



Evaluation and application of tools for the identification of known microRNAs in plants

Qinglian Li^{1,2} , Guanqing Liu¹ , Yu Bao¹ , Yuechao Wu¹ , and Qi You^{1,3,4} 

Manuscript received 15 September 2020; revision accepted 7 February 2021.

¹Key Laboratory of Plant Functional Genomics of the Ministry of Education/Jiangsu Key Laboratory of Crop Genomics and Molecular Breeding/Co-Innovation Center for Modern Production Technology of Grain Crops, College of Agriculture, Yangzhou University, Yangzhou 225009, China

²Jiangsu Xuzhou Sweet Potato Research Center, Xuzhou 221131, China

³State Key Laboratory of Cotton Biology, Anyang, Henan 455000, China

⁴Author for correspondence: youqi@yzu.edu.cn

Citation: Li, Q., G. Liu, Y. Bao, Y. Wu, and Q. You. 2021. Evaluation and application of tools for the identification of known microRNAs in plants. *Applications in Plant Sciences* 9(3): e11414.

doi:10.1002/aps3.11414

MicroRNAs (miRNAs), endogenous non-coding RNA regulators, post-transcriptionally inhibit the expression of their target genes. Several tools have been developed for predicting annotated known miRNAs, but there is no consensus about how to select the most suitable method for any given species. In this study, eight miRNA prediction tools (mirnova, miRPlant, miRDeep-P2, miRExpress, miRkwood, miRDeep2, miR-PREFeR, and sRNAbench) were selected for evaluation. High-throughput small RNA sequencing data from four plant species (including C₃ and C₄ species, and both monocots and dicots, i.e., *Arabidopsis thaliana*, *Oryza sativa*, *Triticum aestivum*, and *Zea mays*) were used for the analysis. The sensitivity, accuracy, area under the curve, consistency, duration, and RAM usage of the known miRNA predictions were evaluated for each tool. The miRNA annotations were obtained using miRBase and sRNAanno. Algorithms, such as random forest, BLAST, and receiver operating characteristic curves, were used to evaluate accuracy. Of the tools evaluated, sRNAbench was found to be the most accurate, miRDeep-P2 was the most sensitive, miRDeep-P2 was the fastest, and miRkwood had the highest memory usage. Due to its large genome size, only three tools were able to successfully predict known miRNAs in wheat (*Triticum aestivum*). Our results enable us to recommend the tool best suited to a variety of researcher needs, which we hope will reduce confusion and enhance future work.

KEY WORDS known miRNAs; random forest; receiver operating characteristic; sRNA-Seq.

MicroRNAs (miRNAs) are a class of non-coding single-stranded RNAs comprising approximately 21 nucleotides, and are found in a large number of plants and other organisms (D'Ario et al., 2017). miRNAs regulate many important biological processes, such as plant development (D'Ario et al., 2017) and the morphogenesis of shoot architecture (Wang et al., 2018), and can be harnessed for crop improvement (Tang and Chu, 2017). In addition, miRNAs influence the interactions between plants and their environments (Song et al., 2019), affecting plant responses to pathogen attack (Islam et al., 2018) and playing an important role in the defense against temperature stress (Megha et al., 2018).

Due to its advantages of rapid speed and low cost, next-generation sequencing has played an important role in the detection of known miRNAs, which are annotated miRNAs in the miRbase or sRNAanno database, in many studies (Le Trionnaire et al., 2011; Moran et al., 2017; Islam et al., 2018; Megha et al., 2018), leading to the development of numerous sequencing software tools for the functional analysis of miRNA data. More than 1000 miRNA bioinformatics tools were used for miRNA identification and target prediction studies between 2003 and 2013 (Chen et al., 2019a). Many research tools

for miRNA analysis are available online, enabling the majority of researchers to access and use them (Shukla et al., 2017). These tools include various types of algorithms and functions, with approximately 77% having been developed for the study of miRNAs in animals rather than plants (Akhtar et al., 2015; Morgado and Johannes, 2019). Additionally, most comprehensive tests and evaluations of these tools have been performed on animals (Li et al., 2012; Bisgin et al., 2018) and are lacking in plants (Srivastava, 2014). In particular, the detection of known miRNAs in species with varying genome sizes has not yet been conducted, making it difficult for researchers to select optimal software (Fig. 1) (Akhtar et al., 2015). Understanding the benefits and drawbacks of using different miRNA analysis programs is essential for improving the efficiency of miRNA studies.

Eight widely used or newly developed miRNA analysis tools were selected for an assessment of their ability to detect known miRNAs in plant species: miRDeep-P2 (Kuang et al., 2019), miRPlant (An et al., 2014), miRExpress (Wang et al., 2009), mirnova (Vitsios et al., 2017), sRNAbench (Barturen et al., 2014), miRDeep2 (Friedländer et al., 2012), miRkwood (Guigon et al., 2019), and miR-PREFeR (Lei and Sun, 2014) (Table 1).

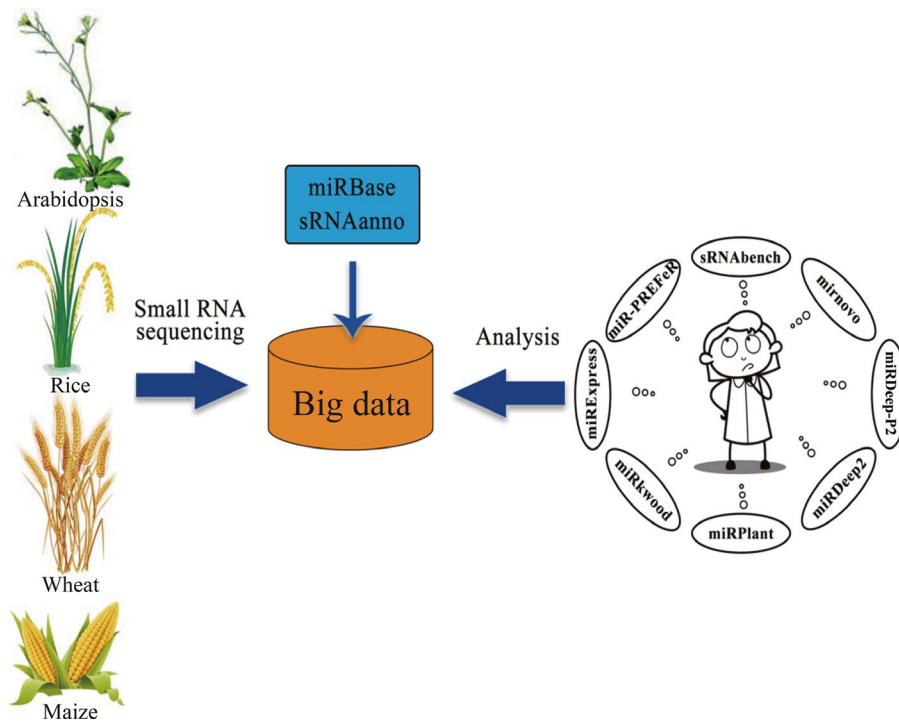


FIGURE 1. The complexity of large data sets and the need for bioinformatics tools.

miRDeep2 was developed using the programming language Perl. The preprocessing of reads by this tool is completed by its mapper. pl script, but their quantification is accomplished by quantifier.pl (Friedländer et al., 2012). sRNAbench is an improved version of miRAnalyzer, with similar functions to miRDeep2 (Hackenberg et al., 2011); it offers several novel features, including genome and library mapping and the analysis of differentially expressed genes (Barturen et al., 2014). miRExpress uses the Smith–Waterman algorithm to perform the alignment; therefore, miRExpress does not require genome mapping, as it maps the reads directly to the known miRNAs in miRbase (Wang et al., 2009; Kozomara et al., 2019). miRDeep-P2, the updated version of miRDeep-P, contains a new filtering strategy

that overhauls the older algorithm (Yang and Li, 2011) and has superior speed in processing next-generation sequencing data (Kuang et al., 2019). miRPlant was the first tool developed for plant miRNA identification that does not require any third-party applications, such as mapping or RNA secondary structure prediction tools. It visualizes the identified miRNAs in a hairpin diagram alongside all the RNA-Seq reads (An et al., 2014). mirnovO is a machine learning-based algorithm that can rapidly identify known miRNAs in animals and plants directly from small RNA (sRNA)-Seq data with or without a reference genome, making it very straightforward and intuitive for users (Vitsios et al., 2017). In addition, miRkwood and miR-PREFeR are used for the prediction of miRNA precursors from sRNA-Seq data. miRkwood is a user-friendly tool that can identify a large diversity of plant miRNAs while avoiding false positives (Guigon et al., 2019), whereas miR-PREFeR, an older tool than miRkwood, uses the expression patterns of RNAs to accurately detect and annotate miRNAs based on plant miRNA criteria (Jha, 2012; Lei and Sun, 2014).

Most previous studies evaluating miRNA analysis software have focused on running time, sensitivity, and accuracy (Li et al., 2012; Srivastava, 2014; Bisgin et al., 2018; Ou et al., 2019). In this study, we not only evaluated the eight tools in terms of these three factors, but also compared the number of known miRNAs predicted by each of them, as well as the maximum memory (random access memory [RAM]) cost when the software is running. To evaluate the accuracy of these tools, we applied the receiver operating characteristic (ROC) curve, which is widely used in different research fields involving animals and plants (Radivojac et al., 2013; Lyu et al., 2018; Zhao et al., 2018). The ROC curve is useful for visualizing the effectiveness of the model by comparing the rate of false positives and true positives. ROC and random forest (RF) assessments were previously combined to evaluate the results of an miRNA study (Zhao

TABLE 1. Summary of the eight miRNA analysis tools evaluated in this study.

Tool	Year	Reference	Platform	Features	Programming language	Citations ^a	Organism
mirnovO	2017	Vitsios et al., 2017	Linux, MAC OS, Web-based	miRNA prediction	Perl, Python, R	13	Plants, animals
miRPlant	2014	An et al., 2014	Linux, MAC OS, Windows	miRNA identification	Java	56	Plants
miRDeep-P2	2018	Kuang et al., 2019	Linux, MAC OS	miRNA prediction	Perl	0	Plants
miRExpress	2009	Wang et al., 2009	Linux, MAC OS	Expression profiles	C++	171	Plants, animals
miRkwood	2019	Guigon et al., 2019	Linux, MAC OS, Web-based	miRNA identification	Perl/C	0	Plants
miRDeep2	2012	Friedländer et al., 2012	Linux, MAC OS	miRNA prediction miRNA identification	Perl	1021	Animals
miR-PREFeR	2014	Lei and Sun, 2014	Linux, MAC OS, Windows	miRNA prediction Next-generation sequencing miRNA identification	Python	46	Plants
sRNAbench	2014	Barturen et al., 2014	Linux, MAC OS, Web-based	miRNA prediction miRNA-Seq Differential expression miRNA identification miRNA prediction	Java	111	Plants, animals
				Next-generation sequencing			

^aThe number of citations of each software tool was determined in June 2019 based on a Google Scholar search.

et al., 2018). We therefore compared the comprehensive evaluation of eight different tools in four different plant species by evaluating the area under the ROC curve (AUC).

The aim of this study was to provide useful information to assist researchers with the selection of the optimal miRNA analysis tool for studying different plant species under varying constraints.

METHODS

Data sets and gene annotations

A total of 20 RNA data sets were obtained from the National Center for Biotechnology Information (NCBI) for four different plants, *Arabidopsis thaliana* L., rice (*Oryza sativa* L.), maize (*Zea mays* L.), and wheat (*Triticum aestivum* L.), representing varying genome sizes, both monocots and dicots, and C_3 and C_4 species (see Data Availability Statement; Appendix 1). The genome sizes of *A. thaliana*, *O. sativa*, *Z. mays*, and *T. aestivum* are 0.12 Gbp (Kaul et al., 2000), 0.37 Gbp (Kawahara et al., 2013), 2.11 Gbp (Jiao et al., 2017), and 14.5 Gbp (IWGSC, 2018), respectively. Six data sets per species, including three wild-type samples and three treatment samples, were selected for *A. thaliana*, *O. sativa*, and *Z. mays*. One wild-type data set and one treatment data set were selected for *T. aestivum*. The known miRNA sequences and annotations from *A. thaliana*, *O. sativa*, and *Z. mays* were downloaded from miRbase (version 22; <http://www.mirbase.org/>). The known miRNA sequences and annotations of *T. aestivum* were downloaded from sRNAanno (Chen et al., 2019b).

Pre-processing of RNA-Seq data

The miRNA adapters were removed from all the RNA-Seq data using cutadapt software (Marcel, 2011). The length distribution of all clean reads in each sample was assessed, and reads with lengths between 18 and 30 bp were kept.

Program implementation

All miRNA sequencing software tools were run on a local server using the default or recommended parameters. The server was equipped with 16 central processing units (CPUs) and 64 GB of RAM. The operating system was CentOS7 (x86_64-bit version).

Prediction system assessment

To evaluate the performance of the software tools, the following measures were calculated: (1) The number of miRNAs predicted using both a BLAST search of the known miRNA sequences and the eight miRNA tools was considered to be the number of true positives (TP). (2) The number of known miRNAs predicted by BLAST but not by the eight miRNA tools was considered to be the number of false negatives (FN). (3) The number of known miRNAs neither predicted by BLAST nor the eight miRNA tools was considered to be the number of true negatives (TN). (4) The number of miRNAs predicted by the eight miRNA tools but not by BLAST was considered to be the number of false positives (FP).

We used the following measures to evaluate the performance of the different software tools:

$$\text{Sensitivity (Sen)} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Normalized Running time} = \frac{\text{Running time}}{\text{Sample size}} \quad (3)$$

$$\text{Specificity (TNR)} = \frac{TN}{FP + TN} \quad (4)$$

For the accuracy evaluation, RF assessment and the ROC curve were used. The two important parameters of RF are $n_{\text{tree}} = 100$ and $m_{\text{try}} = 4$ (Appendix S1). We used BLASTN (Camacho et al., 2009) and the eight miRNA analysis tools to predict the known miRNAs of the four species. The parameters for BLAST were ‘-task blastn-short -evalue 0.01’. To obtain the results of the BLAST alignment, the following filters were used: first, the miRNAs with read mismatches greater than one were removed. Next, the miRNAs that were successfully aligned in the reverse strand were discarded, after which the miRNAs that were mapped to target sequences with homologies of less than 90% were removed. From this set of miRNAs without gaps, those that were uniquely mapped and had a depth greater than five were retained (Appendix S2). For example, SRR1849765 is a sample *O. sativa* data set in this study. First, we BLAST-searched this sample against miRbase and obtained 568 known miRNAs. We then removed the miRNAs with read mismatches greater than one and discarded the miRNAs that were successfully aligned in the reverse strand, resulting in 558 known miRNAs. Next, we removed the miRNAs that were mapped to target sequences with homologies of less than 90% and obtained 302 known miRNAs. We retained the pre-filtered miRNAs without gaps that were uniquely mapped, and finally, we screened all miRNAs with a depth greater than five, resulting in a total of 142 known miRNAs. We compiled the results of known miRNAs predicted by the eight software types for all samples in Appendix S3.

Known miRNAs are annotated miRNAs contained in the miRBase or sRNAanno databases. Unknown miRNAs are miRNAs that are not annotated in the miRbase or sRNAanno databases. True known miRNAs are miRNAs annotated in the miRBase or sRNAanno databases and predicted by BLAST, while false known miRNAs are miRNAs annotated in the miRBase or sRNAanno database but not predicted by BLAST. Both the true known miRNAs and false known miRNAs were combined as the training data. The miRNA results obtained from the eight miRNA analysis tools were used as testing data (Appendix S2). Classification models were built in RF using the training data and applied to the testing data sets for the accuracy evaluation. The ROC curves were plotted using sensitivity and specificity, and the AUCs were calculated to further compare the performance of these models. ROCR (Sing et al., 2005), an R package (R Core Team, 2020), was used to generate the ROC chart. The UpSet plots were drawn with R, while all other pictures were drawn with Excel (Microsoft Corporation, Redmond, Washington, USA) or the online version of ECharts (Apache Software Foundation, Forest Hill, Maryland, USA).

Model establishment

miRkwood and miR-PREFeR were used to predict the known precursor miRNAs. The remaining six kinds of software were used to predict known mature miRNAs, which differ from

precursor miRNA sequences. We analyzed the predicted results of the known precursor miRNAs and known mature miRNAs separately.

The ROC curve is a useful measure for comparing multiple models. The model with the highest AUC value is typically the most optimal. We divided the predicted known miRNAs into mature short-sequence training models and precursor long-sequence training models. At the same time, the BLAST comparison results of each sample were independently constructed to train the models. The prediction results of each sample in all eight software tools were independently constructed to test the models. We found that miRDeep2 and mirnovo identified the fewest known miRNAs in *A. thaliana*, which was insufficient to use RF to create models for miRDeep2 and mirnovo in *A. thaliana*; however, we used RF to construct models for the other six tools using the *A. thaliana* training data. The results of the BLAST analysis were used as testing data to assess each training model. Finally, the test results of six *A. thaliana* samples were used to generate ROC curves (Appendix S4). Similarly, a large number of known miRNAs were identified in *O. sativa* samples by the eight tools, with each generating an ROC curve (Appendix S5). The ROC results for *Z. mays* are also provided in Appendix S6. sRNAbench, miRDeep2, and miRExpress successfully identified multiple known *T. aestivum* miRNAs in the two samples selected for analysis (Appendix S7), which were combined to generate one ROC curve.

RESULTS

Computational time

The computational time required by each software tool to identify the miRNAs in the four species was determined using a single local server (Fig. 2) by selecting one sample from each species (*A. thaliana*: SRR1312898, *O. sativa*: SRR1849770, *Z. mays*: SRR6939404, and *T. aestivum*: SRR5461177). We found that sRNAbench took the least amount of time to predict the miRNAs, while miRPlant and mirnovo took a longer period of computational time. For *A. thaliana*, mirnovo was slowest, while for *O. sativa*, miRPlant was slowest. When performing the miRNA analysis in *Z. mays*, the eight different software tools took varying amounts of time. We successfully used miRDeep2, miRExpress, and sRNAbench to predict the positive known miRNAs in *T. aestivum*. sRNAbench was still fastest, while miRDeep2 was much slower than miRExpress and sRNAbench. Figure 2B shows the normalized running time of the eight tools for predicting the known miRNAs, which were generated using Equation (3). The normalized results better explain the relationship between the time costs of the software and the sizes of the different samples. Overall, sRNAbench was the fastest for the analysis of all four species.

Average number of true positive known mature or precursor miRNAs identified

We next compared the average number of true positive known miRNAs identified by the eight software tools in all samples of the four species (Fig. 3). miRExpress identified the greatest number of mature known miRNAs in *A. thaliana*, *O. sativa*, and *Z. mays*, while sRNAbench identified the greatest number of mature known

miRNAs in *T. aestivum*. Fewer known miRNAs were predicted by miRDeep2 than the other tools in *A. thaliana* and *T. aestivum*. miRkwood identified more known miRNA precursors in *A. thaliana*, *O. sativa*, and *Z. mays* than miR-PREFeR.

Known miRNA comparisons among the three most successful tools

We compared the known miRNAs predicted by the eight software tools in all samples of the four species. Both types of miRNAs detected in one species by each software were counted for comparison. Among the prediction tools, sRNAbench and miRExpress detected the most mature miRNAs, whereas miRkwood identified the most miRNA precursors. The results from each tool were shared with at least one other tool for *A. thaliana*, *O. sativa*, and *Z. mays* (Fig. 4). We next focused on the three tools that ran successfully in all four species: miRExpress, miRDeep2, and sRNAbench (Fig. 4A–D). We believe that the other five tools did not predict the known miRNAs in *T. aestivum* due to the huge size of its genome. These unsuccessful tools require relatively large amounts of memory to predict miRNAs, exceeding the maximum operating memory range of our computer for their analysis of *T. aestivum*. Comparing the predictive results of the three tools that were successful in all species tested will help us to further explore which software is more suitable for the prediction of different plant miRNAs. The results of the three tools were most similar for *A. thaliana* and *O. sativa* (Fig. 4A, B), whereas less consistent results were observed in *Z. mays* (Fig. 4A–C). These three tools were used to predict known miRNAs in *T. aestivum*, but the results were quite inconsistent (Fig. 4D). The known miRNAs detected by miRDeep2 were all included in the miRExpress and sRNAbench databases for three of the four species, with the exception of *T. aestivum*. The results indicated that sRNAbench has the potential to identify more known miRNAs, while miRDeep2 filtered out many true known miRNAs. Thus, miRDeep2 may be not suitable for plant research.

Sensitivity and specificity

High sensitivity (Equation 1), one of the most important criteria for evaluating software operation capacity, results in the prediction of more known miRNAs. As shown in Fig. 5A, sRNAbench, miRExpress, and miRkwood were more highly sensitive than the other tools. The sensitivity of miRPlant was stable across the four species tested. The sensitivities of miR-PREFeR and miRDeep-P2 were higher for *Z. mays* than for *A. thaliana* and *O. sativa*, while the sensitivity of mirnovo was higher for *O. sativa* than for *A. thaliana* and *Z. mays*. The results of the treatment samples (Fig. 5C) showed the same tendency as the results for the wild-type data sets (Fig. 5A).

In addition to sensitivity, specificity (Equation 4) is an important criterion for evaluating software operating results. The higher the specificity of the software, the smaller the number of miRNAs falsely predicted by the software. As shown in Fig. 5B (wild-type data set), the specificities of miRDeep2 and mirnovo were high, and the specificity of mirnovo was stable across the different species. The specificity of miRkwood was the lowest of all the tools evaluated. The specificities for the treatment data sets (Fig. 5D) showed the same tendency as for the wild-type data sets (Fig. 5B).

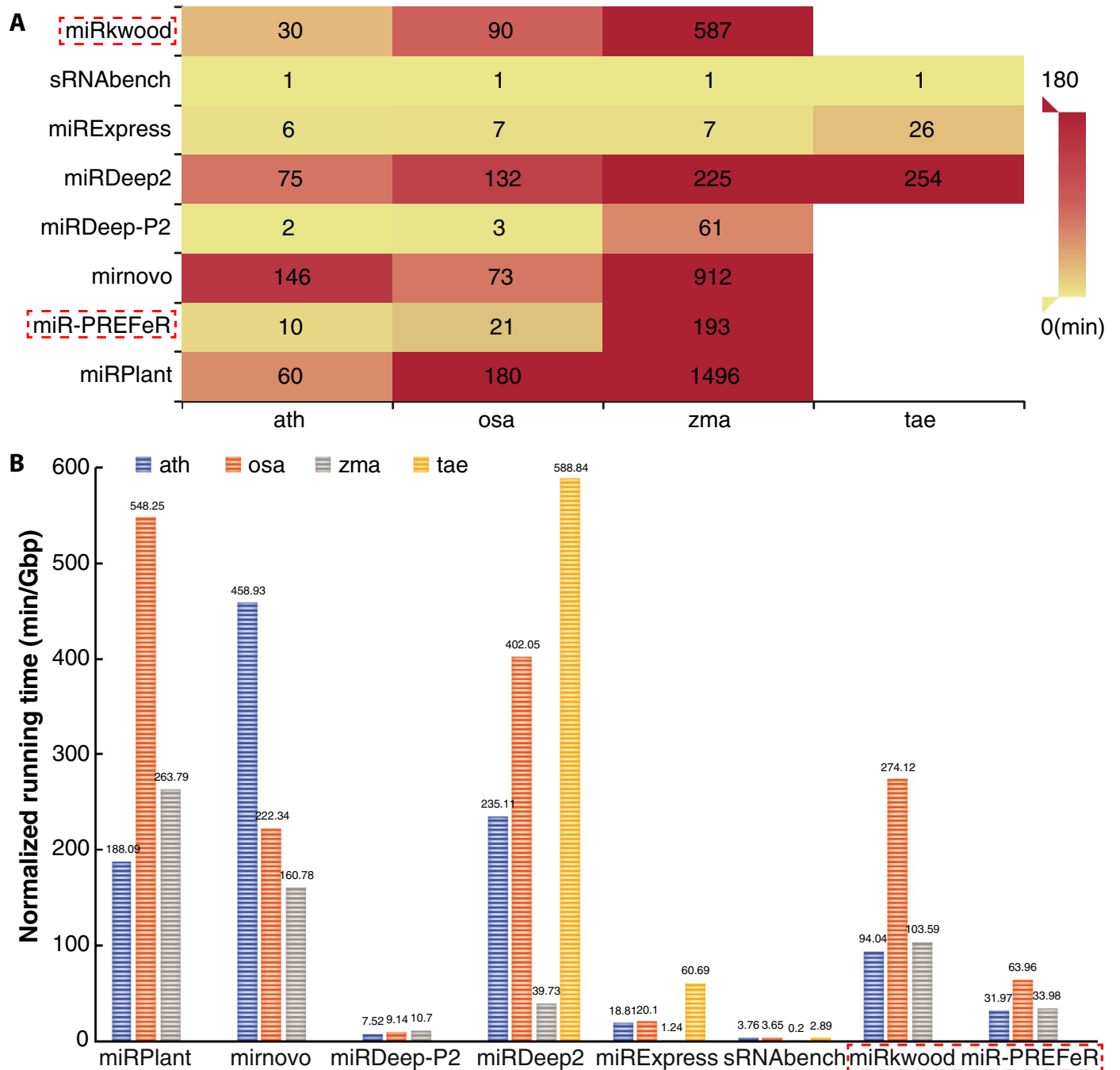


FIGURE 2. Computational time required by each of the eight software tools using the default or recommended settings. The programs in the red dashed boxes are used to predict known miRNA precursors; the rest of the software tools are used to predict known mature miRNAs. (A) Time required for the eight software tools to analyze data sets from the four species (in minutes). Entries are shaded with red to yellow gradients, where red represents the longest period and yellow the shortest. (B) Relative normalized running time required for each software tool. The y-axis displays the running time normalized against the size of the genome (Equation 3). ath, *Arabidopsis thaliana*; osa, *Oryza sativa*; tae, *Triticum aestivum*; zma, *Zea mays*.

Accuracy

Accuracy (Equation 2) is an important consideration when predicting miRNAs. We quantified the accuracy of each tool, as shown in Fig. 6.

When predicting mature miRNAs in *A. thaliana*, miRDeep-P2 had the highest accuracy. When predicting mature miRNAs in *O. sativa* and *Z. mays*, mirnovo had the highest accuracy, while for

T. aestivum, miRDeep2 had the highest accuracy (Fig. 6A). When searching for known miRNA precursors, miRkwood had a higher accuracy rate than miR-PREFeR in all three species (Fig. 6A).

All the accuracy results for the treatment data from the four species are shown in Fig. 6B. In the analysis of the treatment samples, miRkwood again had a higher success rate than miR-PREFeR

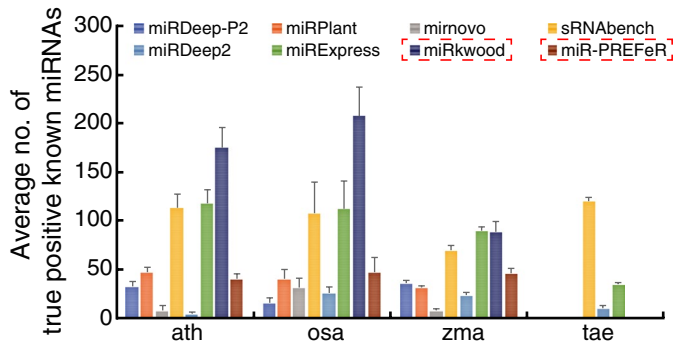


FIGURE 3. Average number of true positive known miRNAs predicted by the eight software tools for the four species. The programs in the red dashed boxes are used to predict known miRNA precursors; the rest of the software tools are used to predict known mature miRNAs. Error bars represent SD. *ath*, *Arabidopsis thaliana*; *osa*, *Oryza sativa*; *tae*, *Triticum aestivum*; *zma*, *Zea mays*.

(Fig. 6B), with a similar accuracy rate for predicting miRNAs in *A. thaliana* and *O. sativa*. miRkwood had a higher accuracy rate when predicting miRNAs in the wild-type samples of *Z. mays* than in the treatment samples. Overall, the accuracy results of wild-type data and treatment data had similar tendencies.

Performance evaluation of the tools across the four species

sRNAbench and miRExpress had the highest AUC values (Fig. 7A, B). The AUC values for miRPlant, miRDeep2, and mirnovo were

all above 0.8 in *O. sativa*, while the AUC value for miRDeep-P2 was below 0.8 for this species. The AUC values for miRkwood were considerably different between the *O. sativa* samples. Finally, miR-PREFeR had the lowest AUC values, although they were stable across the four species (Appendices S4–6). For *T. aestivum*, with its larger genome, the performance of miRDeep2 and miRExpress was poor, although the performance of sRNAbench was better (Appendix S7).

The average AUC values for the four wild-type species data sets are shown in Fig. 7C. Overall, sRNAbench had a relatively high average AUC value for all four species. miRExpress also performed well in all species except *T. aestivum*. Of the eight software tools, miRDeep2 performed worst for *Z. mays*; however, it provided a higher AUC value for *O. sativa* than all but sRNAbench and miRExpress. The average AUC values in the treatment data for the four species are shown in Fig. 7D. The results for the wild-type and treatment data were similar and showed that the most suitable miRNA software for identifying known miRNAs differed between species.

RAM usage

RAM usage also differed dramatically between the eight software tools (Fig. 8). When the genome of the tested species was larger (*A. thaliana* < *O. sativa* < *Z. mays* < *T. aestivum*), sRNAbench, miRPlant, and miR-PREFeR all required larger amounts of memory, while in contrast, the amount of memory required for miRkwood was similar for the analysis of the *A. thaliana*, *O. sativa*, and *Z. mays* samples. This may be explained by the fact that miRkwood was run with the Docker container, which can

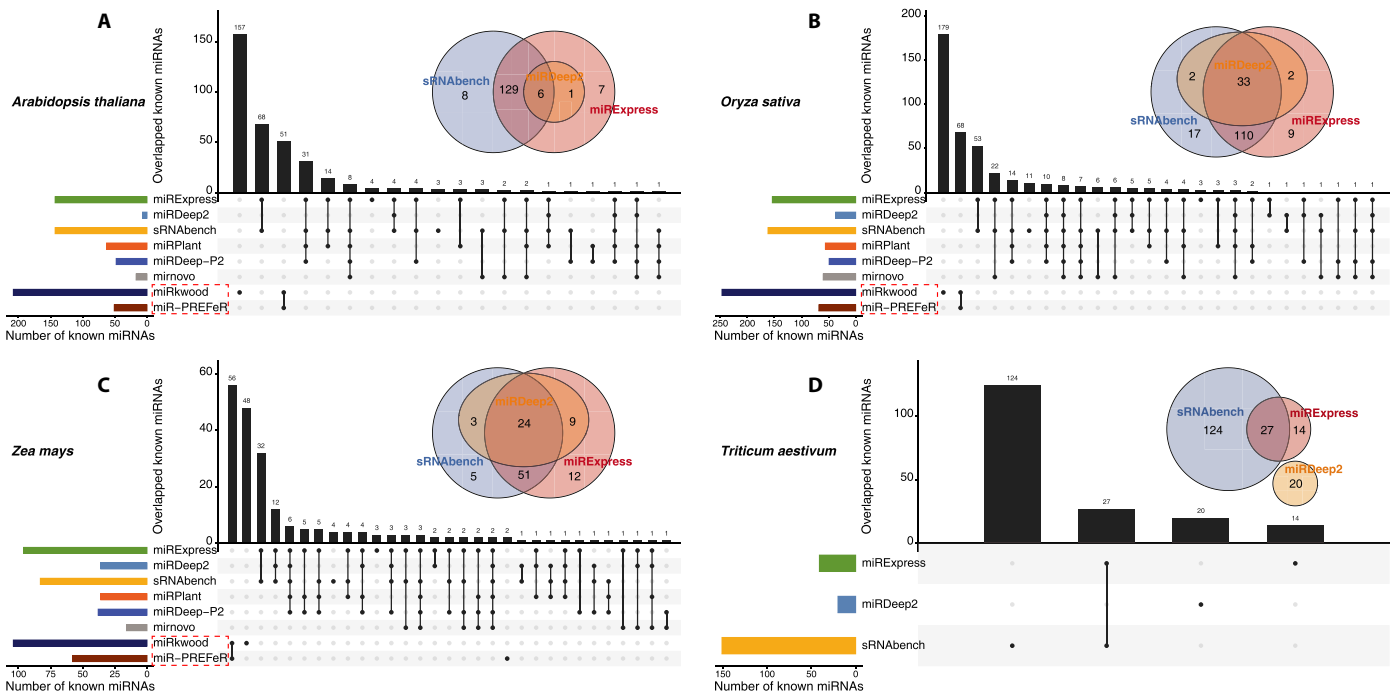


FIGURE 4. UpSet plot and Venn diagrams of the known miRNAs. The programs in the red dashed boxes are used to predict known miRNA precursors; the rest of the software tools are used to predict known mature miRNAs. (A–C) Known miRNAs detected by all tools in *Arabidopsis thaliana* (A), *Oryza sativa* (B), and *Zea mays* (C). (D) Known miRNAs detected by sRNAbench, miRExpress, and miRDeep2 in *Triticum aestivum*. The UpSet plots show all intersections among all tools, while the Venn diagrams show the intersections among sRNAbench, miRExpress, and miRDeep2.

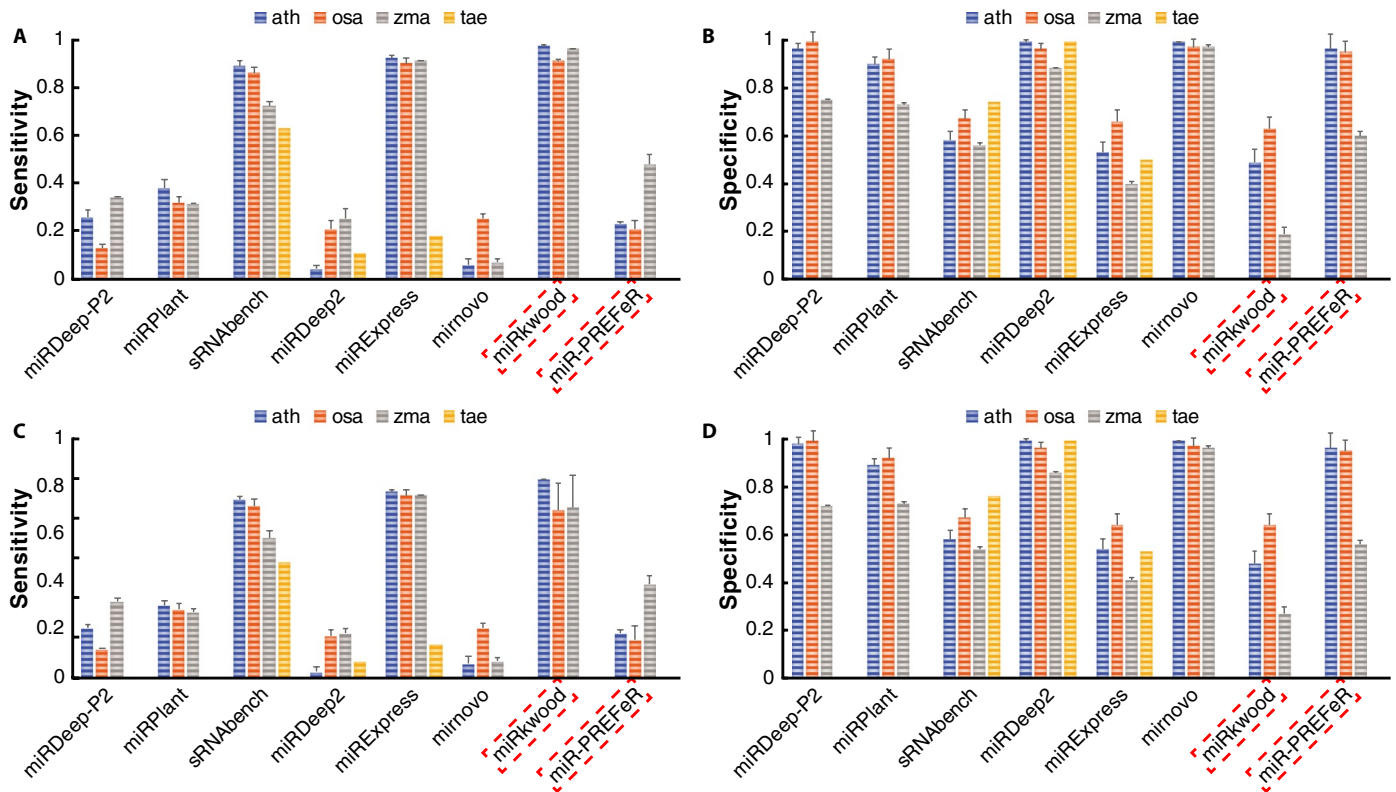


FIGURE 5. Comparison of the sensitivity and specificity of various software tools when predicting known miRNAs. The programs in the red boxes are used to predict known miRNA precursors; the rest of the software tools are used to predict known mature miRNAs. Error bars represent SD. (A) The sensitivity of the eight software tools when analyzing the wild-type samples. (B) The specificity of the eight software tools when analyzing the wild-type samples. (C) The sensitivity of the eight software tools when analyzing the treatment samples. (D) The specificity of the eight software tools when analyzing the treatment samples. ath, *Arabidopsis thaliana*; osa, *Oryza sativa*; tae, *Triticum aestivum*; zma, *Zea mays*.

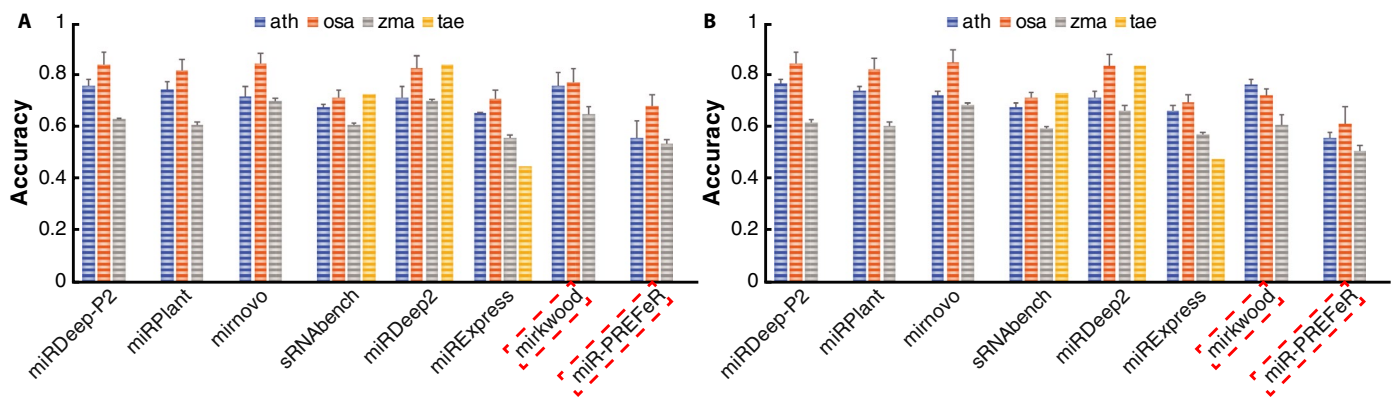


FIGURE 6. Comparison of the accuracy of the various software tools. The programs in the red dashed boxes are used to predict known miRNA precursors; the rest of the software tools are used to predict known mature miRNAs. The charts show the accuracy of the eight software tools in predicting known miRNAs in the four species for wild-type data (A) and treatment data (B). Error bars represent SD. ath, *Arabidopsis thaliana*; osa, *Oryza sativa*; tae, *Triticum aestivum*; zma, *Zea mays*.

influence the calculation of memory usage. Among the six mature miRNA prediction tools (miRPlant, mirnovo, miRDeep-P2, miRDeep2, miRExpress, and sRNAbench), sRNAbench required the most memory for the analyses of *A. thaliana* and *O. sativa*. miRExpress required the least memory for the analysis of all four species. Most of the programs exceeded the maximum computer

memory in their analyses of the large *T. aestivum* genome, and only three programs (miRDeep2, miRExpress, and sRNAbench) successfully predicted the known mature miRNAs in *T. aestivum*. miRDeep2 and sRNAbench required significantly more memory for analyzing *T. aestivum* than for *A. thaliana* and *O. sativa*. The results showed that none of these tools require much memory for

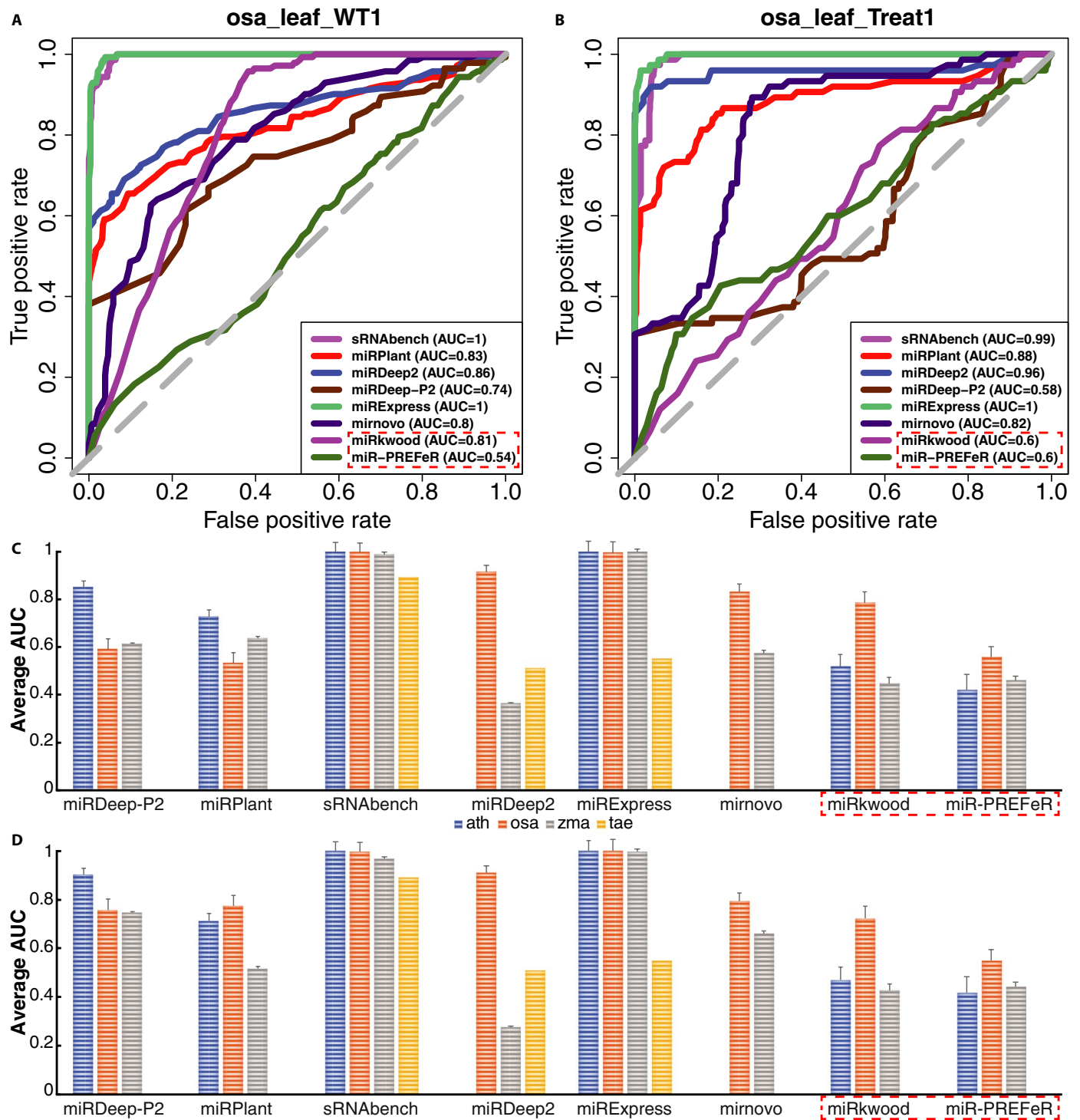


FIGURE 7. The receiver operating characteristic (ROC) curves and area under the curve (AUC) values for the miRNA predictions of the eight software tools. The programs in the red dashed boxes are used to predict known miRNA precursors; the rest of the software tools are used to predict known mature miRNAs. Error bars represent SD. (A) ROC curve of the miRNA predictions using the wild-type *Oryza sativa* data set. (B) ROC curve of the miRNA predictions using the treatment *O. sativa* data set. (C) Average AUC values from wild-type samples of the four species. (D) Average AUC values from the treatment samples of the four species. ath, *Arabidopsis thaliana*; osa, *O. sativa*; tae, *Triticum aestivum*; zma, *Zea mays*.

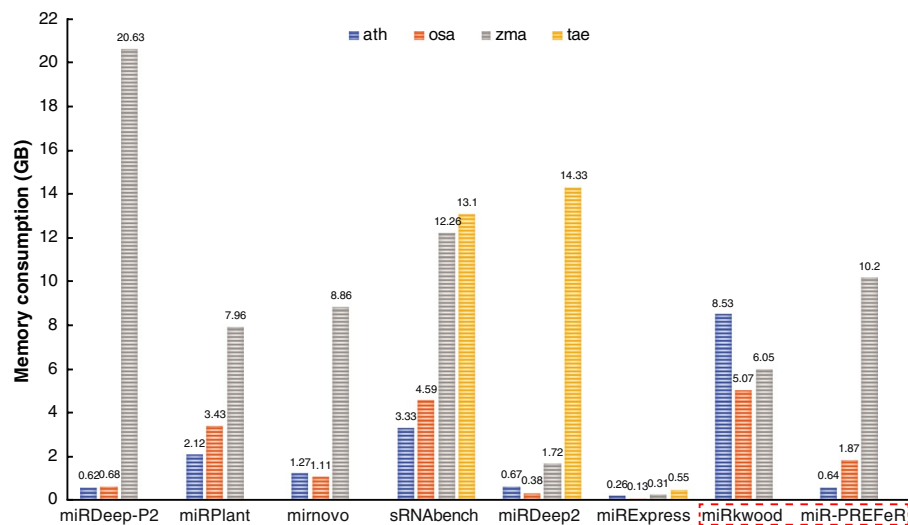


FIGURE 8. RAM usage of the eight software tools when predicting known miRNAs in the four species. Each peak represents the memory cost of a tool when analyzing a specific species. The programs in the red dashed boxes are used to predict known miRNA precursors; the rest of the software tools are used to predict known mature miRNAs. ath, *Arabidopsis thaliana*; osa, *Oryza sativa*; tae, *Triticum aestivum*; zma, *Zea mays*.

the analysis of small genomes, enabling these analyses to be performed using a personal laptop.

Experimental verification of the identified known miRNAs in the four species

We identified numerous known miRNAs using the eight tools, some of which have been verified in previous studies (Baldrich et al., 2015; Feng et al., 2017; Yu et al., 2017; Minow et al., 2018). As shown in Appendix 2, the prediction results of six of the software tools (excluding miRDeep2 and mirnovo) covered most known miRNAs in the *A. thaliana* wild-type data. *Arabidopsis thaliana* has the smallest genome, so most of the analysis tools performed better for this species. For the *O. sativa* leaf wild-type data, both sRNAbench and miRkwood successfully predicted five experimentally verified known miRNAs, while miRExpress predicted four. With the exception of mirnovo, all software tools successfully predicted three experimentally verified known miRNAs in the mature leaf treatment data of *Z. mays*. For *T. aestivum*, which had the largest genome among the four species, only sRNAbench and miRExpress could successfully predict four experimentally verified known miRNAs. In addition, miRkwood could identify more of the experimentally verified known miRNA precursors in *A. thaliana* and *O. sativa* than miR-PREFeR. This is consistent with the experimental results in our study and supports the credibility of these findings.

DISCUSSION

Selecting the appropriate software tool for different tasks

To identify true known miRNAs, the eight tools excluded other RNA fragments by rigorously comparing each read with known rRNA, small conditional RNA (scRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), tRNA, and mRNA sequences (Li et al., 2012). To increase the specificity of its results, miRExpress accepts reads

by aligning sequences with known miRNAs (Chen et al., 2019a). sRNAbench gave the best sensitivity performance when detecting known miRNAs, while miRExpress was second best due to its construction of miRNA expression profiles by aligning sequences with known miRNAs (Chen et al., 2019a). miRExpress predicted the most known mature miRNAs in *A. thaliana*, *O. sativa*, and *Z. mays* compared to other software, and is therefore the optimal software tool for predicting large numbers of known miRNAs.

The number of true positive known miRNAs identified in *Z. mays* by miRDeep2 was very low (Fig. 3), although its average AUC values were higher for *O. sativa* and *T. aestivum* (Fig. 7C, D). When we predicted the known miRNAs in *Z. mays*, miRDeep2 had the lowest AUC value among the eight software tools (Fig. 7C, D); therefore, miRDeep2 does not appear to be suitable for the prediction of known miRNAs in this species. miRDeep2 was developed for use in animal studies, and as the pathway of miRNA maturation in plants is different from that in animals (Du and Zamore, 2005), this tool therefore may not be suitable for predicting known miRNAs in plants.

In addition, several true positive known miRNAs, which were predicted as false negatives by miRDeep2, resulted in the low accuracy of the analysis in *Z. mays*. We therefore retrieved mature miRBase miRNAs, which may be incorrectly predicted by miRDeep2, and reinserted them into the miRDeep2 results to predict the known miRNAs of *Z. mays*. Then, we regenerated the ROC graph of the eight software tools predicting the known miRNAs in four species (Appendix S8). As suspected, the AUC value of miRDeep2 in the six *Z. mays* samples greatly increased.

When analyzing the *A. thaliana* and *O. sativa* data, the calculation speed of miRDeep-P2 was faster than that of miRExpress, placing it second after sRNAbench. In contrast, miRDeep-P2 was much slower than miRExpress when analyzing *Z. mays* data, and it failed to analyze the *T. aestivum* data.

In the known mature miRNA predictions, miRExpress had the best sensitivity performance for *A. thaliana*, *O. sativa*, and *Z. mays*, while sRNAbench had the best sensitivity of all tools when analyzing the *T. aestivum* data. Of the two, miRExpress used less memory than sRNAbench.

Performance evaluation

Various software tools are available for the analysis of miRNA sRNA-Seq data, but selecting the most appropriate tool for a specific purpose still poses a challenge to researchers. miRExpress can be used for the analysis of animals or plants, including *A. thaliana* and humans, while miRDeep2 is more suitable for use with animal data sets, including data from *Caenorhabditis elegans*, fruit fly, human, and mouse. sRNAbench can also be used to predict *C. elegans*, chicken, fruit fly, human, mouse, rat, *Xenopus*, and zebrafish miRNAs. Furthermore, sRNAbench and miRExpress can also identify differentially expressed miRNAs. As early as 2012, Li et al. (2012) compared miRExpress, miRanalyzer (the precursor to sRNAbench), and miRDeep (the precursor to miRDeep2) for human, chicken, and *C. elegans* samples, showing that miRExpress

and miRanalyzer had better predictive capabilities. In a later study, Bisgin et al. (2018) compared the ability of miRExpress, sRNAbench, and miRDeep2 to predict miRNAs in rat liver, revealing that sRNAbench and miRDeep2 are more suited to this task than miRExpress. miRPlant and mirnovo provide a user-friendly interface with improved predictive accuracy for species such as *O. sativa*, although mirnovo can also predict miRNAs in diverse species such as fruit fly, soybean (*Glycine max* (L.) Merr.), and human. miR-PREFeR has previously been used to predict miRNAs in plants including soybean and tomato (*Solanum lycopersicum* L.). miRkwood can identify a large diversity of plant miRNAs with limited false positives, which can be beneficial for species such as *Brassica rapa* L. and soybean.

There are several possible reasons why the eight software tools differed in their ability to predict known miRNAs in the four plant species analyzed. First, the number of annotated known miRNAs varies between the species studied, which could influence the prediction results. Second, the depth of sequencing data differed between the samples, and samples with a low depth may lead to less accurate results. In addition, the same species may have different miRNA expression levels at different growth stages, meaning several miRNAs may not be identified in all stages. Last, different genome sizes may affect the analysis performed by some software tools, such as miRDeep2 and miRkwood.

Based on our findings, we developed the following criteria for selecting the best tool for identifying known miRNAs in plants. First, for researchers with limited access to computer memory, we recommend miRExpress, because its calculation requires less memory while retaining high levels of accuracy. In addition, it can be used for the analysis of species with either large or small genomes. Second, for researchers who have access to sufficient computer memory and want to identify more known miRNAs, we recommend using sRNAbench due to its accuracy and suitability for many species of different genome sizes. For researchers who have access to sufficient computer memory and want to identify known miRNA precursors, we recommend miRkwood, because its calculation time is short and a large number of known miRNA precursors can be obtained. We also recommend different software tools for different data types, as shown in Table 1.

ACKNOWLEDGMENTS

This work was sponsored by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (18KJB180029), State Key Laboratory of Cotton Biology and State Key Laboratory of Cotton Biology Open Fund (CB2021A06), with funding from the Lvyang Golden Phoenix Plan in Yangzhou, a China Postdoctoral Science Foundation Grant (2018M642342), the National Natural Science Foundation of China (grant no. 32000458), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

AUTHOR CONTRIBUTIONS

Q.Y. designed the project. Q.-L.L., G.-Q.L., Y.B., and Y.-C.W. performed the research. Q.-L.L. analyzed the data and conducted the bioinformatics analyses. Q.Y. and Q.-L.L. wrote the article, and all authors approved the final version of the manuscript.

DATA AVAILABILITY

All the sRNA data sets, generated using the Illumina platform (Illumina, San Diego, California, USA), were downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>). The *Arabidopsis thaliana* sRNA-Seq data sets were generated from 12- to 13-d-old seedlings (SRR1312888, SRR1312896, SRR1312897, SRR1312898, SRR1312899, and SRR1312900) (Yu et al., 2017); the *Oryza sativa* sRNA-Seq data sets were generated from 15-d-old leaves (SRR1849765, SRR1849766, SRR1849767, SRR1849768, SRR1849769, and SRR1849760) (Baldrich et al., 2015); the *Zea mays* sRNA-Seq data sets were generated from mature leaves (leaf 7; SRR6939401, SRR6939402, SRR6939403, SRR6939404, SRR6939406, and SRR6939409) (Minow et al., 2018); and the *Triticum aestivum* sRNA-Seq data were generated from flower spikes (SRR5461176 and SRR5461177) (Feng et al., 2017).

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. The results of the random forest parameters. (A) The parameter *mtry* is equal to 4. (B) The dot plot shows the features and their importance; more important features have a higher MeanDecreaseGini value. (C) Determination of the *ntree* value for building the random forest model. The *x*-axis shows the number of trees (*ntree*). The position where the decreased trend of lines becomes flat is ideal, thus the parameter *ntree* is equal to 100.

APPENDIX S2. Flow chart of the generation of the receiver operating characteristic (ROC) curve.

APPENDIX S3. Confusion matrices for all samples.

APPENDIX S4. The receiver operating characteristic (ROC) curves for the predictions of the known miRNAs from *Arabidopsis thaliana* using six software tools. (A–F) ROC curves for samples SRR1312888 (A), SRR1312898 (B), SRR1312896 (C), SRR1312899 (D), SRR1312897 (E), and SRR1312900 (F). The number of known miRNAs predicted by mirnovo and mirdeep2 in *A. thaliana* is too small to draw the corresponding ROC curve, thus those results are not shown here.

APPENDIX S5. The receiver operating characteristic (ROC) curves for the predictions of the known miRNAs from *Oryza sativa* by the eight software tools evaluated. (A–D) ROC curves for samples SRR1849766 (A), SRR1849769 (B), SRR1849767 (C), and SRR1849770 (D).

APPENDIX S6. The receiver operating characteristic (ROC) curves for the predictions of the known miRNAs from *Zea mays* by the eight software tools evaluated. (A–F) ROC curves for samples SRR6939406 (A), SRR6939401 (B), SRR6939403 (C), SRR6939402 (D), SRR6939404 (E), and SRR6939409 (F).

APPENDIX S7. The receiver operating characteristic (ROC) curves for the predictions of the known miRNAs from *Triticum aestivum* by three software tools.

APPENDIX S8. The receiver operating characteristic (ROC) curves following the combination of the known *Zea mays* miRNAs predicted by the eight software tools and the mature miRNA sequences from miRBase. (A–F) ROC curves for samples SRR6939406 (A), SRR6939401 (B), SRR6939403 (C), SRR6939402 (D), SRR6939404 (E), and SRR6939409 (F).

LITERATURE CITED

- Akhtar, M. M., L. Micolucci, M. S. Islam, F. Olivieri, and A. D. Procopio. 2015. Bioinformatic tools for microRNA dissection. *Nucleic Acids Research* 44: 24–44.
- An, J., J. Lai, A. Sajjanhar, M. L. Lehman, and C. C. Nelson. 2014. miRPlant: An integrated tool for identification of plant miRNA from RNA sequencing data. *BMC Bioinformatics* 15: 275.
- Baldrich, P., S. Campo, M. T. Wu, T. T. Liu, Y. I. Hsing, and B. San Segundo. 2015. MicroRNA-mediated regulation of gene expression in the response of rice plants to fungal elicitors. *RNA Biology* 12: 847–863.
- Barturen, G., A. Rueda, M. Hamberg, A. Alganza, R. Lebron, M. Kotsyfakis, B.-J. Shi, et al. 2014. sRNAbench: Profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods in Next Generation Sequencing* 1: 21–31.
- Bisgin, H., B. Gong, Y. Wang, and W. Tong. 2018. Evaluation of bioinformatics approaches for next-generation sequencing analysis of microRNAs with a toxicogenomics study design. *Frontiers in Genetics* 9: 22.
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.
- Chen, L., L. Heikkinen, C. Wang, Y. Yang, H. Sun, and G. Wong. 2019a. Trends in the development of miRNA bioinformatics tools. *Briefings in Bioinformatics* 20: 1836–1852.
- Chen, C., J. Feng, B. Liu, J. Li, L. Feng, X. Yu, J. Zhai, et al. 2019b. sRNAanno: A database repository of uniformly-annotated small RNAs in plants. *BioRxiv* 10.1101/771121 [preprint] [published 16 March 2019]. Available from: <https://www.biorxiv.org/content/10.1101/771121v1> [accessed 27 October 2019].
- D’Ario, M., S. Griffiths-Jones, and M. Kim. 2017. Small RNAs: Big impact on plant development. *Trends in Plant Science* 22: 1056–1068.
- Du, T., and P. D. Zamore. 2005. microPrimer: The biogenesis and function of microRNA. *Development* 132: 4645–4652.
- Feng, N., G. Song, J. Guan, K. Chen, M. Jia, D. Huang, J. Wu, et al. 2017. Transcriptome profiling of wheat inflorescence development from spikelet initiation to floral patterning identified stage-specific regulatory genes. *Plant Physiology* 174: 1779–1794.
- Friedländer, M. R., S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research* 40: 37–52.
- Guigon, L., S. Legrand, J. F. Berthelot, S. Bini, D. Lanselle, M. Benmounah, and H. Touzet. 2019. miRkwood: A tool for the reliable identification of microRNAs in plant genomes. *BMC Genomics* 20: 532.
- Hackenberger, M., N. Rodriguez-Ezpeleta, and A. M. Aransay. 2011. miRAnalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research* 39: W132–W138.
- International Wheat Genome Sequencing Consortium (IWGSC). 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361: <https://doi.org/10.1126/science.aar7191>.
- Islam, W., A. Noman, M. Qasim, and L. D. Wang. 2018. Plant responses to pathogen attack: Small RNAs in focus. *International Journal of Molecular Sciences* 19: 515.
- Jha, A. 2012. miR-BAG: Bagging based identification of microRNA precursors. *PLoS ONE* 7: e45782.
- Jiao, Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, et al. 2017. Improved maize reference genome with single-molecule technologies. *Nature* 546: 524–527.
- Kaul, S., H. L. Koo, J. Jenkins, M. Rizzo, T. Rooney, L. J. Tallon, T. Feldblyum, et al. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.
- Kawahara, Y., M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N Y)* 6: 4.
- Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones. 2019. miRBase: From microRNA sequences to function. *Nucleic Acids Research* 47: D155–D162.
- Kuang, Z., Y. Wang, L. Li, and X. Yang. 2019. miRDeep-P2: Accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics* 35: 2521–2522.
- Le Trionnaire, G., R. T. Grant-Downton, S. Kourmpetli, H. G. Dickinson, and D. Twell. 2011. Small RNA activity and function in angiosperm gametophytes. *Journal of Experimental Botany* 62: 1601–1610.
- Lei, J., and Y. Sun. 2014. miR-PREFeR: An accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics* 30: 2837–2839.
- Li, Y., Z. Zhang, F. Liu, W. Vongsangnak, Q. Jing, and B. Shen. 2012. Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Research* 40: 4298–4305.
- Lyu, C. Q., L. Wang, and J. H. Zhang. 2018. Deep learning for DNase I hypersensitive sites identification. *BMC Genomics* 19: 905.
- Marcel, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10–12. Available from: <https://doi.org/10.14806/ej.17.1.200> [accessed 27 October 2019].
- Megha, S., U. Basu, and N. N. V. Kav. 2018. Regulation of low temperature stress in plants by microRNAs. *Plant Cell and Environment* 41: 1–15.
- Minow, M. A. A., L. M. Avila, K. Turner, E. Ponzoni, I. Mascheretti, F. M. Dussault, L. Lukens, et al. 2018. Distinct gene networks modulate floral induction of autonomous maize and photoperiod-dependent teosinte. *Journal of Experimental Botany* 69: 2937–2952.
- Moran, Y., M. Agron, D. Praher, and U. Technau. 2017. The evolutionary origin of plant and animal microRNAs. *Nature Ecology & Evolution* 1: 0027.
- Morgado, L., and F. Johannes. 2019. Computational tools for plant small RNA detection and categorization. *Briefings in Bioinformatics* 20: 1181–1192.
- Ou, S., W. Su, Y. Liao, K. Chougule, J. R. A. Agda, A. J. Hellinga, C. S. B. Lugo, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* 20: 275.
- R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website <http://www.R-project.org/> [accessed 16 February 2021].
- Radivojac, P., W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Grait, et al. 2013. A large-scale evaluation of computational protein function prediction. *Nature Methods* 10: 221–227.
- Shukla, V., V. K. Varghese, S. P. Kabekkodu, S. Mallya, and K. Satyamoorthy. 2017. A compilation of Web-based research tools for miRNA analysis. *Briefings in Functional Genomics* 16: 249–273.
- Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCr: Visualizing classifier performance in R. *Bioinformatics* 21: 3940–3941.
- Song, X. W., Y. Li, X. F. Cao, and Y. J. Qi. 2019. MicroRNAs and their regulatory roles in plant-environment interactions. *Annual Review of Plant Biology* 70: 489–525.
- Srivastava, P. K. 2014. A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *BMC Genomics* 15: 348.
- Tang, J., and C. Chu. 2017. MicroRNAs in crop improvement: Fine-tuners for complex traits. *Nature Plants* 3: 17077.
- Vitsios, D. M., E. Kentepozidou, L. Quintais, E. Benito-Gutierrez, S. van Dongen, M. P. Davis, and A. J. Enright. 2017. MirNovo: Genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Research* 45: e177.

- Wang, W. C., F. M. Lin, W. C. Chang, K. Y. Lin, H. D. Huang, and N. S. Lin. 2009. miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* 10: 328.
- Wang, B., S. M. Smith, and J. Y. Li. 2018. Genetic regulation of shoot architecture. *Annual Review of Plant Biology* 69: 437–468.
- Yang, X., and L. Li. 2011. miRDeep-P: A computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 27: 2614–2615.
- Yu, Y., L. Ji, B. H. Le, J. Zhai, J. Chen, E. Luscher, L. Gao, et al. 2017. ARGONAUTE10 promotes the degradation of miR165/6 through the SDN1 and SDN2 exonucleases in Arabidopsis. *PLoS Biology* 15: e2001272.
- Zhao, Q., Q. Mao, Z. Zhao, T. Y. Dou, Z. G. Wang, X. Y. Cui, Y. N. Liu, and X. Y. Fan. 2018. Prediction of plant-derived xenomiRs from plant miRNA sequences using random forest and one-dimensional convolutional neural network models. *BMC Genomics* 19: 839.

APPENDIX 1. Deep sequencing data details for the four species.

Species ^a	Genotype	SRA	Tissue	Treatment	Sample size	Genome size
<i>Arabidopsis thaliana</i>	Col-0	SRR1312888	12- to 13-d-old seedlings	Wild type	1.6 Gbp	115 Mbp
	Col-0	SRR1312896	12- to 13-d-old seedlings	Wild type	1.7 Gbp	
	Col-0	SRR1312897	12- to 13-d-old seedlings	Wild type	540.8 Mbp	
	Col-0	SRR1312898	12- to 13-d-old seedlings	AGO10 overexpressor	1.3 Gbp	
	Col-0	SRR1312899	12- to 13-d-old seedlings	AGO10 overexpressor	1.7 Gbp	
	Col-0	SRR1312900	12- to 13-d-old seedlings	AGO10 overexpressor	443.6 Mbp	
<i>Oryza sativa</i>	Nipponbare	SRR1849765	15-d-old leaf	Wild type	202.2 Mbp	373 Mbp
	Nipponbare	SRR1849766	15-d-old leaf	Wild type	206.7 Mbp	
	Nipponbare	SRR1849767	15-d-old leaf	Wild type	181.8 Mbp	
	Nipponbare	SRR1849768	15-d-old leaf	Elicitor treatment for 30 min	138 Mbp	
	Nipponbare	SRR1849769	15-d-old leaf	Elicitor treatment for 30 min	121.3 Mbp	
	Nipponbare	SRR1849770	15-d-old leaf	Elicitor treatment for 30 min	159.4 Mbp	
<i>Zea mays</i>	B73	SRR6939401	Mature leaf (leaf 7)	B73_isolated from induced (id1)_ mature leaves (ML)_small_RNA_replicate1	621.5 Mbp	2.11 Gbp
	B73	SRR6939402	Mature leaf (leaf 7)	B73_isolated from induced (id1) mature leaves (ML)_small_RNA_replicate2	1.1 Gbp	
	B73	SRR6939403	Mature leaf (leaf 7)	B73_wild type (WT)_ mature leaves (ML)_small_RNA_replicate2	700 Mbp	
	B73	SRR6939404	Mature leaf (leaf 7)	B73_wild type (WT)_ mature leaves (ML)_small_RNA_replicate3	1 Gbp	
	B73	SRR6939406	Mature leaf (leaf 7)	B73_wild type (WT)_ mature leaves (ML)_small_RNA_replicate1	731.2 Mbp	
	B73	SRR6939409	Mature leaf (leaf 7)	B73_isolated from induced (id1)_ mature leaves (ML)_small_RNA_replicate3	640.8 Mbp	
<i>Triticum aestivum</i>	Chinese Spring	SRR5461176	Spike (terminal spikelet stage)	Biological replicate 1	676.1 Mbp	14.5 Gbp
	Chinese Spring	SRR5461177	Spike (terminal spikelet stage)	Biological replicate 2	704.3 Mbp	

Note: SRA = Sequence Read Archive.

^aThe PubMed identification number (PMID) of each species is: *Arabidopsis thaliana*, 28231321; *Oryza sativa*, 26083154; *Zea mays*, 29688423; and *Triticum aestivum*, 28515146.

APPENDIX 2. Software-identified known miRNAs in the four species that were experimentally verified.

Species (sample)	Identified miRNA	miRDeep-P2	miRPlant	sRNAbench	miRDeep2	miRExpress	miRkwood	miR-PREFeR	mirnovo
<i>Arabidopsis thaliana</i> (ath_seedling_WT)	ath-miR165	✓	✓	✓		✓	✓	✓	
	ath-miR166		✓	✓		✓	✓	✓	
	ath-miR159	✓	✓	✓		✓	✓	✓	
	ath-miR168	✓	✓	✓		✓	✓	✓	
	ath-miR173			✓		✓	✓		
	ath-miR393	✓	✓	✓		✓	✓	✓	
	ath-miR395	✓	✓	✓		✓	✓	✓	
<i>Oryza sativa</i> (osa_leaf_WT)	osa-miR156		✓	✓	✓	✓	✓	✓	
	osa-miR529			✓		✓	✓	✓	
	osa-miR5078			✓			✓		✓
	osa-miR172		✓	✓	✓	✓	✓		
	osa-miR2863			✓		✓	✓		✓
<i>Zea mays</i> (zma_ML_Treat)	zma-miR399	✓	✓	✓	✓	✓	✓	✓	
	zma-miR156	✓	✓	✓	✓	✓	✓	✓	
	zma-miR166	✓	✓	✓	✓	✓	✓	✓	✓
<i>Triticum aestivum</i> (tae_two_samples)	tae-miR396			✓		✓			
	tae-miR319			✓		✓			
	tae-miR167			✓					
	tae-miR159			✓					