miR-Explore: Predicting MicroRNA Precursors by Class Grouping and Secondary Structure Positional Alignment

Bram Sebastian¹ and Samuel E. Aggrey¹,²

¹Institute of Bioinformatics, University of Georgia, Athens, GA. ²NutriGenomics Laboratory, Department of Poultry Science, University of Georgia, Athens, GA. Corresponding author email: saggrey@uga.edu

Abstract: MicroRNAs (miRNAs) are small noncoding RNAs that regulate gene expressions by targeting the mRNAs especially in the 3′UTR regions. The identification of miRNAs has been done by biological experiment and computational prediction. The computational prediction approach has been done using two major methods: comparative and noncomparative. The comparative method is dependent on the conservation of the miRNA sequences and secondary structure. The noncomparative method, on the other hand, does not rely on conservation. We hypothesized that each miRNA class has its own unique set of features; therefore, grouping miRNA by classes before using them as training data will improve sensitivity and specificity. The average sensitivity was 88.62% for miR-Explore, which relies on within miRNA class alignment, and 70.82% for miR-abela, which relies on global alignment. Compared with global alignment, grouping miRNA by classes yields a better sensitivity with very high specificity for pre-miRNA prediction even when a simple positional based secondary and primary structure alignment are used.

Keywords: miR-explore, chicken, miRNA class alignment, miRNA
Introduction

Mature miRNAs are small 22 nucleotides long, non-coding RNAs. They are expressed in a wide variety of organisms including viruses, plants, and animals.\(^1\)-\(^3\) They have significant role in posttranscriptional control of eukaryotes genome causing degradation of the mRNA transcripts or blocking translation.\(^4\)\(^,\)\(^5\) In most eukaryotes genomes, miRNA genes are transcribed by RNA-polymerase-II to primary miRNA (pri-miRNA).\(^6\) This pri-miRNA is then further processed by RNase-III endonuclease Drosha into precursor miRNA (pre-miRNA), which is 60 to 100 nucleotides long, on average, and forms a stem-loop secondary structure.\(^7\) The pre-miRNA is then processed to form the small double stranded RNAs by the endonuclease Dicer, which also initiates the formation of the RNA-induced silencing complex (RISC).\(^8\) One strand of the double stranded RNA is incorporated with the RISC and targets the mRNA transcripts to block gene expression.\(^9\)

miRNAs have been involved in some critical diseases including heart disease\(^10\) and cancer.\(^11\) In cancer, miRNAs can function as both oncogenes and suppressors.\(^12\) Several classes of miRNAs such as miR-15, let-7, miR-342, miR-223, and miR-107 have been reported to be involved in acute promyelocytic leukemia (APL).\(^13\) The oncogene YES and STAT1, which are responsible in the proliferation of the colon cancer cells, are targeted by miR-145, thereby making this particular miRNA class a colon cancer suppressor.\(^14\) In lung cancer, miR-34 has been shown to have an important role on the PRIMA-1 regulation, which is a small molecule that restores the cancer cell suppression function.\(^15\) The identification of the involvement of miRNAs in several cancers has assisted researchers to develop some potential therapeutics for cancer cure.\(^16\) The significant role of miRNAs in human health has made the continuing discovery of novel miRNAs in the genome important.

Laboratory experiments have been conducted to discover miRNAs by direct cloning and short RNA sequencing. However, it is difficult to identify miRNAs with low levels of expression using only laboratory techniques.\(^17\) Hence, computational predictions have been needed to support the identification of novel miRNAs. There are two major computational approaches for predicting miRNAs: comparative and noncomparative methods. The comparative method relies on the conservation of miRNAs across the genome of organisms. Some examples of computational method using the comparative method are miRScan\(^2,\)\(^18\) miRAlign,\(^19\) ERPIN,\(^20\) and microHarvester.\(^21\) For miRNAs that do not have conserve sequences, non-comparative methods are needed. Some examples of the noncomparative method are triplet-SVM,\(^22\) miPred,\(^23\) PromiR,\(^24\) and miR-abela.\(^25\) The rationale behind the noncomparative methods is to design computer program to learn and distinguish between real pre-miRNA and pseudo pre-miRNA. Hence most of the noncomparative methods are based on the utilization of Support Vector Learning Machine (SVM) and hidden Markov model (HMM). Many of the miRNA prediction algorithms were developed based on broad assumptions.\(^26\) A study by Bram and Aggrey\(^27\) showed marked variability in sensitivity and specificity in predicting chicken pre-miRNA across classes using ERPIN, PromiR, and miR-abela.

In this study, we have developed a comparative method (miR-Explore) to demonstrate that grouping chicken miRNAs by classes increases sensitivity and specificity of the prediction method even when a simple direct positional secondary structure alignment is used. The basic idea of the current approach was to create a consensus structure of the pre-miRNA for each miRNA class and use this consensus structure to perform alignment with the query sequence.

Materials and Methods

Training data

A set of data was taken from the known pre-miRNA in chicken, human, mouse, zebra fish, fugu, worms, frog, chimpanzee, gorilla, platypus, pig, fruit fly, and buffalo, sequences which are available in the microRNA database miRBase.\(^9\) The data selected were based on the available miRNA classes in chicken. For example, if chicken has miR-1 class, then all available known miR-1 class pre-miRNA in the other organisms would be selected in this data set. The training data were taken from 80% of the sequences in the data set that best represents each miRNA class. In order to ascertain a good consensus structure, only classes that have 5 or more known pre-miRNAs were chosen to be part of the training data.

Positive data and negative data

The remaining 20% of the sequences that were not used as the training data were used as a query
Predicting miRNA precursors by class grouping

(positive data) for prediction. The negative data was taken from all coding sequences in human and mouse sequences from the University of California Santa Cruz (UCSC) genome browser. These coding sequences were scanned for hairpin-like structures as a final negative data set. There were 33,932 hairpin-like structures in mouse and 36,662 hairpin-like structures in human. The hairpin-like structures scanning was done by using the program written by Sewer et al for miR-abela. Negative data for comparison with PromiR were 5000 random sequences of length 300 nucleotides from coding genes, tRNA, and rRNA of chicken taken from the UCSC genome browser.

Programming and testing

The training data set was aligned using multiple alignment of RNAs (MARNA) to obtain the consensus structure with 80% primary and secondary structure identity. The identity percentage was defined as the number of the nucleotides that are conserved among all of the training sequences in a particular column of the secondary structure. These consensus structures along with their primary sequence information were then used for alignment with the query sequence. The query sequence was limited to a minimum length of 300 nucleotides. The minimum length of 300 nucleotides was to allow the program to initiate the alignment process. The average length of the consensus sequence was 150 nucleotides as miR-Explore cannot take a query sequence that is shorter than the length of the consensus structure. The alignment was done based on the information of the position of the secondary and primary structure. An alignment data from MARNA provided the consensus secondary structure on the first row along with the primary structure information in the rest of the rows (Fig. 1). A sliding window was created with a length similar to the number of characters including gaps and unpaired nucleotides in the output of the MARNA alignment. The sliding window was used to scan the query sequence starting from position 1. The consensus structures were made up of gaps and hairpin structures, which consist of stem and loop. In Figure 1, the stem starts in positions 1 and 75. The query sequence is a primary sequence without any gaps. Hence the alignment inserts gaps to the query sequence and shifts the gaps to align with the consensus structure and its corresponding nucleotide pairs in the stem. If an alignment matched the secondary structure and the corresponding stem nucleotides in the exact position shown in the MARNA alignment, then it was counted as one, otherwise it was zero. A match was defined as the same nucleotides in a particular column of a consensus secondary structure of the query sequence and the training sequence. The scoring method was a simple 0 and 1 to represent match or mismatch. The maximum score that an alignment can receive is the same as the number of pairs in the consensus secondary structure. Because each miRNA class has its own number of pairs in the secondary structure, the score of each miRNA class was different. This required normalization of the score to enable each miRNA class to attain a standardized score. The standardized score was calculated as standardized score = (NMA/NMC) × 100%, where NMA = number of matched secondary structure and nucleotide stems in the alignment and NMC = number of nucleotide pair in the stem of the consensus structure.

The same scoring system was used for both negative and positive data. Since we used 80% identity in the construction of the consensus structure, and this approach is a direct position to position exact comparison, the program was considered a hit when the standardized score from a query sequence was at least 80%. We compared the current approach (miR-Explore) with miR-abela and PromiR. miR-abela was tested using the same positive and negative data, while PromiR was tested using the negative data from 5000 randomly selected sequences of length 300 nucleotides from chicken coding genes.

![Figure 1. An example of alignment output from MARNA where '(' and ')') indicates the stem in the secondary structure, '.' is a mismatch and '—' is a gap.](image-url)
tRNA, and rRNA. The positive data for PromiR test were taken from the same data set used to test the other two programs, except that only chicken data were used. We tested PromiR with only chicken positive data because this program requires sequences to be inputted individually, which is laborious and time-consuming. Therefore, we tested PromiR against miR-Explore and miR-abela using only chicken data, and miR-Explore was tested against miR-abela using data set across species.

The sensitivity and specificity were calculated.

Sensitivity = TP/(TP + FN) where TP (true positive) is the number of pre-miRNA predicted as pre-miRNA, and FN (false negative) is the number of pre-miRNA predicted as non–pre-miRNA.

Specificity = TN/(TN + FP) where TN (true negative) is the number of non–pre-miRNA predicted as non–pre-miRNA, and FP (false positive) is the number of non–pre-miRNA predicted as pre-miRNA.

For accuracy measurement, we used Mathews Correlation Coefficient (MCC). MCC is a measure of the quality of binary classification. It takes into account true and false positives and negatives. The value of MCC ranges between −1 and +1 where +1 represents a perfect prediction, 0 represents a random prediction, and −1 represents completely inverted predictions. The formula to calculate the MCC is as follows, where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

Results and Discussion

The sensitivity and specificity for miR-Explore, miR-abela, and PromiR using only chicken data are shown in Table 1. The specificity for miR-Explore and miR-abela was 99.99%, and that of PromiR was 99.00%.

The specificity of these programs was high using chicken data. However, there was marked variability in their sensitivity. miR-Explore had 88% sensitivity, whereas miR-abela and PromiR had 78% and 53% sensitivity, respectively. The differences in sensitivity could be due to the fact that both miR-abela and PromiR were developed using mostly human and other mammalian training data whereas miR-Explore included chicken data as part of the training data. The training data in the secondary structure profiling can affect the sensitivity of a program that is dependent on the program. In principle, PromiR, a probabilistic colearning method based on a hidden Markov model, was developed to identify both close and distant homologs. Our results demonstrate that either the chicken is too distant from humans or some of generalized assumptions under which PromiR was developed do not hold. PromiR scans the stem of the stem-loop candidates to determine the signal of the Drosha cleavage site. However, multiple factors govern the Drosha cleavage site. Therefore, it is possible that the PromiR did not capture most of the factors affecting Drosha cleavage in chicken data thereby limiting its sensitivity. Whereas, miR-abela was developed based on Support Vector Machine (SVM), miR-Explore utilized both real pre-miRNA and pseudo pre-miRNA as the training data and used general features of pre-miRNA to train the computer to learn and distinguish between the two.

In general, there are some global features of pre-miRNAs that can be used as data for pre-miRNA prediction programs. These features include but are not limited to the nucleotide length, number of bulges, and minimum free energy. In the most recent miRNA prediction programs, these features have been generalized. The advantage of using these generalized miRNA features is that they are able to predict novel miRNAs that belong to previously unknown classes. Despite of the advantage of using the generalized miRNA features, the results shown in Table 1 suggest that using

---

### Table 1. Sensitivity and specificity of PromiR, miR-Explore and miR-abela in predicting pre-miRNA.

| Species* | PromiR | miR-Explore | miR-abela |
|----------|--------|-------------|-----------|
|          | GGA    | GGA HSA MMU DRE Others Average | GGA HSA MMU DRE Others Average |
| Sensitivity | 53.00 | 91.00 92.00 89.00 95.00 86.00 88.00 | 78.00 78.00 74.00 82.00 65.00 71.00 |
| Specificity | 90.00 | 99.99 99.99 99.99 99.99 99.99 99.99 | 99.99 99.99 99.99 99.99 99.99 99.99 |

Notes: *GGA = chicken; HSA = human; MMU = Mouse; DRE = zebra fish; Others = fugu, worms, frog, chimpanzee, gorilla, platypus, pig, drosophila and buffalo; *average was calculated as the ratio between the total number of predicted pre-miRNA and the total number of pre-miRNA in the test data.
Predicting miRNA precursors by class grouping

global features without adequate training data that can represent all organisms will have reduced sensitivity in identifying new miRNA for species that are not well represented in the training data.

The test results of comparing miR-Explore and miR-abela using an expanded set of positive and negative data are also shown in Table 1. The detailed predictive results of different organisms for miR-Explore and miR-abela are provided in Supplementary tables 1 and 2, respectively. The central hypothesis for this study was that each miRNA class has its own unique set of features; therefore, grouping miRNA by classes before using them as training data will improve sensitivity and specificity. Both miR-Explore and miR-abela are highly efficient in detecting true negative miRNAs. The sensitivity of miR-Explore was higher than miR-abela for every species compared including humans. The average sensitivity was 88.62% for miR-Explore and 70.82% for miR-abela. The calculation of MCC for miR-Explore yield a coefficient of 0.90 whereas miR-abela has a coefficient of 0.75. Compared with global alignment, grouping miRNA by classes yielded a better sensitivity with very high specificity for pre-miRNA prediction even when a simple positional based secondary and primary structure alignment were used. The grouping technique also yielded a higher MCC coefficient compared with the program using generalized features of pre-miRNA. It can be argued that as much as each class has its own unique features, there are other features that are conserved across classes and species. The ability to predict miRNA with increased sensitivity depends on the amount of conserved elements captured in the secondary structure. For example, the training data for miR-21 were from humans, chickens, and mice, yet miR-Explore had 100% sensitivity in predicting miR-21 in other species that were not used in the training data. The features of miR21 could have been conserved well across species. On the other hand, miR125 could be least conserved even within the class and across species, and any predictive algorithm based on global alignment would yield relatively poor sensitivity.

There are some limitations to miR-Explore. First, this approach is dependent on previously known miRNA precursor class and can only predict novel miRNA with that particular class. Second, miR-Explore relies on the conservation of miRNAs within a class and cross species and the availability of known miRNA data. When there are inadequate known miRNA data that can be used to build the consensus structure, the sensitivity may decline. Third, the alignment method of this approach is very simple. We used the positional information of the secondary structure along with the corresponding primary structure nucleotide, yet we were able to achieve a high sensitivity and specificity. Improving the alignment algorithm could improve the sensitivity and specificity.

Conclusion
It is difficult to develop a perfect miRNA prediction program; therefore, to detect miRNA computationally, more than one program may be needed. In this study, we have shown that grouping miRNA by classes before using them as training data will improve sensitivity and specificity. However, this approach can only predict pre-miRNA of known classes. Even though the sensitivity of PromiR and miR-abela may not be as good as that of miR-Explore, as ab initio methods, they have the potential to predict a novel miRNA in unknown classes. Each program has its own strengths and limitations that can complement each other. miR-Explore is a new technique that can contribute to the future discovery of novel miRNA.

Acknowledgement
We thank Dr. Alain Sewer from Université de Lausanne, Switzerland who has provided us with the program to scan for the hairpin-like structures in the human and mouse coding sequences.

Author Contributions
Conceived and designed the experiments: SEA. Analyzed the data: BS. Wrote the first draft of the manuscript: BS. Contributed to the writing of the manuscript: SEA, BS. Agree with manuscript results and conclusions: SEA, BS. Jointly developed the structure and arguments for the paper: SEA, BS. Made critical revisions and approved final version: SEA, BS. All authors reviewed and approved of the final manuscript.

Funding
Author(s) disclose no funding sources.

Competing Interests
Author(s) disclose no potential conflicts of interest.
Disclosures and Ethics
As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References
1. Bartel B, Bartel DP. MicroRNAs: At the root of plant development? Plant Physiol. 2003;132(2):709–17.
2. Lim LP, Glassner ME, Yekta S, Burge CB, Bartel DP. Vertebrate MicroRNA genes. Science. 2003;299(5612):1540–0.
3. Pfeffer S, Zavolan M, Grassner FA, et al. Identification of virus-encoded microRNAs. Science. 2004;304(5671):734–6.
4. Lau NC, Lim LP, Weinstein EG, Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science. 2001;294(5543):858–62.
5. Lee RC, Ambros V. An extensive class of small RNAs in Caenorhabditis elegans. Science. 2001;294(5543):862–4.
6. Lee Y, Kim M, Han JJ, et al. MicroRNA genes are transcribed by RNA polymerase II. Embo J. 2004;23(20):4051–60.
7. Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ. Processing of primary microRNAs by the Microprocessor complex. Nature. 2004;432(7014):231–5.
8. Bernstein E, Caudy AA, Hammond SM, Hannon GJ. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature. 2001;409(6818):363–6.
9. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34:D140–4.
10. Shenouda SK, Alahari SK. MicroRNA function in cancer: oncogene or a tumor suppressor? Cancer and Metastasis Rev. 2009;28(3–4):369–78.
11. Zhao HF, Wang DH, Du WT, Gu DS, Yang RC. MicroRNA and leukemia: Tiny molecule, great function. Crit Rev Oncol Hematol. 2010;74(3):149–55.
12. Gomelsky L, Jacobsen AB, Frankel LB, Wen J, Krogh A, Lund AH. MicroRNA-145 Targets YES and STAT1 in Colon Cancer Cells. PLoS One. 2010;5(1):e8836.
13. Duan WR, Gao L, Wu X, et al. MicroRNA-34a is an important component of PRIMA-1-induced apoptotic network in human lung cancer cells. Int J Cancer. 2010;127(2):313–20.
14. Gregersen LH, Jacobsen AB, Frankel LB, Wen J, Krogh A, Lund AH. MicroRNA-145 Targets YES and STAT1 in Colon Cancer Cells. PLoS One. 2010;5(1):e8836.
15. Wang XW, Zhang J, Gu J, et al. MicroRNA identification based on sequence and structure alignment. Bioinformatics. 2005;21(18):3610–4.
16. Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of Drosophila microRNA genes. Genome Biol. 2003;4(7):R42.
17. Li Y, Lau NC, Weinstein EG, et al. The microRNAs of Caenorhabditis elegans. Genes Dev. 2003;17(9):991–1008.
18. Wang XY, Ying XM, Li WJ. Computational approaches to microRNA precursors in animal genomes. Bioinformatics. 2006;22(3):359–60.
19. Xue CH, Li F, He T, Liu GP, Li YD, Zhang XG. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics. 2005;6:310.
20. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Research. 2007;35:W339–44.
21. Nam JW, Shin KR, Han JJ, Lee Y, Kim VN, Zhang BT. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research. 2005;33(11):3570–3581.
22. Wang XY, Ying XM, Li WJ. Computational approaches to microRNA precursors in animal genomes. Bioinformatics. 2006;22(3):359–60.
23. Sewer A, Paul N, Landgraf P, et al. Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics. 2005;6:267.
24. Zhang BH, Pan XP, Wang QL, Cobb GP, Anderson TA. Computational identification of microRNAs and their targets. Computational Biology and Chemistry. 2006;30(6):395–407.
25. Sebastian B, Aggrey SE. Specificity and sensitivity of PROMIR, ERPIN and MIR-ABELA in predicting pre-microRNAs in the chicken genome. In Silico Biology. 2008;8:377–81.
26. Kuhn RM, Karolchik D, Zweig AS, et al. The UCSC Genome Browser Database: update 2009. Nucleic Acids Research. 2009;37:D755–61.
27. Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics. 2005;21(16):3352–9.
28. Matthews BW. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochimica Et Biophysica Acta. 1975;402(2):442–51.
29. Zeng Y, Yi R, Cullen BR. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. Embo Journal. 2005;24(1):138–48.
30. Hou YY, Ying X-M, Li W-J. Computational approaches to microRNA discovery. Yichuan. 2008;30(6):687–96.
### Table S1. Sensitivity and specificity of miR-Explore in predicting pre-miRNA across species.

| Class     | Specificity | Sensitivity | Chicken | Human | Mouse | Zebra fish | Others | Average |
|-----------|-------------|-------------|---------|-------|-------|------------|--------|---------|
| let-7     | 99.75%      | 2,2         | 2,2     | 2,2   | 2,2   | 38.39      | 97.87% |
| miR1      | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 12.12      | 100.00%|
| miR7      | 100.00%     | 1,2         | 1,2     | 2,2   | 1,2   | 13.16      | 75.00% |
| miR9      | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 14.19      | 81.48% |
| miR10     | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 10.14      | 81.82% |
| miR15     | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 9.10       | 94.44% |
| miR16     | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 8.8        | 100.00%|
| miR17     | 100.00%     | 1,1         | 1,1     | 1,1   | 2,2   | 5.5        | 100.00%|
| miR18     | 100.00%     | 1,1         | 1,1     | 1,1   | 2,2   | 7.7        | 100.00%|
| miR19     | 99.96%      | 2,2         | 2,2     | 2,2   | 2,2   | 10.13      | 85.71% |
| miR20     | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 6.6        | 100.00%|
| miR21     | 100.00%     | 1,1         | 1,1     | 1,1   | 2,2   | 4.4        | 100.00%|
| miR22     | 100.00%     | 1,1         | 1,1     | 1,1   | 2,2   | 4.5        | 90.00% |
| miR23     | 100.00%     | 1,1         | 2,2     | 2,2   | 2,2   | 9.9        | 100.00%|
| miR24     | 100.00%     | 1,1         | 2,2     | 2,2   | 1,2   | 6.6        | 92.31% |
| miR26     | 100.00%     | 1,1         | 2,2     | 2,2   | 2,2   | 6.6        | 100.00%|
| miR27     | 100.00%     | 1,1         | 2,2     | 2,2   | 2,2   | 9.9        | 100.00%|
| miR29     | 99.98%      | 2,2         | 2,2     | 2,2   | 2,2   | 14.17      | 88.00% |
| miR30     | 99.99%      | 1,2         | 2,2     | 2,2   | 2,2   | 16.18      | 84.62% |
| miR31     | 99.99%      | 1,1         | 0,1     | 1,1   | 2,2   | 11.11      | 93.75% |
| miR32     | 100.00%     | 1,1         | 1,1     | N/A   | 2,2   | 3.3        | 100.00%|
| miR33     | 100.00%     | 2,2         | 1,1     | 1,1   | N/A   | 9.9        | 100.00%|
| miR34     | 100.00%     | 1,2         | 1,1     | 2,2   | 0,1   | 14.14      | 90.48% |
| miR92     | 99.81%      | 0,1         | 0,2     | 0,2   | 2,2   | 17.21      | 67.86% |
| miR99     | 100.00%     | 1,1         | 1,1     | 1,1   | 2,2   | 4.5        | 91.67% |
| miR100    | 100.00%     | 1,1         | 1,1     | 1,1   | 2,2   | 8.8        | 100.00%|
| miR101    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 7.8        | 93.75% |
| miR103    | 100.00%     | 2,2         | 2,2     | 2,2   | 1,1   | 7.7        | 100.00%|
| miR106    | 100.00%     | 1,1         | 2,2     | 2,2   | N/A   | 6.6        | 100.00%|
| miR107    | 100.00%     | 1,1         | 1,1     | 1,1   | 1,1   | 4.4        | 100.00%|
| miR122    | 100.00%     | 2,2         | 2,2     | 2,2   | 1,1   | 3.3        | 100.00%|
| miR124    | 99.99%      | 2,2         | 2,2     | 2,2   | 2,2   | 15.15      | 100.00%|
| miR125    | 99.97%      | 0,1         | 0,2     | 0,2   | 0,2   | 0,16       | 0.00%  |
| miR126    | 100.00%     | 1,1         | 1,1     | 1,1   | 1,1   | 3.3        | 100.00%|
| miR128    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 6.6        | 100.00%|
| miR130    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 7.7        | 100.00%|
| miR133    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 11.14      | 86.36% |
| miR135    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 9.10       | 94.44% |
| miR137    | 100.00%     | 1,1         | 1,1     | 1,1   | 2,2   | 5.6        | 90.91% |
| miR138    | 100.00%     | 1,1         | 2,2     | 1,2   | 1,1   | 1.4        | 54.55% |
| miR140    | 100.00%     | 1,1         | 1,1     | 1,1   | 1,1   | 3.3        | 100.00%|
| miR142    | 100.00%     | 1,1         | 1,1     | 1,1   | 1,1   | 4.4        | 100.00%|
| miR144    | 100.00%     | 1,1         | 1,1     | 1,1   | 1,1   | 3.3        | 100.00%|
| miR146    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 5.5        | 100.00%|
| miR147    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 3.3        | 100.00%|
| miR148    | 100.00%     | 1,1         | 2,2     | 2,2   | 1,1   | 4.4        | 100.00%|
| miR153    | 100.00%     | 1,1         | 2,2     | 1,1   | 2,2   | 6.6        | 100.00%|
| miR155    | 100.00%     | 0,1         | 0,1     | 0,1   | 0,1   | 0.2        | 0.00%  |
| miR181    | 100.00%     | 2,2         | 2,2     | 2,2   | 2,2   | 15.18      | 88.46% |
| miR183    | 100.00%     | 1,1         | 1,1     | 1,1   | 1,1   | 2.5        | 66.67% |

(Continued)
### Table S1. (Continued)

| Class     | Specificity | Sensitivity                  |       |       |       |       |       |       | Average |
|-----------|-------------|------------------------------|-------|-------|-------|-------|-------|-------|---------|
|           |             | Chicken                      | Human | Mouse | Zebra fish | Others |       |       |         |
| miR184    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   |       | 9.9   | 100.00%|
| miR187    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 3.3   |       | 100.00%|
| miR190    | 99.99%      | 1.1                          | 1.1   | 1.1   | 1.1   | 7.8   | 3.6   | 75.00% |
| miR193    | 100.00%     | 2.2                          | 1.1   | 1.1   | 2.2   |       | 5.5   | 100.00%|
| miR194    | 100.00%     | 1.1                          | 2.2   | 2.2   | 2.2   |       | 9.9   | 100.00%|
| miR196    | 100.00%     | 2.2                          | 2.2   | 1.1   | 2.2   |       | 8.9   | 94.12% |
| miR199    | 100.00%     | 2.2                          | 2.2   | 2.2   | 2.2   |       | 9.9   | 100.00%|
| miR200    | 100.00%     | 2.2                          | 2.2   | 1.1   | 2.2   |       | 9.9   | 100.00%|
| miR202    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 3.3   |       | 100.00%|
| miR203    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 3.3   |       | 100.00%|
| miR204    | 100.00%     | 2.2                          | 1.1   | 1.1   | 2.2   |       | 5.5   | 100.00%|
| miR205    | 100.00%     | 2.2                          | 1.1   | 1.1   | 1.1   | 4.4   |       | 100.00%|
| miR206    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 3.3   |       | 100.00%|
| miR211    | 100.00%     | 1.1                          | 1.1   | 1.1   | N/A   | 2.2   |       | 100.00%|
| miR214    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 4.4   |       | 100.00%|
| miR215    | 100.00%     | 1.1                          | 1.1   | 1.1   | N/A   | 3.3   |       | 100.00%|
| miR216    | 99.99%      | 2.2                          | 1.1   | 1.1   | 2.2   |       | 4.8   | 85.71% |
| miR217    | 100.00%     | 2.2                          | 1.1   | 1.1   | 1.1   | 4.4   |       | 100.00%|
| miR218    | 100.00%     | 2.2                          | 2.2   | 2.2   | 2.2   | 8.8   |       | 100.00%|
| miR219    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 8.10  |       | 85.71% |
| miR221    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 4.4   |       | 100.00%|
| miR222    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 3.3   |       | 100.00%|
| miR223    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 3.4   |       | 87.50% |
| miR301    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 5.5   |       | 100.00%|
| miR302    | 100.00%     | 0.2                          | 0.2   | 0.2   | N/A   | 0.8   |       | 0.00%  |
| miR365    | 100.00%     | 2.2                          | 2.2   | 2.2   | 2.2   | 3.5   |       | 84.62% |
| miR367    | 100.00%     | 1.1                          | 1.1   | 1.1   | N/A   | 2.2   |       | 100.00%|
| miR375    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 5.5   |       | 100.00%|
| miR383    | 100.00%     | 1.1                          | 1.1   | 1.1   | N/A   | 3.3   |       | 100.00%|
| miR429    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 3.5   |       | 77.78% |
| miR449    | 100.00%     | 0.1                          | 0.1   | 0.1   | N/A   | 0.4   |       | 0.00%  |
| miR451    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 2.2   |       | 100.00%|
| miR454    | 100.00%     | 1.1                          | N/A   | N/A   | 1.1   | 2.2   |       | 100.00%|
| miR455    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 2.2   |       | 100.00%|
| miR460    | 100.00%     | N/A                          | N/A   | N/A   | N/A   | 1.1   |       | 100.00%|
| miR466    | 100.00%     | 0.1                          | N/A   | 0.2   | N/A   | 1.4   |       | 14.29% |
| miR489    | 100.00%     | 1.1                          | 1.1   | 1.1   | 1.1   | 2.2   |       | 100.00%|
| miR490    | 100.00%     | N/A                          | N/A   | N/A   | N/A   | 2.2   |       | 100.00%|
| miR499    | 100.00%     | 1.1                          | 1.1   | 1.1   | N/A   | 3.3   |       | 100.00%|
| miR551    | 100.00%     | 1.1                          | 1.1   | 1.1   | N/A   | 3.3   |       | 100.00%|
| miR1306   | 100.00%     | N/A                          | N/A   | N/A   | N/A   | 1.1   |       | 100.00%|
Table S2. Sensitivity of miR-abela in predicting pre-miRNA across species.

| Class | Sensitivity miR-abela |
|-------|-----------------------|
|       | Chicken  | Human  | Mouse  | Zebra fish | Others | Average |
| let-7 | 2.2      | 0.2    | 1.2    | 2.2        | 26.39  | 65.96%  |
| miR1  | 2.2      | 2.2    | 1.2    | 2.2        | 10.12  | 85.00%  |
| miR7  | 2.2      | 2.2    | 2.2    | 2.2        | 14.16  | 91.67%  |
| miR9  | 2.2      | 2.2    | 2.2    | 2.2        | 16.19  | 88.89%  |
| miR10 | 2.2      | 2.2    | 1.2    | 2.2        | 11.14  | 81.82%  |
| miR15 | 1.2      | 1.2    | 0.2    | 2.2        | 3.9    | 41.18%  |
| miR16 | 2.2      | 2.2    | 2.2    | 1.2        | 7.8    | 87.50%  |
| miR17 | 0.1      | 0.1    | 0.1    | 0.2        | 1.5    | 10.00%  |
| miR18 | 2.2      | 2.2    | 2.2    | 2.2        | 8.13   | 76.19%  |
| miR19 | 2.2      | 2.2    | 2.2    | 2.2        | 10.12  | 85.00%  |
| miR20 | 1.2      | 1.2    | 1.2    | 1.2        | 3.7    | 61.54%  |
| miR21 | 1.1      | 1.1    | 1.1    | 2.2        | 3.4    | 88.89%  |
| miR22 | 1.1      | 1.1    | 1.1    | 2.2        | 3.5    | 80.00%  |
| miR23 | 1.1      | 2.2    | 2.2    | 2.2        | 5.9    | 75.00%  |
| miR24 | 0.1      | 1.2    | 1.2    | 2.2        | 1.6    | 38.46%  |
| miR25 | 1.1      | 2.2    | 2.2    | 2.2        | 3.6    | 100.00% |
| miR26 | 0.1      | 1.2    | 1.2    | 1.2        | 4.9    | 43.75%  |
| miR27 | 2.2      | 2.2    | 2.2    | 2.2        | 13.17  | 84.00%  |
| miR28 | 1.2      | 2.2    | 2.2    | 2.2        | 11.18  | 69.23%  |
| miR31 | 1.1      | 1.1    | 1.1    | 0.2        | 8.11   | 68.75%  |
| miR32 | 1.1      | 1.1    | 1.1    | 1.1        | N/A    | 83.33%  |
| miR33 | 2.2      | 1.1    | 1.1    | 1.1        | N/A    | 89.91%  |
| miR34 | 1.2      | 1.2    | 1.2    | 1.2        | 14.14  | 85.71%  |
| miR38 | 1.1      | 2.2    | 2.2    | 2.2        | 17.21  | 92.86%  |
| miR99 | 1.1      | 1.2    | 1.2    | 1.2        | 5.9    | 33.33%  |
| miR100| 0.1      | 0.1    | 0.1    | 2.2        | 5.8    | 53.85%  |
| miR101| 2.2      | 2.2    | 2.2    | 2.2        | 5.8    | 81.25%  |
| miR103| 2.2      | 1.2    | 2.2    | 1.1        | 1.7    | 50.00%  |
| miR106| 0.1      | 1.2    | 0.2    | N/A        | 4.6    | 45.45%  |
| miR107| 0.1      | 0.1    | 1.1    | 1.1        | 0.4    | 25.00%  |
| miR122| 2.2      | 1.1    | 1.1    | 1.1        | 2.3    | 87.50%  |
| miR124| 2.2      | 2.2    | 1.2    | 1.2        | 11.15  | 73.91%  |
| miR125| 1.1      | 2.2    | 1.2    | 2.2        | 6.16   | 52.17%  |
| miR126| 1.1      | 1.1    | 1.1    | 1.1        | 2.3    | 85.71%  |
| miR128| 2.2      | 2.2    | 2.2    | 2.2        | 3.6    | 78.57%  |
| miR130| 1.2      | 2.2    | 2.2    | 2.2        | 5.7    | 73.33%  |
| miR133| 2.2      | 2.2    | 2.2    | 2.2        | 9.14   | 77.27%  |
| miR135| 2.2      | 2.2    | 2.2    | 0.2        | 9.10   | 83.33%  |
| miR137| 1.1      | 1.1    | 1.1    | 2.2        | 5.6    | 90.91%  |
| miR138| 0.2      | 1.2    | 0.2    | 0.1        | 1.4    | 18.18%  |
| miR140| 1.1      | 1.1    | 1.1    | 1.1        | 3.3    | 100.00% |
| miR142| 1.1      | 1.1    | 0.1    | 1.1        | 3.4    | 75.00%  |
| miR144| 1.1      | 1.1    | 1.1    | 1.1        | 2.3    | 85.71%  |
| miR146| 2.2      | 1.2    | 2.2    | 2.2        | 4.5    | 84.62%  |
| miR147| 0.2      | 1.2    | 0.1    | N/A        | 2.3    | 37.50%  |
| miR148| 0.1      | 1.2    | 2.2    | 0.1        | 1.4    | 40.00%  |
| miR153| 1.1      | 2.2    | 1.1    | 2.2        | 5.6    | 91.67%  |
| miR155| 0.1      | 0.1    | 0.1    | 1.1        | 1.2    | 50.00%  |
| miR181| 1.2      | 1.2    | 1.2    | 1.2        | 5.18   | 34.62%  |
| miR183| 0.1      | 0.1    | 0.1    | 0.1        | 3.5    | 33.33%  |
| miR184| 1.1      | 1.1    | 1.1    | 1.1        | 7.9    | 84.62%  |
| miR187| 1.1      | 0.1    | 0.1    | 0.1        | 0.3    | 14.29%  |
| miR190| 1.1      | 1.1    | 1.1    | 1.1        | 6.7    | 90.91%  |

(Continued)
Table S2. (Continued)

| Class   | Sensitivity miR-abela |       |       |       |       |       |
|---------|-----------------------|-------|-------|-------|-------|-------|
|         |                      | Chicken | Human | Mouse | Zebra fish | Others | Average    |
| miR193  |                      | 1.2    | 1.1   | 0.1   | 2.2   | 4.6   | 66.67%     |
| miR194  |                      | 1.1    | 1.2   | 2.2   | 2.2   | 4.5   | 75.00%     |
| miR196  |                      | 2.2    | 2.2   | 2.2   | 2.2   | 6.9   | 100.00%    |
| miR199  |                      | 2.2    | 2.2   | 1.2   | 2.2   | 3.9   | 58.82%     |
| miR200  |                      | 2.2    | 2.2   | 2.2   | 2.2   | 7.9   | 88.24%     |
| miR202  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 3.3   | 100.00%    |
| miR203  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 3.3   | 100.00%    |
| miR204  |                      | 2.2    | 1.1   | 1.1   | 1.1   | 4.5   | 100.00%    |
| miR205  |                      | 2.2    | 1.1   | 1.1   | 1.1   | 3.4   | 100.00%    |
| miR206  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 2.3   | 85.71%     |
| miR211  |                      | 1.1    | 1.1   | 1.1   | N/A   | 2.2   | 100.00%    |
| miR214  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 4.4   | 100.00%    |
| miR215  |                      | 1.1    | 1.1   | 1.1   | N/A   | 3.3   | 100.00%    |
| miR216  |                      | 0.2    | 0.1   | 0.1   | 0.2   | 5.8   | 42.86%     |
| miR217  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 3.4   | 87.50%     |
| miR218  |                      | 2.2    | 2.2   | 2.2   | 2.2   | 5.8   | 81.25%     |
| miR219  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 5.10  | 64.29%     |
| miR221  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 3.4   | 87.50%     |
| miR222  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 3.3   | 100.00%    |
| miR223  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 4.4   | 100.00%    |
| miR301  |                      | 1.1    | 0.1   | 0.1   | 0.1   | 0.5   | 11.11%     |
| miR302  |                      | 1.2    | 2.2   | 1.2   | N/A   | 8.8   | 85.71%     |
| miR306  |                      | 1.2    | 2.2   | 1.2   | 1.2   | 3.5   | 61.54%     |
| miR367  |                      | 0.1    | 1.1   | 1.1   | N/A   | 0.2   | 40.00%     |
| miR375  |                      | 1.1    | 0.1   | 0.1   | 1.1   | 3.5   | 55.56%     |
| miR383  |                      | 0.1    | 0.1   | 0.1   | N/A   | 0.3   | 0.00%      |
| miR429  |                      | 1.1    | 0.1   | 1.1   | 1.1   | 4.5   | 77.78%     |
| miR449  |                      | 1.1    | 1.1   | 1.1   | N/A   | 2.4   | 71.43%     |
| miR451  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 2.2   | 100.00%    |
| miR454  |                      | 1.1    | N/A   | N/A   | 1.1   | 1.2   | 75.00%     |
| miR455  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 2.2   | 100.00%    |
| miR460  |                      | N/A    | N/A   | N/A   | 1.1   | 1.1   | 100.00%    |
| miR466  |                      | 0.1    | N/A   | 1.2   | N/A   | 4.4   | 71.43%     |
| miR489  |                      | 1.1    | 1.1   | 1.1   | 1.1   | 2.2   | 100.00%    |
| miR490  |                      | N/A    | N/A   | N/A   | N/A   | 0.2   | 0.00%      |
| miR499  |                      | 1.1    | 1.1   | N/A   | N/A   | 2.3   | 80.00%     |
| miR551  |                      | 1.1    | 1.1   | 1.1   | N/A   | 3.3   | 100.00%    |
| miR1306 |                      | N/A    | N/A   | N/A   | N/A   | 0.1   | 0.00%      |