SCIENTIFIC REP♦RTS

OPEN

miPIE: NGS-based Prediction of miRNA Using Integrated Evidence

R. J. Peace, M. Sheikh Hassani & J. R. Green

Methods for the de novo identification of microRNA (miRNA) have been developed using a range of sequence-based features. With the increasing availability of next generation sequencing (NGS) transcriptome data, there is a need for miRNA identification that integrates both NGS transcript expression-based patterns as well as advanced genomic sequence-based methods. While miRDeep2 does examine the predicted secondary structure of putative miRNA sequences, it does not leverage many of the sequence-based features used in state-of-the-art de novo methods. Meanwhile, other NGS-based methods, such as miRanalyzer, place an emphasis on sequence-based features without leveraging advanced expression-based features reflecting miRNA biosynthesis. This represents an opportunity to combine the strengths of NGS-based analysis with recent advances in de novo sequence-based miRNA prediction. We here develop a method, microRNA Prediction using Integrated Evidence (miPIE), which integrates both expression-based and sequence-based features to achieve significantly improved miRNA prediction performance. Feature selection identifies the 20 most discriminative features, 3 of which reflect strictly expression-based information. Evaluation using precision-recall curves, for six NGS data sets representing six diverse species, demonstrates substantial improvements in prediction performance compared to three methods: miRDeep2, miRanalyzer, and mirnovo. The individual contributions of expression-based and sequence-based features are also examined and we demonstrate that their combination is more effective than either alone.

MicroRNAs (miRNA) are short non-coding RNAs, typically 18–25 nts, which modulate post-transcriptional expression of messenger RNA (mRNA) transcripts[1]. As such, miRNA play a central role in cellular regulation. It has been estimated that 60–90% of all mammalian mRNAs may be targeted by miRNAs[2]. Through comparative expression analyses and gain- and loss-of-function experiments, it has been shown that miRNA regulate the expression of proteins involved in biological development[3], cell differentiation[4], apoptosis[5], cell cycle control[6], stress response[7], and disease pathogenesis[8]. Studies have also shown that miRNA play a role in cellular adaptation to severe environmental stresses such as freezing, dehydration and anoxia[9–11]. For these reasons, the ability to discover novel miRNA is of great importance.

It is believed that most miRNA share a similar biogenesis mechanism. RNA transcripts known as pri-miRNA contain one or more hairpin structures of approximately 70–120 nt in length, known as pre-miRNA. Endonucleases (Drosha and Dicer in animals; DCL1 in plants) process the pri-miRNA in order to form duplexes containing one or more mature miRNA sequences. Ultimately, mature miRNA are incorporated into the RNA-induced silencing complex (miRISC), where a miRNA guides the associated RISC proteins to a targeted mRNA strand. The RISC proteins anneal to the target mRNA and either promote degradation or repress translation of the mRNA[12].

Computational miRNA discovery techniques can be broadly categorized as either *de novo* miRNA prediction or expression-based (NGS-based) miRNA prediction[13]. In *de novo* prediction, putative pre-miRNA sequences believed to form miRNA-like hairpins are extracted from genomic data sets, and these sequences are classified based on the presence or absence of qualities such as structural stability, sequence motifs typical of miRNA, and structural robustness[14]. The advantage of *de novo* miRNA prediction is that only genomic sequence is required as input, not transcriptomic data. A disadvantage of such techniques is that they are ignorant of the actual expression of the candidate pre-miRNA region and must therefore consider a far greater number of putative miRNA which may never be expressed. Recent advances in *de novo* sequence-based miRNA prediction have been derived primarily through the application of new pattern classification techniques to the miRNA prediction problem[15], and the introduction of new classes of classification features[16]. However, high class imbalance (1:1000 or higher) within genomic data sets limits the effectiveness of *de novo* classifiers on actual datasets, in spite of high

Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada. Correspondence and requests for materials should be addressed to J.R.G. (email: jrgreen@sce.carleton.ca)

Data set	Organism	Accession	# Reads
hsa	H. sapiens	GSM1820470[41]	38210937
mmu	M. musculus	GSM1528810	54947527
dme	D. melanogaster	GSM1123781[42]	18723989
bta	B. Taurus	GSE74879[43]	43164654
gga	G. Gallus	GSM2095817	27937224
eca	E. Caballus	GSE100852	42178766

Table 1. NGS data sets examined in this article.

performance often reported on small test data sets with artificially balanced frequencies of positive and negative exemplars[17].

In expression-based prediction, data are collected from next-generation sequencing (NGS) experiments. These data represent the sequence and relative abundance of all expressed RNA in a sample, including RNA arising from microRNA (true positives) and other sources including mRNA degradation products and other ncRNA. Predictions of novel miRNA are made from NGS data by seeking patterns of read depth (proxy for transcript abundance in the cell) indicative of processing by Drosha and Dicer endonuclease activity[18]. These techniques also often examine the strength of the miRNA:miRNA* duplex corresponding to the mature miRNA and miRNA* regions within a putative pre-miRNA region[19]. Expression-based techniques for miRNA prediction have seen success in recent years[20–25], which can be explained in part by the lower class imbalance present in NGS data sets. The number of false positives in a typical NGS experiment is on the order of tens of thousands[18,26], whereas one expects tens of millions of miRNA-like structures in a typical genome[17]. Expression-based methods need only evaluate expressed regions, whereas *de novo* methods must evaluate all putative regions capable of forming hairpin structures. Furthermore, methods such as miRDeep2[27] often filter by transcript abundance, considering only the most highly expressed regions as a means to reduce their computational runtime.

Considering both *de novo* and expression-based miRNA prediction techniques, multiple categories of sequence- and expression-based features have been explored, where each may provide independent support for the prediction of miRNA within NGS data sets. State-of-the-art expression-based miRNA prediction techniques, however, only leverage a limited set of these lines of evidence. MiRDeep2[27] predicts miRNA based on the match between expression levels of NGS read data and expected Dicer processing, the stability of the miRNA:miRNA* duplex, and the significance of the minimum free energy of the pre-miRNA hairpin. MiRanalyzer[28] predicts miRNA based on secondary structure features, total read depth within the pre-miRNA region, expression of expected Dicer products, and minimum free energy features. These two methods have emerged as standards within the field of expression-based miRNA prediction[29].

Recently, Vitsios *et al*. have introduced mirnovo[30], a miRNA prediction method that combines both expression- and sequence-based features using a decision forest classifier. Since mirnovo can effectively predict miRNA without requiring a reference genome, it is particularly well-suited to newly sequenced and under-explored species lacking a high quality reference genome. While their method is conceptually similar to our approach, the primary emphasis of their study is the extraction of sequence-based features directly from NGS data without requiring a reference genome and not to complete a systematic development of a fully integrated feature set.

Although previous methods, such as miRDeep2 and miRanalyzer, have integrated a limited subset of expression- and sequence-based features, we here improve on the state-of-the-art performance of expression-based miRNA prediction by integrating the full range of sequence-based and expression-based features to create a novel miRNA predictor. We refer to this new method as miPIE (miRNA Prediction using Integrated Evidence). Our predictor is built using rigorous machine learning techniques, and tested using the metrics of recall and precision, which are directly applicable to real-world miRNA prediction. Additionally, unlike previous methods for expression-based miRNA prediction[27,28], all features used in our experiment are invariant to experiment size (total read depth). As NGS technology improves and read depths continue to increase, it is important that all features have this property in order for predictors to be effective on future data sets.

In summary, this paper makes the following contributions: (1) we collect an extensive set of both expression- and sequence-based features from the literature; (2) we apply feature selection to this set to create an integrated feature set; (3) we develop a decision forest to create miPIE, a novel miRNA predictor that leverages both sequence- and expression-based features; and (4) we use meaningful performance metrics of recall and precision to evaluate miPIE and demonstrate its superior performance relative to the state of the art.

Methods

Data set selection.
Sample data were selected from the NCBI GEO database, using a query consisting of the keywords "small RNA" and an organism name. Samples were selected for the following criteria: Extracted molecule is "total RNA", no infections or knockouts present in the cell, and size fractionation selection is for "small RNA". Samples GSE100852 and GSE74879 were collected using the Illumina HiSeq2500 instrument, sample GSM2095817 was collected using Illumina Genome Analyzer 2, sample GSM1901968 was collected using Illumina HiSeq1000 and all other samples were collected using the Illumina HiSeq2000 instrument. Table 1 describes the data sets that were retrieved for this experiment.

A flowchart illustrating the complete classification pipeline is included in Fig. 1. As shown in the top portion of the figure, we require five data sources to create training and test sets for each species: NGS expression data,
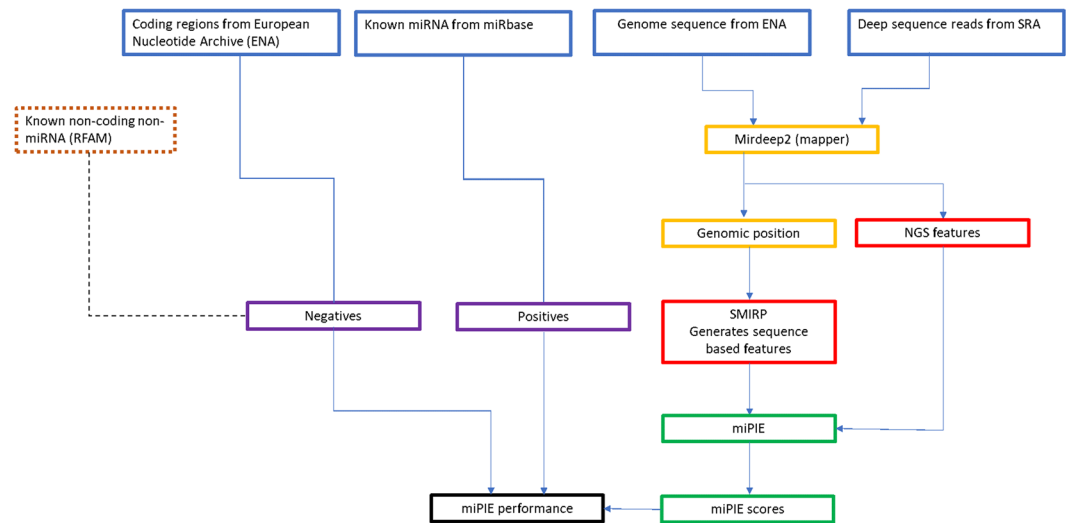
Figure 1. miPIE Pipeline flowchart.

Data set	#of positive genomic regions	#of negative genomic regions
hsa	167	609
mmu	384	872
dme	110	97
bta	341	683
gga	193	104
eca	364	228

Table 2. Number of samples in positive and negative classification data sets derived from each NGS experiment data set.

known miRNAs, genome data, known coding and known functional non-coding RNA. The remaining steps in Fig. 1 are discussed below.

Referring to the purple "Negatives" and "Positives" sets in Fig. 1, for each of the six species samples, the following procedure was performed in order to develop positive and negative training sets. The miRDeep2[27] pre-processing algorithm (as implemented in "mapper.pl") was applied to all data sets (see top-right of Fig. 1). This tool maps each read stack with at least four reads to the reference genome. Putative pre-miRNA regions are extracted (−10/+70 nt and −70/+10 nt windows based on locally maximal read stacks) and the secondary structure was computed to check for hairpin structures. Running the main script (mirDeep2.pl) resulted in a set of candidate pre-miRNA, each represented by a pre-miRNA sequence, pre-miRNA structure, and the set of reads which map to the sequence (mature, miRNA*, and loop regions). As detailed in Section 2.2, both genomic- and expression-based features are extracted for each candidate pre-miRNA region (see red boxes in Fig. 1).

For each sample, all candidate pre-miRNA which were matched to known high-confidence miRNA from miRBase 21.0[31] using the miRDeep2 "quantifier.pl" algorithm were selected as true positive samples for training and test. Each candidate mature miRNA not identified as miRNA in the previous step (labelled as predicted miRNA) was then aligned to the respective species' coding region data. Alignment was performed using bowtie[32]. All candidate mature miRNA which aligned to a coding region with at most two mismatches ("-v 2" bowtie parameter) were selected as negative samples for training and test. Coding region data was retrieved from the Ensembl sequence FTP database[33]. Table 2 lists the sizes of the final data sets used for this experiment.

Feature set selection. In this study, we examine a set of 223 sequence- and expression-based features. These features incorporate several distinct lines of evidence that have been shown to have predictive power for the classification of miRNA. Of these features, 215 are derived from the feature vector of the sequence-based method HeteroMiRPred[16], which in turn gathered these features from a number of methods dating back to 2005. These features all pertain to pre-miRNA sequence and structure and include [a] minimum free energy (MFE)-derived features, including z-features which encapsulate the significance of the RNA structure relative to those of permuted sequences; [b] sequence/structure triplet features and dinucleotide sequence motifs; and [c] structural robustness features which reflect the ability of the precursor structure to maintain its stability through addition or removal of nucleotides.

Eight expression-based features are added to these sequence-based features. These features are:

[1] Percentage of mature miRNA nts which are paired
[2] Number of pairs in lower stem (outside of mature and miRNA* regions)
[3] Percentage of RNA-seq reads in region which are inconsistent with Dicer processing
[4] Percentage of RNA-seq reads from the loop region which match Dicer processing
[5] Percentage of RNA-seq reads from the mature miRNA which match Dicer processing
[6] Percentage of RNA-seq reads from the miRNA* region which match Dicer processing
[7] Percentage of RNA-seq reads which match Dicer processing
[8] Total number of reads in the precursor region, normalized to experiment size

Here, a match between a read and expected Dicer processing is identical to the definition used by the miRDeep2 study[27]. A match occurs when a read which maps to a miRNA sequence overlaps the mature, miRNA*, or loop portions of a miRNA with at most 2-nt difference between starting positions on the 5′ ends of the sequences, and at most 5-nt difference between terminating positions on the 3′ ends of the sequences.

These expression-based features provide miRNA classification methods with additional independent lines of evidence for miRNA prediction which are not available from strictly sequence-based feature vectors. Features [1] and [2] provide information on the mature and lower stem regions of the miRNA, while features [3] through [8] provide information regarding the expression pattern within the miRNA region. While the number of expression-based features examined in this study is far less than the number of sequence-based features, it is greater than the number of expression-based features used by both the miRDeep2 scoring algorithm (total of 3 features: read count within mature region, miRNA* or loop regions; presence of miRNA* reads matching dicer processing; ratio of reads in the pre-miRNA region which are consistent with Dicer processing)[27] and the miRanalyzer scoring algorithm (total of 2 features: total read count; and expression of miRNA*)[28]. The remaining features used in these two methods actually pertain to the sequence and/or secondary structure of the putative pre-miRNA or homology to known miRNA. Of the sequence-based features described above, miRDeep2 limits its sequence-based features to those of type [a]. Conversely, miRanalyzer is largely focused more on sequence-based features (features of types [a], [b] and [c] are all used), while limiting its use of expression-based features. The mirnovo[30] method can operate strictly from NGS data without requiring a reference genome. In this case, the sequence-based features are focused on GC and AU content in different portions of the pre-miRNA (i.e. type [b] features). When provided with a reference genome, mirnovo employs a variety of sequence-based features covering all three types [a-c].

The final miPIE feature set was determined using the correlation-based feature subset selection method[34] as implemented in the Weka package[35]. This algorithm iteratively selects features through a greedy forward search to maximize correlation between each feature and the class assignment (miRNA vs. pseudo-miRNA), while minimizing inter-correlation between all selected features. The final feature vector is presented in section 3.1 of this article. Feature selection was performed on the mmu data set, and the resulting feature vector was subsequently employed across all data sets. As a result, performance results for the *mmu* data set represent some optimization using the test set data. However, this does not appear to have biased the *mmu* results, since performance is sustained across the other five species.

Classification pipeline.   A step-by-step tutorial describing how to implement our pipeline using sample data for the *dme* species is provided on the website (http://github.com/jrgreen7/miPIE). All miRNA classification in this experiment was performed using a random forest classifier of 500 trees. Trees were built according to the default parameters of the SKLearn random forest library[36]. Previous studies have demonstrated that random forest classification outperforms competing classifier types for the prediction of miRNA[17,37]. The miRanalyzer and mirnovo methods (see comparison in section 3.3 below) also employ random forest classifiers. Within individual data sets, 10-fold cross validation (10CV) was used to estimate classification performance. Within each fold, the SMOTE algorithm[38] was used to oversample the minority class of each training data set to parity with the majority class. Oversampling was performed only on training data sets; class imbalance within each test set was unchanged. When determining performance across data sets, a single classifier was trained using the training data set, and this classifier was used to predict all samples from the hold-out data set.

Performance metrics.   For each data set, the miRDeep2, miRanalyzer, and mirnovov prediction algorithms were run with default parameters over the datasets. Performance is measured by comparing each method's predictions to the known positive and negative data sets. Performance for all methods is measured using a precision-recall curve (PR-curve). Test set class imbalance is unaltered and represents that of real-world data, as each test set represents the total amount of positive and negative data recovered and processed from an actual NGS experiment. Precision is defined as $Pr = \frac{TP}{TP+FP}$ and recall as $Re = \frac{TP}{TP+FN}$.

Achievable recall at 75% precision (Re@Pr75) and recall at 90% precision (Re@Pr90) are used as summary statistics for all three methods. These statistics represent the recall rate achievable at an acceptable degree of precision for experimental validation (75%), and the percentage of miRNA which are correctly classified with very high confidence (90%), respectively. For the mirnovo[30] method, a full precision-recall curve is output as an image from which Re@Pr75 and Re@Pr90 were estimated. For miRDeep2, area under the precision-recall curve was also computed; however, this was not possible for miRanalyzer since it reports performance at a single decision threshold (representing a single point in the precision-recall space) nor for mirnovo since the precision-recall curve is provided as an image file.

To compute the statistical significance of the observed differences in performance between miPIE and the other three methods, a permutation test was conducted. For the six scores from the six species data sets, we computed the average observed difference. A distribution of expected average differences was then computed under

Data set	Re@Pr75			Re@Pr90		
	all	Seq	Exp	all	Seq	Exp
mmu	0.977	0.958	0.958	0.948	0.904	0.885
hsa	0.915	0.915	0.836	0.879	0.849	0.545
dme	0.955	0.982	0.954	0.891	0.918	0.864
bta	0.924	0.891	0.891	0.928	0.912	0.969
gga	0.993	0.953	0.995	0.798	0.757	0.648
eca	0.985	0.953	0.995	0.953	0.929	0.887
Average	0.958	0.942	0.938	0.900	0.878	0.800

Table 3. Comparison of performance of the integrated miPIE feature set, relative to the performance of similarly trained classifiers trained using only sequence- and only expression-based features.

every possible permutation of swapping the scores between the two methods within each dataset. By permuting the scores within each data set, we enforce the null hypothesis that there is no difference between the two methods. The percentage of randomized differences that were observed to be greater or equal to the actual observed differences in score between the two methods provides the p-value for the observed difference in scores.

Results and Discussion

Final feature set. The final feature set selected by the correlation feature subset algorithm contains 20 features, which represent seven different classes of evidence for the prediction of miRNA. The selected sequence-based features related to three types defined in section 2.2, including [a] MFE-derived features {MFE3, dH, Tm, Tm/loop}, [b] sequence/structure triplet features and dinucleotide sequence motifs {"C((.", "T.((", "T..(", "CG", "GA", {Probpair 2, 3, 7, 9, 19, and 94}}, and [c] structural robustness features {SC*absZG, SC/1dp}. Expression features [1] Percentage of mature miRNA nts which are paired, [5] Percentage of RNA-seq reads from the mature miRNA which match Dicer processing, and [6] Percentage of RNA-seq reads from the miRNA* region which match Dicer processing were selected from the eight original features enumerated in section 2.2. The fact that automated feature selection arrived at a heterogeneous feature set including both sequence- and expression-based feature supports our hypothesis that an integration of multiple lines of evidence will lead to increased classification performance. By only optimizing the feature set using one of our six datasets, we are implementing a highly conservative validation approach. Optimizing the feature subset over each species will likely lead to further improvements in performance. Descriptions of all features employed by the miPIE classifier are available in Supplemental Table 1.

Combining sequence- and expression-based features. In order to demonstrate our hypothesis that the predictive power of our method is a result of combining multiple lines of evidence for miRNA prediction, we repeated our 10CV classification pipeline using two subsets of features present in our full original feature set: i) the sequence feature subset containing 20 optimal features selected from the set of all sequence-based features available to our classifier; ii) the expression feature subset, containing the eight expression-based features examined in our study. These feature sets were used to train and test predictors for each of the six data sets described in the methods section. Results of these experiments are shown in Table 3, along with the performance of the miPIE classifier built using the integrated set (i.e. both sequence- and expression-based features). As demonstrated in Table 3, in almost all cases, the performance is improved when sequence and structure features are combined, relative to the use of either class of features alone. In a small number of cases, the combination of features, underperforms one of the individual feature sets alone. The reason for these discrepancies is unclear, however, taken as a whole, these results strongly suggest that the combination of sequence- and expression-based features is preferable to either individual feature set alone.

Performance increase over existing methods. Here we demonstrate the performance increase that our methods achieve over existing state-of-the-art methods for expression-based miRNA prediction. Our method is compared against the miRDeep2, miRanalyzer, and mirnovo methods, over the six data sets described in the methods section.

Figure 2 shows the performance of our method, miRDeep2, and miRanalyzer over the six data sets. While it was possible to create a continuous PR-curve for miRDeep2 by measuring the Pr and Re at various decision thresholds, this was not possible for miRanalyzer since it produces binary predictions without associated probability scores. Therefore, miRanalyzer's performance is illustrated as a single point in the P-R space, reflecting its performance at the default decision threshold. As can be seen in these figures, miPIE consistently outperforms both existing methods, particularly for decision thresholds corresponding to high precision (90%).

Table 4 summarizes the performance increase of miPIE over the miRDeep2 method. On average, our methods increase the number of high-confidence (Pr $\geq$ 90%) miRNA detected by 16%. This improvement is somewhat diminished to 4% at the more permissive decision threshold of Pr $\geq$ 75%; however, it is noted that the observed increase in Re@75 is limited by a saturation effect, as both methods are approaching perfect recall at this level of precision. Additionally, miPIE demonstrates an increased performance compared to miRDeep2 over all six species when looking at area under the precision-recall curve, yielding an average increase of 7.57%. When a paired permutation test for significant differences in means is applied to all three performance metrics, significant (p $=$ 0.016) differences are observed in all cases.
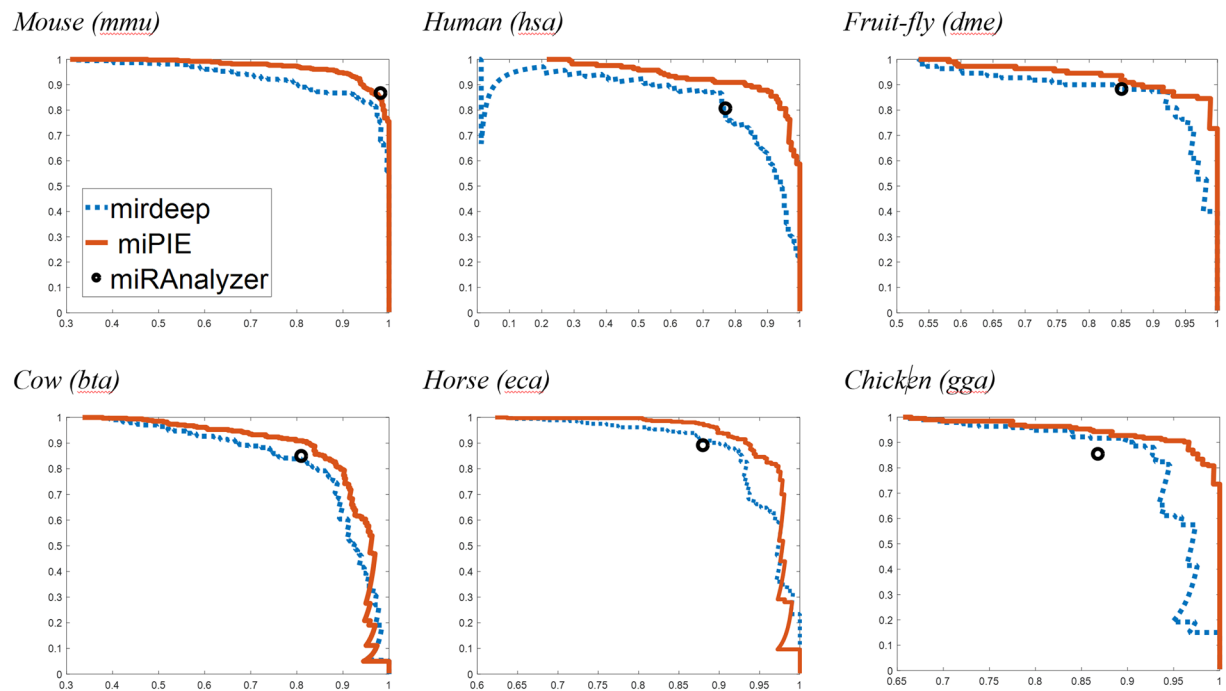
Figure 2. Performance of miPIE, miRDeep2, and miRanalyzer across six data sets. miPIE performance is estimated through 10-fold cross-validation. miRanalyzer produced binary prediction values, so only a single precision level is represented. miPIE outperforms miRDeep2 and miRanalyzer on all six data sets, with the possible exception of miRanalyzer's performance on *mmu*. In all plots, the y-axis represents precision while the x-axis is recall.

Data set	Re@Pr75 (p = 0.016)		Re@Pr90 (p = 0.016)		AUPRC (p = 0.016)	
	miRDeep2	miPIE	miRDeep2	miPIE	miRDeep2	miPIE
mmu	0.930	0.977 (+5.1%)	0.867	0.948 (+9.3%)	0.813	0.898 (+10.5%)
hsa	0.867	0.909 (+4.8%)	0.598	0.873 (+46.0%)	0.948	0.978 (+3.08%)
dme	0.909	0.955 (+5.1%)	0.873	0.891 (+2.1%)	0.931	0.960 (+3.11%)
bta	0.860	0.924 (+7.4%)	0.604	0.798 (+32.1%)	0.873	0.918 (+5.12%)
gga	0.964	0.985 (+2.2%)	0.902	0.927 (+2.8%)	0.940	0.976 (+3.80%)
eca	0.975	0.997 (+2.3%)	0.897	0.940 (+4.8%)	0.952	0.966 (+1.50%)
Average	0.918	0.958 (+4.0%)	0.790	0.896 (+16.0%)	0.882	0.949 (+7.57%)

Table 4. Summary of results comparing miPIE with the state of the art miRDeep2 method, on six NGS data sets. miPIE outperforms miRDeep2 by 16% and 4%, at the 90% and 75% precision thresholds, respectively, and by 7.57% when performance is measured using area under the precision-recall curve.

The original miRDeep2 training data included the three species human (*hsa*), mouse (*mmu*), and fruit-fly (*dme*). Considering this fact, we would have expected a higher performance difference in the three data sets not included in miRDeep2's training data. However, no consistent reduction in miRDeep2's performance was observed over these new species. The highest performance difference was observed in the human data set. It should be noted that known miRNA data from three of these species (*mouse, human, and fruit-fly*) were used to develop both miRDeep2 and miRanalyzer, while the other three (*bta, gga, and eca*) were not used by either.

In order to compare our method with the miRanalyzer method, which provides only binary classification results at a single decision threshold value, Table 5 describes the relative recall rates of miRanalyzer and miPIE at the precision level achieved by the miRanalyzer classifier on each data set. On average, our method increases recall rate by 6.90% relative to miRanalyzer predictions at the stated precision levels. The only dataset on which miRanalyzer performance approaches that of miPIE is mmu. This may be explained by apparent similarities between this dataset and that used to train miRanalyzer (NGS series GSE20384), as they both contain mouse testes samples.

As a final comparison with the state of the art, we here compare miPIE to mirnovo[30], a recently developed method that also employs an integrated feature set and decision forest classification models. For each species, the pre-trained species-specific model was used where possible, while the universal animal model was used for *gga* and *eca*. The method outputs a precision-recall curve as an image file and these curves are included in

Data set	Precision level	miRanalyzer recall rate	miPIE recall rate
mmu	0.982	0.866	0.841 (−2.9%)
hsa	0.770	0.806	0.909 (+12.8%)
dme	0.851	0.882	0.927 (+5.1%)
bta	0.810	0.849	0.910 (+7.2%)
gga	0.868	0.854	0.943 (+10.4%)
eca	0.880	0.891	0.970 (+8.9%)
Average	0.821	0.840	0.897 (+6.90%)

Table 5. Summary of results comparing miPIE with miRanalyzer using six NGS data sets. When operating at miRanalyzer's precision threshold, miPIE outperforms miRanalyzer by 6.9% on average *(p = 0.046)*.

Data set	Re@Pr75 *(p = 0.016)*		Re@Pr90 *(p = 0.016)*	
	mirnovo	miPIE	mirnovo	miPIE
mmu	0.78	0.977	0.70	0.948
hsa	0.51	0.909	0.37	0.873
dme	0.20	0.955	0.12	0.891
bta	0.76	0.924	0.61	0.798
gga	0.76	0.985	0.36	0.927
eca	0.62	0.997	0.08	0.940
Avg.	0.61	0.96	0.37	0.90

Table 6. Summary of results comparing miPIE with the mirnovo method on six NGS data sets.

Data set	Same-species 10CV	Cross-species Training Dataset						
		all	*mmu*	*hsa*	*dme*	*bta*	*gga*	*eca*
mmu	0.948	**0.930**	—	0.878	0.857	0.927	0.820	0.929
hsa	0.873	0.770	0.594	—	0.578	0.769	**0.794**	0.649
dme	0.891	0.782	0.764	0.854	—	0.781	**0.881**	0.809
bta	0.798	**0.700**	0.528	0.669	0.150	—	0.420	0.636
gga	0.927	**0.903**	0.586	0.901	0.891	0.876	—	0.633
eca	0.940	**0.939**	0.920	0.902	0.840	0.939	0.871	—

Table 7. Recall achievable at a precision of at least 90% (Re@Pr90) for 6 test datasets using our method trained over the following datasets: *10CV* = 10-fold-cross-validation within test species dataset; *all* = combination of 5 datasets, excluding test set; 10CV = 10-fold cross-validation over test dataset; *mmu* = mouse dataset; *hsa* = human dataset; *dme* = fruit-fly dataset; *bta* = cow dataset; *gga* = chicken dataset; *eca* = horse dataset.

Supplementary Fig. 2 for direct comparison with miPIE. From these curves, we estimated the Re@Pr75 and Re@Pr90 performance metrics (to 2 significant digits) and summarize them in Table 6 below. For all six species, miPIE consistently outperforms mirnovo on both performance metrics *(p = 0.016)*.

Generalization across experiments. Here we demonstrate the ability of our method to generalize across NGS data sets within and across species. For each of our six data sets, we compare the results of a 10CV experiment (where training and testing data arise from the same NGS experiment) with classification of the same data set using models trained on each of the other five data sets independently, as described in the methods section. Finally, for each hold-out data set, a training set was built using the combination of all other data sets (labeled *all*). Supplemental Fig. 1 shows the PR-curves for this experiment. Table 7 summarizes this performance for all six datasets. While some decrease in performance is observed when training data arises from a different experiment than the test data, the performance of classifiers trained using all experiments except the test experiment is consistently strong (curve labelled *all* in Supplementary Fig. 1).

Table 7 lists the Re@Pr90 for each combined training and test set for each training set. From these results, we see that miPIE generalizes well to hold-out experimental data sets when data sets from multiple NGS experiments are combined. Our method, when trained using all available data sets except the test species, outperforms miRDeep2 on five of six data sets (see Table 4 for miRDeep2 results). Average increase in Re@Pr90 between our method when trained in this manner and the miRDeep2 method is 6.0%. The combined *all* data sets perform as well or better than single experiment data sets in four of six experiments. Fruit-fly is the sole dataset that does not seem to benefit from multi-species pooling of training data. Interestingly, experiments within the same species are not necessarily preferred here (e.g. the top-performing training set for human is the chicken dataset). This result

demonstrates that the miPIE method generalizes well to hold-out data sets across experiments and across species when training data is pooled from multiple training experiments.

One caveat with the 10CV results presented in sections 3.2 and 3.3 is that the 10CV protocol may permit highly similar miRNA (e.g. from same family) to appear in both training and testing subsets. This could lead to overoptimistic evaluation metrics in both the present study and for previously reported methods. However, the cross-species results summarized in Table 7 appear to allay this concern, as performance is largely sustained when data from different species are used for training and testing of the method.

To demonstrate miPIE's ability to predict novel miRNA at high confidence, the final *gga* model was applied to the entire chicken genome. Using a decision threshold of 0.9 resulted in 71 high-confidence novel miRNA predictions (see Supplemental Table 2). Applying a conservative threshold of score levels 10 or 9 (out of a range [−10,10]) to miRDeep2 predictions over the *gga* genome resulted in 46 high-confidence novel miRNA predictions (see Supplemental Table 3). Of the 71 miPIE novel predictions, 27 were uniquely predicted by miPIE (see Supplementary Table 4), while only 2 of the 27 miRDeep2 novel predictions were unique (see Supplemental Table 5). The intersection of miPIE and miRDeep2 predictions results in 44 novel miRNA, each representing novel testable hypotheses (see Supplemental Table 6). A table of all the novel predictions made by miPIE for the 6 test species is presented in Supplemental Table 7.

Conclusions

In this study, we introduce miPIE, a classification method for NGS-based miRNA prediction that integrates both sequence- and expression-based features and employs a rigorous pattern classification approach. All features used by miPIE are independent of the read count of the NGS experiment, such that the performance of this method will remain consistent as NGS technology continues to develop. miPIE is compared with two existing state-of-the-art methods using the metrics of precision and recall, which are directly applicable to the end users of miRNA prediction software in that it answers the two questions: "Of the actual miRNA in my sample, what percentage will be identified" (recall) and "Of all the predicted novel miRNA, what percentage will correspond to actual miRNA?" (precision). At high precision levels (90%), miPIE increases recall by 16% relative to the popular miRDeep2 method. When examining all achievable precision levels, the area under the precision-recall curve is consistently higher for miPIE than for miRDeep2. Furthermore, miPIE increases recall by 6.9% relative to the miRanalyzer algorithm, at the precision levels reported by miRanalyzer. Lastly, miPIE outperforms a recent and conceptually-similar method, mirnovo, achieving higher achievable recall at both 90% and 75% precision levels. For all three comparisons, all performance differences were statistically significant (p < 0.05), as determined by a paired exact permutation test.

The negative data used in this study come from protein coding regions, since these regions are believed to exclude miRNA. To confirm that miPIE is effective for arbitrary pseudo-miRNA regions, we added all available non-coding RNA (tRNA, snoRNA, snRNA, rRNA) data from Rfam[39] to our negative data sets and re-ran the experiments. Performance was sustained over these broader negative data, confirming that miPIE has not simply learned to recognize protein-coding regions as being negative.

The primary avenue for future improvement of miPIE, and of expression-based miRNA prediction as a whole, is the development of strong training data sets that combine data from multiple NGS experiments. As evidenced in section 3.4 of this study, miPIE's generalization performance increases with the incorporation of multiple training data sets, a result which is consistent with that of the miRanalyzer experiment[28]. Additionally, increasing the quality of training data has proven successful in the field of *de novo* miRNA prediction[40]. With the ever-increasing availability of NGS data across myriad species, it will be feasible to create larger training data sets incorporating more species. Once training data has been curated from many species, approaches such as used in the species-specific miRNA pipeline[17] can be applied to NGS-based miRNA prediction data sets, thereby increasing prediction performance on non-model species. We anticipate that users will gain additional insight by running several methods. Combination strategies such as the union or intersection of the predicted miRNA will likely increase sensitivity or specificity, respectively. Furthermore, it is suggested that novel features be developed in both the expression and sequence space, and that feature selection be repeated periodically to incorporate the newest and most effective features from both fields. Lastly, miRDeep2's decision to examine only the most abundant transcripts will limit the overall recall of the method. With the increase in precision achievable using miPIE, it may be possible to examine a greater number of putative pre-miRNA, while still limiting the expected number of false positive predictions. Lastly, future studies will examine the possibility of applying multi-view co-training to miRNA prediction. By doing so, each of the expression and sequence-based feature sets can be used as independent views to create distinct views of the problem. Classifiers trained on each view can then be used to boost the performance of the other view.

Availability of Data and Material

Our method is available as an open source project at http://github.com/jrgreen7/miPIE. All datasets used in this study are available on NCBI-GEO (see Table 1 for accession numbers).

References

1. Humphreys, D. T., Westman, B. J., Martin, D. I. K. & Preiss, T. MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proc. Natl. Acad. Sci. USA* **102**, 16961–6 (2005).
2. Miranda, K. C. *et al*. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–17 (2006).
3. La Torre, A., Georgi, S. & Reh, T. A. Conserved microRNA pathway regulates developmental timing of retinal neurogenesis. *Proc. Natl. Acad. Sci. USA* **110**, E2362–70 (2013).
4. Le, M. T. N. *et al*. MicroRNA-125b promotes neuronal differentiation in human cells by repressing multiple targets. *Mol. Cell. Biol.* **29**, 5290–305 (2009).

5. Körner, C. *et al.* MicroRNA-31 sensitizes human breast cells to apoptosis by direct targeting of protein kinase C epsilon (PKCepsilon). *J. Biol. Chem.* **288**, 8750–61 (2013).
6. Iwasaki, Y. W. *et al.* Global microRNA elevation by inducible Exportin 5 regulates cell cycle entry. *RNA* **19**, 490–7 (2013).
7. Maistrovski, Y., Biggar, K. K. & Storey, K. B. HIF-1α regulation in mammalian hibernators: role of non-coding RNA in HIF-1α control during torpor in ground squirrels and bats. *J. Comp. Physiol. B.* **182**, 849–59 (2012).
8. Kowarsch, A., Marr, C., Schmidl, D., Ruepp, A. & Theis, F. J. Tissue-specific target analysis of disease-associated microRNAs in human signaling pathways. *PLoS One* **5**, e11154 (2010).
9. Biggar, K. K., Kornfeld, S. F., Maistrovski, Y. & Storey, K. B. MicroRNA regulation in extreme environments: differential expression of microRNAs in the intertidal snail Littorina littorea during extended periods of freezing and anoxia. *Genomics. Proteomics Bioinformatics* **10**, 302–9 (2012).
10. Biggar, K. K. & Storey, K. B. Evidence for cell cycle suppression and microRNA regulation of cyclin D1 during anoxia exposure in turtles. *Cell Cycle* **11**, 1705–13 (2012).
11. Wu, C.-W., Biggar, K. K. & Storey, K. B. Dehydration mediated microRNA response in the African clawed frog Xenopus laevis. *Gene* **529**, 269–75 (2013).
12. Bartel, D. P., Lee, R. & Feinbaum, R. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116**, 281–297 (2004).
13. Kleftogiannis, D. *et al.* Where we stand, where we are moving: Surveying computational techniques for identifying miRNA genes and uncovering their regulatory role. *J. Biomed. Inform.* **46**, 563–73 (2013).
14. Gomes, C. P. C. *et al.* A Review of Computational Tools in microRNA Discovery. *Front. Genet.* **4**, 81 (2013).
15. Gudyś, A., Szcześniak, M., Sikora, M. & Makalowska, I. HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics* **14**, 83 (2013).
16. Lertampaiporn, S., Thammarongtham, C., Nukoolkit, C., Kaewkamnerdpong, B. & Ruengjitchatchawalya, M. Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. *Nucleic Acids Res.* **41**, e21 (2013).
17. Peace, R. J., Biggar, K. K., Storey, K. B. & Green, J. R. A framework for improving microRNA prediction in non-human genomes. *Nucleic Acids Res.*, 10.109/nar/gkv698; See Sample Contribution 1 (2015).
18. Friedländer, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407–15 (2008).
19. Mapleson, D., Moxon, S., Dalmay, T. & Moulton, V. MirPlex: A Tool for Identifying miRNAs in High-Throughput sRNA Datasets Without aGenome. *J. Exp. Zool. Part B Mol. Dev. Evol.*, https://doi.org/10.1002/jez.b.22483 (2013).
20. Agharbaoui, Z. *et al.* An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress. *BMC Genomics* **16**, 339 (2015).
21. Yin, L. *et al.* Discovering novel microRNAs and age-related nonlinear changes in rat brains using deep sequencing. *Neurobiol. Aging* **36**, 1037–1044 (2015).
22. Cowled, C. *et al.* Characterisation of novel microRNAs in the Black flying fox (Pteropus alecto) by deep sequencing. *BMC Genomics* **15**, 682 (2014).
23. Stocks, M. B. *et al.* The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* **28**, 2059–61 (2012).
24. Paicu, C. *et al.* miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. *Bioinformatics* **33**, 2446–2454 (2017).
25. Zhang, H., Vieira Resende e Silva, B. & Cui, J. miRDis: a Web tool for endogenous and exogenous microRNA discovery based on deep-sequencing data analysis. *Brief. Bioinform.* bbw140, https://doi.org/10.1093/bib/bbw140 (2017).
26. Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J. M. & Aransay, A. M. miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.* **37** (2009).
27. Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).
28. Hackenberg, M., Rodríguez-Ezpeleta, N. & Aransay, A. M. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.* **39**, W132–8 (2011).
29. Williamson, V. *et al.* Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Brief. Bioinform.* **14**, 36–45 (2013).
30. Vitsios, D. M. *et al.* Mirnovo: Genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Res.* **45**, e177–e177 (2017).
31. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, 152–157 (2011).
32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
33. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43**, D662–669 (2014).
34. Hall, M. Correlation-based Feature Selection forMachine Learning. *Methodology* **21i195-i20**, 1–5 (1999).
35. Hall, M. *et al.* The WEKA data mining software. *ACM SIGKDD Explor. Newsl.* **11**, 10 (2009).
36. Pedregosa, F., Varoquaux, G., Weiss, R. & Brucher, M. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
37. Gudyś, A. *et al.* HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics* **14**, 83 (2013).
38. Chawla, N. & Bowyer, K. SMOTE: synthetic minority over-sampling technique. *J. Artificial Intell. Res.* **16**, 321–357 (2002).
39. Nawrocki, E. P. *et al.* Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
40. Wei, L. *et al.* Improved and Promising Identification of Human MicroRNAs by Incorporating a High-quality Negative Set. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1–12, 5C261F39-0F1D-4AE0-923F-CEF4DA8E06AE (2013).
41. Vongrad, V. *et al.* HIV-1 RNAs are Not Part of the Argonaute 2 Associated RNA Interference Pathway in Macrophages. *PLoS One* **10**, e0132127 (2015).
42. Shpiz, S., Ryazansky, S., Olovnikov, I., Abramov, Y. & Kalmykova, A. Euchromatic transposon insertions trigger production of novel Pi- and endo-siRNAs at the target sites in the drosophila germline. *PLoS Genet.* **10**, e1004138 (2014).
43. Navakanitworakul, R. *et al.* Characterization and Small RNA Content of Extracellular Vesicles in Follicular Fluid of Developing Bovine Antral Follicles. *Sci. Rep.* **6**, 25486 (2016).

Acknowledgements

Author Contributions

R.P. and J.G. conceived of the study. R.P. and M.S.H. developed the algorithms and conducted the experiments. R.P. wrote the initial draft and all authors reviewed and approved of the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-38107-z.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.