0486
Cis-reg	4002	-2.6887 ± 0.0308	-2.3364 ± 0.0280	-1.9053 ± 0.0203	-1.5172 ± 0.0172	-0.8439 ± 0.0142	-0.7928 ± 0.0139	-0.8336 ± 0.0142	-0.7878 ± 0.0140	-0.8206 ± 0.0147	-0.7788 ± 0.0143	-0.7851 ± 0.0148	-0.7452 ± 0.0145
Cis-reg|frameshift	808	-5.6222 ± 0.0477	-3.7443 ± 0.0357	-4.4470 ± 0.0359	-2.3964 ± 0.0200	-1.1436 ± 0.0158	-1.1970 ± 0.0155	-1.1579 ± 0.0160	-1.1044 ± 0.0146	-0.9865 ± 0.0192	-1.0716 ± 0.0187	-0.9768 ± 0.0197	-0.9303 ± 0.0181
Cis-reg|IRES	1201	-0.7674 ± 0.0353	-1.0895 ± 0.0293	-0.5451 ± 0.0192	-0.6451 ± 0.0149	-0.1924 ± 0.0250	-0.2063 ± 0.0252	-0.2027 ± 0.0250	-0.2300 ± 0.0251	-0.2134 ± 0.0256	-0.2208 ± 0.0259	-0.2121 ± 0.0257	-0.2296 ± 0.0260
Cis-reg|riboswitch	917	-1.5838 ± 0.0452	-1.4806 ± 0.0446	-1.1569 ± 0.0282	-1.0231 ± 0.0261	-0.8469 ± 0.0293	-0.8163 ± 0.0294	-0.8585 ± 0.0294	-0.8513 ± 0.0293	-0.8139 ± 0.0309	-0.7884 ± 0.0309	-0.8086 ± 0.0312	-0.8030 ± 0.0309
Cis-reg|thermoregulator	21	-1.0551 ± 0.2108	-1.0754 ± 0.2211	-0.8443 ± 0.1263	-0.7904 ± 0.1349	-0.5827 ± 0.1521	-0.5791 ± 0.1523	-0.5961 ± 0.1533	-0.6004 ± 0.1542	-0.4561 ± 0.1723	-0.4496 ± 0.1725	-0.4511 ± 0.1754	-0.4460 ± 0.1753
Gene	480	-2.9702 ± 0.0842	-2.8100 ± 0.0851	-1.9501 ± 0.0521	-1.7827 ± 0.0510	-1.0260 ± 0.0391	-1.0098 ± 0.0394	-1.0379 ± 0.0392	-1.0335 ± 0.0395	-1.0127 ± 0.0410	-1.0017 ± 0.0412	-1.0122 ± 0.0412	-1.0082 ± 0.0415
Gene|antisense	147	-4.0852 ± 0.1258	-4.0585 ± 0.1283	-2.9472 ± 0.0900	-2.6765 ± 0.0829	-1.5501 ± 0.0404	-1.5317 ± 0.0409	-1.5473 ± 0.0399	-1.5387 ± 0.0403	-1.5408 ± 0.0466	-1.5220 ± 0.0469	-1.5117 ± 0.0456	-1.4990 ± 0.0460
Gene|ribozyme	561	-3.0964 ± 0.0704	-2.7927 ± 0.0706	-1.9182 ± 0.0392	-1.6665 ± 0.0376	-0.7666 ± 0.0347	-0.7312 ± 0.0346	-0.7737 ± 0.0348	-0.7588 ± 0.0346	-0.7567 ± 0.0361	-0.7347 ± 0.0355	-0.7492 ± 0.0364	-0.7450 ± 0.0357
Gene|rRNA	1010	-2.0655 ± 0.0551	-2.0126 ± 0.0523	-1.3108 ± 0.0298	-1.2051 ± 0.0268	-0.6742 ± 0.0296	-0.6618 ± 0.0296	-0.6858 ± 0.0298	-0.6943 ± 0.0295	-0.6424 ± 0.0302	-0.6329 ± 0.0301	-0.6406 ± 0.0305	-0.6491 ± 0.0302
Gene|snRNA	28	-2.0909 ± 0.2613	-1.3712 ± 0.3083	-1.6055 ± 0.1806	-1.0729 ± 0.2076	-0.6335 ± 0.1771	-0.5674 ± 0.1695	-0.6180 ± 0.1789	-0.5995 ± 0.1699	-0.6270 ± 0.1735	-0.6108 ± 0.1597	-0.5832 ± 0.1759	-0.6090 ± 0.1609
Gene|snRNA|guide|CD-box	1050	-0.8113 ± 0.0397	-0.7089 ± 0.0360	-0.7209 ± 0.0270	-0.6465 ± 0.0244	-0.2189 ± 0.0292	-0.2236 ± 0.0286	-0.2146 ± 0.0295	-0.2497 ± 0.0286	-0.1810 ± 0.0298	-0.1952 ± 0.0291	-0.1512 ± 0.0302	-0.1886 ± 0.0294
Gene|snRNA|guide|HACA-box	419	-2.3694 ± 0.0745	-1.7490 ± 0.0780	-1.6445 ± 0.0499	-1.2434 ± 0.0489	-0.9913 ± 0.0497	-0.9265 ± 0.0494	-0.9997 ± 0.0499	-0.9567 ± 0.0493	-0.9621 ± 0.0532	-0.9169 ± 0.0519	-0.9546 ± 0.0537	-0.9275 ± 0.0523
Gene|snRNA|splicing	250	-2.6848 ± 0.1171	-2.4767 ± 0.1097	-1.7286 ± 0.0667	-1.4502 ± 0.0567	-1.0036 ± 0.0604	-0.9687 ± 0.0610	-1.0112 ± 0.0606	-0.9999 ± 0.0601	-1.0018 ± 0.0632	-0.9681 ± 0.0636	-0.9933 ± 0.0636	-0.9783 ± 0.0631
Gene|sRNA	233	-2.7470 ± 0.1222	-2.7672 ± 0.1234	-1.7773 ± 0.0712	-1.6417 ± 0.0664	-0.9773 ± 0.0589	-0.9675 ± 0.0589	-0.9771 ± 0.0592	-0.9903 ± 0.0584	-1.0182 ± 0.0608	-1.0073 ± 0.0612	-0.9991 ± 0.0611	-1.0064 ± 0.0607
Gene|tRNA	1114	-1.8663 ± 0.0281	-1.7570 ± 0.0289	-1.4794 ± 0.0193	-1.3739 ± 0.0189	-0.5740 ± 0.0237	-0.5524 ± 0.0239	-0.5770 ± 0.0238	-0.5804 ± 0.0238	-0.5223 ± 0.0257	-0.5109 ± 0.0257	-0.5019 ± 0.0260	-0.5050 ± 0.0260
Intron	146	-3.7603 ± 0.1402	-3.6841 ± 0.1513	-2.7426 ± 0.0976	-2.5026 ± 0.0982	-1.3073 ± 0.0531	-1.2842 ± 0.0534	-1.3177 ± 0.0533	-1.3065 ± 0.0530	-1.2483 ± 0.0558	-1.2290 ± 0.0559	-1.2424 ± 0.0564	-1.2335 ± 0.0560
mRNAs	31	-0.7223 ± 0.2089	0.1021 ± 0.1625	-0.4770 ± 0.1098	-0.0830 ± 0.0845	-0.1894 ± 0.1503	-0.1434 ± 0.1486	-0.1907 ± 0.1504	-0.1680 ± 0.1487	-0.1126 ± 0.1518	-0.0994 ± 0.1492	-0.1017 ± 0.1516	-0.1055 ± 0.1496
Pseudo hairpins	8494	-0.6493 ± 0.0121	-0.2347 ± 0.0114	-0.5606 ± 0.0073	-0.3373 ± 0.0067	-0.1058 ± 0.0113	-0.0756 ± 0.0112	-0.1052 ± 0.0114	-0.1044 ± 0.0112	-0.0444 ± 0.0117	-0.0385 ± 0.0114	-0.0208 ± 0.0118	-0.0364 ± 0.0114

(*Counts*) Number of sequences being investigated. Values are stated as mean ± standard error. MS, Mononucleotide Shuffling; DS, Dinucleotide Shuffling; ZM, Zero-order Markov Model; FM, First-order Markov Model.

Table B.3: Statistical comparison between pre-miRs, ncRNAs, mRNA, and pseudo hairpins based on *zP*, and *zF* based on four sequence randomization algorithms.

		zP				*zF*			
Datasets	*Counts*	*MS*	*DS*	*ZM*	*FM*	*MS*	*DS*	*ZM*	*FM*
Arthropoda	171	2.4736 ± 0.0653	2.4309 ± 0.0686	2.2904 ± 0.0560	2.2112 ± 0.0579	0.7107 ± 0.0912	0.6432 ± 0.0910	0.5025 ± 0.0791	0.4001 ± 0.0736
Nematoda	189	2.4392 ± 0.0807	2.4022 ± 0.0819	2.1992 ± 0.0673	2.1076 ± 0.0661	1.2007 ± 0.0675	1.1643 ± 0.0660	1.0738 ± 0.0651	1.0329 ± 0.0628
Vertebrata	1203	2.4911 ± 0.0287	2.3364 ± 0.0301	2.3065 ± 0.0246	2.1516 ± 0.0251	0.1902 ± 0.0340	0.1859 ± 0.0333	0.1359 ± 0.0330	0.1427 ± 0.0323
Viridiplantae	606	2.9329 ± 0.0449	2.7807 ± 0.0461	2.6133 ± 0.0376	2.4634 ± 0.0383	0.3538 ± 0.1870	0.5419 ± 0.2134	0.1306 ± 0.1470	0.2984 ± 0.1693
Viruses	72	2.6924 ± 0.0921	2.6297 ± 0.0922	2.4721 ± 0.0760	2.3915 ± 0.0754	-0.0844 ± 0.0335	0.0750 ± 0.0351	-0.1823 ± 0.0289	-0.0562 ± 0.0295
Cis-reg	4002	1.3687 ± 0.0197	1.2727 ± 0.0185	1.2527 ± 0.0160	1.1631 ± 0.0141	-0.1601 ± 0.0469	-0.1078 ± 0.0479	-0.2165 ± 0.0434	-0.1643 ± 0.0441
Cis-reg|frameshift	808	1.8580 ± 0.0237	1.4936 ± 0.0218	1.8881 ± 0.0207	1.4944 ± 0.0157	0.7519 ± 0.0732	0.6479 ± 0.0723	0.6347 ± 0.0692	0.5327 ± 0.0682
Cis-reg|IRES	1201	-0.0329 ± 0.0298	0.1285 ± 0.0291	0.1392 ± 0.0241	0.2594 ± 0.0231	1.1140 ± 0.1503	1.0258 ± 0.1485	0.8554 ± 0.1252	0.7597 ± 0.1218
Cis-reg|riboswitch	917	0.4080 ± 0.0365	0.3818 ± 0.0362	0.5064 ± 0.0292	0.4811 ± 0.0285	1.5069 ± 0.0682	1.5169 ± 0.0692	1.2329 ± 0.0614	1.1410 ± 0.0600
Cis-reg|thermoregulator	21	0.7628 ± 0.2207	0.8579 ± 0.2135	0.8156 ± 0.1827	0.9010 ± 0.1667	0.1460 ± 0.0809	0.0892 ± 0.0815	0.0623 ± 0.0758	0.0149 ± 0.0748
Gene	480	0.8078 ± 0.0495	0.8192 ± 0.0495	0.8420 ± 0.0414	0.8754 ± 0.0403	-0.0795 ± 0.1732	0.0456 ± 0.1812	-0.1270 ± 0.1624	-0.0042 ± 0.1698
Gene|antisense	147	1.4284 ± 0.0751	1.4366 ± 0.0715	1.3817 ± 0.0618	1.3731 ± 0.0582	-0.5938 ± 0.0065	-0.5623 ± 0.0067	-0.6124 ± 0.0057	-0.5726 ± 0.0058
Gene|ribozyme	561	0.8520 ± 0.0431	0.7607 ± 0.0440	0.8343 ± 0.0342	0.7654 ± 0.0337	0.7107 ± 0.0912	0.6432 ± 0.0910	0.5025 ± 0.0791	0.4001 ± 0.0736
Gene|rRNA	1010	0.8805 ± 0.0330	0.8847 ± 0.0329	0.8612 ± 0.0256	0.8387 ± 0.0252	1.2007 ± 0.0675	1.1643 ± 0.0660	1.0738 ± 0.0651	1.0329 ± 0.0628
Gene|snRNA	28	0.8315 ± 0.1820	0.4112 ± 0.1759	0.8631 ± 0.1418	0.5505 ± 0.1294	0.1902 ± 0.0340	0.1859 ± 0.0333	0.1359 ± 0.0330	0.1427 ± 0.0323
Gene|snRNA|guide|CD-box	1050	0.5020 ± 0.0334	0.4050 ± 0.0335	0.6217 ± 0.0278	0.5643 ± 0.0274	0.3538 ± 0.1870	0.5419 ± 0.2134	0.1306 ± 0.1470	0.2984 ± 0.1693
Gene|snRNA|guide|HACA-box	419	0.5749 ± 0.0454	0.3727 ± 0.0479	0.6545 ± 0.0377	0.5069 ± 0.0385	-0.0844 ± 0.0335	0.0750 ± 0.0351	-0.1823 ± 0.0289	-0.0562 ± 0.0295
Gene|snRNA|splicing	250	0.4583 ± 0.0708	0.5303 ± 0.0692	0.5381 ± 0.0574	0.5607 ± 0.0546	-0.1601 ± 0.0469	-0.1078 ± 0.0479	-0.2165 ± 0.0434	-0.1643 ± 0.0441
Gene|sRNA	233	0.7139 ± 0.0700	0.7771 ± 0.0701	0.7326 ± 0.0546	0.7518 ± 0.0529	0.7519 ± 0.0732	0.6479 ± 0.0723	0.6347 ± 0.0692	0.5327 ± 0.0682
Gene|tRNA	1114	0.8293 ± 0.0281	0.8075 ± 0.0281	0.9282 ± 0.0235	0.9367 ± 0.0225	1.1140 ± 0.1503	1.0258 ± 0.1485	0.8554 ± 0.1252	0.7597 ± 0.1218
Intron	146	1.6381 ± 0.0778	1.6300 ± 0.0780	1.5693 ± 0.0661	1.5258 ± 0.0637	1.5069 ± 0.0682	1.5169 ± 0.0692	1.2329 ± 0.0614	1.1410 ± 0.0600
mRNAs	31	0.3438 ± 0.2004	-0.0620 ± 0.1970	0.3700 ± 0.1576	0.0849 ± 0.1532	0.1460 ± 0.0809	0.0892 ± 0.0815	0.0623 ± 0.0758	0.0149 ± 0.0748
Pseudo hairpins	8494	0.5399 ± 0.0103	0.3444 ± 0.0105	0.6197 ± 0.0080	0.4970 ± 0.0079	-0.0795 ± 0.1732	0.0456 ± 0.1812	-0.1270 ± 0.1624	-0.0042 ± 0.1698

(*Counts*) Number of sequences being investigated. Values are stated as mean ± standard error. MS, Mononucleotide Shuffling; DS, Dinucleotide Shuffling; ZM, Zero-order Markov Model; FM, First-order Markov Model.

Table B.4: The correlation coefficients, 95th percentile, and *p*-values for pre-miRs using Mononucleotide Shuffling algorithm.

$C_p(f, g)$	*Length*	$MFEI_2$	$MFEI_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.3777	-0.0366	-0.0784	-0.0567	0.0394	0.2737	0.2424	-0.4389	-0.2470	0.0148	0.0085	0.0988	-0.0932
$MFEI_2$	6.76E-77	-0.0296	0.5484	-0.0535	-0.2937	0.5478	0.3374	0.3401	-0.8925	0.4418	0.2240	0.2366	-0.2070	-0.8288
$MFEI_1$	8.36E-02	3.76E-176	-0.0064	0.3589	-0.5960	0.5644	0.4323	0.4228	-0.4084	**0.9192**	0.5042	0.4885	-0.5936	-0.4453
%G+C	2.02E-04	1.14E-02	4.28E-69	62.3790	0.0701	-0.5437	0.0166	0.0364	-0.1909	0.2596	0.2028	0.1884	-0.0601	-0.0648
dP	7.25E-03	7.98E-46	1.30E-215	8.91E-04	0.4000	-0.6030	-0.3244	-0.2649	0.0515	-0.5436	-0.2878	-0.2377	**0.9013**	0.0538
dG	6.25E-02	1.11E-175	1.01E-188	1.23E-172	4.89E-222	-0.2937	0.3972	0.3745	-0.1964	0.6065	0.2944	0.2929	-0.4934	-0.3448
dQ	8.69E-40	8.85E-61	1.05E-102	4.33E-01	4.43E-56	1.40E-85	0.2885	**0.9829**	-0.2230	0.4257	**0.9444**	**0.9290**	-0.3441	-0.1315
dD	2.44E-31	8.42E-62	7.14E-98	8.48E-02	2.70E-37	1.57E-75	0.00E+00	0.0984	-0.2400	0.4251	**0.9396**	**0.9545**	-0.2971	-0.1620
dF	3.50E-106	0.00E+00	8.37E-91	7.86E-20	1.47E-02	6.50E-21	1.17E-26	1.01E-30	0.3820	-0.2594	-0.1319	-0.1515	0.0006	0.8292
zG	1.68E-32	1.06E-107	0.00E+00	7.70E-36	1.44E-172	2.69E-225	2.33E-99	4.94E-99	8.88E-36	-1.8302	0.5474	0.5325	-0.6583	-0.4068
zQ	4.85E-01	6.82E-27	7.81E-145	3.20E-22	5.17E-44	4.62E-46	0.00E+00	0.00E+00	3.68E-10	1.99E-175	-0.5625	**0.9844**	-0.3876	-0.1195
zD	6.86E-01	6.85E-30	9.42E-135	2.42E-19	3.70E-30	1.37E-45	0.00E+00	0.00E+00	5.62E-13	2.34E-164	0.00E+00	-0.3737	-0.3316	-0.1508
zP	2.78E-06	4.15E-23	1.87E-213	4.41E-03	0.00E+00	7.75E-138	2.61E-63	6.64E-47	9.77E-01	1.52E-278	3.09E-81	1.13E-58	4.2017	0.0269
zF	9.80E-06	0.00E+00	1.38E-109	2.16E-03	1.09E-02	1.37E-63	4.19E-10	1.21E-14	0.00E+00	4.47E-90	1.38E-08	7.24E-13	2.03E-01	1.3094

$C_s(f, g)$	*Length*	$MFEI_2$	$MFEI_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.4177	0.0087	-0.0162	-0.0836	0.0175	0.1887	0.1679	-0.5274	-0.1333	-0.0281	-0.0258	0.0209	-0.1057
$MFEI_2$	7.28E-190	-0.0296	0.3772	-0.0689	-0.2149	0.4190	0.3124	0.3092	-0.7060	0.2867	0.1452	0.1614	-0.1283	-0.5533
$MFEI_1$	5.42E-01	8.09E-158	-0.0064	0.2446	-0.4185	0.3975	0.3022	0.2865	-0.2376	0.7732	0.3429	0.3278	-0.4063	-0.2725
%G+C	2.55E-01	1.03E-06	2.25E-67	62.3790	0.0245	-0.3586	0.0258	0.0277	-0.1374	0.1562	0.1775	0.1579	-0.0365	-0.0388
dP	4.24E-09	2.66E-52	4.79E-193	8.24E-02	0.4000	-0.4024	-0.2048	-0.1539	0.0334	-0.3727	-0.1560	-0.1190	0.7354	0.0148
dG	2.19E-01	3.87E-194	6.60E-175	1.70E-142	1.54E-178	-0.2937	0.2502	0.2357	-0.0917	0.4447	0.1470	0.1530	-0.3294	-0.2152
dQ	3.22E-40	8.83E-109	6.25E-102	6.76E-02	1.15E-47	1.92E-70	2.88E-01	0.8927	-0.2253	0.2783	0.7257	0.7197	-0.2132	-0.1117
dD	3.49E-32	1.59E-106	9.29E-92	4.95E-02	1.29E-27	1.18E-62	0.00E+00	0.0984	-0.2332	0.2681	0.7194	0.7613	-0.1667	-0.1368
dF	1.05E-271	0.00E+00	1.40E-57	2.34E-20	2.47E-02	6.70E-10	5.89E-52	1.84E-55	0.3820	-0.1166	-0.0926	-0.1115	-0.0285	0.5982
zG	6.77E-21	5.71E-92	0.00E+00	1.67E-28	1.60E-153	1.84E-218	9.80E-87	1.44E-80	4.08E-15	-1.8302	0.3672	0.3494	-0.4490	-0.2341
zQ	4.83E-02	6.85E-25	8.95E-131	2.57E-36	2.17E-28	1.80E-25	0.00E+00	0.00E+00	4.49E-10	1.21E-149	-0.5625	0.8913	-0.2345	-0.0625
zD	6.96E-02	2.31E-30	1.12E-119	4.31E-29	3.47E-17	1.97E-27	0.00E+00	0.00E+00	6.05E-14	1.16E-135	0.00E+00	-0.3737	-0.1864	-0.0951
zP	1.41E-01	8.84E-20	9.77E-183	9.71E-03	0.00E+00	8.91E-121	1.20E-51	2.99E-32	5.50E-02	1.12E-222	3.70E-62	6.32E-40	4.2017	-0.0313
zF	1.02E-13	0.00E+00	2.76E-83	5.94E-03	2.93E-01	1.31E-52	2.33E-15	3.00E-22	0.00E+00	5.81E-62	9.29E-06	1.48E-11	2.66E-02	1.3094

$C_k(f, g)$	*Length*	$MFEI_2$	$MFEI_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.5843	0.0136	-0.0244	-0.1230	0.0254	0.2765	0.2471	-0.6788	-0.1973	-0.0365	-0.0339	0.0328	-0.1581
$MFEI_2$	2.38E-205	-0.0296	0.5299	-0.1018	-0.3162	0.5720	0.4496	0.4458	-0.8518	0.3971	0.2111	0.2350	-0.1884	-0.7362
$MFEI_1$	5.19E-01	1.76E-162	-0.0064	0.3509	-0.5911	0.5597	0.4326	0.4115	-0.3260	**0.9253**	0.4862	0.4662	-0.5787	-0.3946
%G+C	2.48E-01	1.38E-06	6.04E-66	62.3790	0.0363	-0.5068	0.0376	0.0408	-0.1924	0.2265	0.2567	0.2288	-0.0541	-0.0577
dP	5.17E-09	3.11E-53	3.14E-211	8.61E-02	0.4000	-0.5698	-0.3026	-0.2282	0.0475	-0.5352	-0.2325	-0.1776	**0.9049**	0.0221
dG	2.30E-01	6.67E-195	5.31E-185	1.62E-146	4.43E-193	-0.2937	0.3630	0.3430	-0.1291	0.6129	0.2173	0.2260	-0.4729	-0.3150
dQ	1.33E-40	6.12E-112	7.07E-103	7.50E-02	1.13E-48	9.35E-71	0.2885	**0.9837**	-0.3128	0.3967	0.8929	0.8910	-0.3131	-0.1659
dD	1.58E-32	6.99E-110	2.52E-92	5.36E-02	7.35E-28	6.60E-63	0.00E+00	0.0984	-0.3237	0.3839	0.8891	**0.9162**	-0.2466	-0.2027
dF	1.16E-302	0.00E+00	1.25E-56	3.95E-20	2.45E-02	8.69E-10	4.65E-52	8.18E-56	0.3820	-0.1537	-0.1265	-0.1528	-0.0407	0.7564
zG	4.12E-21	1.64E-85	0.00E+00	1.82E-27	2.81E-166	2.44E-231	2.28E-85	1.28E-79	2.58E-13	-1.8302	0.5136	0.4913	-0.6282	-0.3370
zQ	8.40E-02	5.34E-24	2.63E-133	4.63E-35	6.84E-29	2.32E-25	0.00E+00	0.00E+00	1.86E-09	3.94E-151	-0.5625	**0.9831**	-0.3429	-0.0926
zD	1.08E-01	1.71E-29	2.69E-121	5.24E-28	2.46E-17	2.40E-27	0.00E+00	0.00E+00	3.54E-13	1.62E-136	0.00E+00	-0.3737	-0.2741	-0.1406
zP	1.20E-01	2.37E-19	1.51E-200	1.05E-02	0.00E+00	3.13E-125	3.73E-52	2.07E-32	5.38E-02	2.54E-246	7.45E-63	6.58E-40	4.2017	-0.0459
zF	5.10E-14	0.00E+00	2.22E-84	6.28E-03	2.95E-01	8.61E-53	2.71E-15	3.31E-22	0.00E+00	1.27E-60	1.13E-05	2.33E-11	2.97E-02	1.3094

Three tables of Pearson correlation coefficients *Cp*, Spearman-rank *Cs* (ranks-based) and Kendall's *Ck* (relative ranks-based). (*Upper diagonal*) Correlation coefficients $C(f, g)$. $|C| \leq 1.0$, 1.0 for trend identical, -1.0 for perfect opposite, and 0.0 for complete independence. *Bold, 0.9* $\leq |C|$ strongly correlated, *0.4* $\leq |C| <$ *0.9 moderately, and* $|C| <$ *0.4 weakly;* (*Diagonal*) 95th percentile; (*Lower diagonal*) two-tailed *p*-values using the Student's t distribution for *Cp*. two-tailed *p*-values using the large-sample approximations for *Cs* and *Ck*. The pair(s) of variables with $Cp > 0$ ($Cp < 0$) and *p*-value < 0.001 tend to increase together (one variable decreases while the other increases).

Table B.5: The correlation coefficients, 95th percentile, and *p*-values for pre-miRs using Dinucleotide Shuffling algorithm.

$C_p(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.3777	-0.0366	-0.0784	-0.0567	0.0394	0.2737	0.2424	-0.4389	-0.2180	0.0183	0.0098	0.0814	-0.0884
MFEI$_2$	6.76E-77	-0.0296	0.5484	-0.0535	-0.2937	0.5478	0.3374	0.3401	-0.8925	0.4157	0.2147	0.2279	-0.1888	-0.8092
MFEI$_1$	8.36E-02	3.76E-176	-0.0064	0.3589	-0.5960	0.5644	0.4323	0.4228	-0.4084	0.8833	0.4905	0.4813	-0.5677	-0.4343
%G+C	2.02E-04	1.14E-02	4.28E-69	62.3790	0.0701	-0.5437	0.0166	0.0364	-0.1909	0.2613	0.2059	0.1970	-0.0756	-0.0637
dP	7.25E-03	7.98E-46	1.30E-215	8.91E-04	0.4000	-0.6030	-0.3244	-0.2649	0.0515	-0.5410	-0.2862	-0.2379	0.8867	0.0464
dG	6.25E-02	1.11E-175	1.01E-188	1.23E-172	4.89E-222	-0.2937	0.3972	0.3745	-0.1964	0.5748	0.2794	0.2784	-0.4598	-0.3364
dQ	8.69E-40	8.85E-61	1.05E-102	4.33E-01	4.43E-56	1.40E-85	0.2885	**0.9829**	-0.2230	0.4823	**0.9387**	**0.9221**	-0.3730	-0.1336
dD	2.44E-31	8.42E-62	7.14E-98	8.48E-02	2.70E-37	1.57E-75	0.00E+00	0.0984	-0.2400	0.4791	**0.9348**	**0.9486**	-0.3237	-0.1640
dF	3.50E-106	0.00E+00	8.37E-91	7.86E-20	1.47E-02	6.50E-21	1.17E-26	1.01E-30	0.3820	-0.2459	-0.1279	-0.1486	-0.0074	0.8195
zG	1.59E-25	2.31E-94	0.00E+00	2.59E-36	1.36E-170	2.84E-197	6.51E-131	5.93E-129	3.25E-32	-1.4415	0.5998	0.5862	-0.6668	-0.3767
zQ	3.86E-01	8.99E-25	5.58E-136	7.18E-23	1.62E-43	1.85E-41	0.00E+00	0.00E+00	1.24E-09	4.68E-219	-0.5185	**0.9846**	-0.4207	-0.1125
zD	6.44E-01	8.74E-28	2.72E-130	4.74E-21	3.26E-30	3.71E-41	0.00E+00	0.00E+00	1.52E-12	5.75E-207	0.00E+00	-0.3807	-0.3646	-0.1450
zP	1.14E-04	1.98E-19	2.10E-191	3.39E-04	0.00E+00	1.15E-117	6.87E-75	8.20E-56	7.28E-01	2.83E-288	8.09E-97	2.00E-71	4.1119	0.0030
zF	2.78E-05	0.00E+00	9.35E-104	2.56E-03	2.82E-02	1.95E-60	2.16E-10	5.54E-15	0.00E+00	1.78E-76	9.21E-08	5.33E-12	8.87E-01	1.3811

$C_s(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.4177	0.0087	-0.0162	-0.0836	0.0175	0.1887	0.1679	-0.5274	-0.1114	-0.0212	-0.0211	0.0102	-0.0973
MFEI$_2$	7.28E-190	-0.0296	0.3772	-0.0689	-0.2149	0.4190	0.3124	0.3092	-0.7060	0.2778	0.1371	0.1534	-0.1211	-0.5369
MFEI$_1$	5.42E-01	8.09E-158	-0.0064	0.2446	-0.4185	0.3975	0.3022	0.2865	-0.2376	0.7201	0.3241	0.3121	-0.3907	-0.2646
%G+C	2.55E-01	1.03E-06	2.25E-67	62.3790	0.0245	-0.3586	0.0258	0.0277	-0.1374	0.1604	0.1767	0.1589	-0.0466	-0.0327
dP	4.24E-09	2.66E-52	4.79E-193	8.24E-02	0.4000	-0.4024	-0.2048	-0.1539	0.0334	-0.3728	-0.1520	-0.1148	0.7186	0.0143
dG	2.19E-01	3.87E-194	6.60E-175	1.70E-142	1.54E-178	-0.2937	0.2502	0.2357	-0.0917	0.4153	0.1313	0.1380	-0.3062	-0.2117
dQ	3.22E-40	8.83E-109	6.25E-102	6.76E-02	1.15E-47	1.92E-70	0.2885	0.8927	-0.2253	0.3108	0.7139	0.7099	-0.2291	-0.1148
dD	3.49E-32	1.59E-106	9.29E-92	4.95E-02	1.29E-27	1.18E-62	0.00E+00	0.0984	-0.2332	0.2986	0.7080	0.7487	-0.1815	-0.1395
dF	1.05E-271	0.00E+00	1.40E-57	2.34E-20	2.47E-02	6.70E-10	5.89E-52	1.84E-55	0.3820	-0.1194	-0.0896	-0.1084	-0.0274	0.5785
zG	4.59E-15	1.74E-86	0.00E+00	5.64E-30	1.36E-153	8.33E-191	9.91E-108	1.89E-99	9.00E-16	-1.4415	0.3996	0.3796	-0.4591	-0.2254
zQ	1.36E-01	2.27E-22	5.48E-117	5.42E-36	5.21E-27	1.22E-20	0.00E+00	0.00E+00	1.65E-09	7.65E-177	-0.5185	0.8916	-0.2539	-0.0587
zD	1.38E-01	1.34E-27	1.16E-108	1.91E-29	4.20E-16	1.27E-22	0.00E+00	0.00E+00	2.99E-13	8.26E-160	0.00E+00	-0.3807	-0.2045	-0.0916
zP	4.72E-01	8.41E-18	4.17E-169	9.57E-04	0.00E+00	1.21E-104	2.13E-59	6.97E-38	6.55E-02	8.56E-233	1.54E-72	1.04E-47	4.1119	-0.0365
zF	7.66E-12	0.00E+00	1.31E-78	2.04E-02	3.13E-01	5.72E-51	3.90E-16	4.69E-23	0.00E+00	1.37E-57	3.12E-05	8.21E-11	9.63E-03	1.3811

$C_k(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.5843	0.0136	-0.0244	-0.1230	0.0254	0.2765	0.2471	-0.6788	-0.1642	-0.0270	-0.0275	0.0162	-0.1457
MFEI$_2$	2.38E-205	-0.0296	0.5299	-0.1018	-0.3162	0.5720	0.4496	0.4458	-0.8518	0.3910	0.1995	0.2235	-0.1775	-0.7185
MFEI$_1$	5.19E-01	1.76E-162	-0.0064	0.3509	-0.5911	0.5597	0.4326	0.4115	-0.3260	0.8869	0.4612	0.4455	-0.5561	-0.3836
%G+C	2.48E-01	1.38E-06	6.04E-66	62.3790	0.0363	-0.5068	0.0376	0.0408	-0.1924	0.2307	0.2555	0.2306	-0.0688	-0.0486
dP	5.17E-09	3.11E-53	3.14E-211	8.61E-02	0.4000	-0.5698	-0.3026	-0.2282	0.0475	-0.5337	-0.2261	-0.1716	0.8892	0.0212
dG	2.30E-01	6.67E-195	5.31E-185	1.62E-146	4.43E-193	-0.2937	0.3630	0.3430	-0.1291	0.5776	0.1948	0.2046	-0.4416	-0.3102
dQ	1.33E-40	6.12E-112	7.07E-103	7.50E-02	1.13E-48	9.35E-71	0.2885	**0.9837**	-0.3128	0.4430	0.8842	0.8833	-0.3366	-0.1704
dD	1.58E-32	6.99E-110	2.52E-92	5.36E-02	7.35E-28	6.60E-63	0.00E+00	0.0984	-0.3237	0.4276	0.8798	**0.9078**	-0.2683	-0.2065
dF	1.16E-302	0.00E+00	1.25E-56	3.95E-20	2.45E-02	8.69E-10	4.65E-52	8.18E-56	0.3820	-0.1603	-0.1223	-0.1485	-0.0391	0.7361
zG	5.13E-15	8.88E-83	0.00E+00	1.87E-28	3.27E-165	1.26E-199	2.30E-108	2.54E-100	2.32E-14	-1.4415	0.5574	0.5335	-0.6398	-0.3270
zQ	2.02E-01	1.47E-21	2.02E-118	1.00E-34	2.25E-27	1.33E-20	0.00E+00	0.00E+00	6.33E-09	3.56E-183	-0.5185	**0.9832**	-0.3708	-0.0860
zD	1.94E-01	9.34E-27	1.04E-109	2.00E-28	2.87E-16	1.31E-22	0.00E+00	0.00E+00	1.59E-12	4.83E-165	0.00E+00	-0.3807	-0.3008	-0.1347
zP	4.45E-01	2.58E-17	3.97E-182	1.12E-03	0.00E+00	1.28E-107	1.70E-60	2.89E-38	6.43E-02	2.75E-258	5.30E-74	4.31E-48	4.1119	-0.0539
zF	4.13E-12	0.00E+00	1.89E-79	2.15E-02	3.15E-01	3.61E-51	4.65E-16	5.32E-23	0.00E+00	5.18E-57	4.56E-05	1.54E-10	1.07E-02	1.3811

Three tables of Pearson correlation coefficients *Cp*, Spearman-rank *Cs* (ranks-based) and Kendall's *Ck* (relative ranks-based). (*Upper diagonal*) Correlation coefficients $C(f, g)$. $|C| \leq 1.0$, 1.0 for trend identical, -1.0 for perfect opposite, and 0.0 for complete independence. *Bold, 0.9* $\leq |C|$ strongly correlated, *0.4* $\leq |C| <$ *0.9 moderately, and* $|C| <$ *0.4 weakly;* (*Diagonal*) 95th percentile; (*Lower diagonal*) two-tailed *p*-values using the Student's t distribution for *Cp*. two-tailed *p*-values using the large-sample approximations for *Cs* and *Ck*. The pair(s) of variables with $Cp > 0$ ($Cp < 0$) and *p*-value < 0.001 tend to increase together (one variable decreases while the other increases).

Table B.6: The correlation coefficients, 95th percentile, and *p*-values for pre-miRs using Zero-order Markov Model algorithm.

$C_p(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.3777	-0.0366	-0.0784	-0.0567	0.0394	0.2737	0.2424	-0.4389	-0.1602	0.0172	0.0087	0.0750	-0.0935
MFEI$_2$	6.76E-77	-0.0296	0.5484	-0.0535	-0.2937	0.5478	0.3374	0.3401	-0.8925	0.4701	0.2283	0.2405	-0.2250	-0.8157
MFEI$_1$	8.36E-02	3.76E-176	-0.0064	0.3589	-0.5960	0.5644	0.4323	0.4228	-0.4084	**0.9704**	0.5056	0.4906	-0.6296	-0.4441
%G+C	2.02E-04	1.14E-02	4.28E-69	62.3790	0.0701	-0.5437	0.0166	0.0364	-0.1909	0.3849	0.1922	0.1750	-0.0969	-0.0597
dP	7.25E-03	7.98E-46	1.30E-215	8.91E-04	0.4000	-0.6030	-0.3244	-0.2649	0.0515	-0.5565	-0.2953	-0.2478	**0.9384**	0.0583
dG	6.25E-02	1.11E-175	1.01E-188	1.23E-172	4.89E-222	-0.2937	0.3972	0.3745	-0.1964	0.5294	0.3060	0.3077	-0.4964	-0.3492
dQ	8.69E-40	8.85E-61	1.05E-102	4.33E-01	4.43E-56	1.40E-85	0.2885	**0.9829**	-0.2230	0.4231	**0.9475**	**0.9317**	-0.3394	-0.1305
dD	2.44E-31	8.42E-62	7.14E-98	8.48E-02	2.70E-37	1.57E-75	0.00E+00	0.0984	-0.2400	0.4194	**0.9423**	**0.9569**	-0.2895	-0.1609
dF	3.50E-106	0.00E+00	8.37E-91	7.86E-20	1.47E-02	6.50E-21	1.17E-26	1.01E-30	0.3820	-0.3303	-0.1309	-0.1485	0.0215	0.8123
zG	2.37E-14	1.37E-123	0.00E+00	4.63E-80	2.05E-182	3.99E-162	5.16E-98	3.35E-96	3.38E-58	-1.3010	0.5364	0.5210	-0.6525	-0.4227
zQ	4.15E-01	6.82E-28	9.22E-146	4.41E-20	2.40E-46	8.79E-50	0.00E+00	0.00E+00	5.03E-10	3.25E-167	-0.5731	**0.9840**	-0.3769	-0.1192
zD	6.82E-01	7.57E-31	4.58E-136	7.06E-17	1.07E-32	2.43E-50	0.00E+00	0.00E+00	1.61E-12	3.59E-156	0.00E+00	-0.3648	-0.3234	-0.1509
zP	3.80E-04	4.12E-27	9.33E-248	4.27E-06	0.00E+00	9.14E-140	1.49E-61	1.63E-44	3.10E-01	5.89E-272	1.50E-76	9.97E-56	3.7370	0.0465
zF	9.33E-06	0.00E+00	6.02E-109	4.71E-03	5.78E-03	2.86E-65	5.59E-10	1.81E-14	0.00E+00	8.07E-98	1.53E-08	6.97E-13	2.76E-02	1.0831

$C_s(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.4177	0.0087	-0.0162	-0.0836	0.0175	0.1887	0.1679	-0.5274	-0.0611	-0.0263	-0.0255	0.0000	-0.1074
MFEI$_2$	7.28E-190	-0.0296	0.3772	-0.0689	-0.2149	0.4190	0.3124	0.3092	-0.7060	0.3154	0.1509	0.1672	-0.1454	-0.5530
MFEI$_1$	5.42E-01	8.09E-158	-0.0064	0.2446	-0.4185	0.3975	0.3022	0.2865	-0.2376	0.8793	0.3465	0.3328	-0.4418	-0.2713
%G+C	2.55E-01	1.03E-06	2.25E-67	62.3790	0.0245	-0.3586	0.0258	0.0277	-0.1374	0.2557	0.1683	0.1471	-0.0605	-0.0399
dP	4.24E-09	2.66E-52	4.79E-193	8.24E-02	0.4000	-0.4024	-0.2048	-0.1539	0.0334	-0.3840	-0.1629	-0.1276	0.8075	0.0147
dG	2.19E-01	3.87E-194	6.60E-175	1.70E-142	1.54E-178	-0.2937	0.2502	0.2357	-0.0917	0.3753	0.1580	0.1665	-0.3337	-0.2132
dQ	3.22E-40	8.83E-109	6.25E-102	6.76E-02	1.15E-47	1.92E-70	0.2885	0.8927	-0.2253	0.2862	0.7353	0.7268	-0.2079	-0.1117
dD	3.49E-32	1.59E-106	9.29E-92	4.95E-02	1.29E-27	1.18E-62	0.00E+00	0.0984	-0.2332	0.2736	0.7277	0.7690	-0.1601	-0.1368
dF	1.05E-271	0.00E+00	1.40E-57	2.34E-20	2.47E-02	6.70E-10	5.89E-52	1.84E-55	0.3820	-0.1850	-0.0928	-0.1103	-0.0087	0.6003
zG	1.71E-05	6.65E-111	0.00E+00	1.98E-73	7.11E-163	3.69E-156	1.17E-91	7.56E-84	1.27E-35	-1.3010	0.3676	0.3507	-0.4478	-0.2521
zQ	6.43E-02	9.45E-27	1.96E-133	7.89E-33	8.97E-31	3.69E-29	0.00E+00	0.00E+00	4.25E-10	5.56E-150	-0.5731	0.8887	-0.2234	-0.0655
zD	7.33E-02	1.82E-32	3.07E-123	1.87E-25	1.63E-19	3.49E-32	0.00E+00	0.00E+00	1.12E-13	1.25E-136	0.00E+00	-0.3648	-0.1793	-0.0987
zP	**1.00E+00**	5.85E-25	1.11E-215	1.78E-05	0.00E+00	7.30E-124	3.11E-49	7.52E-30	5.57E-01	1.66E-221	1.43E-56	4.56E-37	3.7370	-0.0107
zF	4.03E-14	0.00E+00	1.46E-82	4.69E-03	2.99E-01	1.07E-51	2.25E-15	3.04E-22	0.00E+00	1.48E-71	3.31E-06	2.51E-12	4.49E-01	1.0831

$C_k(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.5843	0.0136	-0.0244	-0.1230	0.0254	0.2765	0.2471	-0.6788	-0.0914	-0.0339	-0.0336	0.0020	-0.1604
MFEI$_2$	2.38E-205	-0.0296	0.5299	-0.1018	-0.3162	0.5720	0.4496	0.4458	-0.8518	0.4419	0.2192	0.2430	-0.2140	-0.7350
MFEI$_1$	5.19E-01	1.76E-162	-0.0064	0.3509	-0.5911	0.5597	0.4326	0.4115	-0.3260	**0.9793**	0.4912	0.4731	-0.6225	-0.3934
%G+C	2.48E-01	1.38E-06	6.04E-66	62.3790	0.0363	-0.5068	0.0376	0.0408	-0.1924	0.3660	0.2436	0.2132	-0.0898	-0.0591
dP	5.17E-09	3.11E-53	3.14E-211	8.61E-02	0.4000	-0.5698	-0.3026	-0.2282	0.0475	-0.5491	-0.2424	-0.1905	**0.9470**	0.0224
dG	2.30E-01	6.67E-195	5.31E-185	1.62E-146	4.43E-193	-0.2937	0.3630	0.3430	-0.1291	0.5287	0.2331	0.2455	-0.4790	-0.3126
dQ	1.33E-40	6.12E-112	7.07E-103	7.50E-02	1.13E-48	9.35E-71	0.2885	**0.9837**	-0.3128	0.4089	0.8997	0.8963	-0.3058	-0.1658
dD	1.58E-32	6.99E-110	2.52E-92	5.36E-02	7.35E-28	6.60E-63	0.00E+00	0.0984	-0.3237	0.3924	0.8951	**0.9211**	-0.2371	-0.2023
dF	1.16E-302	0.00E+00	1.25E-56	3.95E-20	2.45E-02	8.69E-10	4.65E-52	8.18E-56	0.3820	-0.2495	-0.1265	-0.1508	-0.0129	0.7563
zG	1.47E-05	8.70E-108	0.00E+00	5.29E-72	9.72E-177	1.38E-161	4.76E-91	2.28E-83	3.84E-33	-1.3010	0.5161	0.4944	-0.6279	-0.3634
zQ	1.08E-01	8.97E-26	1.90E-136	1.25E-31	2.44E-31	4.82E-29	0.00E+00	0.00E+00	1.89E-09	8.86E-153	-0.5731	**0.9824**	-0.3276	-0.0966
zD	1.12E-01	1.78E-31	2.30E-125	1.86E-24	9.36E-20	4.01E-32	0.00E+00	0.00E+00	7.16E-13	1.86E-138	0.00E+00	-0.3648	-0.2642	-0.1454
zP	**9.26E-01**	1.27E-24	1.21E-240	2.08E-05	0.00E+00	6.61E-129	9.92E-50	5.32E-30	5.43E-01	5.11E-246	3.39E-57	4.21E-37	3.7370	-0.0152
zF	2.24E-14	0.00E+00	7.79E-84	5.13E-03	2.89E-01	5.51E-52	2.83E-15	3.91E-22	0.00E+00	6.39E-71	4.58E-06	4.60E-12	4.72E-01	1.0831

Three tables of Pearson correlation coefficients *Cp*, Spearman-rank *Cs* (ranks-based) and Kendall's *Ck* (relative ranks-based). (*Upper diagonal*) Correlation coefficients $C(f, g)$. $|C| \leq 1.0$, 1.0 for trend identical, -1.0 for perfect opposite, and 0.0 for complete independence. *Bold, 0.9* $\leq |C|$ strongly correlated, *0.4* $\leq |C| <$ *0.9 moderately, and* $|C| <$ *0.4 weakly;* (*Diagonal*) 95th percentile; (*Lower diagonal*) two-tailed *p*-values using the Student's t distribution for *Cp*. two-tailed *p*-values using the large-sample approximations for *Cs* and *Ck*. The pair(s) of variables with $Cp > 0$ ($Cp < 0$) and *p*-value < 0.001 tend to increase together (one variable decreases while the other increases).

Table B.7: The correlation coefficients, 95th percentile, and *p*-values for pre-miRs using First-order Markov Model algorithm.

$C_p(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.3777	-0.0366	-0.0784	-0.0567	0.0394	0.2737	0.2424	-0.4389	-0.1578	0.0208	0.0101	0.0641	-0.0880
MFEI$_2$	6.76E-77	-0.0296	0.5484	-0.0535	-0.2937	0.5478	0.3374	0.3401	-0.8925	0.4137	0.2179	0.2309	-0.1919	-0.7784
MFEI$_1$	8.36E-02	3.76E-176	-0.0064	0.3589	-0.5960	0.5644	0.4323	0.4228	-0.4084	**0.9075**	0.4930	0.4841	-0.5859	-0.4301
%G+C	2.02E-04	1.14E-02	4.28E-69	62.3790	0.0701	-0.5437	0.0166	0.0364	-0.1909	0.3349	0.1979	0.1851	-0.0902	-0.0592
dP	7.25E-03	7.98E-46	1.30E-215	8.91E-04	0.4000	-0.6030	-0.3244	-0.2649	0.0515	-0.5626	-0.2912	-0.2463	**0.9177**	0.0469
dG	6.25E-02	1.11E-175	1.01E-188	1.23E-172	4.89E-222	-0.2937	0.3972	0.3745	-0.1964	0.5241	0.2892	0.2922	-0.4662	-0.3372
dQ	8.69E-40	8.85E-61	1.05E-102	4.33E-01	4.43E-56	1.40E-85	0.2885	**0.9829**	-0.2230	0.4721	**0.9417**	**0.9252**	-0.3634	-0.1304
dD	2.44E-31	8.42E-62	7.14E-98	8.48E-02	2.70E-37	1.57E-75	0.00E+00	0.0984	-0.2400	0.4644	**0.9374**	**0.9514**	-0.3104	-0.1612
dF	3.50E-106	0.00E+00	8.37E-91	7.86E-20	1.47E-02	6.50E-21	1.17E-26	1.01E-30	0.3820	-0.2710	-0.1266	-0.1451	-0.0027	0.7941
zG	5.85E-14	2.26E-93	0.00E+00	7.52E-60	2.67E-187	2.28E-158	8.60E-125	2.68E-120	5.21E-39	-1.0477	0.5781	0.5641	-0.6670	-0.3799
zQ	3.24E-01	1.68E-25	1.40E-137	3.17E-21	4.84E-45	2.03E-44	0.00E+00	0.00E+00	1.82E-09	4.85E-200	-0.5572	**0.9841**	-0.4021	-0.1095
zD	6.32E-01	1.71E-28	4.94E-132	1.01E-18	2.54E-32	2.42E-45	0.00E+00	0.00E+00	5.06E-12	1.65E-188	0.00E+00	-0.3779	-0.3481	-0.1437
zP	2.39E-03	4.96E-20	1.08E-206	1.88E-05	0.00E+00	2.71E-121	6.46E-71	2.93E-51	8.98E-01	1.63E-288	7.14E-88	7.73E-65	3.6154	0.0103
zF	2.99E-05	0.00E+00	1.33E-101	5.09E-03	2.64E-02	1.02E-60	5.84E-10	1.65E-14	0.00E+00	7.48E-78	2.04E-07	8.35E-12	6.26E-01	1.0988

$C_s(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.4177	0.0087	-0.0162	-0.0836	0.0175	0.1887	0.1679	-0.5274	-0.0590	-0.0226	-0.0232	-0.0054	-0.0992
MFEI$_2$	7.28E-190	-0.0296	0.3772	-0.0689	-0.2149	0.4190	0.3124	0.3092	-0.7060	0.2965	0.1407	0.1574	-0.1314	-0.5352
MFEI$_1$	5.42E-01	8.09E-158	-0.0064	0.2446	-0.4185	0.3975	0.3022	0.2865	-0.2376	0.7672	0.3276	0.3165	-0.4134	-0.2609
%G+C	2.55E-01	1.03E-06	2.25E-67	62.3790	0.0245	-0.3586	0.0258	0.0277	-0.1374	0.2181	0.1694	0.1485	-0.0576	-0.0328
dP	4.24E-09	2.66E-52	4.79E-193	8.24E-02	0.4000	-0.4024	-0.2048	-0.1539	0.0334	-0.3918	-0.1560	-0.1217	0.7789	0.0127
dG	2.19E-01	3.87E-194	6.60E-175	1.70E-142	1.54E-178	-0.2937	0.2502	0.2357	-0.0917	0.3734	0.1408	0.1509	-0.3131	-0.2083
dQ	3.22E-40	8.83E-109	6.25E-102	6.76E-02	1.15E-47	1.92E-70	0.2885	0.8927	-0.2253	0.3044	0.7211	0.7162	-0.2153	-0.1147
dD	3.49E-32	1.59E-106	9.29E-92	4.95E-02	1.29E-27	1.18E-62	0.00E+00	0.0984	-0.2332	0.2905	0.7151	0.7564	-0.1666	-0.1393
dF	1.05E-271	0.00E+00	1.40E-57	2.34E-20	2.47E-02	6.70E-10	5.89E-52	1.84E-55	0.3820	-0.1599	-0.0877	-0.1050	-0.0181	0.5805
zG	3.28E-05	2.85E-98	0.00E+00	6.40E-54	1.91E-169	1.38E-154	2.06E-103	2.99E-94	5.06E-27	-1.0477	0.3807	0.3634	-0.4622	-0.2390
zQ	1.12E-01	1.78E-23	1.54E-119	3.26E-33	2.23E-28	1.67E-23	0.00E+00	0.00E+00	3.53E-09	1.02E-160	-0.5572	0.8891	-0.2342	-0.0600
zD	1.02E-01	6.02E-29	1.19E-111	6.55E-26	6.96E-18	9.78E-27	0.00E+00	0.00E+00	1.55E-12	1.41E-146	0.00E+00	-0.3779	-0.1891	-0.0937
zP	7.03E-01	1.16E-20	4.54E-189	4.39E-05	0.00E+00	2.70E-109	1.16E-52	3.42E-32	2.22E-01	6.49E-236	5.17E-62	4.76E-41	3.6154	-0.0204
zF	2.94E-12	0.00E+00	1.72E-76	2.01E-02	3.69E-01	2.13E-49	4.03E-16	5.36E-23	0.00E+00	1.64E-64	2.05E-05	2.93E-11	1.48E-01	1.0988

$C_k(f, g)$	*Length*	*MFEI*$_2$	*MFEI*$_1$	*%G+C*	*dP*	*dG*	*dQ*	*dD*	*dF*	*zG*	*zQ*	*zD*	*zP*	*zF*
Length	174.4500	0.5843	0.0136	-0.0244	-0.1230	0.0254	0.2765	0.2471	-0.6788	-0.0875	-0.0286	-0.0302	-0.0069	-0.1483
MFEI$_2$	2.38E-205	-0.0296	0.5299	-0.1018	-0.3162	0.5720	0.4496	0.4458	-0.8518	0.4190	0.2048	0.2288	-0.1933	-0.7159
MFEI$_1$	5.19E-01	1.76E-162	-0.0064	0.3509	-0.5911	0.5597	0.4326	0.4115	-0.3260	**0.9157**	0.4660	0.4514	-0.5841	-0.3788
%G+C	2.48E-01	1.38E-06	6.04E-66	62.3790	0.0363	-0.5068	0.0376	0.0408	-0.1924	0.3120	0.2452	0.2158	-0.0854	-0.0488
dP	5.17E-09	3.11E-53	3.14E-211	8.61E-02	0.4000	-0.5698	-0.3026	-0.2282	0.0475	-0.5569	-0.2324	-0.1819	**0.9259**	0.0193
dG	2.30E-01	6.67E-195	5.31E-185	1.62E-146	4.43E-193	-0.2937	0.3630	0.3430	-0.1291	0.5265	0.2087	0.2232	-0.4511	-0.3057
dQ	1.33E-40	6.12E-112	7.07E-103	7.50E-02	1.13E-48	9.35E-71	0.2885	**0.9837**	-0.3128	0.4352	0.8896	0.8884	-0.3171	-0.1701
dD	1.58E-32	6.99E-110	2.52E-92	5.36E-02	7.35E-28	6.60E-63	0.00E+00	0.0984	-0.3237	0.4170	0.8853	**0.9131**	-0.2472	-0.2059
dF	1.16E-302	0.00E+00	1.25E-56	3.95E-20	2.45E-02	8.69E-10	4.65E-52	8.18E-56	0.3820	-0.2171	-0.1201	-0.1439	-0.0258	0.7360
zG	3.34E-05	5.47E-96	0.00E+00	8.92E-52	9.13E-183	4.62E-160	3.25E-104	5.16E-95	2.60E-25	-1.0477	0.5351	0.5135	-0.6441	-0.3468
zQ	1.76E-01	1.21E-22	3.44E-121	4.98E-32	7.47E-29	1.76E-23	0.00E+00	0.00E+00	1.17E-08	3.23E-166	-0.5572	**0.9826**	-0.3437	-0.0882
zD	1.53E-01	5.14E-28	5.89E-113	4.93E-25	4.08E-18	1.08E-26	0.00E+00	0.00E+00	7.62E-12	5.14E-151	0.00E+00	-0.3779	-0.2793	-0.1376
zP	7.42E-01	2.59E-20	3.60E-205	5.14E-05	0.00E+00	9.54E-113	1.54E-53	1.54E-32	2.21E-01	7.85E-263	3.81E-63	1.94E-41	3.6154	-0.0302
zF	1.73E-12	0.00E+00	2.31E-77	2.08E-02	3.61E-01	1.10E-49	5.16E-16	6.88E-23	0.00E+00	2.46E-64	2.93E-05	6.00E-11	1.53E-01	1.0988

(Three tables of Pearson correlation coefficients *Cp*, Spearman-rank *Cs* (ranks-based) and Kendall's *Ck* (relative ranks-based). (*Upper diagonal*) Correlation coefficients $C(f, g)$. $|C| \leq 1.0$, 1.0 for trend identical, -1.0 for perfect opposite, and 0.0 for complete independence. *Bold, 0.9* $\leq |C|$ strongly correlated, *0.4* $\leq |C| <$ *0.9 moderately, and* $|C| <$ *0.4 weakly;* (*Diagonal*) 95th percentile; (*Lower diagonal*) two-tailed *p*-values using the Student's t distribution for *Cp*. two-tailed *p*-values using the large-sample approximations for *Cs* and *Ck*. The pair(s) of variables with $Cp > 0$ ($Cp < 0$) and *p*-value < 0.001 tend to increase together (one variable decreases while the other increases).

Appendix C.

Supplemental for Chapter 5

Continue on next page.

Table C.1: The prediction performances of *miPred* evaluated on the pre-miR datasets TR-H, TE-H, and IE-NH.

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 virus (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 virus (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).

Table C.2: The prediction performances of *miPred-NBC* evaluated on the pre-miR datasets TR-H, TE-H, and IE-NH.

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 virus (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 virus (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).

Table C.3: The prediction performances of *Triplet-SVM* evaluated on the pre-miR datasets TR-H, TE-H, and IE-NH.

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 virus (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 virus (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				

†, *Triplet-SVM* 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).

Table C.4: The prediction performances of *Triplet-SVM-NBC* evaluated on the pre-miR datasets TR-H, TE-H, and IE-NH.

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 virus (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 virus (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 C.5: The mean sensitivity and specificity of *miPred*, *miPred-NBC*, *Triplet-SVM*, and *Triplet-SVM-NBC* evaluated on the non-human pre-miR dataset IE-NH categorized by genus of pre-miRs.

		miPred		*miPred-NBC*			*Triplet-SVM*‡		*Triplet-SVM-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. ‡, *Triplet-SVM* 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 C.6: The prediction performances of *miPred*, *miPred-NBC*, *Triplet-SVM*, and *Triplet-SVM-NBC* evaluated on the non pre-miR datasets IE-NC and IE-M.

				miPred		*miPred-NBC*		*Triplet-SVM*[‡]			*Triplet-SVM-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

				miPred		*miPred-NBC*		*Triplet-SVM*[‡]				*Triplet-SVM-NBC*
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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
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

				miPred		*miPred-NBC*		*Triplet-SVM*[‡]			*Triplet-SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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
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

				miPred		*miPred-NBC*		*Triplet-SVM*[‡]			*Triplet-SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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
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 (fiz 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

				miPred		*miPred-NBC*		*Triplet-SVM*[‡]			*Triplet-SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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
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

				miPred		*miPred-NBC*		*Triplet-SVM*[‡]			*Triplet-SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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
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

				miPred		*miPred-NBC*			*Triplet-SVM*[‡]		*Triplet-SVM-NBC*	
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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
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

				miPred		*miPred-NBC*		*Triplet-SVM*[‡]				
Accession	*Type*[†]	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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
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

				miPred		*miPred-NBC*		*Triplet-SVM*‡			*Triplet-SVM-NBC*	
Accession	*Type*†	*Class*	*N*	*TN*	*%SP*	*TN*	*%SP*	*N*	*TN*	*%SP*	*TN*	*%SP*
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). ‡, *Triplet-SVM* 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 C.7: The mean specificity of *miPred*, *miPred-NBC*, *Triplet-SVM*, and *Triplet-SVM-NBC* evaluated on the non pre-miR dataset IE-NC categorized by classes of ncRNAs.

		miPred	*miPred-NBC*		*Triplet-SVM*†	*Triplet-SVM-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

†, *Triplet-SVM* 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 C.8: F1 and F2 scores for features of *miPred* and *Triplet-SVM*, sorted by descending F1 scores.

		miPred			*Triplet-SVM*†			
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	–

†, *Triplet-SVM* 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 C.9: 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 C.10: Putative viral-encoded pre-miRs in four viruses.

S	*SP*	*L*	*Epstein 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	UAACAAACCCGUGGGGGGUCUUAGUGGAAGUGACGUGCUGUGAAU**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**CCGACCGCACUUGCAUGCGGCCGGUGGUCCUGCGGGGGGUGACGGUC (*ebv-miR-p17*)
+	103504	95	UGGCAGAGCUUUCACCGGUGGAACUCGUGACAGAUGUCUACGCCACC**cua**GGCAUCGUGGAGAUCAUCGACGAGCUCUACCGGAGCAGUCGCCUG (*ebv-miR-p18*)
+	68250	94	CUGAGUGUGGGGGCCAUACGAGGCCUUCACUGGCCCUGUGGCC**aaggcuca**GGACGUGGGGGGCCGUUGAGGCCCACGUUGUCUGCUCGGUAGCAG (*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*	*Kaposi sarcoma-associated herpesvirus* (KSHV; U75698.1); 5' → 3'
−	119293	93	UCCAGUAGGUAUACCCAGCUGGGUCUACCCAGCUGCGUAAACCC**cgcu**GCGUAAACACAGCUGGGUAUACGCAGCUGCGUAAACCCGGCUGGG (‡*kshv-miR-K12-9*)

−	119273	94	UGGGUCUACCCAGCUGCGUAAACCCCGCUGCGUAAACACAGCUGGG**uaua**CGCAGCUGCGUAAACCCGGCUGGGUAAAUCCAGCUGUAAUUCUA ([A]*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	AGGACGCUGACGUUGGGGCCCCCGUAAGGACGUCGGCGAUCGUC**ucg**GCGCUGUCGCCACUCGUACAAAAAAUAACCCUUACUGUCAGCGCCU (*kshv-miR-p5*)
−	9635	95	GCCGUGAUCUCGUUGGCCACAAAGUGGAAGCUGUCCUCGUGGGUAGU**cugg**AUGGAGCGCGGGAAGGUUUUCCACAGUGCCAGCGGACACACGGC (*kshv-miR-p6*)
−	77286	95	GUAUAUCUUUGUUUUUUCAAGUUUGUGGACGAGGUGGUCCAUGCAUA**gac**UGGCAUGUGAUUACUCGCACAAGCGCUGACGAAAGCUAUGGUUUU (*kshv-miR-p7*)
−	121856	95	GCAGGGUGCGGUGCUGCCCAGGACGGCCGGAUGCGGGCGAUUACAG**gaaa**CUGGGUGUAAGCUGUACAUAAUCCCCGGCAGCACCUGUUUCCUGC (*kshv-miR-p8*)
−	108693	94	UGGAGUGUGGAUGGUGAUAUGGUCUCCUGGGCCUGGCCGGCCACCG**ugu**CUAUGGUGGCCAACAGGAGGCCGGCCUGCUCCGCUUCCUGUACCA (*kshv-miR-p9*)
−	120342	95	GCGCAUAUUGGCGUUGAGCGCCACCGGACGGGGAUUUAUGCUGUA**ucu**UACUACCAUGAUCCCAUGUUGCUGGCGCUCACGGCCCGUGUGCCAGC (*kshv-miR-K12-7*)

S	*SP*	*L*	*Mouse 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*	*Human 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	AAAGGACGACCCGUCUCCCCCCGCACCCGGGUUUUUUC**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**gauggG**UGGCUCUCGUCUCGGGGGACCACAGGGGGAGGGGGGUAA (*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**cuuuccaggugu uuuc**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*)

Putative viral-encoded pre-miRs having maximum length (≤ 95 nucleotides), minimum size of terminal loop (≥ 3 nucleotides), 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 *Epstein barr virus*, *Kaposi sarcoma-associated herpesvirus*, *Mouse γ-herpesvirus 68 strain WUMS*, and *Human cytomegalovirus strain AD169*; sorted in descending *miPred* scores. *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 nucleotides (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).

Appendix D.

Supplemental for Chapter 6

Continue on next page.

Table D.1: Distribution of concatamers, small RNAs, non-annotated small RNAs (candidate miRNAs), candidate pre-miRs, putative pre-miRs, and putative miRNAs.

Libraries	Concatamers	Small RNAs		Non-annotated small RNAs (candidate miRNAs)		Candidate pre-miRs	Putative pre-miRs	Putative miRNAs
		Non unique	*Unique*	*Non unique*	*Unique*			
ATE	1536	2494	1953	1362	1262	2004	682	14
AOV	1632	5870	2523	1294	1211	818	142	11
5WT	1440	3002	2167	1514	1283	3977	2882	16
5WO	1432	1990	1414	1010	844	827	102	9
5WMB	1536	2917	1991	1479	1224	2075	513	19
5WFB	2880	2743	1743	1809	1140	3747	1881	9
Total	10456	19016	11791	8468	6964	13448	6202	78

ATE, adult testis; *AOV*, adult ovary; *5WT*, 35 days post fertilization juvenile testis; *5WO*, 35 days post fertilization juvenile ovary; *5WMB*, 35 days post fertilization juvenile male brain; *5WFB*, 35 days post fertilization juvenile female brain.

Table D.2: Raw expression profiles of 780 small RNAs matching 88 known miRNAs and two novel miRNAs expressed across six miRNA Libraries.

MicroRNAs	*Adult Testis (ATE)*	*Adult Ovary (AOV)*	*Juvenile Testis (5WT)*	*Juvenile Ovary (5WO)*	*Juvenile Male Brain (5WMB)*	*Juvenile Female Brain (5WFB)*
dre-let-7a	17.450	12.050	31.800	15.983	30.400	17.433
dre-let-7b	8.667	6.333	18.333	12.833	16.000	9.833
dre-let-7c	15.533	9.800	31.183	19.233	27.567	16.367
dre-let-7d	15.467	9.600	30.033	18.833	26.467	15.567
dre-let-7e	12.983	10.450	22.183	11.650	19.433	11.700
dre-let-7f	15.117	11.383	27.133	13.983	23.233	13.767
dre-let-7g	16.950	11.883	28.133	13.983	23.233	14.267
dre-let-7h	3.000	2.000	5.000	4.500	9.333	5.833
dre-let-7i	0.000	2.000	0.000	1.000	0.333	2.333
dre-let-7j	3.833	1.500	7.200	1.000	1.000	1.900
dre-miR-101a	1.000	1.000	0.500	0.000	1.500	0.000
dre-miR-101b	1.000	1.000	0.500	0.000	0.500	0.000
dre-miR-122	0.000	2.000	0.000	3.000	0.000	0.000
dre-miR-124	0.000	0.000	0.000	0.000	0.000	8.000
dre-miR-125a	0.000	0.000	0.000	0.000	1.000	0.000
dre-miR-125b	2.500	0.000	2.500	4.000	6.000	4.000
dre-miR-125c	0.500	0.000	0.500	0.000	0.000	0.000
dre-miR-126	3.000	2.000	3.000	1.000	2.000	0.000
*dre-miR-126**	2.000	1.000	2.000	0.000	3.000	0.000
dre-miR-128	0.000	0.000	0.000	0.000	1.000	0.000
dre-miR-130a	1.000	1.000	0.333	0.000	1.000	1.500
dre-miR-130b	0.000	0.500	0.833	0.000	0.000	2.000
dre-miR-130c	1.000	1.500	0.833	0.000	1.000	2.500
dre-miR-132	0.000	0.500	0.000	0.000	0.000	0.000
dre-miR-138	0.000	0.000	0.000	0.000	0.000	1.000
dre-miR-139	0.000	0.000	0.000	1.000	1.000	0.000
*dre-miR-140**	0.000	0.000	0.000	1.000	0.000	0.000
dre-miR-141	0.000	0.000	0.000	0.000	0.500	0.000
dre-miR-142a-3p	5.000	10.333	12.000	7.000	2.000	1.000
dre-miR-142a-5p	5.000	9.333	9.000	7.000	2.000	1.000
dre-miR-142b-5p	0.000	0.333	0.000	0.000	0.000	0.000
dre-miR-143	37.000	30.000	78.000	49.000	75.000	31.000
dre-miR-144	0.000	0.000	1.000	0.000	0.000	0.000
dre-miR-145	0.000	0.000	2.000	2.000	1.000	0.000
dre-miR-146a	2.000	2.000	1.000	0.000	6.000	1.000
dre-miR-146b	1.000	4.000	0.000	1.000	0.000	0.000
dre-miR-150	3.000	1.000	4.000	3.000	7.000	4.000
dre-miR-17a	0.333	0.333	0.667	0.000	0.667	0.333
dre-miR-194a	0.000	0.000	0.500	0.000	0.000	0.000
dre-miR-194b	0.000	0.000	0.500	0.000	0.000	0.000
dre-miR-196a	1.000	0.000	0.500	0.000	0.000	0.000
dre-miR-196b	1.000	0.000	0.500	0.000	0.000	0.000
dre-miR-199	0.000	0.000	2.000	0.000	0.000	0.000
*dre-miR-199**	0.000	2.000	2.000	2.000	4.000	0.000
dre-miR-19a	0.250	2.917	0.000	0.250	0.750	0.750
dre-miR-19b	0.250	2.917	0.000	0.250	0.750	0.750
dre-miR-19c	0.250	2.917	0.000	0.250	0.750	0.750
dre-miR-19d	0.250	2.250	0.000	0.250	0.750	0.750
dre-miR-200a	0.000	0.000	0.000	0.000	0.500	0.000
*dre-miR-202**	3.000	3.000	6.000	1.000	0.000	0.000
dre-miR-204	0.000	0.000	1.000	0.000	0.000	0.000
dre-miR-20a	0.333	0.333	0.667	0.000	0.667	0.333
*dre-miR-20a**	0.000	0.000	0.000	0.000	0.000	1.000
dre-miR-20b	0.333	0.333	0.667	0.000	0.667	0.333
dre-miR-210	0.000	0.000	0.000	1.000	0.000	0.000
dre-miR-212	0.000	0.500	0.000	0.000	0.000	0.000
dre-miR-214	1.000	1.000	0.000	1.000	1.000	0.000
dre-miR-221	0.000	0.000	0.000	1.000	0.000	0.000
dre-miR-222	1.000	0.000	0.000	0.000	1.000	0.000
dre-miR-24	1.000	0.000	1.000	1.000	0.000	0.000
dre-miR-25	8.000	2.000	14.000	6.000	17.000	8.000
dre-miR-26a	0.000	0.000	0.000	0.000	0.500	0.000
dre-miR-26b	0.000	0.000	0.000	0.000	0.500	0.000
dre-miR-27b	0.000	0.500	0.833	0.000	0.000	0.333

MicroRNAs	*Adult Testis (ATE)*	*Adult Ovary (AOV)*	*Juvenile Testis (5WT)*	*Juvenile Ovary (5WO)*	*Juvenile Male Brain (5WMB)*	*Juvenile Female Brain (5WFB)*
dre-miR-27c	0.000	0.500	0.833	0.000	0.000	0.333
dre-miR-27d	0.000	0.000	0.333	0.000	0.000	0.333
dre-miR-29a	1.000	0.000	2.000	0.000	1.000	0.000
dre-miR-29b	0.000	0.000	0.000	0.000	1.000	0.000
dre-miR-301a	0.000	1.000	0.000	0.000	0.333	0.000
dre-miR-301b	0.000	0.000	0.000	0.000	0.833	0.500
dre-miR-301c	0.000	0.000	0.000	0.000	0.833	0.500
dre-miR-30a	0.000	0.000	0.000	0.000	0.500	0.000
dre-miR-30b	0.000	0.000	0.000	1.000	1.000	0.000
dre-miR-30c	0.000	0.000	0.000	0.000	1.000	0.000
dre-miR-30d	0.000	0.000	0.000	0.000	0.500	0.000
*dre-miR-30e**	0.000	0.000	0.000	1.000	1.000	0.000
dre-miR-31	0.000	0.000	1.000	1.000	1.000	2.000
dre-miR-34	0.000	0.000	0.000	0.000	1.000	0.000
dre-miR-430c	0.000	1.000	0.000	0.000	0.000	0.000
dre-miR-456	0.000	1.000	0.000	0.000	0.000	5.000
dre-miR-457a	0.000	0.000	0.000	0.000	0.000	1.000
*dre-miR-459**	0.000	0.000	0.000	1.000	0.000	0.000
dre-miR-489	1.000	0.000	1.000	0.000	0.000	2.000
dre-miR-735	0.000	0.000	0.000	1.000	0.000	0.000
dre-miR-7a	4.000	0.000	2.500	2.500	0.500	0.500
dre-miR-7b	4.000	0.000	2.500	2.500	0.500	0.500
dre-miR-92a	0.000	0.500	1.500	0.000	0.000	0.500
dre-miR-92b	0.000	0.500	1.500	0.000	0.000	0.500
dre-miR-N1	1.000	0.000	0.000	0.000	0.000	0.000
dre-miR-N2	0.000	0.000	0.000	0.000	0.000	1.000

The counts of small RNAs matching several known miRNAs are equally divided between them.

Bibliography

Abrahante,J.E. *et al*. (2003) The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev. Cell*, **4**, 625-637.

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

Adams,M.D. *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651-1656.

Ahmed,R. and Duncan,R.F. (2004) Translational regulation of Hsp90 mRNA. AUG-proximal 5'-untranslated region elements essential for preferential heat shock translation. *J. Biol. Chem.*, **279**, 49919-49930.

Alex,P. *et al*. (1990) Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, **11**, 430-452.

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.

Ambros,V. (2001) microRNAs: tiny regulators with great potential. *Cell*, **107**, 823-826.

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

Ambros,V. *et al*. (2003b) MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr. Biol.*, **13**, 807-818.

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350-355.

Ampatzis,K. and Dermon,C.R. (2007) Sex differences in adult cell proliferation within the zebrafish (Danio rerio) cerebellum. *Eur. J. Neurosci.*, **25**, 1030-1040.

Anthony,A.M. and Peter,M.W. (2005) Plant and animal microRNAs: similarities and differences. *Functional & Integrative Genomics*, **V5**, 129-135.

Ason,B. *et al*. (2006) Differences in vertebrate microRNA expression. *Proc. Natl. Acad. Sci. USA*, **103**, 14385-14389.

Banerjee,D. and Slack,F. (2002) Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *Bioessays*, **24**, 119-129.

Barash,D. (2003) Deleterious mutation prediction in the secondary structure of RNAs. *Nucl. Acids Res.*, **31**, 6578-6584.

Barash,D. (2004a) Second eigenvalue of the Laplacian matrix for predicting RNA conformational switch by mutation. *Bioinformatics*, **20**, 1861-1869.

Barash,D. (2004b) Spectral Decomposition for the Search and Analysis of RNA Secondary Structure. *J. Comp. Biol.*, **11**, 1169-1174.

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

Beckwith,J. (1996) The operon: an historical account. In Neidhardt,F.C. and Curtiss,R. (eds), *Escherichia coli and Salmonella cellular and molecular biology*. American Society for Microbiology Press, Washington, D.C, pp. 1227-1330.

Benson,D.A. *et al*. (2005) GenBank. *Nucl. 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.

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

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

Boguski,M.S. *et al*. (1993) dbEST--database for "expressed sequence tags". *Nat. Genet.*, **4**, 332-333.

Bonen,L. and Vogel,J. (2001) The ins and outs of group II introns. *Trends Genet.*, **17**, 322-331.

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.

Borchert,G.M. *et al*. (2006) RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.*, **13**, 1097-1101.

Bottoni,A. *et al*. (2005) miR-15a and miR-16-1 down-regulation in pituitary adenomas. *J. Cell Physiol.*, **204**, 280-285.

Bracht,J. *et al*. (2004) Trans-splicing and polyadenylation of let-7 microRNA primary transcripts. *RNA*, **10**, 1586-1594.

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.

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

Brenner,S. *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotech.*, **18**, 630-634.

Brown,J.R. and Sanseau,P. (2005) A computational view of microRNAs and their targets. *Drug Discovery Today*, **10**, 595-601.

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. *et al*. (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA*, **99**, 15524-15529.

Care,A. *et al*. (2007) MicroRNA-133 controls cardiac hypertrophy. *Nat. Med.*, **13**, 613-618.

Cech,T.R. (1990) Self-Splicing of Group I Introns. *Annu. Rev. Biochem.*, **59**, 543-568.

Chan,C.S. *et al*. (2005) Revealing Posttranscriptional Regulatory Elements Through Network-Level Conservation. *PLoS Comput Biol.*, **1**, e69.

Chang,C. and Lin,C. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen,C.Z. *et al*. (2004) MicroRNAs Modulate Hematopoietic Lineage Differentiation. *Science*, **303**, 83-86.

Chen,P.Y. *et al*. (2005) The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes Dev.*, **19**, 1288-1293.

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.

Clote,P. (2005) RNALOSS: a web server for RNA locally optimal secondary structures. *Nucl. Acids Res.*, **33**, W600-W604.

Clote,P. *et al*. (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578-591.

Coward,E. (1999) Shufflet: shuffling sequences while conserving the k-let counts. *Bioinformatics*, **15**, 1058-1059.

Cui,C. *et al*. (2006) Prediction and Identification of Herpes Simplex Virus 1-Encoded MicroRNAs. *J. Virol.*, **80**, 5499-5508.

Cullen,B.R. (2004a) Transcription and processing of human microRNA precursors. *Mol. Cell.*, **16**, 861-865.

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

Cullen,B.R. (2004b) Derivation and function of small interfering RNAs and microRNAs. *Virus Res.*, **102**, 3-9.

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

Delihas,N. and Forst,S. (2001) MicF: an antisense RNA gene involved in response of Escherichia coli to global stress factors. *J. Mol. Biol.*, **313**, 1-12.

Devlin,R.H. and Nagahama,Y. (2002) Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture*, **208**, 191-364.

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

Doench,J.G. *et al*. (2003) siRNAs can function as miRNAs. *Genes Dev.*, **17**, 438-442.

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.

Du,T. and Zamore,P.D. (2005) microPrimer: the biogenesis and function of microRNA. *Development*, **132**, 4645-4652.

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

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

Eder,M. and Scherr,M. (2005) MicroRNA and Lung Cancer. *N. Engl. J. Med.*, **352**, 2446-2448.

Elbashir,S.M. *et al*. (2001a) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494-498.

Elbashir,S.M. *et al*. (2001b) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.*, **15**, 188-200.

Elmen,J. *et al*. (2007) Antagonism of microRNA-122 in mice by systemically administered LNA-antimiR leads to up-regulation of a large set of predicted target mRNAs in the liver. *Nucl. Acids Res.* gkm1113.

Erdmann,V.A. *et al*. (2000) Non-coding, mRNA-like RNAs database Y2K. *Nucl. Acids Res.*, **28**, 197-200.

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

Filipowicz,W. *et al*. (2005) Post-transcriptional gene silencing by siRNAs and miRNAs. *Curr. Opin. Struct. Biol.*, **15**, 331-341.

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

Flynt,A.S. *et al*. (2007) Zebrafish miR-214 modulates Hedgehog signaling to specify muscle cell fate. *Nat. Genet.*, **39**, 259-263.

Franco-Zorrilla,J.M. *et al*. (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.*, **39**, 1033-1037.

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.

Gan,H.H. *et al*. (2003) Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucl. Acids Res.*, **31**, 2926-2943.

Giraldez,A.J. *et al*. (2006) Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, **312**, 75-79.

Giraldez,A.J. *et al*. (2005) MicroRNAs Regulate Brain Morphogenesis in Zebrafish. *Science*, **308**, 833-838.

Gray,N.K. and Wickens,M. (1998) Control of Translation Initiation in Animals. *Annu. Rev Cell Dev Biol.*, **14**, 399-458.

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

Gregory,R.I. and Shiekhattar,R. (2005) MicroRNA Biogenesis and Cancer. *Cancer Res.*, **65**, 3509-3512.

Grey,F. *et al*. (2005) Identification and characterization of human cytomegalovirus-encoded microRNAs. *J. Virol.*, **79**, 12095-12099.

Griffiths-Jones,S. (2004) The microRNA Registry. *Nucl. Acids Res.*, **32**, D109-D111.

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

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

Grivna,S.T. *et al*. (2006) A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.*, **20**, 1709-1714.

Grundhoff,A. *et al*. (2006) A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA*, **12**, 733-750.

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

Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244-251.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev. Genet*, **5**, 522-531.

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

Hesselberth,J.R. and Ellington,A.D. (2002) A (ribo) switch in the paradigms of genetic regulation. *Nat. Struct. Biol.*, **9**, 891-893.

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

Hou,Y. *et al*. (2003) Efficient remote homology detection using local structure. *Bioinformatics*, **19**, 2294-2301.

Houbaviy,H.B. *et al*. (2003) Embryonic stem cell-specific MicroRNAs. *Dev. Cell*, **5**, 351-358.

Houwing,S. *et al*. (2007) A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell*, **129**, 69-82.

Huang,J. *et al*. (2007) Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+ T lymphocytes. *Nat. Med.*, **13**, 1241-1247.

Huttenhofer,A. *et al*. (2005) Non-coding RNAs: hope or hype? *Trends Genet.*, **21**, 289-297.

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

Iorio,M.V. *et al*. (2005) MicroRNA Gene Expression Deregulation in Human Breast Cancer. *Cancer Res.*, **65**, 7065-7070.

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

Jiang,J. *et al*. (2005) Real-time expression profiling of microRNA precursors in human cancer cell lines. *Nucl. Acids Res.*, **33**, 5394-5403.

Jiang,P. *et al*. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**, W339-W344.

Jin,G. *et al*. (2006) Primary transcripts and expressions of mammal intergenic microRNAs detected by mapping ESTs to their flanking sequences. *Mammalian Genome*, **V17**, 1033-1041.

Johansson,J. *et al*. (2002) An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. *Cell*, **110**, 551-561.

John,B. *et al*. (2004) Human MicroRNA Targets. *PLoS Biol.*, **2**, e363.

Johnson,S.M. *et al*. (2005) RAS is regulated by the let-7 microRNA family. *Cell*, **120**, 635-647.

Johnston,R.J. and Hobert,O. (2003) A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature*, **426**, 845-849.

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.

Kapsimali,M. *et al*. (2007) MicroRNAs show a wide diversity of expression profiles in the developing and mature central nervous system. *Genome Biol.*, **8**, R173.

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

Katz,L. and Burge,C.B. (2003) Widespread Selection for Local RNA Secondary Structure in Coding Regions of Bacterial Genes. *Genome Res.*, **13**, 2042-2051.

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

Kitagawa,J. *et al*. (2003) Analysis of the conformational energy landscape of human snRNA with a metric based on tree representation of RNA structures. *Nucl. Acids Res.*, **31**, 2006-2013.

Klein,R.J. *et al*. (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. USA*, **99**, 7542-7547.

Kloosterman,W.P. *et al*. (2007) Targeted Inhibition of miRNA Maturation with Morpholinos Reveals a Role for miR-375 in Pancreatic Islet Development. *PLoS Biol.*, **5**, e203.

Kloosterman,W.P. *et al*. (2006) Cloning and expression of new microRNAs from zebrafish. *Nucl. Acids Res.*, **34**, 2558-2569.

Kloosterman,W.P. *et al*. (2004) Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucl. Acids Res.*, **32**, 6284-6291.

Krichevsky,A.M. *et al*. (2003) A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, **9**, 1274-1281.

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.

Lagos-Quintana,M. *et al*. (2002) Identification of Tissue-Specific MicroRNAs from Mouse. *Curr. Biol.*, **12**, 735-739.

Lai,E.C. (2003) RNA sensors and riboswitches: self-regulating messages. *Curr. Biol.*, **13**, R285-R291.

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

Landgraf,P. *et al*. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell.*, **129**, 1401-1414.

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

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

Lecellier,C.H. *et al*. (2005) A Cellular MicroRNA Mediates Antiviral Defense in Human Cells. *Science*, **308**, 557-560.

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.

Lee,Y. *et al*. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415-419.

Lee,Y. *et al*. (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.*, **21**, 4663-4670.

Lewis,B.P. *et al*. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.

Li,S.C. *et al*. (2006) Bioinformatic discovery of microRNA precursors from human ESTs and introns. *BMC Genomics*, **7**, 164.

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.

Li,X. and Zhang,Y.Z. (2005) Computational detection of microRNAs targeting transcription factor genes in Arabidopsis thaliana. *Comput. Biol. Chem.*, **29**, 360-367.

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.

Lin,S.L. *et al*. (2006) Intronic MicroRNA (miRNA). *J Biomed. Biotechnol.*, **2006**, 26818.

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. *Nucl. Acids Res.*, **25**, 955-964.

Lu,C. *et al*. (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res.*, **16**, 1276-1288.

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

Ma,L. *et al*. (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*, **449**, 682-688.

Maeda,N. *et al*. (2006) Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs. *PLoS Genet.*, **2**, e62.

Mallory,A.C. and Vaucheret,H. (2004) MicroRNAs: something important between the genes. *Curr. Opin. Plant Biol.*, **7**, 120-125.

Mandal,M. and Breaker,R.R. (2004) Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.*, **5**, 451-463.

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.

Mattick,J.S. and Makunin,I.V. (2005) Small regulatory RNAs in mammals. *Hum. Mol. Genet.*, **14**, R121-R132.

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

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

Michael,M.Z. *et al*. (2003) Reduced Accumulation of Specific MicroRNAs in Colorectal Neoplasia. *Mol. Cancer Res.*, **1**, 882-891.

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.

Mishima,Y. *et al*. (2006) Differential Regulation of Germline mRNAs in Soma and Germ Cells by Zebrafish miR-430. *Curr. Biol.*, **16**, 2135-2142.

Missal,K. *et al*. (2006) Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. *J. Exp. Zoolog. B Mol Dev. Evol.*, **306**, 379-392.

Missal,K. *et al*. (2005) Non-coding RNAs in Ciona intestinalis. *Bioinformatics*, **21**, ii77-ii78.

Moss,E.G. *et al*. (1997) The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell*, **88**, 637-646.

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

Murchison,E.P. and Hannon,G.J. (2004) miRNAs on the move: miRNA biogenesis and the RNAi machinery. *Curr. Opin. Cell Biol.*, **16**, 223-229.

Nam,J.W. *et al*. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucl. Acids Res.*, **34**, W455-W458.

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

Ng,K.L.S. and Mishra,S.K. (2007a) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321-1330.

Ng,K.L.S. and Mishra,S.K. (2007b) Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification. *RNA*, **13**, 170-187.

Ng,P. *et al*. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Method*, **2**, 105-111.

Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, **29**, 11-17.

Ohler,U. *et al*. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, **10**, 1309-1322.

Olsen,P.H. and Ambros,V. (1999) The lin-4 Regulatory RNA Controls Developmental Timing in Caenorhabditis elegans by Blocking LIN-14 Protein Synthesis after the Initiation of Translation. *Dev. Biol.*, **216**, 671-680.

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.

Pedersen,J.S. *et al*. (2006) Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol.*, **2**, e33.

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

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

Pfeffer,S. *et al*. (2004) Identification of Virus-Encoded MicroRNAs. *Science*, **304**, 734-736.

Poy,M.N. *et al*. (2004) A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, **432**, 226-230.

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

Puerta-Fernandez,E. *et al*. (2003) Ribozymes: recent advances in the development of RNA tools. *FEMS Microbiol. Rev.*, **27**, 75-97.

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.

Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583-605.

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

Ro,S. *et al*. (2007) Cloning and expression profiling of testis-expressed microRNAs. *Dev. Biol.*, **311**, 592-602.

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

Samols,M.A. *et al*. (2005) Cloning and Identification of a MicroRNA Cluster within the Latency-Associated Region of Kaposi's Sarcoma-Associated Herpesvirus. *J. Virol.*, **79**, 9301-9305.

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

Schattner,P. (2002) Searching for RNA genes using base-composition statistics. *Nucl. Acids Res.*, **30**, 2076-2082.

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. *Nucl. Acids Res.*, **27**, 1578-1584.

Sempere,L. *et al*. (2004) Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol.*, **5**, R13.

Sewer,A. *et al*. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267.

Shiraki,T. *et al*. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA*, **100**, 15776-15781.

Sleutels,F. *et al*. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**, 810-813.

Smalheiser,N. (2003) EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. *Genome Biol.*, **4**, 403.

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

Soukup,J.K. and Soukup,G.A. (2004) Riboswitches exert genetic control through metabolite-induced conformational change. *Curr. Opin. Struct. Biol.*, **14**, 344-349.

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

Stern-Ginossar,N. *et al*. (2007) Host Immune System Gene Targeting by a Viral miRNA. *Science*, **317**, 376-381.

Stilgenbauer,S. *et al*. (1998) Expressed sequences as candidates for a novel tumor suppressor gene at band 13q14 in B-cell chronic lymphocytic leukemia and mantle cell lymphoma. *Oncogene*, **16**, 1891-1897.

Stormo,G.D. (2003) New tricks for an old dogma: riboswitches as cis-only regulatory systems. *Mol. Cell*, **11**, 1419-1420.

Storz,G. *et al*. (2005) An abundance of RNA regulators. *Annu. Rev. Biochem.*, **74**, 199-217.

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

Sudarsan,N. *et al*. (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, **9**, 644-647.

Suh,M.R. *et al*. (2004) Human embryonic stem cells express a unique set of microRNAs. *Dev. Biol.*, **270**, 488-498.

Sullivan,C.S. and Ganem,D. (2005) MicroRNAs and viral infection. *Mol. Cell.*, **20**, 3-7.

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.

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

Takada,S. *et al*. (2006) Mouse microRNA profiles determined with a new and sensitive cloning method. *Nucl. Acids Res.*, **34**, e115.

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

The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636-640.

Tijsterman,M. and Plasterk,R.H. (2004) Dicers at RISC; the mechanism of RNAi. *Cell*, **117**, 1-3.

Tinoco,J.I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271-281.

Uchida,D. *et al*. (2002) Oocyte apoptosis during the transition from ovary-like tissue to testes during sex differentiation of juvenile zebrafish. *J. Exp. Biol.*, **205**, 711-718.

Vapnik,V. (1998) Statistical learning theory. Wiley-Interscience.

Vitreschak,A.G. *et al*. (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.*, **20**, 44-50.

Vogel,J. *et al*. (2003) RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucl. Acids Res.*, **31**, 6435-6443.

von Hofsten,J. and Olsson,P.E. (2005) Zebrafish sex determination and differentiation: Involvement of FTZ-F1 genes. *Reprod. Biol. Endocrinol.*, **3**, 63.

Wallace,B.M.N. and Wallace,H. (2003) Synaptonemal complex karyotype of zebrafish. *Heredity*, **90**, 136-140.

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

Wang,X.G. and Orban,L. (2007) Anti-Mullerian hormone and 11 beta-hydroxylase show reciprocal expression to that of aromatase in the transforming gonad of zebrafish males. *Dev. Dyn.*, **236**, 1329-1338.

Washietl,S. and Hofacker,I.L. (2004) Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics. *J. Mol. Biol.*,

342, 19-30.

Washietl,S. *et al*. (2005a) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotech.*, **23**, 1383-1390.

Washietl,S. *et al*. (2005b) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**, 2454-2459.

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

Wienholds,E. *et al*. (2005) MicroRNA Expression in Zebrafish Embryonic Development. *Science*, **309**, 310-311.

Wienholds,E. *et al*. (2003) The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nat. Genet.*, **35**, 217-218.

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

Winkler,W.C. *et al*. (2001) The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs. *RNA*, **7**, 1165-1172.

Workman,C. and Krogh,A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids Res.*, **27**, 4816-4822.

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

Xia,T. *et al*. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochem.*, **37**, 14719-14735.

Xiao,C. *et al*. (2007) The XIST noncoding RNA functions independently of BRCA1 in X inactivation. *Cell*, **128**, 977-989.

Xie,X. *et al*. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338-345.

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.

Yamashita,A. *et al*. (1998) RNA-assisted nuclear transport of the meiotic regulator Mei2p in fission yeast. *Cell*, **95**, 115-123.

Yang,J.H. *et al*. (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucl. Acids Res.*, **34**, 5112-5123.

Yekta,S. *et al*. (2004) MicroRNA-Directed Cleavage of HOXB8 mRNA. *Science*, **304**, 594-596.

Ying,S.Y. and Lin,S.L. (2005) Intronic microRNAs. *Biochem. Biophy. Res. Comm.*, **326**, 515-520.

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

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

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

Zeng,Y. *et al*. (2003) MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc. Natl. Acad. Sci. USA*, **100**, 9779-9784.

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

Zhang,B.H. *et al*. (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res.*, **15**, 336-360.

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

Zikopoulos,B. *et al*. (2001) Cell genesis in the hypothalamus is associated to the sexual phase of a hermaphrodite teleost. *Neuroreport.*, **12**, 2477-2481.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9**, 133-148.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.*, **31**, 3406-3415.