https://doi.org/10.1093/plphys/kiac538 PLANT PHYSIOLOGY 2023: 191: 894–903

Plant Physiology®

Research Article

miRador: a fast and precise tool for the prediction of plant miRNAs

Reza K. Hammond [1,2], Pallavi Gupta [3,4], Parth Patel [1,2] and Blake C. Meyers [4,5,*]

1 Center for Bioinformatics and Computational Biology, University of Delaware, Newark, Delaware 19714, USA
2 Delaware Biotechnology Institute, University of Delaware, Newark, Delaware 19714, USA
3 MU Institute for Data Science and Informatics, University of Missouri, Columbia, Columbia, Missouri 65211, USA
4 Donald Danforth Plant Science Center, St. Louis, Missouri 63132, USA
5 Division of Plant Science and Technology, University of Missouri, Columbia, 52 Agriculture Lab, Columbia, Missouri 65211, USA

*Author for correspondence: bmeyers@danforthcenter.org (B.C.M.)
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/General-Instructions) is Blake C. Meyers (bmeyers@danforthcenter.org).

Abstract

Plant microRNAs (miRNAs) are short, noncoding RNA molecules that restrict gene expression via posttranscriptional regulation and function in several essential pathways, including development, growth, and stress responses. Accurately identifying miRNAs in populations of small RNA sequencing libraries is a computationally intensive process that has resulted in the misidentification of inaccurately annotated miRNA sequences. In recent years, criteria for miRNA annotation have been refined with the aim to reduce these misannotations. Here, we describe miRador, a miRNA identification tool that utilizes the most up-to-date, community-established criteria for accurate identification of miRNAs in plants. We combined target prediction and Parallel Analysis of RNA Ends (PARE) data to assess the precision of the miRNAs identified by miRador. We compared miRador to other commonly used miRNA prediction tools and found that miRador is at least as precise as other prediction tools while being substantially faster than other tools. miRador should be broadly useful for the plant community to identify and annotate miRNAs in plant genomes.

Introduction

Eukaryotic genomes have evolved to encode diverse classes of small noncoding RNA (sRNA) molecules that function in partially overlapping epigenetic silencing pathways. It is believed that sRNAs evolved as a means of defense against RNA viral infections and for silencing transposable elements, and that they later adapted to regulate the expression of endogenous genes (Borges and Martienssen, 2015; Chen et al., 2018). In plants, microRNAs (miRNAs) are a subclass of sRNAs that function to regulate gene expression via posttranscriptional gene silencing, operating in several pathways important to plants, including development, growth, and stress responses. The miRNA biogenesis and miRNA-induced silencing pathways have been extensively studied in Arabidopsis (*Arabidopsis thaliana*), and, until date, these pathways have been found to be well conserved in all land plants that have been examined.

Given the importance of miRNAs in gene regulation, a set of standards was created for the accurate annotation of miRNAs in the organisms in which they are identified. The first effort to define standards for miRNA annotation was published in 2003 and primarily relied on a combination of evidence of both expression and biogenesis (Ambros et al., 2003); criteria for plant miRNA annotation were not explicitly defined, separate from animals. Evidence of expression included, at that time, the accumulation of candidate miRNA in gel blots and the identification of a candidate miRNA in a library of cDNAs made from size-fractionated RNAs. Evidence of biogenesis included the prediction of a fold-back miRNA precursor and the mature sequence mapping entirely to a single arm of that hairpin, conservation of the candidate

Received June 28, 2022. Accepted November 1, 2022. Advance access publication November 28, 2022

miRNA and its predicted precursor secondary structure, and the detection of increased precursor accumulation in *dicer* mutants. By today's standards, it is understood that these requirements alone are insufficient to properly classify miRNAs. In particular, the accumulation of a candidate miRNA does not differentiate it from the other many classes of sRNAs; conservation with a known miRNA assumes that the sequence is required to be a miRNA in the newly studied organism and that the original annotation was correct; in addition, reduced accumulation in *dicer* mutants ignores the partial redundancy of plant *DICER-LIKE* genes (Axtell and Meyers, 2018).

In 2008, another community effort was made to redefine miRNA annotation standards for plants, at that point integrating observations from the then new, deep sequencing technologies in order to reduce false-positive miRNA annotations. These criteria required, for validation of a candidate miRNA, that the miRNA:miRNA* duplex is identified on a hairpin precursor (Meyers et al., 2008; Axtell and Meyers, 2018). Those requirements were summarized into eight specific rules that largely served as the basis of numerous first-generation plant miRNA prediction tools. Subsequently, in 2018, these rules were updated to reflect the changes in understanding of plant miRNAs and their biogenesis that resulted from the massive amounts of data that had accumulated over a decade's worth of sequencing, genomics, and analysis (Table 1). The motivation for making changes to the criteria was to leverage the increased understanding of miRNA biogenesis to further minimize false-positive miRNA annotations, by employing stricter annotation requirements for candidate miRNAs. This update did more than just make the requirements stricter, however; some rules were relaxed to prevent false negatives in miRNA predictions (Axtell and Meyers, 2018).

With the development of these more recent rules, however, it became imperative to develop a computational tool to improve plant miRNA predictions by implementing and enforcing the rules. For this reason, we developed a plant miRNA prediction tool, miRador, that utilizes updated rules to create what we assert is one of the most precise plant miRNA prediction tools available today. In comparison to more commonly used plant miRNA prediction tools, we found that miRador is faster and at least as precise as existing plant miRNA prediction tools. We also developed functionality within miRador to provide users with important conservation insights into predicted miRNAs that other tools lack. In conjunction with sPARTA, a target prediction and validation tool (Kakrana et al., 2014), we found that we can generate high-quality miRNA predictions in a variety of plant species.

Table 1 Past and current criteria for plant miRNA annotations

	2008 Criteria	2018 Criteria
1	One or more miRNA:miRNA* duplexes with two-nucleotide 3′ overhangs	Add requirements that exclude secondary stems or large loops (larger than five nucleotides) in the miRNA:miRNA* duplex and limit precursor length to 300 nucleotides
2	Confirmation of both the mature miRNA and its miRNA*	Disallow confirmation by blot; sRNA-seq only
3	miRNA:miRNA* duplex contains ≤4 mismatched bases	Up to five mismatched positions, only three of which are nucleotides in asymmetric bulges
4	The duplex has at most one asymmetric bulge containing at most two bulged nucleotides	Up to five mismatched positions, only three of which are nucleotides in asymmetric bulges
5	≥75% of reads from exact miRNA or miRNA*	Include one-nucleotide positional variants of miRNA and miRNA* when calculating precision
6	Replication suggested but not required	Required; novel annotations should meet all criteria in at least two sRNA-seq libraries (biological replicates)
7	Homologs, orthologs, and paralogs can be annotated without expression data, provided all criteria met for at least one locus in at least one species	Homology-based annotations should be noted as provisional, pending actual fulfillment of all criteria by sRNA-seq
8	miRNA length not an explicit consideration	No RNAs <20 nucleotide or >24 nucleotides should be annotated as miRNAs. Annotations of 23- or 24-nucleotide miRNAs require extremely strong evidence.

Note: 2008 criteria from Meyers et al., 2008; 2018 criteria from Axtell and Meyers, 2018.

Results

Description of the miRador miRNA prediction strategy

A flowchart depicting the pipeline is depicted in Figure 1. Upon initiating a miRNA prediction run with miRador, the application utilizes the user-provided genome file to identify inverted repeats within each chromosome using einverted (Rice et al., 2000). einverted has several scoring parameters that can be manually set by the user, or the user can choose from three preset miRador options to generate a series of inverted repeats. These inverted repeats will serve as a base set of candidate precursor miRNAs and are stored into a Python dictionary for subsequent analysis.

Small RNA libraries, which can be provided as FASTA, FASTQ, or tag count (i.e. unique sequence and their read count) files, are independently processed from start to finish. Prior to mapping a library to a genome, sRNA sequences are read into another Python dictionary with sRNA sequences as keys and their read count as the attached value. This dictionary is then used to create a FASTA file containing only unique reads from the library. As this file contains only unique reads, mapping time is substantially reduced as each sRNA read will only be mapped to the genome once. Libraries are mapped to the genome with bowtie v1 and

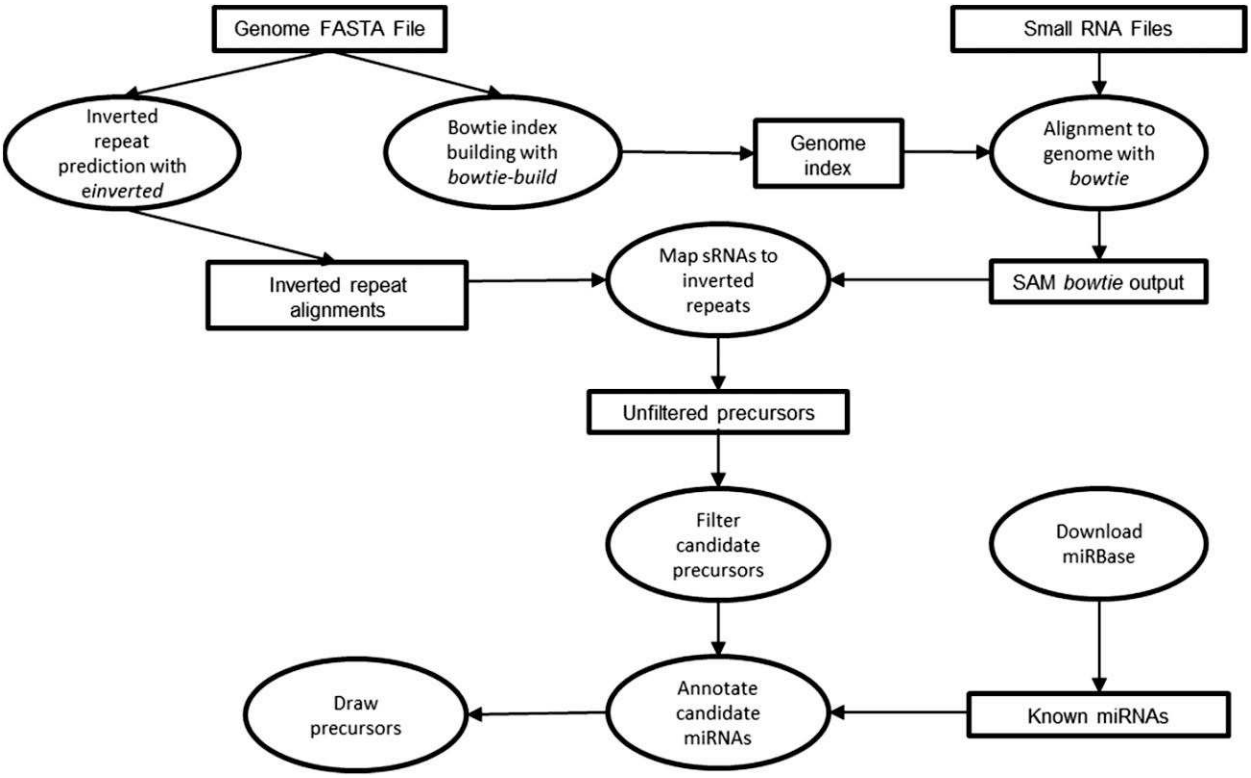

Figure 1 Pipeline of miRador. miRador requires two sets of input files: (1) a genome FASTA file, and (2) sequenced small RNA files. Processing begins with the prediction of inverted repeats on each chromosome with einverted. Small RNA sequences are subsequently mapped to the genome for alignment against inverted repeats, which are the initial set of "candidate precursor miRNAs." These are then filtered utilizing the 2018 plant miRNA annotation criteria. Finally, these candidate miRNAs are annotated utilizing known plant miRNAs from miRBase, and replication requirements are enforced if a candidate miRNA is unique and previously identified.

output in SAM format for immediate processing (Langmead et al., 2009).

This mapped file is parsed into nested Python dictionaries with sRNA positional coordinates as keys and a list of sequences that map to that position as the attached value. The use of dictionaries (Python's built-in implementation of hash tables) is critical due to the data structure's speed at accessing stored data when the index of the stored value is not known. Unlike lists, dictionaries can be queried on a string of characters, including numbers, to find their stored value. The average complexity of a search for a value in an array is $O(n)$, whereas the average complexity of a search of a value in a hash table is $O(1)$. The real-world impact of this design is reflected in the lookup times for cumulative millions of mapped read locations for which sRNAs need to be processed for mapping to inverted repeats. The mapping process is completed with the normalization of reads as reads per million (RPM).

For an inverted repeat to be considered a candidate precursor, a small RNA must map to both arms of the precursor. Thus, miRador iterates through each position on each arm of the inverted repeat to identify any sRNA that lies, in its entirety, on an arm of the inverted repeat. The previously created dictionaries are utilized to identify all sRNAs that map to each inverted repeat. Under the current criteria of plant miRNA annotation, no miRNA can be confirmed without a corresponding miRNA* (where "miRNA*", read as miRNA-star, is the complement in the processed duplex). Thus, any inverted repeat without an sRNA mapping to both arms is immediately removed from the analysis. The remaining inverted repeats are analyzed to identify two sequences on opposite arms that could complete a miRNA:miRNA* duplex at this precursor. The criteria for a miRNA:miRNA* duplex are the following: (1) 2-nt 3' overhangs on the alignment of the candidate miRNA and miRNA*, (2) up to five mismatched positions, only three of which may be nucleotides in asymmetric bulges (G-U pairing assessed ½ a mismatch), and (3) at least 75% of read abundance mapping to the precursor miRNA are from 1-nt positional variants of miRNA and miRNA*. miRador then assesses the alignment parameters by identifying the sRNA sequences on the predicted inverted repeats. If the two alignment criteria are met, then each 1-nt positional variants of both the candidate miRNA and the candidate miRNA* are identified to pool their abundances. If these abundances exceed 75% of the read abundance mapping to this inverted repeat, and the candidate miRNA has an abundance of at least 3 RPM per hit to the genome, then the miRNA:miRNA* duplex will be classified as a candidate miRNA within the library and will be analyzed further in the final step.

Upon the completion of prediction for each library, each confirmed miRNA:miRNA* duplex on the precursor miRNA is drawn utilizing RNAFold (Kerpedjiev et al., 2015). The alignments generated by RNAFold can differ from the einverted alignment due to differences in the scoring systems of the two tools.

Among the features of miRador is its annotation component that classifies candidate miRNAs. There are five classifications to which a candidate miRNA might be assigned: (1) known, (2) identical to known, (3) additional member of existing family, (4) conserved outside the species of study, and (5) unique and previously unidentified. When miRador initiates, it will automatically download all plant miRNAs that have been annotated in the selected version of miRBase, if it has not been downloaded already. In the annotation step, each candidate miRNA is analyzed for sequence similarity to any known miRNA via Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990; Altschul et al., 1997; Camacho et al., 2009). If the species being analyzed exists in miRBase, the General Feature Format (GFF) file is downloaded to determine the locations at which miRNAs and their precursors have been previously identified. A miRNA identified by miRador can only be classified as "known" if the miRNA was identified at the same position at which it exists in miRBase (the terms given in quotation marks here are part of the output of miRador). If the sequence is identical to a known miRNA, but not at the same location of any known miRNA, it is classified as "identical to a known miRNA." A sequence is classified as a "new member of an existing family" if there are five or fewer differences to any known miRNA in the organism of study (a difference is referred to as a gap, mismatch, or bulge, while a G-U wobble is given a half point penalty). BLAST analysis may not find a similar miRNA within the organism of study, but it might identify a match in another plant. If a miRNA is identified as having five or fewer differences to any known miRNA outside of this organism, it will be classified as "conserved outside the organism of study." This classification and "new member of an existing family" within the organism of study are not mutually exclusive, though the classification as a unique and previously undiscovered member within the organism is given precedence and will appear first in the output file. Finally, if none of these classifications fit the candidate miRNA, it will be classified as "novel," that is, unique and previously undescribed. Given the capacity of this function and its general lack of dependency on miRador's execution structure, we made this function available to run as a standalone tool to be used with the results of two of the most used miRNA prediction tools, ShortStack and miRDeep-P2. This tool is available in a GitHub repository: https://github.com/rkweku/mirnaAnnotation.

The final step of miRador filters candidate miRNAs by ensuring unique and previously unidentified miRNA families are predicted independently in multiple sRNA libraries. Upon assignment of the classification of a candidate miRNA in the annotation step, miRador will determine the number of libraries in which the candidate miRNA was predicted, if the candidate miRNA was not conserved within the organism of study. If the number of libraries predicting this candidate miRNA both exceeds one and exceeds 10% of the number of libraries provided for prediction, the candidate miRNA is confirmed by miRador. By requiring a candidate miRNA to be present in at least 10% of libraries, we ensure that miRador does not predict false positives when operating with numerous libraries.

Assessing the predictive capabilities of miRador

To test miRador, we first predicted miRNAs across 21 individual Arabidopsis seedling and flower libraries. miRador identified 228 total miRNAs, 131 of which were already present in miRBase (Table 2). One miRNA was found to be identical to an already known miRNA but at a different position of the genome, 45 were classified as additional members of existing miRNA families, 19 were classified as conserved miRNAs that are known in other organisms, and 32 were classified as unique and previously unidentified. From these results, we were able to determine precision utilizing known miRNAs that are annotated in miRBase. In the case of Arabidopsis, miRador had a precision of 0.575.

We next extended our analysis to 30 rice (*Oryza sativa*) libraries. From these data, miRador identified 391 candidate miRNAs, 252 of which were unique and previously unidentified (Table 3). Of the remaining predicted miRNAs, 76 were known, 2 were identical to a known miRNA at a different position, 45 were classified as additional members of existing families, and 17 were conserved outside of rice. In the case of rice, far more unknown miRNAs were predicted, and therefore the precision here was 0.194.

Subsequent predictions with 44 maize (*Zea mays*) anther, seedling, and tassel libraries identified 173 maize miRNAs (Table 4). miRador identified few miRNAs that were unique and previously identified for maize (16 total). However, nearly half of the total predictions made by miRador were known mature miRNA sequences derived from previously identified *MIRNA* genes (16) as well as additional members of known miRNA families (60). Unlike the Arabidopsis and rice predictions, there were very few predictions of unique and

Table 2 Arabidopsis miRNAs predicted by each tool

Category	miRador	ShortStack	miRDeep-P2
Known	131	60	124
Identical to known miRNAs at different positions	1	0	3
Additional member of existing family	45	7	27
Conserved outside this organism	19	9	5
Unique and previously unidentified (aka "novel")	32	6	8
Precision—sPARTA	0.338	0.488	0.558
Precision—miRBase miRNAs	0.575	0.732	0.743

Number of miRNAs predicted by each miRNA prediction tool in a dataset of 27 Arabidopsis seedling and flower small RNA libraries.

Table 3 Rice miRNAs predicted by each tool

Category	miRador	ShortStack	miRDeep-P2
Known	76	38	111
Identical to known miRNAs at different positions	2	0	3
Additional member of existing family	44	7	141
Conserved outside this organism	17	1	6
Unique and previously unidentified (aka "novel")	252	38	178
Precision—sPARTA	0.373	0.190	0.292
Precision—miRBase miRNAs	0.194	0.452	0.253

Number of miRNAs predicted by each miRNA prediction tool in a dataset of 30 rice anther small RNA libraries.

Table 4 Maize miRNAs predicted by each tool

Category	miRador	ShortStack	miRDeep-P2
Known	81	18	95
Identical to known miRNAs at different positions	16	4	23
Additional member of existing family	60	13	133
Conserved outside this organism	7	0	3
Unique and previously unidentified (aka "novel")	9	6	65
Precision—sPARTA	0.595	0.610	0.611
Precision—miRBase miRNAs	0.757	0.439	0.298

Number of miRNAs predicted by each miRNA prediction tool in a dataset of maize anther, seedling, and tassel small RNA libraries.

previously identified miRNAs—only 9 in total. From these results, we found that miRador had a precision of 0.757.

Comparison to other miRNA prediction tools

To better assess the value of miRador's predictive capabilities, we compared its performance with those of two commonly used plant miRNA prediction tools: ShortStack and miRDeep-P2. ShortStack (Shahid and Axtell, 2014) is a comprehensive analytical tool that classifies mapped sRNA sequencing data. miRDeep-P2 is exclusively a plant miRNA prediction tool and was published as an update to the popular miRDeep-P (Kuang et al., 2018). We used each of the three prediction tools to predict miRNAs in the same Arabidopsis, rice, and maize libraries described above. An instance of miRDeep-P2 performs predictions on a single library, so its predictions across multiple serialized runs were merged to create a single set of predicted miRNAs for all libraries within a dataset. Additionally, miRDeep-P2 utilizes an abundance cutoff, but it bypasses this cutoff if the candidate miRNA differs by up to 1 nucleotide from any miRBase miRNA. Since we wish to address the merits of each of these prediction tools using their default behavior for all candidate miRNAs, we removed this bypass in our assessment. ShortStack and miRDeep-P2 miRNA predictions were also annotated using the miRNA annotation component from miRador for consistency. This approach also has the added benefit of enforcing consistent replication requirements for predictions from miRDeep-P2 as there is no utility for this built into miRDeep-P2, due to its capability to only process single libraries.

miRBase is a repository for miRNA annotations that have been identified in the peer-reviewed literature (Kozomara et al., 2018), though it is important to note that it is not a gatekeeper in enforcing the quality of miRNA annotations (Axtell and Meyers, 2018). Outdated verification criteria and poor-quality sequencing libraries have both been implicated as a cause of improper annotation and limited verification of miRNAs within miRBase (Taylor et al., 2014; Ludwig et al., 2017). Despite these limitations of miRBase, previous miRNA prediction tools have used miRBase miRNAs as a source of true-positive miRNAs (Shahid and Axtell, 2014; Kuang et al., 2018). In addition to assessing the precision of the predictions with miRBase miRNAs, we opted to determine the validity of a candidate miRNA as demonstrated through evidence of cleavage of targets, using PARE (Parallel Analysis of RNA Ends) libraries (German et al., 2009). This allowed us to determine a better set of true and active miRNAs within each prediction set. This method of identifying true miRNAs has the added benefit of assessing the predictability of previously undescribed miRNAs, a useful feature given that these prediction tools exist largely for that purpose—to identify novel miRNAs. We exported predictions made by miRador to evaluate the evidence of mRNA cleavage facilitated by miRador-predicted miRNAs via target prediction and PARE validation with sPARTA (Kakrana et al., 2014). We also tested the validity of this method by comparing the precision of miRNAs predicted by each of the miRNA prediction tools in comparison to an empirical distribution from 100 iterations of 100 randomly sampled 20- to 24-nt sRNAs from each set of sRNA sequencing libraries.

We found that miRador and miRDeep-P2 both identified a similar number of known Arabidopsis miRNAs, 131 and 124, respectively, while ShortStack identified 60, approximately half as many. While miRador found the most known miRNAs, it also predicted far more unique and previously unidentified miRNAs than the other two tools. In terms of precision, miRador was the lowest at 0.338 while miRDeep-P2 was the highest at 0.558 (Table 2). Among these precision values, we found that all three prediction tools had substantially higher precision as compared to the randomly sampled 20- to 24-nt sRNAs (Figure 2A). We also explored the overlap between these predictions to determine if any of these prediction tools were encapsulated by one of the others. In the case of these Arabidopsis libraries, we found that only five of the total 82 miRNAs identified by ShortStack were exclusive to it (Figure 3A). While we did observe some overlap between the predictions made by both miRador and miRDeep-P2, 53.9% of miRador's predictions were not observed by miRDeep-P2, and conversely 37.1% of miRDeep-P2's predictions were not observed by miRador (Table 2).

Further predictions conducted in the 30 rice libraries resulted in slightly contrasting results to the case of Arabidopsis. Here,

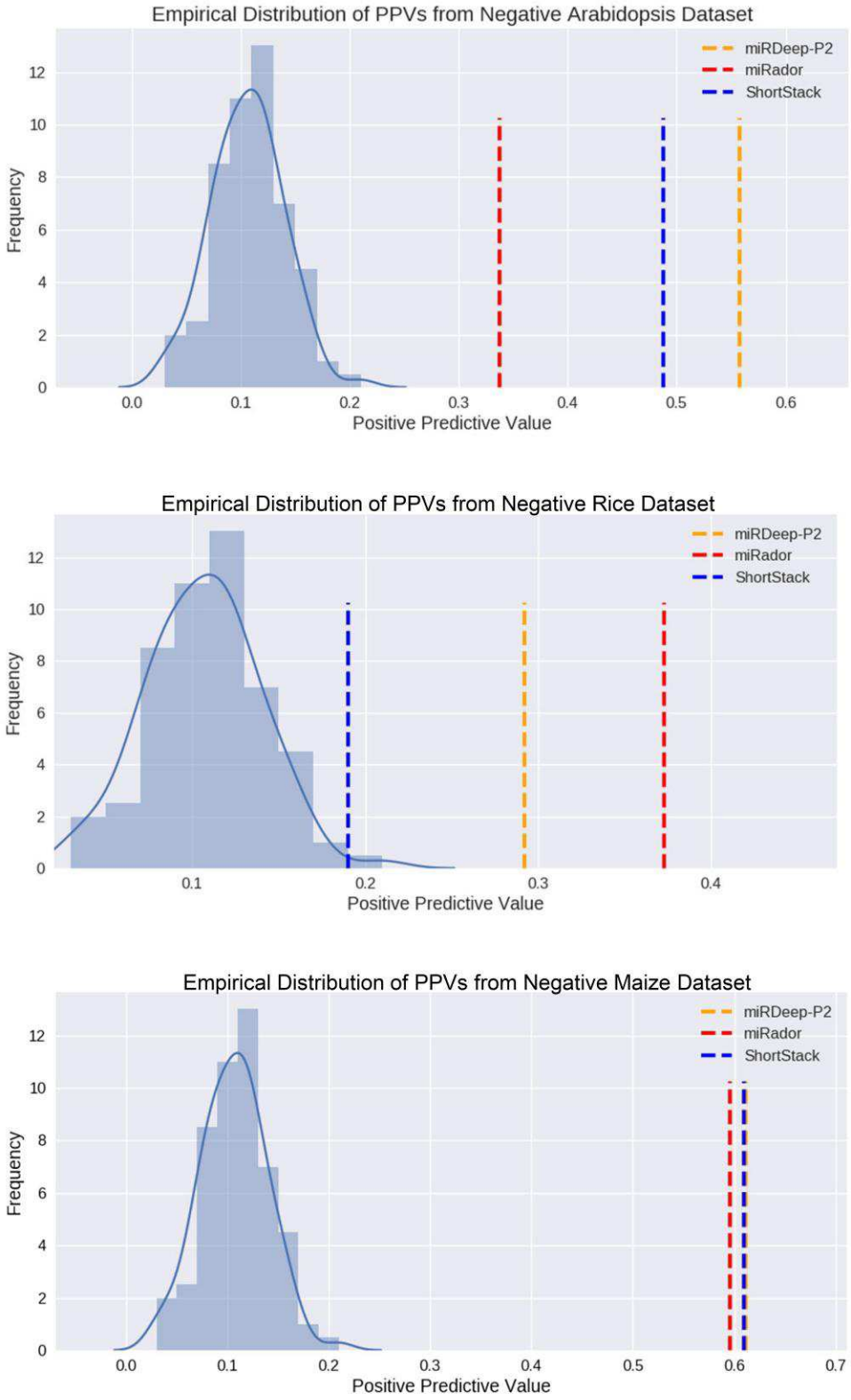

Figure 2 Overlap across the three prediction tools for miRNAs predicted from three plant species. In each panel, the number of distinct candidate miRNAs that are found by each tool, and those predictions that were found in common by the different prediction tools, are indicated as a Venn diagram. A, The number of distinct candidate Arabidopsis miRNAs that were found by each tool, and those predictions that were commonly found by the different prediction tools. B, The number of distinct candidate rice miRNAs that were found by each tool, and those predictions that were commonly found by the different prediction tools. C, The number of distinct candidate maize miRNAs that were found by each tool, and those predictions that were commonly found by the different prediction tools.

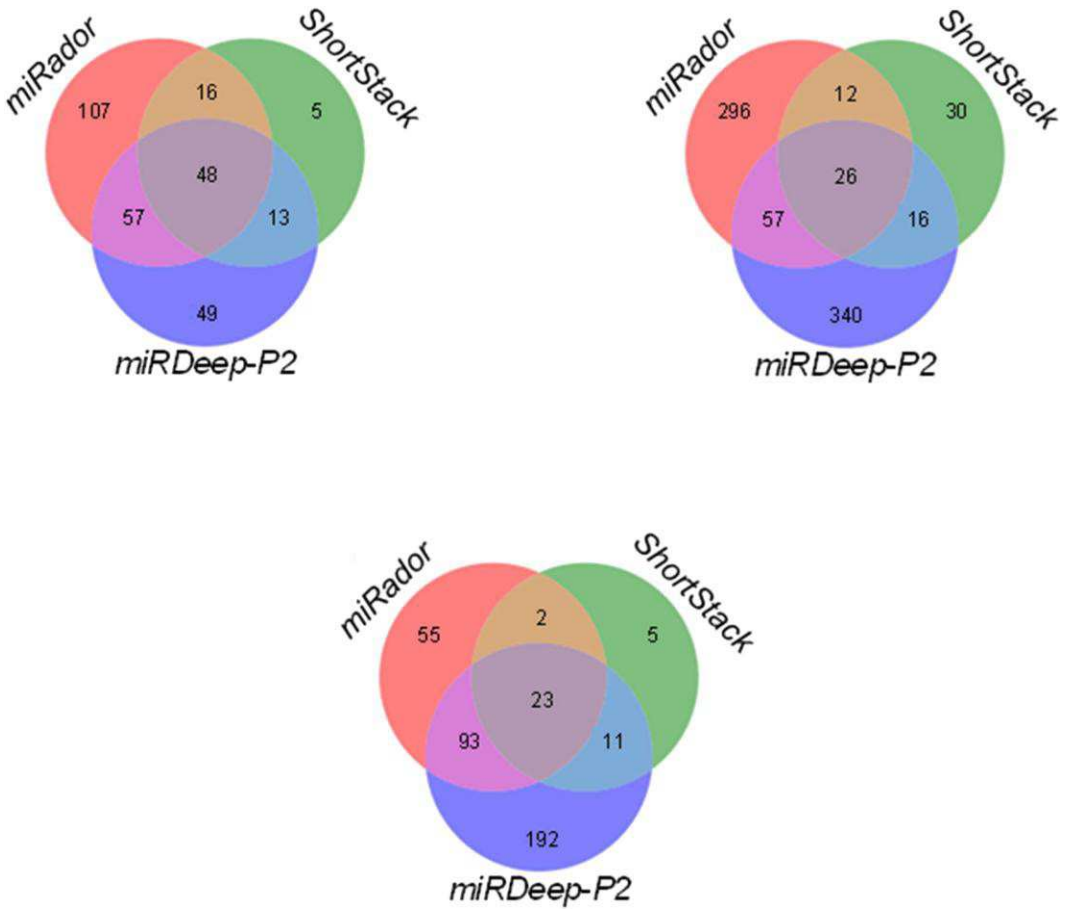

Figure 3 Empirical distribution of precision of negative datasets in comparison to miRNA prediction tools. An empirical distribution of positive predictive values was generated from 100 randomly selected 20- to 24-nt sRNAs separately from our Arabidopsis, rice, and maize datasets. Precision was determined utilizing PARE evidence at predicted targets of the randomly sampled sRNA sequences. A, Empirical distribution of positive predictive values of Arabidopsis-negative datasets in comparison to those of three miRNA prediction tools. B, Empirical distribution of positive predictive values of rice-negative datasets in comparison to those of three miRNA prediction tools. C, Empirical distribution of positive predictive values of maize-negative datasets in comparison to those of three miRNA prediction tools.

miRDeep-P2 predicted the most (439) miRNAs, 178 of which were novel. ShortStack predicted the fewest miRNAs (84), while miRador predicted 357 miRNAs. In the rice libraries, we observed that each tool predicted several miRNAs that no other tool predicted (Figure 3B). The precision of PARE-validated candidate miRNAs predicted by these tools varied greatly; ShortStack only had a precision of 0.190, miRDeep-P2 had a precision of 0.292, and miRador had a precision of 0.373 (Table 3). We also observe that while both miRador and miRDeep-P2 outperform the randomly sampled 20- to 24-nt rice sRNAs, ShortStack however does not (Figure 2B).

We next compared predictions made in 44 maize sRNA libraries and found different trends than in the case of Arabidopsis and rice. Here, miRDeep-P2 predicted the most miRNAs at 319, miRador predicted 228, and ShortStack predicted 41 (Table 4). In this case, however, the precision of PARE-validated candidate miRNAs was nearly identical in each of the three tools, with miRDeep-P2 at the highest at 0.611 and miRador at the lowest at 0.595. The precision of each prediction tool in this case was far beyond the randomly sampled 20- to 24-nt maize sRNAs (Figure 2C).

Finally, we compared the runtimes of each tool with several Arabidopsis, rice, maize, and wheat (*Triticum aestivum*) libraries averaged across three separate runs (Figure 4). We utilized miRador in its sequential run-mode to compare single core executions with ShortStack and miRDeep-P2. Of note, however, is that ShortStack is a tool that discovers and annotates sRNA clusters while also identifying *MIRNA* genes. Given this multifunctionality of ShortStack, we note that a comparison to its runtime is not necessarily one-to-one. However, given that we have utilized ShortStack as a comparison of predictability and it is a commonly utilized tool for predicting miRNAs, we have opted to include its runtimes in our comparison. Each run was performed on a server with an Intel Xeon E5-4620 processor with 256 GB RAM.

In our analysis of runtime performance, we predicted miRNAs in several libraries from four organisms of diverse

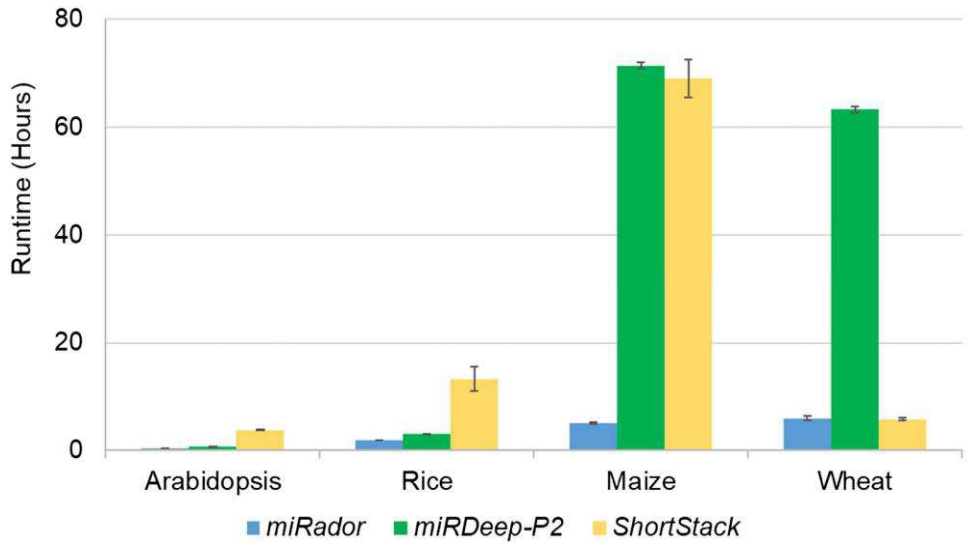

Figure 4 Runtime comparison of miRador to miRDeep-P2 and ShortStack. Comparison of execution times of miRador, miRDeep-P2, and ShortStack across 21 Arabidopsis in three separate runs (116 Mb genome size), 30 rice (364 Mb), 44 maize (2.1 Gb), and 7 wheat (17 Gb) sRNA libraries. Execution was run using default parameters on the same server. Error bars represent standard error between runtimes.

genome sizes: Arabidopsis: 116 MB (21 libraries), rice: 364 MB (30 libraries), maize: 2.1 GB (44 libraries), and wheat: 17 GB (7 libraries). Given that each tool utilizes bowtie to map libraries to the genome and both ShortStack and miRDeep-P2 require bowtie indices to be created prior to execution, we did not include the optional bowtie-build step of miRador in these runtimes. As mentioned before, miRDeep-P2 only processes one library per instance, so its runtime is the summation of multiple serialized across all test libraries. In our assessment of Arabidopsis, rice, and maize libraries, we observed that miRador was the fastest tool. Notably, however, the quick runtime of ShortStack in its prediction of wheat miRNAs is relative to its runtime in the other organisms. ShortStack's runtime may be beneficial when there are fewer libraries as an input, and potentially led to it being the fastest tool at predicting wheat miRNAs at 5.77 h. miRador was only slightly longer at 5.97 h while miRDeep-P2 was the slowest at 63.27 h. The runtime of miRador for predicting maize miRNAs was particularly strong, as its predictions across 44 sRNA libraries completed in an average of 5.05 h, whereas both other tools took over 3 d to complete. Overall, we found that the runtime of miRador was, in our opinion, impressively quick. It outperformed the other tools in nearly every tested case, and in the case of wheat, it was nearly as fast as ShortStack. Its demonstrated scalability, supporting an ability to predict miRNAs in large genomes with many input libraries. Additionally, miRador has the ability to utilize multiple cores to improve prediction times, which can enable far larger analyses than those that we tested.

Discussion

In this paper, we describe miRador, a plant miRNA prediction tool that utilizes the most recent community-developed plant miRNA annotation criteria. Unlike previous studies, we utilized PARE libraries to assess the quality of miRNA predictions beyond just the sequences that exist in miRBase, giving a better representation of the predictability of miRNA prediction tools. In addition to its strong predictive capabilities, we showed that miRador is faster than existing tools without compromising predictive efficiency. Additionally, we developed an annotation function to annotate miRNAs with respect to their presence in miRBase, and similarity to known miRNAs, and we exported this function for use by other miRNA prediction tools.

In assessing the utility of miRador for miRNA predictions, we used PARE to identify the candidate miRNAs that have evidence of cleavage at their predicted targets, as identified by sPARTA. We utilized sRNA and PARE libraries from Arabidopsis, rice, and maize to identify the precision of miRador, ShortStack, and miRDeep-P2. ShortStack consistently identified the fewest miRNAs, though its predictions were not completely encapsulated by the predictions of the other two tools. miRDeep-P2 and miRador had large overlaps in all organisms, but each tool also uniquely predicted numerous miRNAs. Overall, our findings suggest that each tool may be used to predict miRNAs with similar precision to one another, but none of these tools are all encapsulating—that is, none identify the broadest and most complete set of candidate miRNAs. Thus, there may be utility in running more than one tool when predicting miRNAs in a set of sRNA libraries.

In our attempt to validate candidate miRNA activity by target prediction and authentication with PARE, we utilized PARE from corresponding tissues from which the sRNA data were acquired. Surprisingly, despite utilizing well-staged, low-input sRNA and PARE sequencing libraries generated in

triplicates from small amounts of tissue (Jiang et al., 2020), we found that this dataset showed the poorest precision of the three datasets. Our belief that these tissue-specific datasets may provide cleaner results with less noise in the PARE data than what was found in the Arabidopsis and maize libraries ultimately did not prove to be true. We even found that the precision of candidate miRNAs predicted by ShortStack did not differ from randomly sampled 20- to 24-nt sRNAs. As with precision of PARE-validated candidate miRNAs, precision when utilizing miRBase miRNAs as true positives varied among the three tools depending on the dataset. There appear to be cases where each tool could be viable depending on the input dataset (Tables 2–4).

Although we are confident in the quality of miRador, it is not without limitations. In its identification of candidate miRNA genes, miRador first searches for inverted repeats in a genome assembly using einverted. This worked well using high-quality genome builds with which we performed the tests, but miRador may miss miRNAs when predicting in newly sequenced genomes comprised of several disconnected scaffolds and contigs. Presumably, as increasingly complete, long-read-based genome assemblies become the norm, this weakness will be mitigated. We also largely utilized the community-established guidelines for plant miRNA annotations; these guidelines could be fine-tuned even further with machine learning to minimize false positives and false negatives. These additional methods would also allow for confidence scores to be assigned to the resulting predictions. We were aware of these limitations when building miRador, and thus it was largely developed with modular functions such that adjustments to its prediction filters can be made without overhauling the entire tool.

We recognize there is a substantial lack of overlap of candidate miRNAs predicted by each of the three tools tested. We hypothesize that there are fundamental differences in the implementation of these pipelines that result in these differences. As discussed previously, miRador starts with a set of candidate precursors based on the predicted inverted repeats utilizing the reference genome. miRDeep-P2 maps sRNAs to the genome first and then predicts potential precursor sequences from the reference using aligned reads as a guideline. ShortStack exists as a multipurpose sRNA gene annotation tool largely using clusters of sRNAs to predict different types of sRNA genes, including miRNAs. Given that no tool performed best across the board, we believe there are merits to each method and users may find maximal utility when combining the results across multiple prediction tools.

Despite these limitations, we assert that we have developed a tool that is as precise and sensitive as other plant miRNA prediction tools while being far faster. miRador is highly scalable, ensuring its ability to predict miRNAs with large genomes with many sRNA libraries. The additional annotation component of miRador, which has been exported for use by other prediction tools, provides users great insights into the status of their predicted miRNAs as either known or not. Altogether, miRador is a highly capable, standalone, plant miRNA prediction and annotation tool.

Materials and methods

Software and data availability

A description of this pipeline and overall algorithm are described in the "Results" section above. The entire miRador pipeline is available on GitHub: https://github.com/rkweku/miRador. Detailed installation and usability information are included in the README file. To improve the user experience, we've included a conda environment file that can be setup with anaconda or miniconda, and we have also included test data that can be utilized as a model when running de novo analyses.

We utilized public Arabidopsis (*Arabidopsis thaliana*), rice (*Oryza sativa*), maize (*Zea mays*), and wheat (*Triticum aestivum*) sRNA and PARE datasets for miRNA prediction and validation. These libraries, their GEO accession numbers, and sequencing information are listed in Supplemental Table 1.

Accession numbers

See Supplemental Table 1 for accession numbers of data used in this study.

Supplemental data

The following materials are available in the online version of this article.

Supplemental Table S1. Library Information.

Funding

This work was supported by funding provided by the National Science Foundation awards #1754097 and 2130883 to B.C.M., and resources provided by the Donald Danforth Plant Science Center and the University of Missouri—Columbia.

Acknowledgments

We would like to thank Joanna Friesner and Michael Axtell for comments on the manuscript. We thank members of the Meyers lab for helpful discussions.

Author contributions

Work conceived and designed by R.K.H and B.C.M.; software written and tested by R.K.H. with some components contributed by P.G. and P.P.; R.K.H and B.C.M. wrote the manuscript with contributions from other authors.

Conflict of interest statement. None declared.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Molecular Biol **215**(3): 403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**(17): 3389–3402

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, et al. (2003) A uniform system for microRNA annotation. RNA **9**(3): 277–279

Axtell MJ, Meyers BC (2018) Revisiting criteria for plant microRNA annotation in the era of big data. Plant Cell **30**(2): 272–284

Borges F, Martienssen RA (2015) The expanding world of small RNAs in plants. Nat Rev Mol Cell Biol **16**(12): 727–741

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics **10**(1): 421

Chen C, Zeng Z, Liu Z, Xia R (2018) Small RNAs, emerging regulators critical for the development of horticultural traits. Hortic Res **5**(1): 63

German MA, Luo S, Schroth G, Meyers BC, Green PJ (2009) Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. Nat Protoc **4**(3): 356–362

Jiang P, Lian B, Liu C, Fu Z, Shen Y, Cheng Z, Qi Y (2020) 21-nt phasiRNAs direct target mRNA cleavage in rice Male germ cells. Nat Commun **11**(1): 5191

Kakrana A, Hammond R, Patel P, Nakano M, Meyers BC (2014) sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. Nucleic Acids Res **42**(18): e139–e139

Kerpedjiev P, Hofacker IL, Hammer S (2015) Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. Bioinformatics **31**(20): 3377–3379

Kozomara A, Birgaoanu M, Griffiths-Jones S (2018) miRBase: from microRNA sequences to function. Nucleic Acids Res **47**(D1): D155–D162

Kuang Z, Wang Y, Li L, Yang X (2018) miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. Bioinformatics **35**(14): 2521–2522

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10**(3): 25

Ludwig N, Becker M, Schumann T, Speer T, Fehlmann T, Keller A, Meese E (2017) Bias in recent miRBase annotations potentially associated with RNA quality issues. Sci Rep **7**(1): 5162

Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington J, Chen X, Green PJ, et al. (2008) Criteria for annotation of plant microRNAs. Plant Cell **20**(12): 3186–3190

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet **16**(6): 276–277

Shahid S, Axtell MJ (2014) Identification and annotation of small RNA genes using ShortStack. Methods **67**(1): 20–27

Taylor RS, Tarver JE, Hiscock SJ, Donoghue PCJ (2014) Evolutionary history of plant microRNAs. Trends Plant Sci **19**(3): 175–182