scientific reports

Check for updates

OPEN

P-TarPmiR accurately predicts plant-specific miRNA targets

Victoria Ajila[1], Laura Colley[1], Dave T. Ste-Croix[2], Nour Nissan[3,4], Ashkan Golshani[4], Elroy R. Cober[3], Benjamin Mimee[2], Bahram Samanfar[3,4] & James R. Green[1]✉

microRNAs (miRNAs) are small non-coding ribonucleic acids that post-transcriptionally regulate gene expression through the targeting of messenger RNA (mRNAs). Most miRNA target predictors have focused on animal species and prediction performance drops substantially when applied to plant species. Several rule-based miRNA target predictors have been developed in plant species, but they often fail to discover new miRNA targets with non-canonical miRNA–mRNA binding. Here, the recently published TarDB database of plant miRNA–mRNA data is leveraged to retrain the TarPmiR miRNA target predictor for application on plant species. Rigorous experiment design across four plant test species demonstrates that animal-trained predictors fail to sustain performance on plant species, and that the use of plant-specific training data improves accuracy depending on the quantity of plant training data used. Surprisingly, our results indicate that the complete exclusion of animal training data leads to the most accurate plant-specific miRNA target predictor indicating that animal-based data may detract from miRNA target prediction in plants. Our final plant-specific miRNA prediction method, dubbed P-TarPmiR, is freely available for use at http://ptarpmir.cu-bic.ca. The final P-TarPmiR method is used to predict targets for all miRNA within the soybean genome. Those ranked predictions, together with GO term enrichment, are shared with the research community.

microRNAs (miRNAs) are a class of short non-coding ribonucleic acids (RNAs) 20 to 24 nucleotides in length that achieve post-transcriptional gene expression regulation[1–3]. miRNAs are created through a multi-step process that includes the formation of pre-miRNA (precursor miRNA) sequences before the final processing step creating mature miRNA[2,4]. miRNAs regulate gene expression through the ribonucleoprotein complex (RISC)[3,4]. The miRNA-RISC complex binds to its target messenger RNA (mRNA) inducing its silencing or degradation through translational repression which can be coupled with mRNA decay and RISC-catalyzed endonucleolytic mRNA cleavage[2,4,5].

Biochemical assays including western blots, microarrays, next-generation sequencing, and quantitative polymerase chain reaction have been used to successfully determine miRNA targets at the gene level[6]. However, these techniques are unable to determine the exact binding site of miRNA within the mRNA[7]. Other experimental techniques, such as HITS-CLIP and PAR-CLIP can identify specific target sequences[8]. Cross-linking ligation and sequencing hybrids (CLASH) is a high throughput experimental approach that simultaneously identifies miRNA target sequences and the corresponding miRNA[9]. Cross-linking and immunoprecipitation (CLIP) is a high throughput profiling data technique that identifies transcript targets associated with functional RISC complexes; however, the discovery of miRNA target sites does not guarantee functional target suppression[10]. High-quality high-throughput data from wet-lab experimentation are essential for the improvement of computational miRNA target prediction methods[10]. Experimental observations have led to widely used miRNA target prediction rules that describe important features in a probable target[1].

Given the complexity and cost of experimental techniques, several computational methods for predicting miRNA targets have been developed. The computational methods can be organized into ab initio, machine learning, and hybrid methods[1]. The ab initio methods are designed to apply experimentally-derived rules that examine sequence complementarity in the seed region and other features including the accessibility of the target site, AU content, folding energy, conservation, perfect pairing of the miRNA 5′ end, and low GC content in the target site[1,8]. TargetScan and miRanda are the two most widely used ab initio miRNA target prediction tools[8]. MiRanda[9] uses an estimated sequence complementarity score, sequence conservation, and free energy values to

[1]Department of Systems and Computer Engineering, Carleton University, Ottawa K1S 5B6, Canada. [2]Saint-Jean-sur-Richelieu Research and Development Center, Agriculture and Agri-Food Canada, Saint-Jean-sur-Richelieu J3B 7B5, Canada. [3]Ottawa Research and Development Center, Agriculture and Agri-Food Canada, Ottawa K1A 0C6, Canada. [4]Department of Biology, Carleton University, Ottawa K1S 5B6, Canada. ✉email: jrgreen@sce.carleton.ca

predict target sites[1,8]. TargetScan[10] looks for perfect seed matches to form a candidate target list then uses site-type, local AU enrichment, and other features to calculate a target score for each candidate[1].

In both plants and animals, miRNAs regulate gene expression by controlling regulatory genes; however, there are many differences between the two kingdoms concerning miRNA biogenesis, miRNA–mRNA binding, and method of mRNA control[11]. Plant miRNA typically require much higher sequence complementarity than animals in the seed region[11]. Additionally, homology-based searches of similar miRNA–mRNA relationships in similar species are much more successful in plants than in animals[11]. Notably, the location of miRNA binding sites on mRNA is different between plants and animals. Animal miRNA bind in the 3′ UTR region of the mRNA and can exhibit multiplicity, where one mRNA can have many miRNA binding sites and one miRNA can target multiple mRNAs[4,11]. Whereas a plant miRNA binds to the target gene's open reading frame and there is typically only one binding site per mRNA[2,4,11].

Plant-specific ab initio miRNA target predictors have also been developed. psRNATarget[12] is a method that uses the RNAup algorithm[13] and a modified Smith–Waterman algorithm to find high-confidence miRNA targets[14]. Other plant-specific algorithms like Targetfinder[15], TAPIR[16], and Target-align[17] use the FASTA or Smith–Waterman algorithm accompanied with scoring methods to discover high confidence interactions[14]. The combination of psRNATarget and Targetfinder has resulted in improved performance[14]. Plant miRNA targeting was initially thought to be simple since high seed region complementarity is a requirement for effective gene silencing; however, deviations from the experimentally defined rules have been reported[18].

Traditionally, machine learning algorithms apply a classifier trained on features extracted from experimentally verified data to filter candidate predictions arising from ab initio algorithms[1]. Some of these machine learning methods include RFMirTarget[19] (a random forest classifier), MultiMiTar[20] (an SVM classifier), NBmiRTar[21] (a hybrid Naïve Bayes classifier), MiRAW[22] (deep learning), DeepMirTar[6] (stacked denoising autoencoders), and MiRDTL[23] (convolutional neural networks).

TarPmiR is a random-forest-based approach that integrates six conventional features with seven new features to predict miRNA target sites[8]. TarPmiR first extracts candidate target sites using miRanda[9] or other ab initio methods, then extracts 13 features for prediction[21,24]. These features include folding energy, seed match accessibility, AU content, stem conservation, flanking conservation, m/e motif, the total number of paired positions, the length of the target mRNA region, the length of the largest consecutive pairings, the position of the largest consecutive pairings relative to the 5′ end of miRNA, the number of paired positions at the miRNA 3′ end, the difference between the number of paired positions in the seed region and that in the miRNA 3′ end[8]. The algorithm was developed through a thorough validation and feature selection process which determined the best machine learning model and most important features[8]. TarPmiR performed better than both TargetScan and miRanda, two of the most commonly used miRNA target prediction tools when tested on two datasets from the human HEK293 cell line, a mouse dataset, and a general human dataset[8]. Their results also suggest that miRanda and TargetScan do not accurately predict non-seed-matching binding sites[9].

miRNA–mRNA interactions, predicted by the methods mentioned above and others, are aggregated in several databases, including EIMMo, DIANA-microT, Microcosm, Microrna.org, MirDB, PITA, TargetScan, miRWalk-predictive, and TargetSpy[1] all of which contain stricly animal interactions. Experimentally validated miRNA–mRNA pairs can be found in repositories such as miRWalk[25], miRecords[26], TarBase[27], miRTarBase[28], and starBase[29]. Of these databases, only TarBase and miRTarbase list plant interactions in addition to animal interactions.

Most machine learning miRNA–mRNA target prediction algorithms are based on training data derived from organisms in the *Animalia* kingdom. However, a small number of methods are amenable to fine-tuning or retraining using data from species closely related to the target species. Specifically, the TarPmiR method can be adapted to extract 11 of the 13 required features from custom datasets of miRNA–mRNA interactions with known binding sites. However, the retraining of machine learning models requires a significant amount of miRNA–mRNA interaction data, which has thus far been limited for most plant species. TarDB[30] is a newly released database containing tens of thousands of high-confidence plant miRNA–mRNA interactions. Although the database was originally created for biologists to use for manual homology-based target analysis, we here demonstrate that it can be used to retrain a machine learning method to create a highly accurate plant-specific miRNA target prediction pipeline.

In this study, TarPmiR, a state-of-the-art animal-based miRNA target predictor is modified and retrained for use on plants. TarDB, a new plant miRNA–mRNA database is used for the first time to train a miRNA target predictor. Negative miRNA–mRNA interaction examples are derived from positive miRNA–mRNA interaction examples to form comprehensive training and evaluation datasets. Rigorous experiment design is used to demonstrate that the inclusion of plant interaction data, and the complete exclusion of animal interaction data, significantly improves miRNA target prediction performance across four plant species. Our experiments also determine that a large amount of plant interaction data is required to significantly improve prediction performance. Our final method, dubbed P-TarPmiR, is available for use at http://ptarpmir.cu-bic.ca. The final predictor is applied to all miRNA in the soybean genome and ranked targets are shared with the research community at https://doi.org/10.5683/SP3/LOD4E3. GO term enrichment analysis is completed among all predicted gene targets for each miRNA, in an effort to elucidate the function of each miRNA.

Results

TarPmiR is a miRNA target predictor traditionally trained on the Human CLASH dataset. In this study, four classifiers with the same model architecture as TarPmiR but different training data were trained and tested in four different experiments. The experiments were designed to ascertain the effect of including plant interaction data in the training dataset on the plant miRNA target prediction performance of the classifier. Four classifiers

of varying proportions of plant interaction data were trained, including an animal-based classifier, a multi-kingdom classifier with minimal plant data, a multi-kingdom classifier significantly augmented with plant data, and a strictly plant-based classifier. Each experiment consisted of applying a classifier to the four different test sets. These test sets consisted of positive and negative miRNA–mRNA interaction examples from one of the four most represented organisms in the TarDB dataset: *Glycine max* (*gma*), *Oryza sativa* (*osa*), *Populus trichocarpa* (*ptc*), and *Brachypodium distachyon* (*bdi*).

Rigorous experiments were designed to ensure low sequence homology between the training data of the training sets and the test sets. This was used to simulate the case where an unannotated species is analyzed. For each plant species, it was ensured that none of the training data was similar to any known interaction in the test data. All miRNA included in the TarDB dataset were first clustered by CD-HIT using a sequence identity threshold of 70%.

The miRNA and mRNA sequence data of the interactions in the *H. sapiens*, *A. thaliana*, and TarDB datasets were retrieved. Some mRNA listed in the datasets could not be retrieved due to their removal from current genome annotations. The negative target sites were extracted from the retrievable mRNA and the features were extracted from all positive and negative interactions (see "Methods" section for further detail). The composition of the *H. sapiens* dataset, the *A. thaliana* dataset, and the TarDB dataset are listed in Table 1.

Four classifiers with the same model architecture as TarPmiR but differing training sets were trained and tested in four plant species. As described in the "Methods" section, training data from each experiment excluded miRNA–mRNA interactions similar to miRNA from the test species. In each experiment, the test dataset contained representative miRNA from one of the four target organisms (*gma*, *osa*, *ptc*, and *bdi*).

In the first experiment, the "Human" classifier was trained on the "Human" training set comprising only *H. sapiens* data. The "Human+ath" classifier was trained on the "Human+ath" dataset consisting of the *H. sapiens* and *A. thaliana* datasets. The four "Human+Plant" classifiers were trained on the H. sapiens dataset augmented with TarDB plant training data for each plant species. Finally, the "Plant" classifiers were trained using only plant training data from TarDB. The composition of the training sets and test sets for the four test species in the four experiments are listed in Table 2. Notably, there is a drastic decrease between the size of the TarDB dataset shown in Table 1 and the training datasets available for the "Plant" classifiers. TarDB contains many cross-species conserved miRNA targets, which results in a large reduction of the training dataset when all interactions involving a miRNA with 70% similarity or larger are removed[30]

Table 3 summarizes the performance of the four classifiers on the test sets in terms of area under the Precision–Recall curve (AUC), recall (Re), Precision (Pr), and accuracy (ACC), where the latter three metrics were evaluated at a confidence threshold of 0.5. Classifiers that included a large number of plant interactions (i.e., "Human+Plant" and "Plant") performed the best in terms of AUC, recall, precision, and accuracy. Fig. 1 compares the average of each performance metric of the four classifiers.

Figure 2 contains the Precision–Recall curves of the four classifiers over all experiments. As more plant interaction data are included in the training sets of the classifiers, the AUC of the Precision–Recall curve increases. Notably, the AUC on the *ptc* test set of all the experiments was greater than other test sets (except the plant-only case). ANOVA tests ($p < 0.05$) found that the AUC, precision, and accuracy of the classifier results listed in Table 2 were statistically significantly different. Post hoc paired t-tests showed that the performance of the "Human+Plant" and "Plant" classifiers was significantly different from the "Human" and "Human+ath" classifiers. However, the performance of the "Human" and "Human+ath" classifiers was not significantly different from each other. Conversely, the "Human+Plant" and "Plant" classifiers were significantly different from each other, except for the AUC and recall performance metrics. Figure 3. displays the density plots for the classifiers on the *gma* test set. Here, a wider separation between negative and positive scores is desirable. In line with the results

Training set	*H. sapiens* dataset	*A. thaliana* dataset	TarDB dataset
Positive	17,187	68	40,483
Negative	12,198	53	36,940

Table 1. The composition of the miRNA–mRNA interaction datasets used to develop model training sets, reporting the number of interacting (positive) and non-interacting (negative) miRNA:mRNA sequence pairs.

Training set	Human training set	Human+*ath* training dataset	Human+Plant training set				Plant training set				Test set			
			gma	*bdi*	*osa*	*ptc*	*gma*	*bdi*	*osa*	*ptc*	*gma*	*bdi*	*osa*	*ptc*
Positive	17,187	17,255 (68 from *ath*)	22,078 (4891 from TarDB)	25,285 (8098 from TarDB)	23,123 (5936 from TarDB)	23,098 (5911 from TarDB)	4891	8098	5936	5911	3939	1664	1237	2761
Negative	12,198	12,251 (53 from *ath*)	16,527 (4329 from TarDB)	19,572 (7374 from TarDB)	17,494 (5296 from TarDB)	17,638 (5440 from TarDB)	4329	7374	5296	5440	3635	1616	1174	2540

Table 2. The composition of the training sets and test sets used to train four classifiers (Human, Human+ath, Human+Plant, and Plant) for application on four test sets (*Glycine max*, *Oryza sativa*, *Populus trichocarpa*, and *Brachypodium distachyon*).

	Human				Human+ath				Human+Plant				Plant only			
Org./exp.	AUC	Re	Pr	Acc	AUC	Re	Pr	Acc	AUC	Re	Pr	Acc	AUC	Re	Pr	Acc
gma	0.934	0.992	0.697	0.772	0.937	0.991	0.695	0.769	0.997	0.998	0.788	0.859	1.000	0.986	0.995	0.990
bdi	0.934	0.996	0.663	0.741	0.939	0.997	0.660	0.738	0.998	0.999	0.744	0.825	0.999	0.996	0.993	0.994
osa	0.932	1.000	0.677	0.755	0.928	0.997	0.679	0.757	0.998	0.999	0.720	0.800	0.999	0.997	0.963	0.978
ptc	0.959	0.999	0.682	0.757	0.964	0.998	0.703	0.779	0.998	1.000	0.804	0.873	0.996	0.996	0.987	0.991
Average	0.939	0.996	0.680	0.756	0.942	0.996	0.684	0.761	0.998	0.999	0.764	0.839	0.999	0.993	0.984	0.988

Table 3. Performance of the four (Human, Human+ath, Human+Plant, and Plant) classifiers in each of the four test plant species (*Glycine max*, *Oryza sativa*, *Populus trichocarpa*, and *Brachypodium distachyon*).

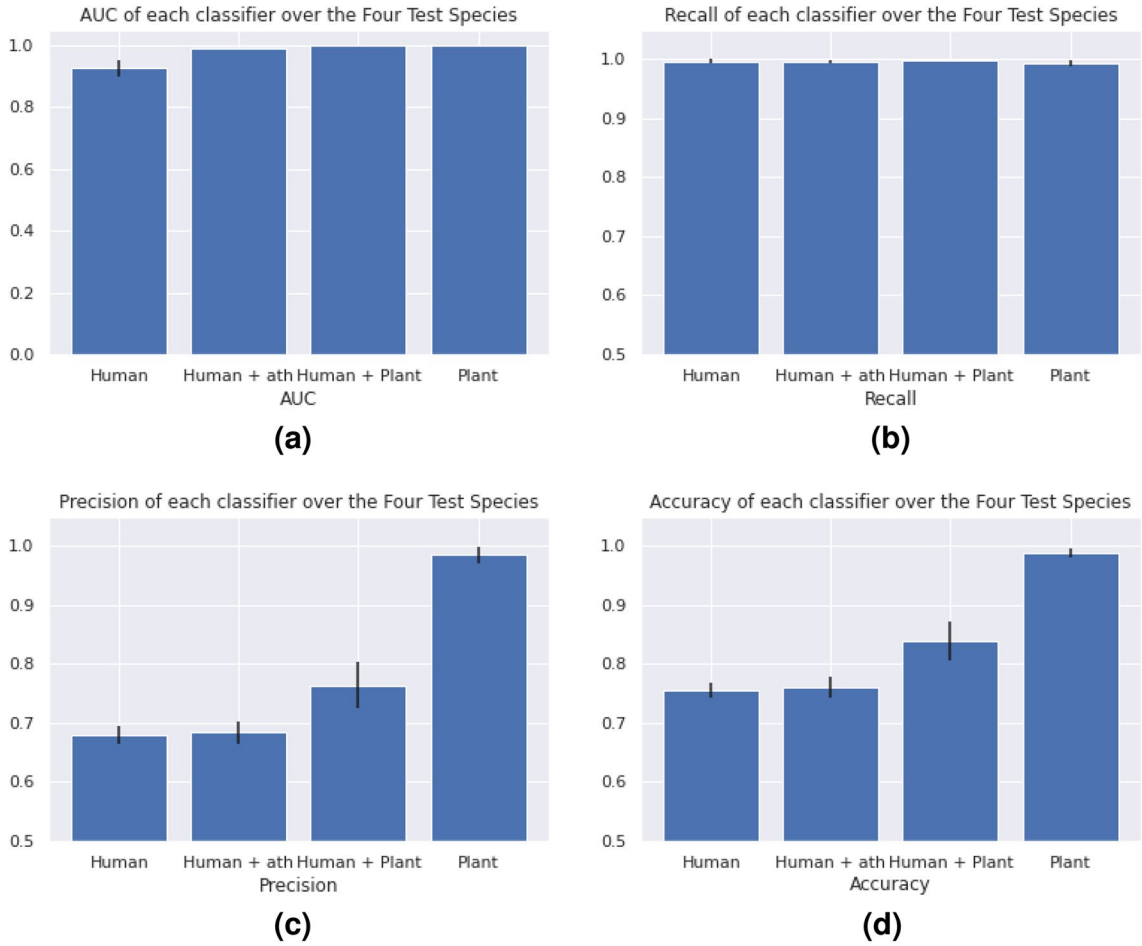


Figure 1. The average value and standard deviation of the performance metrics of each classifier (Human, Human+ath, Human+Plant, and Plant) over the four test species (*Glycine max*, *Oryza sativa*, *Populus trichocarpa*, and *Brachypodium distachyon*). The performance metrics included are area under the Precision–Recall curve (AUC), recall (Re), Precision (Pr), and accuracy (ACC).

in Table 3 and Fig. 2, as more plant interaction data are included in the training sets, the separation between the positive and negative test data increases.

Web server. A user-friendly web server was developed to perform miRNA target prediction using the P-TarPmiR algorithm. The web server allows users to upload miRNA and target files or copy text into the available text boxes for prediction (Fig. 4). Sequence length limits are imposed to limit the strain on the external remote compute resource. The typical run time for a submission of maximum size is 4 h. The time in the queue is dependent on the job load experienced by the remote compute resource.

Once the job is complete, prediction results are displayed and available for download in CSV format. In addition to the target prediction confidence, the results include the target region location and sequence (Fig. 5).

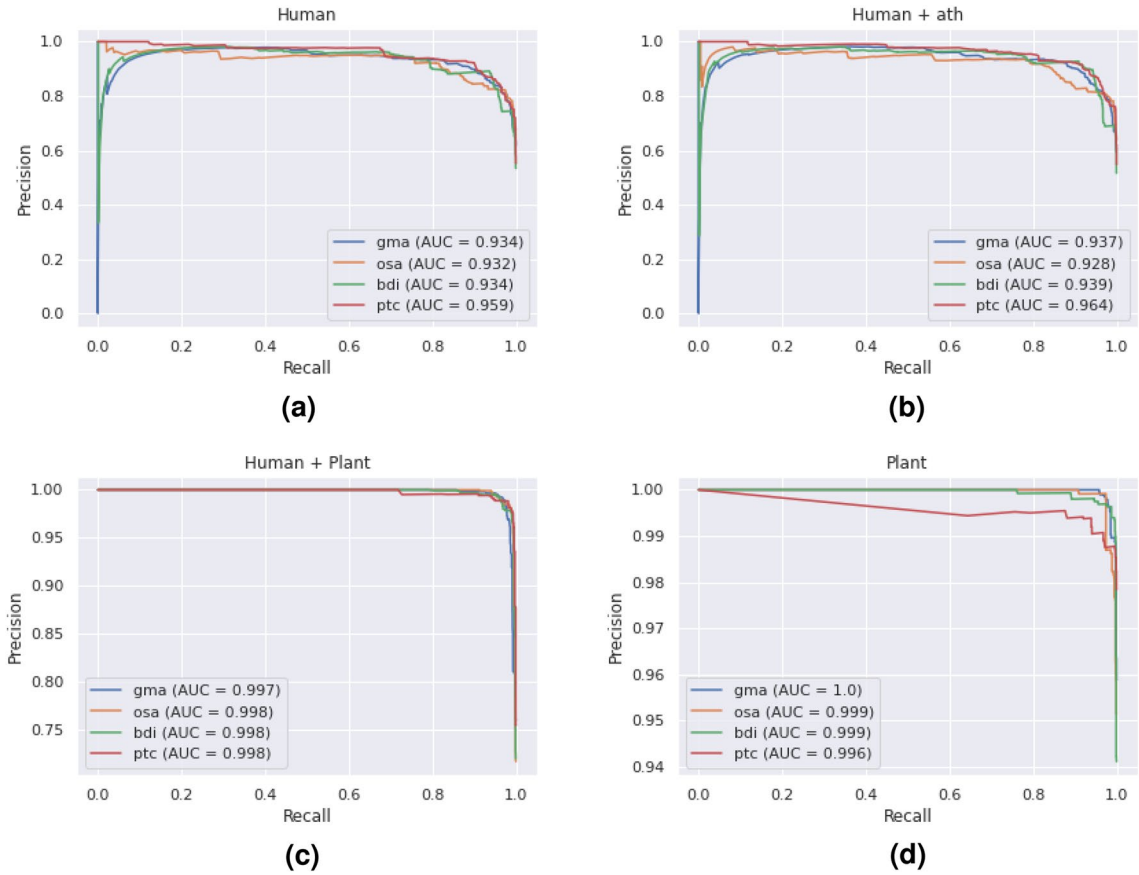


Figure 2. The Precision–Recall curves of each of the four classifiers (Human, Human+ath, Human+Plant, and Plant) for the four plant test species (*Glycine max*, *Oryza sativa*, *Populus trichocarpa*, and *Brachypodium distachyon*).

We have also made available the code and installation instructions such that users can install and run PTarPMir locally at https://github.com/GreenCUBIC/PTarPmiR.

P-TarPmiR applied to soybean genome. P-TarPmiR trained on all plant data described in Table 1 was applied to available soybean miRNA and mRNA. Following fivefold cross-validation across all known soybean interactions, a confidence threshold of 0.98 was selected, which represents a recall and precision of 0.90 ± 0.01 and 1.0 ± 0.00 respectively. This threshold was applied to the target predictions to arrive at a list of high-confidence interactions for each soybean miRNA. GO term enrichment was computed among the high-confidence mRNA targets predicted for each miRNA, using the PANTHER package[31,32]. This resulted in 3 tables of enriched GO terms for each miRNA, one for each annotation set (biological process, molecular function, and cellular component). All target predictions and their relative confidence levels and GO term enrichment results are available at https://doi.org/10.5683/SP3/LOD4E3

Discussion

We have leveraged a newly released large database of plant-specific miRNA–mRNA interactions to retrain a state-of-the-art predictor to achieve near-perfect performance over four plant test species. We have demonstrated that plant-specific predictors are more effective than cross-kingdom or multi-kingdom classifiers.

In our study, we demonstrated that the inclusion of plant interaction data in the training data resulted in a statistically significant difference in the performance of the classifier. "Human+Plant" and "Plant" classifiers performed significantly better than the two other classifiers that were not trained on a large amount of plant interaction data. Moreover, we demonstrated that the classifier trained strictly on plant interaction data can result in a consistent increase in performance over the "Human+Plant" classifiers even with the reduction in the size of the training dataset.

This study has shown that the inclusion of species-specific data increases the precision of the target predictions while maintaining a high recall. Over four test species, the "Human" classifier was applied to four different plant interaction datasets where the experimental design used here ensured that there was no significant sequence similarity between the training and test sets. Thus resulting in an average recall of 0.996 and average precision of 0.680. The "Human" classifier was able to predict high-confidence miRNA-mRNA interactions with high recall, but the precision was lower than the classifiers trained using plant-based data. These findings are in line with the

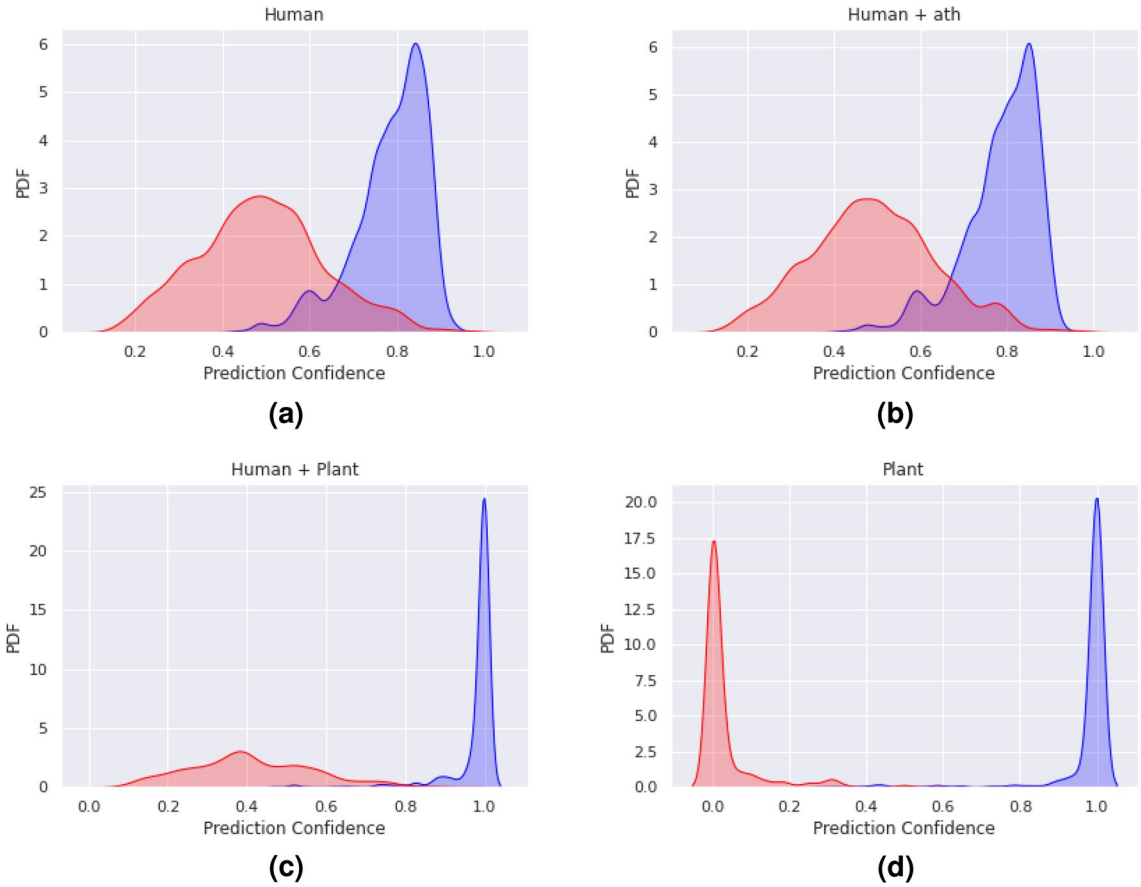


Figure 3. Density plots of the prediction scores of the four classifiers (Human, Human+ath, Human+Plant and Plant) on the *gma* test set. Here, prediction scores for negative test samples are shown in red, while positive test sample scores are shown in blue. A stronger classifier will lead to greater discriminability between the scores generated from positive and negative test miRNA:mRNA pairs.

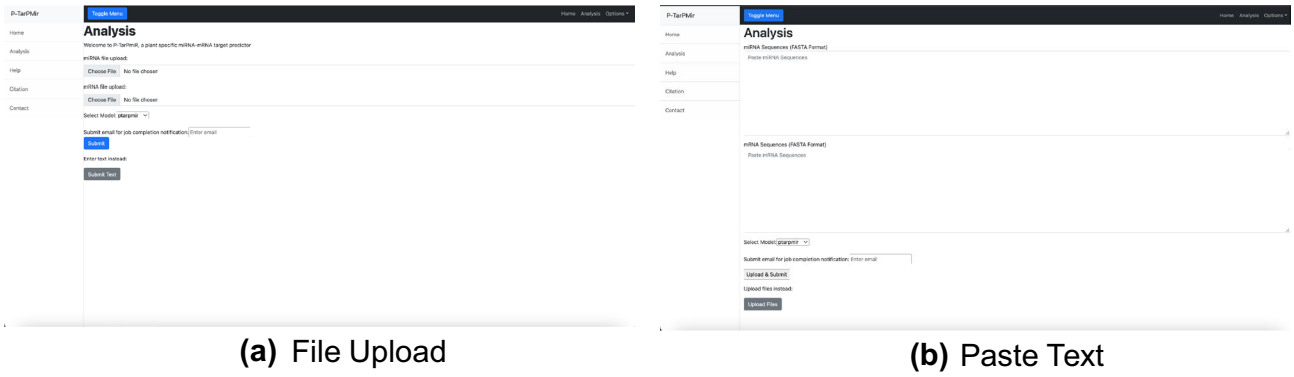
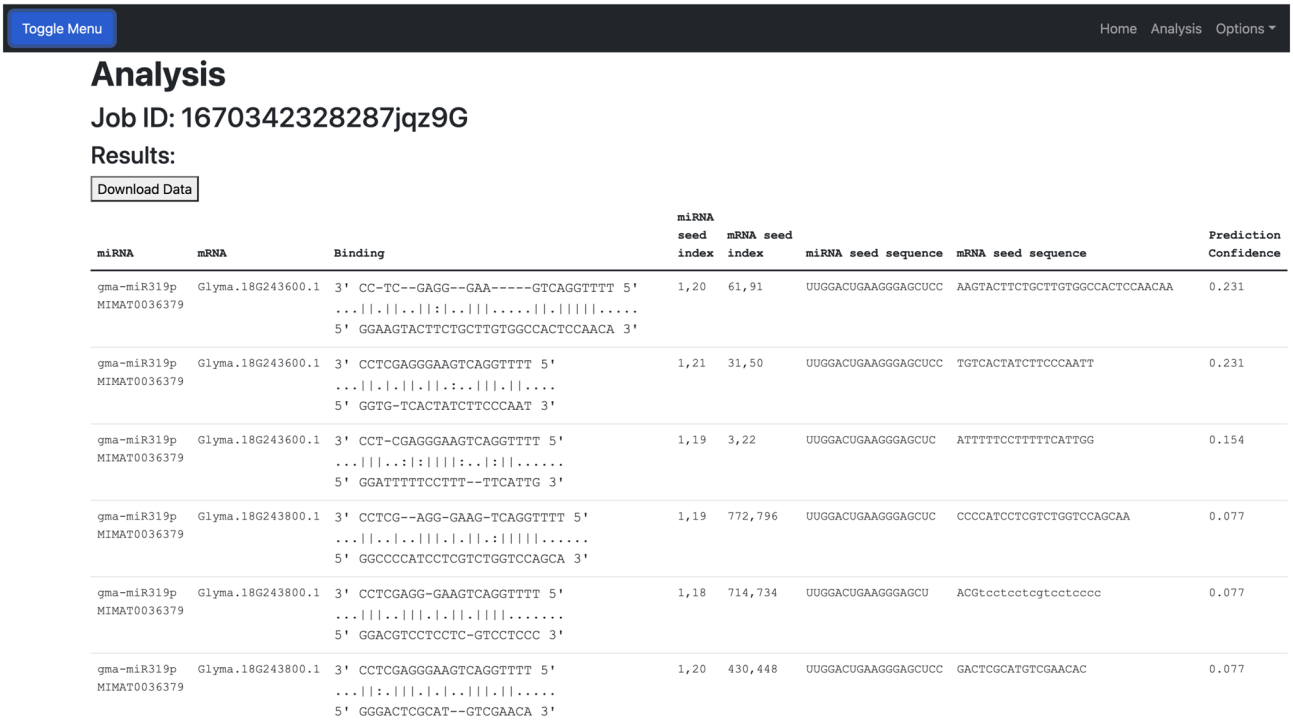


Figure 4. Screenshots of the P-TarPmiR web server job submission page, (**a**) for file upload and (**b**) for direct text input of sequences. Both types of submissions include the ability to add an email so the web server can notify the user when the job is complete. This functionality is particularly useful for large jobs.

original TarPmiR manuscript, where Ding et al. reported a significant drop in performance when independent test data were used to evaluate their method[8]. The lower precision observed in this study between the "Human" classifier and the "Human+Plant" classifier could be a measure of the generalizability of the "Human" classifier on a cross-kingdom case. In Ref.[33], we have previously demonstrated the value of species-specific training sets for miRNA discovery. The present study reinforces this finding but for the case of miRNA target prediction.

The training and test sets exhibited some class imbalance which would affect the precision, accuracy, and AUC scores; however, the recall performance metric would be unaffected by the imbalance. Along with the other metrics, the recall scores of the classifier increased between the animal-based classifier and the plant-based



Toggle Menu Home Analysis Options

Analysis

Job ID: 1670342328287jqz9G

Results:

Download Data

miRNA	mRNA	Binding	miRNA seed index	mRNA seed index	miRNA seed sequence	mRNA seed sequence	Prediction Confidence
gma-miR319p MIMAT0036379	Glyma.18G243600.1	3' CC-TC--GAGG--GAA-----GTCAGGTTTT 5' ...||.||..||:|..|||.....||.|||||..... 5' GGAAGTACTTCTGCTTGTGGCCACTCCAACA 3'	1,20	61,91	UUGGACUGAAGGGAGCUCC	AAGTACTTCTGCTTGTGGCCACTCCAACAA	0.231
gma-miR319p MIMAT0036379	Glyma.18G243600.1	3' CCTCGAGGGAAGTCAGGTTTT 5' ...||.|.||.||.:..|||.||.... 5' GGTG-TCACTATCTTCCCAAT 3'	1,21	31,50	UUGGACUGAAGGGAGCUCC	TGTCACTATCTTCCCAATT	0.231
gma-miR319p MIMAT0036379	Glyma.18G243600.1	3' CCT-CGAGGGAAGTCAGGTTTT 5' ...|||..:|:|||:..|:||..... 5' GGATTTTTCCTTT--TTCATTG 3'	1,19	3,22	UUGGACUGAAGGGAGCUC	ATTTTTCCTTTTTCATTGG	0.154
gma-miR319p MIMAT0036379	Glyma.18G243800.1	3' CCTCG--AGG-GAAG-TCAGGTTTT 5' ...||..|..|||.|.||.:|||||..... 5' GGCCCCATCCTCGTCTGGTCCAGCA 3'	1,19	772,796	UUGGACUGAAGGGAGCUC	CCCCATCCTCGTCTGGTCCAGCAA	0.077
gma-miR319p MIMAT0036379	Glyma.18G243800.1	3' CCTCGAGG-GAAGTCAGGTTTT 5' ...|||..|||.|.||.||||..... 5' GGACGTCCTCCTC-GTCCTCCC 3'	1,18	714,734	UUGGACUGAAGGGAGCU	ACGtcctcctcgtcctcccc	0.077
gma-miR319p MIMAT0036379	Glyma.18G243800.1	3' CCTCGAGGGAAGTCAGGTTTT 5' ...||:.|||.|.|..|||.||..... 5' GGGACTCGCAT--GTCGAACA 3'	1,20	430,448	UUGGACUGAAGGGAGCUCC	GACTCGCATGTCGAACAC	0.077

Figure 5. A screenshot of example results of a job submission on P-TarPmiR web server. The results page includes the predicted binding of the seed location, the index of the predicted seed location on the miRNA and mRNA, the miRNA and mRNA seed sequences, and the prediction confidence of the miRNA–mRNA pair.

classifier. When the methods are deployed to examine miRNA targets within a complete genome, we expect the class imbalance to be much higher since most miRNA–mRNA pairs will not represent actual interacting pairs.

psRNATarget[12], a commonly used plant-specific *ab initio* miRNA target predictor reported an average recall of 0.431, an average precision of 1.0, and an average accuracy of 0.732 on our test sets. Relative to psRNATarget, our "Plant only" P-TarPmiR predictor dramatically improved prediction recall (0.993) with minimal reduction in precision (0.984). Similarly, PTarPmiR outperforms TAPIR and Targetfinder in terms of recall. TAPIR and Targetfinder resulted in average recalls of 0.264 and 0.374 and average precision of 0.999 and 1.0 respectively on our test sets.

To explore the ability of our plant-specific PTarPMir model to recover atypical miRNA–mRNA interactions, a subset of miRNA–mRNA interactions was extracted from TarDB to test the performance of psRNATarget and PTarPmiR, specifically on non-canonical interactions. 371 interactions were determined to be atypical since they do not follow the definition of a canonical interaction[34]. Among this subset, psRNATarget resulted in a recall of only 0.108 at the default threshold of an expectation of 3, while PTarPmR resulted in a recall of 0.958 at a similar threshold. The high overall precision and low general recall highlight the fact that ab initio miRNA target predictors can only recover canonical interactions, as discussed by Dai et al.[12].

Conclusion

In this paper, we adapted TarPmiR, an animal miRNA target predictor, for use on plants. TarDB, a new plant miRNA–mRNA interaction database, was used to create plant-specific training sets for the miRNA target predictor. We demonstrated that an animal-based target predictor cannot adequately perform on plant data. We determined that a significant amount of plant interaction data could significantly improve the target predictor. Surprisingly, we discovered that a plant-only dataset consistently performed better than the multi-kingdom training sets. P-TarPmiR, the final plant-based miRNA target predictor, is available for use at ptarpmir.cu-bic.ca.

Future work will examine the use of a reciprocal perspective (RP) to improve plant-specific miRNA target predictions. Although RP, a cascaded semi-supervised machine learning method, was first developed to enhance protein–protein interaction prediction, it has shown great promise in other pairwise prediction tasks, including miRNA target prediction in animals[35].

Future work will also apply P-TarPmiR to soybean to discover miRNA that may play a role in early flowering and resistance to pathogens. miRNA-mediated gene regulation plays an important role in many animal and plant processes. Plants can alter their gene expression in response to stressors[36,37]. However, several studies have indicated that, in addition to intra-species gene regulation, miRNA can be transmitted between species and inhibit another species' gene expression[38–42]. Inter-species miRNA targeting has been reported in several plant–pathogen relationships where pathogen miRNA target host genes or host miRNA target pathogen genes[43–48].

In soybean, the differential expression of many miRNAs has been linked to the presence of the soybean cyst nematode (SCN)[37,49]. Additionally, the differential expression of exocyst genes in soybean has been tied to the facilitation or suppression of SCN parasitism in the plant[50]. Soybean is a major legume crop in North America,

resulting in billions of dollars of revenue[51]. SCN, a highly specialized plant-parasitic nematode, is a major pathogen of soybean worldwide, causing both significant yield and grain quality losses[51]. While no direct evidence has been identified thus far, the possibility of cross-kingdom interaction between SCN miRNAs and soybean mRNA was recently investigated using a predictor-based approach[52]. Future work will examine animal-based, plant-based, and multi-kingdom classifiers to determine which approach is most useful for cross-kingdom host-pathogen miRNA target prediction. More broadly, PTarPmiR could also be applied to other plant species to further elucidate plant gene regulation as it relates to all kinds of biological processes, such as development, yield maximization, and stress adaptations[36].

Methods

Data retrieval. All the plant miRNA–mRNA interactions listed in the TarDB database[30] were downloaded. TarDB is a recently released database of high-confidence plant miRNA–mRNA interactions, including binding site information, as determined by miRNA-triggered phasiRNA loci, cross-species conserved targets, and degradome/PARE (Parallel Analysis of RNA Ends)[30]. The miRNA and mRNA sequences of 42,692 interactions could be retrieved. Additionally, 70 of the miRNA and mRNA listed in the *Arabidopsis thaliana* miRNA–mRNA interactions from German et al. were also retrieved[53]. The *A. thaliana* data serves as the source of plant interaction data for the minimally augmented multi-kingdom classifier discussed further in "Discussion" section. Lastly, the miRNA and mRNA of 18,514 of the *Homo sapiens* miRNA–mRNA interactions listed in the Human CLASH dataset originally used to train TarPmiR were also retrieved miRBase and NCBI GenBank[54]. These three sources resulted in the TarDB, *A. thaliana*, and *H. sapiens* positive datasets, respectively.

Negative miRNA–mRNA interaction examples were created using the methodology described in Ding et al. to form the corresponding TarDB, *A. thaliana*, and *H. sapiens* negative sets[8]. In brief, the negative examples were selected by examining potential negative sites on an mRNA from a documented positive interaction that did not overlap with the positive site and had a similar CG dinucleotide frequency to the positive site. For each mRNA, the negative site exhibiting the lowest folding energy was used as the final negative exemplar[8].

Feature extraction. The TarPmiR[8] software package was modified to extract 11 features from the miRNA and target site pairs, including folding energy, seed match, accessibility, AU content, m/e motif, the total number of paired positions, length of the target mRNA region, length of the largest consecutive pairings, the position of the largest consecutive pairings relative to the 5′ end of the miRNA, the number of paired positions at the miRNA 3′ end and the difference between the number of paired positions in the seed region and in the miRNA 3′ end. TarPmiR utilizes the miRanda software[9] to find and extract features from seed regions on the miRNA and mRNA sequences.

Training set development and experimental set-up. For each test species, the following steps were used to create the training sets and the test sets, such that none of the training data shared significant sequence identity with any of the evaluation data from the test species. The plant training sets were used to train one of the four classifiers during the four experiments. The test set was used to determine the performance of each classifier. For each plant test species, the training set contained all examples from the TarDB dataset excluding those involving miRNA exhibiting greater than 70% sequence identity with any miRNA from the test species. All TarDB examples from the target organism formed the test set. Examples with miRNA sharing sequence identity with any examples from the *A. thaliana* dataset miRNA were also excluded from the test set. Test examples with duplicate features were also removed.

All classifiers used Random Forest models with 13 trees (from Ding et al.). The first experiment replicated the original TarPmiR classifier where the Random Forest classifier was trained on the Human CLASH positive and negative data. The second experiment consisted of training the model on a training set including the *H. sapiens* and the *A. thaliana* positive and negative sets. A third experiment trained the model on the *H. sapiens* dataset along with the plant positive and negative datasets for that experiment. The training sets in experiments two and three contain different proportions of plant interaction data. This will test how much plant data is required to perform well. In the fourth experiment, the model was trained on only the plant training sets in each experiment.

Web server. The miRNA Target prediction web server was developed using the Node.js Express framework. Common JavaScript libraries were used to develop a user-friendly interface. The web server runs with the support of Digital Research Alliance of Canada, a remote cloud-based compute resource that allows the submission of multiple concurrent jobs. The web server was containerized using Docker to ensure the portability and scalability of the web server. The web server is freely available for use at ptarpmir.cu-bic.ca

P-TarPmiR applied to soybean. A model trained on the entirety of the TarDB database was applied to all 756 mature miRNA available for soybean (*Glycine max*) in miRBase[55] and all 88647 transcripts available in version Wm82.a2 of soybean from Soybase[56]. The threshold to determine high-confidence interactions was determined using fivefold cross-validation of the soybean interactions extracted above. GO term enrichment analysis was applied to all high-confidence interactions predicted for each mature miRNA using PANTHER[31,32].

Data availability

All sequence and interaction data used to train and test the methods are available from public repositories (see details in the manuscript).

Received: 14 September 2022; Accepted: 29 December 2022
Published online: 06 January 2023

References

1. Tabas-Madrid, D. *et al.* Improving miRNA–mRNA interaction predictions. *BMC Genom.* **15**, 1–12 (2014).
2. O'Brien, J., Hayder, H., Zayed, Y. & Peng, C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol.* **9**, 402 (2018).
3. Jones-Rhoades, M. W. & Bartel, D. P. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* **14**, 787–799 (2004).
4. Shukla, G. C., Singh, J. & Barik, S. MicroRNAs: Processing, maturation, target recognition and regulatory functions. *Mol. Cell. Pharmacol.* **3**, 83 (2011).
5. Dai, X., Zhuang, Z. & Zhao, P. X. Computational analysis of mirna targets in plants: Current status and challenges. *Brief. Bioinform.* **12**, 115–121 (2011).
6. Wen, M., Cong, P., Zhang, Z., Lu, H. & Li, T. DeepMirTar: A deep-learning approach for predicting human miRNA targets. *Bioinformatics* **34**, 3781–3787 (2018).
7. Liu, W. & Wang, X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol.* **20**, 1–10 (2019).
8. Ding, J., Li, X. & Hu, H. TarPmiR: A new approach for microRNA target site prediction. *Bioinformatics* **32**, 2768–2775 (2016).
9. Enright, A. *et al.* MicroRNA targets in drosophila. *Genome Biol.* **4**, 1–27 (2003).
10. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mrnas. *Elife* **4**, e05005 (2015).
11. Millar, A. A. & Waterhouse, P. M. Plant and animal microRNAs: Similarities and differences. *Funct. Integr. Genom.* **5**, 129–135 (2005).
12. Dai, X., Zhuang, Z. & Zhao, P. X. psRNATarget: A plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* **46**, W49–W54 (2018).
13. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Chem. Mon.* **125**, 167–188 (1994).
14. Srivastava, P. K., Moturu, T. R., Pandey, P., Baldwin, I. T. & Pandey, S. P. A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *BMC Genom.* **15**, 1–15 (2014).
15. Fahlgren, N. *et al.* High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLoS ONE* **2**, e219 (2007).
16. Bonnet, E., He, Y., Billiau, K. & Van de Peer, Y. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* **26**, 1566–1568 (2010).
17. Xie, F. & Zhang, B. Target-align: A tool for plant microRNA target identification. *Bioinformatics* **26**, 3002–3003 (2010).
18. Li, Z., Xu, R. & Li, N. MicroRNAs from plants to animals, do they define a new messenger for communication? *Nutr. Metab.* **15**, 1–21 (2018).
19. Mendoza, M. R. *et al.* RFMirTarget: Predicting human microRNA target genes with a random forest classifier. *PLoS ONE* **8**, e70153 (2013).
20. Mitra, R. & Bandyopadhyay, S. MultiMiTar: A novel multi objective optimization based mirna-target prediction method. *PLoS ONE* **6**, e24583 (2011).
21. Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C. & Showe, M. K. Naïve bayes for microRNA target predictions-machine learning for microRNA targets. *Bioinformatics* **23**, 2987–2992 (2007).
22. Pla, A., Zhong, X. & Rayner, S. miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLoS Comput. Biol.* **14**, e1006185 (2018).
23. Cheng, S. *et al.* MiRTDL: A deep learning approach for miRNA target prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **13**, 1161–1169 (2015).
24. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol. Cell* **27**, 91–105 (2007).
25. Sticht, C., De La Torre, C., Parveen, A. & Gretz, N. miRWalk: An online resource for prediction of microRNA binding sites. *PLoS ONE* **13**, e0206239 (2018).
26. Xiao, F. *et al.* miRecords: An integrated resource for microRNA–target interactions. *Nucleic Acids Res.* **37**, D105–D110 (2009).
27. Sethupathy, P., Corda, B. & Hatzigeorgiou, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA* **12**, 192–197 (2006).
28. Huang, H.-Y. *et al.* miRTarBase update 2022: An informative resource for experimentally validated miRNA–target interactions. *Nucleic Acids Res.* **50**, D222–D230 (2022).
29. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H. & Yang, J.-H. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42**, D92–D97 (2014).
30. Liu, J. *et al.* TarDB: An online database for plant miRNA targets and mirna-triggered phased sirnas. *BMC Genom.* **22**, 1–12 (2021).
31. Mi, H. *et al.* Protocol update for large-scale genome and gene function analysis with the panther classification system (v. 14.0). *Nat. Protoc.* **14**, 703–721 (2019).
32. Thomas, P. D. *et al.* Panther: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022).
33. 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.* **43**, e138 (2015).
34. Seok, H., Ham, J., Jang, E.-S. & Chi, S. W. microRNA target recognition: Insights from transcriptome-wide non-canonical interactions. *Mol. Cells* **39**, 375 (2016).
35. Kyrollos, D. G., Reid, B., Dick, K. & Green, J. R. RPmirDIP: Reciprocal perspective improves miRNA targeting prediction. *Sci. Rep.* **10**, 1–13 (2020).
36. Ivashuta, S. *et al.* Regulation of gene expression in plants through miRNA inactivation. *PLoS ONE* **6**, e21330 (2011).
37. Li, X. *et al.* Identification of soybean microRNAs involved in soybean cyst nematode infection by deep sequencing. *PLoS ONE* **7**, e39650 (2012).
38. Mathur, M., Nair, A. & Kadoo, N. Plant–pathogen interactions: MicroRNA-mediated trans-kingdom gene regulation in fungi and their host plants. *Genomics* **112**, 3021–3035 (2020).
39. Zhou, G., Zhou, Y. & Chen, X. New insight into inter-kingdom communication: Horizontal transfer of mobile small RNAs. *Front. Microbiol.* **8**, 768 (2017).
40. Zeng, J. *et al.* Cross-kingdom small RNAs among animals, plants and microbes. *Cells* **8**, 371 (2019).
41. Chen, X., Liang, H., Zhang, J., Zen, K. & Zhang, C.-Y. Secreted microRNAs: A new form of intercellular communication. *Trends Cell Biol.* **22**, 125–132 (2012).
42. Liang, H., Zen, K., Zhang, J., Zhang, C.-Y. & Chen, X. New roles for microRNAs in cross-species communication. *RNA Biol.* **10**, 367–370 (2013).
43. Choy, E.Y.-W. *et al.* An Epstein–Barr virus-encoded microRNA targets puma to promote host cell survival. *J. Exp. Med.* **205**, 2551–2560 (2008).
44. Samols, M. A. *et al.* Identification of cellular genes targeted by kshv-encoded microrrnas. *PLoS Pathog.* **3**, e65 (2007).

45. Mayoral, J. G. *et al.* Wolbachia small noncoding RNAs and their role in cross-kingdom communications. *Proc. Natl. Acad. Sci.* **111**, 18721–18726 (2014).
46. Weiberg, A. *et al.* Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science* **342**, 118–123 (2013).
47. Wang, B. *et al. Puccinia striiformis* f. sp. tritici microRNA-like RNA 1 (pst-milr1), an important pathogenicity factor of pst, impairs wheat resistance to pst by suppressing the wheat pathogenesis-related 2 gene. *New Phytol.* **215**, 338–350 (2017).
48. Cui, C. *et al.* A fungal pathogen deploys a small silencing RNA that attenuates mosquito immunity and facilitates infection. *Nat. Commun.* **10**, 1–10 (2019).
49. Tian, B. *et al.* Genome-wide identification of soybean microRNA responsive to soybean cyst nematodes infection by deep sequencing. *BMC Genom.* **18**, 1–13 (2017).
50. Sharma, K. *et al.* Exocyst components promote an incompatible interaction between *Glycine max* (soybean) and *Heterodera glycines* (the soybean cyst nematode). *Sci. Rep.* **10**, 1–22 (2020).
51. Bradley, C. A. *et al.* Soybean yield loss estimates due to diseases in the United States and Ontario, Canada, from 2015 to 2019. In *Plant Health Progress, PHP–01* (2021).
52. Barnes, S. N. *Molecular Mechanisms Governing Plant Parasitic Nematode Signaling and Host Parasitism*. Ph.D. thesis, Iowa State University (2018).
53. German, M. A. *et al.* Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.* **26**, 941–946 (2008).
54. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by clash reveals frequent noncanonical binding. *Cell* **153**, 654–665 (2013).
55. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. mirbase: From microrna sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
56. Grant, D., Nelson, R. T., Cannon, S. B. & Shoemaker, R. C. Soybase, the usda-ars soybean genetics and genomics database. *Nucleic Acids Res.* **38**, D843–D846 (2010).

Author contributions

V.A. and J.R.G. conceived the experiments, B.S. identified the plant-specific interaction data, V.A. conducted the experiments, V.A and J.R.G. analysed the results. All authors provided input on the web server design and reviewed the manuscript.

Funding

The funding was provided by Natural Sciences and Engineering Research Council of Canada (RGPIN-2021-04184).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.R.G.

Reprints and permissions information is available at www.nature.com/reprints.

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