Journal of Visualized Experiments

# A Bioinformatics Pipeline to Accurately and Efficiently Analyze the MicroRNA Transcriptomes in Plants

--Manuscript Draft--

| | |
|---|---|
| **Article Type:** | Methods Article - JoVE Produced Video |
| **Manuscript Number:** | JoVE59864R1 |
| **Full Title:** | A Bioinformatics Pipeline to Accurately and Efficiently Analyze the MicroRNA Transcriptomes in Plants |
| **Section/Category:** | JoVE Biology |
| **Keywords:** | microRNA (miRNA); plant; sRNA-seq; miRDeep-P2 (miRDP2); Next Generation Sequencing; plant miRNA criteria; miRDeep-P (miRDP) |
| **Corresponding Author:** | Xiaozeng Yang, Ph.D<br>Beijing Academy of Agriculture and Forestry Sciences<br>Beijing, Beijing CHINA |
| **Corresponding Author's Institution:** | Beijing Academy of Agriculture and Forestry Sciences |
| **Corresponding Author E-Mail:** | yangxiaozeng@baafs.net.cn |
| **Order of Authors:** | Ying Wang |
| | Zheng Kuang |
| | Lei Li |
| | Xiaozeng Yang, Ph.D |
| **Additional Information:** | |
| **Question** | **Response** |
| Please indicate whether this article will be Standard Access or Open Access. | Standard Access (US$2,400) |
| Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations. | Beijing, China |

August 26st, 2019

Dear Dr. Steindel,

Sorry for the late resubmission since our first author had a serious family issue which delayed the whole process.

Thank you very much for reviewing our manuscript entitled “A Bioinformatics Pipeline to Accurately and Efficiently Analyze the MicroRNA Transcriptomes in Plants”. We have carefully examined the comments from the editor and reviewers, and were able to address all the issues raised in the revised manuscript and the accompanying user’s manual. I hope you will find these revisions satisfactory.

Thank you for your consideration.

Xiaozeng Yang, Ph.D.
Professor of Computational Biology
Beijing Academy of Agriculture and Forestry Sciences

## TITLE:

# A Bioinformatics Pipeline to Accurately and Efficiently Analyze the MicroRNA Transcriptomes in Plants

## AUTHORS AND AFFILIATIONS:

Ying Wang[1,2]*, Zheng Kuang[1,2]*, Lei Li[2], Xiaozeng Yang[1]

[1]Beijing Key Laboratory of Agricultural Genetic Resources and Biotechnology, Beijing Agro-biotechnology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China
[2]State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, School of Advanced Agricultural Sciences and School of Life Sciences, Peking University, Beijing, China

*These authors contributed equally.

Email addresses of co-authors:
Ying Wang (ying.wang87@pku.edu.cn)
Zheng Kuang (kuangzheng@pku.edu.cn)

Corresponding author:
Lei Li (lei.li@pku.edu.cn)
Xiaozeng Yang (yangxiaozeng@baafs.net.cn)

## KEYWORDS:

microRNA (miRNA), plant, sRNA-seq, miRDeep-P2 (miRDP2), Next generation sequencing, plant miRNA criteria, miRDeep-P (miRDP)

## SUMMARY:

A bioinformatics pipeline, namely miRDeep-P2 (miRDP2 for short), with updated plant miRNA criteria and an overhauled algorithm, could accurately and efficiently analyze microRNA transcriptomes in plants, especially for species with complex and large genomes.

## ABSTRACT:

MicroRNAs (miRNAs) are 20- to 24-nucleotide (nt) endogenous small RNAs (sRNAs) extensively existing in plants and animals that play potent roles in regulating gene expression at the post-transcriptional level. Sequencing sRNA libraries by Next Generation Sequencing (NGS) methods has been widely employed to identify and analyze miRNA transcriptomes in the last decade, resulting in a rapid increase of miRNA discovery. However, two major challenges arise in plant miRNA annotation due to increasing depth of sequenced sRNA libraries as well as the size and complexity of plant genomes. First, many other types of sRNAs, in particular, short interfering RNAs (siRNAs) from sRNA libraries, are erroneously annotated as miRNAs by many computational tools. Second, it becomes an extremely time-consuming process for analyzing miRNA transcriptomes in plant species with large and complex genomes. To overcome these challenges, we recently upgraded miRDeep-P (a popular tool for miRNA transcriptome analyses) to miRDeep-P2 (miRDP2 for short) by employing a new filtering strategy, overhauling the scoring algorithm and incorporating newly updated plant miRNA annotation criteria. We tested miRDP2 against sequenced sRNA populations in five representative plants with increasing genomic complexity, including Arabidopsis, rice, tomato, maize and wheat. The results indicate that miRDP2 processed these tasks with very high efficiency. In addition, miRDP2 outperformed other prediction tools regarding sensitivity and accuracy. Taken together, our results demonstrate miRDP2 as a fast and accurate tool for analyzing plant miRNA transcriptomes, therefore a useful tool in helping the community better annotate miRNAs in plants.

## INTRODUCTION:

One of the most exciting discoveries in the last two decades in biology is the proliferating role of sRNA species in regulating diverse functions of the genome[1]. In particular, miRNAs constitute an important class of 20- to 24-nt sRNAs in eukaryotes, and mainly function at post-transcriptional level as prominent gene regulators throughout life cycle development stages as well as in stimulus and stress responses[2,3]. In plants, miRNAs arise from primary transcripts called pri-miRNAs, which are generally transcribed by RNA polymerase II as individual transcription units[4,5]. Processed by evolutionarily conserved cellular machinery (Drosha RNase III in animals, DICER-like in plants), pri-miRNAs are excised into the immediate miRNA precursors, pre-miRNAs, which contain sequences forming intra-molecular stem-loop structures[6,7]. Pre-miRNAs are then processed into double-stranded intermediates, namely miRNA duplexes, consisting of the functional strand, mature miRNA, and the less frequently functional partner, miRNA*[2,8]. After loaded into the RNA-induced silencing complex (RISC), the mature miRNAs could recognize their mRNA targets based on sequence complementarity, resulting in a negative regulatory function[2,8]. miRNAs could either destabilize their target transcripts or prevent target translation but the former manner is dominated in plants[8,9].

Since the fortuitous discovery of the first miRNA in the nematode *Caenorhabditis elegans*[10,11], much research has been committed to miRNA identification and its functional analysis, especially after the availability of NGS method. The wide application of the NGS method has greatly promoted the utilization of computational tools that were designed to capture the unique feature of miRNAs, such as the stem-loop structure of precursors and their preferential accumulation of sequence reads on mature miRNA and miRNA*. As a result, researchers have achieved remarkable success in identifying miRNAs in diverse species. Based on a previously described probability model[12], we developed miRDeep-P[13], which was the first computational tool for discovering plant miRNAs from NGS data. miRDeep-P was specifically aimed at conquering the challenges of decoding plant miRNAs featuring more variable precursor length and large paralogous families[13-15]. After its release, this program has been downloaded thousands of times and used to annotate miRNA transcriptomes in more than 40 plant species[16]. Propelled by NGS-based tools like miRDeep-P, there has been a dramatic increase in the number of registered miRNAs in the public miRNA repository miRBase[17], where over 38,000 miRNA items are currently hosted (release 22.1) in comparison to only ~500 miRNA items (release 2.0) in 2008[18].

However, two new challenges have arisen from plant miRNA annotation. First, high ratios of false-positives have heavily impacted the quality of plant miRNA annotations[16,19] for the following reasons: 1) a deluge of endogenous short interfering RNAs (siRNAs) from NGS sRNA libraries were erroneously annotated as miRNAs due to lacking of a stringent miRNA annotation criteria; 2) for species without a priori miRNA information, false-positives predicted based on NGS data are difficult to eliminate. Using miRBase as an example, Taylor et al.[20] found one third of plant miRNA entries in the public repository[21] (release 21) lacked convincing supporting evidence and even three-fourths of plant miRNA families were questionable. Second, it becomes an extremely time-consuming process for predicting plant miRNAs with large and complex genomes[16]. To overcome these challenges, we updated miRDeep-P by adding a new filtering strategy, overhauling the scoring algorithm and integrating new criteria for plant miRNA annotation, and released the new version miRDP2. In addition, we tested miRDP2 using NGS sRNA datasets with gradually increasing genome sizes: Arabidopsis, rice, tomato, maize and wheat. Compared to other five widely used tools and its old version, miRDP2 parsed these sRNA data and analyzed miRNA transcriptomes faster with improved accuracy and sensitivity.

## Contents of the miRDP package

The miRDP2 package consists of six documented Perl scripts that should be run sequentially by the prepared bash script. Of the six scripts, three (*convert_bowtie_to_blast.pl*, *filter_alignments.pl*, and *excise_candidate.pl*) are inherited from miRDeep-P. The other scripts are modified from the original version. Functions of the six scripts are described in the following:

*preprocess_reads.pl* filters input reads, including reads that are too long or too short (<19 nt or >25 nt), and reads correlated with Rfam ncRNA sequences, as well as reads with RPM (Reads Per Million) less than 5. The script then retrieves reads correlated to known miRNA mature sequences. The input files are original reads in FASTA/FASTQ format and bowtie2 output of reads mapping to miRNA and ncRNA sequences.

The formula for calculating RPM is as the following:

$$\text{RPM of a miRNA} = \frac{\text{Number of reads mapped to a miRNA (mature part)} \times 10^6}{\text{Total number of mapped reads from a given library}}$$

*convert_bowtie_to_blast.pl* changes the bowtie format into BLAST-parsed format. BLAST-parsed format is a custom tabular separated format derived from standard NCBI BLASToutput format.

*filter_alignments.pl* filters the alignments of deep sequencing reads to a genome. It filters partial alignments as well as multi-aligned reads (user-specified frequency cutoff). The basic input is a file in BLAST-parsed format.

*excise_candidate.pl* cuts out potential precursor sequences from a reference sequence using aligned reads as guidelines. The basic input is a file in BLAST-parsed format and a FASTA file. The output is all potential precursor sequences in FASTA format.

*mod-miRDP.pl* needs two input files, signature file and structure file, which is modified from the core miRDeep-P algorithm by changing the scoring system with plant specific parameters. The input files are dot-bracket precursor structure file and reads distribution signature file.

*mod-rm_redundant_meet_plant.pl* needs three input files: chromosome_length, precursors and original_prediction generated by mod-miRDP.pl. It generates two output files, non-redundant predicted file and predicted file filtered by newly updated plant miRNA criteria. Details on the format of output file are described in section 1.4.

## PROTOCOL:

### 1. Installation and testing

1.1. Download required dependencies: Bowtie2[22] and RNAfold[23]. Compiled packages are recommended.

1.1.1. Download Bowtie2, a read mapping tool, from its home site (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml).

1.1.2. Download RNAfold, a tool of the Vienna package used to predict RNA secondary structure, from http://www.tbi.univie.ac.at/~ivo/RNA/.

1.1.3. Before installing miRDP2, ensure that these two dependencies are correctly installed, and customize the bash environment file (e.g., .bashrc) to set a correct PATH for these two dependencies.

NOTE: Other mapping tools such as Bowtie[24] are also suitable to miRDP2; either Bowtie or Bowtie2 can be used after version 1.1.3.

1.2. To download the miRDP2 package, go to https://sourceforge.net/projects/mirdp2/files/latest_version/ and fetch the tarball files.

1.3. Before installing miRDP2, make sure that Perl is in the PATH. To install miRDP2, extract all contents of the downloaded tarball file into one folder (command lines as in 1.4.2), and then set the folder path into the PATH.

NOTE: A computer or computing node with at least 8 GB RAM and 100 GB storage are recommended to run miRDP2.

1.4. Test the MiRDP2 pipeline.

1.4.1. To test whether miRDP2 has been correctly installed, use the test data and the expected output found in https://sourceforge.net/projects/mirdp2/files/TestData/. Test data contain one formatted GSM sequencing file and one *Arabidopsis thaliana* genome file.

1.4.2. Move all downloaded files to the current working directory:

```
mv miRDP2-v*.tar.gz TestData.tar.gz ncRNA_rfam.tar.gz <user_selected_folder>
cd <user_selected_folder>
```

1.4.3. Extract the compressed tarball files:

```
tar –xvzf miRDP2-v*.tar.gz
tar –xvzf TestData.tar.gz
tar –xvzf ncRNA_rfam.tar.gz
```

1.4.4. Build the Arabidopsis genome reference index:

```
bowtie2-build -f ./TestData/TAIR10_genome.fa ./TestData/TAIR10_genome
```

1.4.5. Build the ncRNA reference index:

```
bowtie2-build -f ./ncRNA_rfam.fa ./1.1.3/script/index/rfam_index
```

1.4.6. Run the miRDP2 pipeline:

```
bash ./1.1.3/miRDP2-v1.1.3_pipeline.bash –g ./TestData/TAIR10_genome.fa -i ./ TestData/TAIR10_genome –f ./TestData/GSM2094927.fa –o .
```

NOTE: Linux commands used are in bold and italic fonts, with command line options in italics. *indicates the version of miRDP2 (the current version is 1.1.3). The bowtie2-build command should take roughly 10 minutes, and the miRDP2 pipeline should finish within several minutes

1.5. Check testing outputs.

1.5.1. Note that a folder named ‘GSM2094927-15-0-10’ is automatically generated in *<user_selected_folder>*, containing all intermediate files and results.

1.5.2. Check that the tab-delimited output file GSM2094927-15-0-10_filter_P_prediction, the final output of predicted miRNAs, contains columns that indicate chromosome id, strand direction, representative reads id, precursor id, mature miRNA location, precursor location, mature sequence, and precursor sequence. Note the additional bed file derived from this file to facilitate further analysis.

1.5.3. Check the file “progress_log”, which provides information about finished steps, and the files “script_log” and “script_err”, that contain program output and warnings.

NOTE: Currently, we have tested miRDP2 on two Linux platforms, including CentOS release 6.5 on a cluster server, and Cygwin 2.6.0 on PC Windows system, and miRDP2 should work on similar systems that support Perl.

## 2. **Identifying novel miRNAs**

2.1. Before running the pipeline, ensure that the input reads are preprocessed into proper format.

NOTE: The new version 1.1.3 of miRDP2 can accept original FASTQ format files as inputs, although the process of formatting reads is carried out as in previous versions.

2.1.1. First, remove adapters from the 5’ and 3’ ends of the deep sequencing reads (if present).

2.1.2. Second, parse the deep sequencing reads into FASTA format.

2.1.3. Third, remove redundancy such that reads with identical sequence are represented with a single and unique FASTA entry.

2.1.4. Finally, ensure that all of the FASTA identifiers are unique. Each sequence identifier must end with a '_x' and an integer, indicating the copy number of the exact sequence that was retrieved in the deep sequencing datasets. One way to ensure unique FASTA identifier is to include a running number in the ID. For reference, see the file GSM2094927.fa in the test data (https://sourceforge.net/projects/mirdp2/files/TestData/).

2.1.5. See the following for examples of correctly formatted reads:

```
>read0_x29909
TTTGGATTGAAGGGAGCTCTA
>read1_x36974
TTCCACAGCTTTCTTGAACTG
>read2_x32635
TTCCACAGCTTTCTTGAACTT
```

2.2. Build reference indices.

2.2.1. For the genome reference, to save time, download Bowtie2 index files from the iGenomes website (https://support.illumina.com/sequencing/sequencing_software/igenome.html) if the genome sequences of the species of interest have been indexed. Otherwise, users index reference sequences and keep the index file for a while till the project is finished since the genome sequence might need to be re-indexed. Details on how to index a genome reference are included in bowtie2 manual (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml).

2.2.2. Another non-miRNA ncRNA index is also needed to filter out noisy sequences from other non-coding RNA fragments. The file is a collection of main ncRNA sequences from Rfam, including rRNA, tRNA, snRNA, and snoRNA. To build this index, please refer to part 1.4, as the index should be placed and named correctly, i.e. <miRDP2_version>/script/index/rfam_index.

2.3. Run miRDP2.

2.3.1. To use miRDP2 to detect new miRNAs from deep sequencing data, run the bash script in the package to start the analysis pipeline (An example can be found in step 1.4):

```
<path_to_miRDP2_folder>/miRDP2-v*.*_pipeline.bash –g <genome_file> -i <path_to_index/index_prefix> -f <seq_file > -o <output_folder>
```

where * indicates the version of the pipeline bash script. There are three parameters that can be modified: 1) the number of different locations a read could be mapped to, 2) the mismatch number for running bowtie2, and 3) the threshold of RPM (Reads Per Million). Modify these using the –L, -M, and –R options, respectively. A detailed explanation is in section 3.1.

2.4. Check the miRDP2 outputs.

2.4.1. Note that the output folder will be automatically generated under <output_folder>, and named '<seq_file_name>-15-0-10'; the last 3 numbers indicate the values (default in this case) for parameters 1, 2, and 3, respectively. The file <seq_file_name>_filter_P_prediction contains information of the final predicted miRNAs satisfying the newly updated plant miRNA annotation criteria. Details on the format of output file are described in part 1.4.

## 3. Modifications and caution using miRDP2

### 3.1. Parameters that can be modified

3.1.1. Use the '-L' option to set the limit of how many locations a read could be mapped to (parameter 1). Read mapping to too many sites are possibly associated with repeat sequences, and are not likely to miRNAs. The default setting is 15. For specific species, if there are miRNA families with many members, the first parameter may be increased manually to adapt to the genome landscape.

3.1.2. Use the '-M' option to set the allowed mismatches for bowtie (parameter 2). The default setting is 0.

3.1.3. Use the '-R' option to set the threshold for reads potentially corresponding to mature miRNAs (parameter 3). To reduce time consumption and false-positives, filter reads by RPM. Only reads exceeding a certain RPM threshold may represent mature sequences of miRNAs rather than background noise, and would be kept for further analysis. The default setting is 10 RPM.

3.1.4. Note that changing these parameters can potentially affect performance and time consumption. In general, an increase of parameter 1 and 2 and a decrease of parameter 3 would generate a less stringent result and longer running time and vice versa.

### 3.2. **Redundancy and miRNA***

3.2.1. Note that the output miRNAs from miRDP2 may differ from the known miRNAs. We found that this is mainly due to one of two reasons: heterogeneity of the mature miRNAs or the relative abundance of miRNA and miRNA*. We found that this does not impact the optimal length selection of precursors and the profiling of known miRNA genes.

## REPRESENTATIVE RESULTS:

The miRNA annotation pipeline, miRDP2, described herein is applied to 10 public sRNA-seq libraries from 5 plant species with gradually increased genome length, including *Arabidopsis thaliana*, *Oryza sativa* (rice), *Solanum lycopersicum* (tomato), *Zea mays* (maize) and *Triticum aestivum* (wheat) (**Figure 1A**). Overall, for each species, 2 representative sRNA libraries from different tissues (collapsed into unique reads, details in the protocol section) and their indexed genome sequences are processed as two inputs (**Table 1**). Five miRNA computational prediction tools (miRDeep-P[13], miRPlant[25], miR-PREFeR[26], miRA[27], miReNA[28]) were selected to make the comparison.

### Running time test

To compare the runtime and performance of miRDP2 and other five tools, we installed five tools (miRDP2, miRDeep-P, miR-PREFeR, miRA, and miReNA) in a cluster server with Cent OS release 6.5 system. These programs were run with the same input files, hardware and resource (details in **Supplementary File 1**). Especially, miRPlant is controlled from a GUI written in Java and was not able to run on the server. Instead, we tested miRPlant on a PC with Windows 10 while we have also tested miRDP2 and miRDeep-P on this PC (details in **Supplementary File 1**).

For small genome species as *Arabidopsis thaliana*, *Oryza sativa*, and *Solanum lycopersium*, all the programs ran properly. However, for large genomes species such as *Zea mays* and *Triticum aestivum* (including *Solanum lycopersium* for miRA), some of the programs depleted all computing resources and broke down halfway. For instance, miReNA, miRA, and miR-PREFeR failed to generate results, probably due to memory deficiency while dealing with large sam files or intermediate files. In particular, miRPlant temporary files consumed too much space, and the result was not able to run on the PC when dealing with large genome species. miRDP2 finished these prediction processes in a very short time, from minutes to hours (**Figure 1B**). Thus, compared to its old version and other tools, the running time of miRDP2 was markedly shortened.

### Sensitivity and accuracy test

Since miRNAs in Arabidopsis are intensively studied, we made use of known miRNAs in Arabidopsis in miRBase[21] (release 22.1) to evaluate miRDP2, and made the comparison with other tools. As previously reported[19,26], the following formulas are employed to calculate sensitivity and accuracy:

$$\text{Sensitivity} = \frac{\text{Known expressed miRNAs No.}}{\text{Total expressed miRNAs No.}}$$

$$\text{Accuracy} = \frac{\text{Predicted Known miRNAs No.}}{\text{Expressed Known miRNAs No.}}$$

Known miRNAs are those annotated in miRBase. A miRNA is designated as expressed if the mature sequences have more than 5 RPM, and ≥75% reads on the precursor mapped to mature and star miRNA sequences. Two sequenced sRNA libraries from Arabidopsis (**Table 1**) were used to make the test. miRDP2 (**Figure 1C,D**) performed better in both sensitivity and accuracy compared to other tools.

Taken together, these results demonstrate that miRDP2 is a fast and accurate tool for analyzing the miRNA transcriptome in plants.

## FIGURE AND TABLE LEGENDS:

**Figure 1: Performance of miRDP2.** (**A**) Genome size (in Gb) of *Arabidopsis thaliana* (*Ath*), *Oryza sativa* (*Osa*), *Solanum lycopersicum* (*Sly*), *Zea mays* (*Zma*), *Triticum aestivum* (*Tae*). (**B-D**) Comparison of runtime, sensitivity and accuracy of miRDP2 and other five tools. Two dots corresponding to each tool indicate two tests were made by each tool. This figure has been adapted from Kuang et al.[16].

**Table 1. Genomes and sRNA libraries used for testing miRDP2 and other tools.** This table has been adapted from Kuang et al.[16].

**Supplementary File 1: Comparison of runtime, sensitivity and accuracy of miRDP2 and other five tools.**

**Supplementary File 2: Examples of authentic miRNAs with bifurcate structure in loops.**

**Supplementary File 3: Updated criteria for plant miRNA annotation and criteria for 23-nt and 24-nt miRNAs.**

**Supplementary File 4: Diagram of the workflow of miRDP2.**

## DISCUSSION:

With the advent of NGS, a large number of miRNA loci have been identified from an ever-increasing amount of sRNA sequencing data in diverse species[29,30]. In the centralized community database miRBase[21], the deposited miRNA items have increased almost 100 times in the last decade. However, in comparison to miRNAs in animals, plant miRNAs have many unique features that make the identification/annotation more complicated[13,14].

First, the precursors of plant miRNAs are more variable in length and structure (**Supplementary File 2**)[16]. Not like the relatively uniform length of animal miRNA precursors around 70-90 nt, the length of plant precursors vary by several folds and could reach several hundred nts[13,31]. This difference introduces a lot of uncertainty when predicting the secondary structure of miRNA precursors even though a cutoff of precursor length is usually set arbitrarily such as not exceeding 300 nt[19] (this parameter was embedded in miRDP2, and experienced users of miRDP2 could adjust this by themselves). In addition, conserved plant miRNA families tend to have more members, and the length variation of these members is also often significant[13]. This is the reason why miRDP2 has the parameter –L, which indicates the potential largest miRNA families in member size. Together, the heterogeneity of plant miRNA precursors raises many difficulties for their accurate annotation.

Second, the noise or false-positives introduced by siRNAs is hard to eliminate. Alongside miRNAs, NGS methods also produce a deluge of siRNAs in the sequenced sRNA libraries. Even though siRNAs could be separated from miRNAs by their biogenesis and functions[32,33], it is extremely difficult to distinguish them based on sequencing data and mining tools. The public databases such as miRBase, argued by many researchers, have deteriorated sharply by the large number of false-positives siRNAs, which are erroneously annotated as miRNAs[20,31]. Thus, refined tools with a new and strict set of criteria for plant miRNA annotation like the newly updated criteria[25] (**Supplementary File 3**) are highly desired in the miRNA annotation pipeline/process.

Last but not least, the computational time for parsing sRNA libraries has increased exponentially when the same method is transplanted from a small size genome species to a large size one. The computational tools such as miRDeep-P[13] and miR-PREFeR[26], by capturing and quantifying the signature distribution of sRNA reads along miRNA precursors, have become two popular methods and are widely used to annotate miRNAs. The mapping strategy, the process of excising precursor candidates and subsequent secondary structure prediction demand considerable computing time[16]. When these tools are employed to parse the data from small size genomes like Arabidopsis to large ones like maize, the data processing time is increased from hours to days even weeks (**Figure 1B**), resulting in frequent collapse of the process. An innovation on the foregoing limitations is thus urgently in need.

Our new miRDP2[16] program, updated from miRDeep-P[13], is designed to overcome the challenges mentioned above (**Supplementary File 4**). In this program, we employed a new filtering strategy, optimized the scoring algorithm, and incorporated newly updated plant miRNA annotation criteria. As a result of these new features, the running time was markedly shortened when tested using ten sRNA libraries from five plant species with increasing genome size. Additionally, compared to other tools, miRDP2 displayed superior performance in both sensitivity and accuracy (**Figure 1**). Taken together, these results demonstrate that miRDP2 is a fast and accurate tool for analyzing the miRNA transcriptomes in plants.

It should be cautioned that the current understanding on miRNA characteristics might limit the performance of any computational tools. Even the newly updated miRNA annotation criteria are based on a limited set of well-studied examples. The deduced information is thus only empirical. In fact, unique features of miRNAs have been shown to exist in different plant species or lineages[3]. In addition, characteristics such as the structures of upstream and downstream regions of the miRNA/miRNA* duplex also play critical roles in miRNA biogenesis[34,35], which are not taken into account in current annotation tools. With the accumulation of well-studied examples in more plant species, it is likely that even more advanced annotation tools are developed in the future that can capture more subtle distinctions and classify miRNAs with a higher degree of accuracy than current methods. A promising new miRNA annotation direction is to incorporate machine learning approaches[36] as the quality of training datasets and annotation criteria continually evolve.

## ACKNOWLEDGMENTS:

This work has been supported by Beijing Academy of Agriculture and Forestry Sciences (KJCX201917, KJCX20180425, and KJCX20180204) to XY and National Natural Science Foundation of China (31621001) to LL.

## DISCLOSURES:

The authors have nothing to disclose.

## REFERENCES

1 Ghildiyal, M., Zamore, P. D. Small silencing RNAs: an expanding universe. *Nature Reviews Genetics.* **10** (2), 94-108 (2009).

2 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell.* **136** (2), 215-233 (2009).

3 Moran, Y., Agron, M., Praher, D., Technau, U. The evolutionary origin of plant and animal microRNAs. *Nature Ecology Evolution.* **1** (3), 27 (2017).

4 Xie, Z. et al. Expression of Arabidopsis MIRNA genes. *Plant Physiology.* **138** (4), 2145-2154 (2005).

5 Zhao, X., Zhang, H., Li, L. Identification and analysis of the proximal promoters of microRNA genes in Arabidopsis. *Genomics.* **101** (3), 187-194 (2013).

6 Bologna, N. G., Mateos, J. L., Bresso, E. G., Palatnik, J. F. A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159. *EMBO JOURNAL.* **28** (23), 3646-3656 (2009).

7 Rogers, K., Chen, X. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell.* **25** (7), 2383-2399 (2013).

8 Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. *Cell.* **136** (4), 669-687 (2009).

9 Iwakawa, H. O., Tomari, Y. The Functions of MicroRNAs: mRNA Decay and Translational Repression. *Trends in Cell Biology.* **25** (11), 651-665 (2015).

10 Lee, R. C., Feinbaum, R. L., Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell.* **75** (5), 843-854 (1993).

11 Wightman, B., Ha, I., Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell.* **75** (5), 855-862 (1993).

12 Friedlander, M. R. et al. Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology.* **26** (4), 407-415 (2008).

13 Yang, X., Li, L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics.* **27** (18), 2614-2615 (2011).

14 Meyers, B. C. et al. Criteria for annotation of plant MicroRNAs. *Plant Cell.* **20** (12), 3186-3190 (2008).

15 Yang, X., Zhang, H., Li, L. Global analysis of gene-level microRNA expression in Arabidopsis using deep sequencing data. *Genomics.* **98** (1), 40-46 (2011).

16 Kuang, Z., Wang, Y., Li, L., Yang, X. miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics.* 10.1093/bioinformatics/bty972 (2018).

17 Kozomara, A., Birgaoanu, M., Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Research.* **47** (D1), D155-D162 (2019).

18 Griffiths-Jones, S., Saini, H. K., van Dongen, S., Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Research.* **36** (Database issue), D154-158 (2008).

19 Axtell, M. J., Meyers, B. C. Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. *Plant Cell.* **30** (2), 272-284 (2018).

20 Taylor, R. S., Tarver, J. E., Hiscock, S. J., Donoghue, P. C. Evolutionary history of plant microRNAs. *Trends in Plant Science.* **19** (3), 175-182 (2014).

21 Kozomara, A., Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research.* **42** (Database issue), D68-73 (2014).

22 Langmead, B., Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods.* **9** (4), 357-359 (2012).

23 Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology.* **6** 26 (2011).

24 Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology.* **10** (3), R25 (2009).

25 An, J., Lai, J., Sajjanhar, A., Lehman, M. L., Nelson, C. C. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. *BMC Bioinformatics.* **15** 275 (2014).

26 Lei, J., Sun, Y. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics.* **30** (19), 2837-2839 (2014).

27 Evers, M., Huttner, M., Dueck, A., Meister, G., Engelmann, J. C. miRA: adaptable novel miRNA identification in plants using small RNA sequencing data. *BMC Bioinformatics.* **16** 370 (2015).

28 Mathelier, A., Carbone, A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics.* **26** (18), 2226-2234 (2010).

29 Zhu, Q. H. et al. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Research.* **18** (9), 1456-1465 (2008).

30 Fahlgren, N. et al. MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. *Plant Cell.* **22** (4), 1074-1089 (2010).

31 Fromm, B. et al. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annual Review of Genetics.* **49** 213-242 (2015).

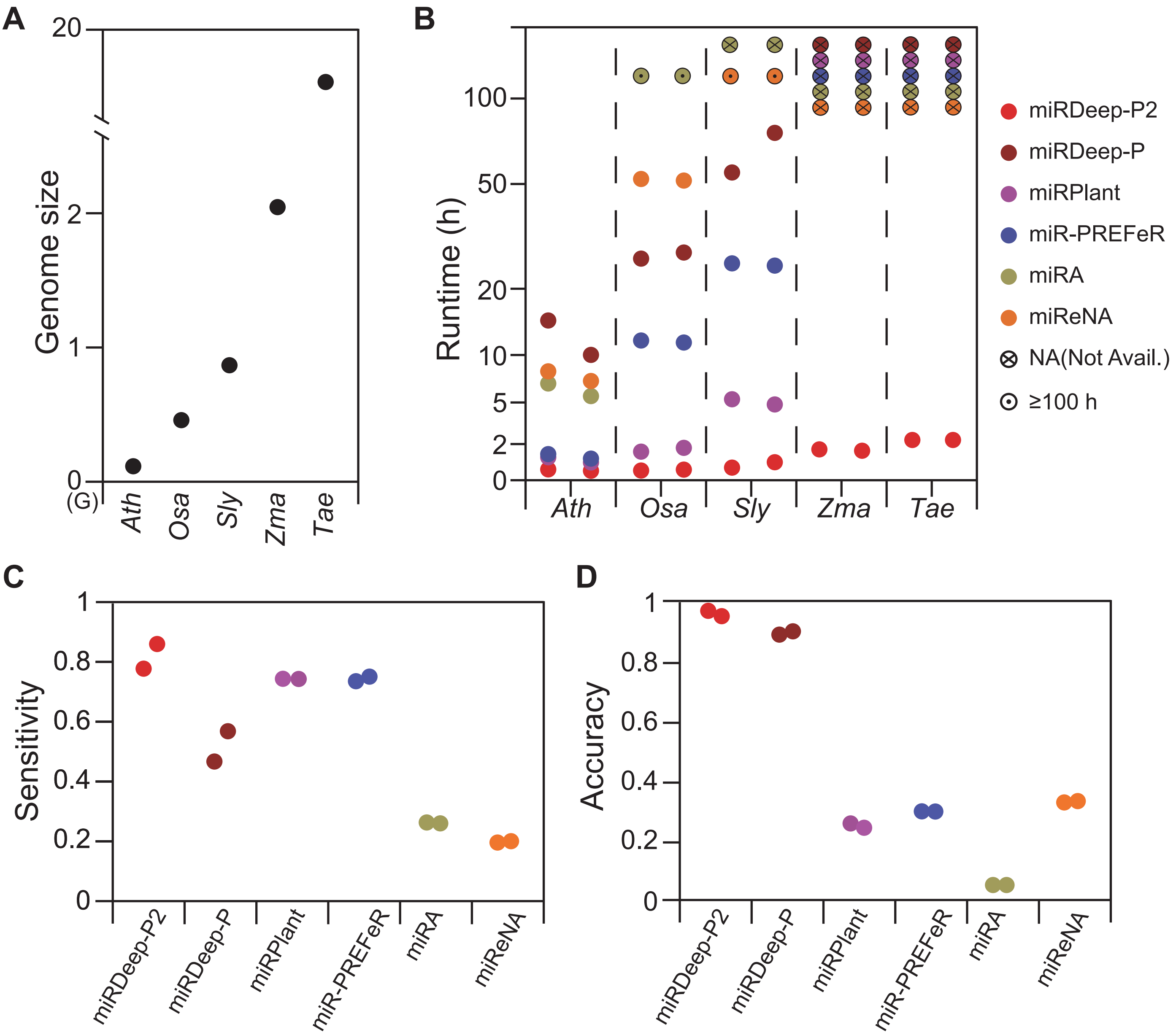
32 Blevins, T. et al. Identification of Pol IV and RDR2-dependent precursors of 24 nt siRNAs guiding de novo DNA methylation in Arabidopsis. *Elife.* **4** e09591 (2015).

33 Zhai, J. et al. A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. *Cell.* **163** (2), 445-455 (2015).

34 Werner, S., Wollmann, H., Schneeberger, K., Weigel, D. Structure determinants for accurate processing of miR172a in Arabidopsis thaliana. *Current Biology.* **20** (1), 42-48 (2010).

35 Mateos, J. L., Bologna, N. G., Chorostecki, U., Palatnik, J. F. Identification of microRNA processing determinants by random mutagenesis of Arabidopsis MIR172a precursor. *Current Biology.* **20** (1), 49-54 (2010).

36 Vitsios, D. M. et al. Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Research.* **45** (21), e177 (2017).


A
20
Genome size
2
1
0
(G)
Ath
Osa
Sly
Zma
Tae
B
Runtime (h)
100
50
20
10
5
2
0
Ath
Osa
Sly
Zma
Tae
miRDeep-P2
miRDeep-P
miRPlant
miR-PREFeR
miRA
miReNA
NA(Not Avail.)
≥100 h
C
Sensitivity
1
0.8
0.6
0.4
0.2
0
miRDeep-P2
miRDeep-P
miRPlant
miR-PREFeR
miRA
miReNA
D
Accuracy
1
0.8
0.6
0.4
0.2
0
miRDeep-P2
miRDeep-P
miRPlant
miR-PREFeR
miRA
miReNA

| **Species (abb.)** | **Genome** | **sRNA libraries** | | | | |
|---|---|---|---|---|---|---|
| | **version** | **Library ID** | **File size** | **Total reads** | **Unique reads** | **Tissue** |
| *Arabidopsis thaliana* (*Ath*) | version 10 | GSM2094927 | 24.9 Mb | 40.5M | 9.7M | Adult leaf |
| | | GSM2412287 | 29.5 Mb | 45.1M | 11.1M | Leaf |
| *Oryza sativa* (*Osa*) | version 7 | GSM2883136 | 44.2 Mb | 54.9M | 16.3M | Seedling |
| | | GSM3030848 | 34.7 Mb | 49.1M | 13.0M | Flagleaf |
| *Solanum lycopersicum* (*Sly*) | version 3 | GSM1213985 | 205.4 Mb | 161.5M | 58.0M | Leaf |
| | | GSM1976413 | 118.5 Mb | 139.3M | 46.2M | Root |
| *Zea mays* (*Zma*) | version 4 | GSM1277437 | 158.4 Mb | 266.1M | 60.5M | Seedling |
| | | GSM1428531 | 144.1 Mb | 172.5M | 56.3M | Seed |
| *Triticum aestivum* (*Tae*) | iwgsc 1 | GSM1294660 | 76.1 Mb | 59.2M | 29.6M | Shoot |
| | | GSM1294661 | 113.6 Mb | 84.0M | 44.0M | Leaf |

| Name of Material/Equipment | Company | Catalog Number | Comments/Description |
|---|---|---|---|
| Computer/computing node | N/A | N/A | Perl is required; at least 8 GB RAM and 100 GB st |

torage are recommended

jove JOURNAL OF VISUALIZED EXPERIMENTS

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

# ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article: A Bioinformatics Pipeline to Accurately and Efficiently Analyze the MicroRNA Transcriptomes in Plants

Author(s): Ying Wang, Zheng Kuang, Lei Li, Xiaozeng Yang

Item 1: The Author elects to have the Materials be made available (as described at http://www.jove.com/publish) via:

☒ Standard Access ☐ Open Access

Item 2: Please select one of the following items:

☒ The Author is **NOT** a United States government employee.

☐ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

☐ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

## ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of The Journal of Visualized Experiments; "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and(c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this **Section 6** is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Protection of the Work.** The Author(s) authorize JoVE to take steps in the Author(s) name and on their behalf if JoVE believes some third party could be infringing or might infringe the copyright of either the Author's Article and/or Video.

9. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

10. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

11. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole

# ARTICLE AND VIDEO LICENSE AGREEMENT

discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

12. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

13. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication of the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

14. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement is required per submission.

**CORRESPONDING AUTHOR**

| | |
|---|---|
| Name: | Xiaozeng Yang |
| Department: | Beijing Agro-biotechnology Research Center |
| Institution: | Beijing Academy of Agriculture and Forestry Sciences |
| Title: | A Bioinformatics Pipeline to Accurately and Efficienty Analyze the MicroRNA Transcriptomes in plants |
| Signature: | |
| Date: | 02/20/2019 |

Please submit a **signed** and **dated** copy of this license by one of the following three methods:

1. Upload an electronic version on the JoVE submission site
2. Fax the document to +1.866.381.2236
3. Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02140

**Editorial comments:**

General:
1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues.
We have corrected all typos and grammatical errors figured out by reviewers in the revised manuscript. In addition, we went over the revised manuscript several times and ensured there are no more errors.

2. Please revise lines 69-79 and 92-99 to avoid overlap with previously-published text.
We thank the Editor for pointing this out. We have revised these two parts.

Protocol:
1. Everything in the protocol (except for the introductory ethics statement and possibly command line steps) should be in a numbered step (in the imperative tense and of no more than 4 sentences), header, or 'Note'; please revise accordingly.
Following the rules the Editor mentioned above, we have edited several sections considering each of them harboring too much steps. Meanwhile, each new section was assigned a numbered step and we revised all content accordingly.

2. There is a 10 page limit for the Protocol, but there is a 2.75 page limit for filmable content. Please highlight 2.75 pages or less of the Protocol (including headers and spacing) that identifies the essential steps of the protocol for the video, i.e., the steps that should be visualized to tell the most cohesive story of the Protocol. Remember that non-highlighted Protocol steps will remain in the manuscript, and therefore will still be available to the reader.
We have followed the above rules and made corresponding changes. Please check the revised manuscript.

3. Please add more details to your protocol steps. Please ensure you answer the "how" question, i.e., how is the step performed? Alternatively, add references to published material specifying how to perform the protocol action. If revisions cause a step to have more than 2-3 actions and 4 sentences per step, please split into separate steps or substeps.
We thank the Editor for these suggestions. We have carefully followed these suggestions and made corresponding changes. More details are in the revised manuscript.

Specific Protocol steps:
1. 1.3: Can you elaborate a bit on ensuring perl is in the PATH (especially for users who may be less familiar with this concept)?
We edited the sentence to describe how to set the PATH.

2. 2.1: Please explain a bit more on how to fulfill all these conditions (including with

references if necessary).
We thank the Editor for suggesting us to make this clearer. We have updated miRDP2 to version 1.1.3, where we provide a new option for the format of input files. By this option, miRDP2 can handle different format of input files such as Fastq and Fasta. The formatting process becomes much easier.

3. 2.2: Please explain how to download index files and index a reference sequence (possibly including references).
We have added an explanation on downloading index files. In addition, in part 1.4, we added more comments including how to index a reference sequence.

4. 3: Please convert to a series of steps or move elsewhere (e.g., the introduction).
We moved this section to the end of PROTOCOL considering this is a summary of this pipeline.

Figures:

1. Please obtain explicit copyright permission to reuse any figures/tables from a previous publication. Explicit permission can be expressed in the form of a letter from the editor or a link to the editorial policy that allows reprints. Please upload this information as a .doc or .docx file to your Editorial Manager account.

We thank the Editor for pointing this out. We have obtained the copyright permission to reuse all figures/tables from previous journal *Bioinformatics*. We have uploaded this document to the Editorial Manager account

2. Figure 1B: ‘100 h’, not ‘100h’ (include a space).
We have corrected this in the revised manuscript

3. Figure 1C, D: Please explain these a little more- what are the 2 dots for each program?
The 2 dots indicate two tests were taken in each program. The explanation has been added into Figure legend.

Discussion:
1. As we are a methods journal, please revise the Discussion to explicitly cover the following in detail in 3–6 paragraphs with citations:
a) Critical steps within the protocol
b) Any modifications and troubleshooting of the technique
We thank the Editor for suggesting to revise the Discussion. We have followed your suggestion and revised the discussion, including how to associate the parameters with knowledge on plant miRNAs, and potential modifications and improvements in the future.

References:
1. Please do not abbreviate journal titles.

We thank the Editor for pointing this out. We have employed Endnote to format all reference items with JoVE template.

Table of Materials:
1. Please ensure the Table of Materials has information on all materials and equipment used, especially those mentioned in the Protocol (including advised computational specifications).
We thank the Editor for pointing this out. Since our method is a software pipeline, materials such as chemical reagent were not used, all tested NGS datasets were included in the package of our pipeline.

**Reviewers' comments:**

**Reviewer #1:**

Manuscript Summary:

The authors describe their pipeline for analysis of sRNA libraries to detect miRNA sequences in plants. This pipeline was published in Bioinformatics in 2018 as a two-page article including a supplementary manual, which is the basis for this manuscript.
A reader has to invest some effort to understand the "pipeline" because information on what to type, which program is doing what, which options are available, what are irrelevant output files, etc., is spread somehow over the manuscript. IMHO, a figure showing the pipeline of programs including their in- and output might be quite helpful for a reader.

We really appreciate all the effort and comments the Reviewer made. As the Reviewer suggested that a supplementary file 4 showing the pipeline of program has been added.

Major Concerns:

-Line 105: Dependencies: Move the note (line 125--127) to the end of this section. Add an additional note on further requirements like minimum of RAM (in dependence of genome and sRNA library size) and minimum of disk space. Any reader/potential user of miRDP2 would like to know about these prerequisites prior to an installation.
We thank the Reviewer for suggesting us to make this clearer. As the Reviewer suggested, we have made changes accordingly.

-Line 106: BowTie and the Vienna package are available, for example, as Ubuntu packages; this should be mentioned because using a package is much easier than compiling/installing from source.
We thank the Reviewer for pointing this out. We totally agree with the Reviewer that using a compiled package is much easier than compiling/installing from source. A brief

explanation was added.

-Line 121: Mention that miRDP2-v1.1.1.tar.gz extracts to the directory 1.1.1/ and that the command snippets given in the manuscript respect this.
A command line has been added in the revised manuscript.

-Line 137ff: This example works also for Windows? Even for a Linux newbie some further explanations of the commands might be helpful.
We thank the Reviewer for suggesting to provide comments and explanations of the commands. We have added comments for almost each one.

-Line 135: A link to ncRNA_r fam.tar.gz (https://sourceforge.net/projects/mirdp2/files/latest_version/) is missing in the manuscript but this file is necessary for the test.
We thank the Reviewer for catching this. A hyperlink has been added.

-Line 149ff: What is the content and meaning of the further 25(?) files in the output directory?
We did not get the exact point the Reviewer raised. Just guess the Reviewer would like to know the content in the output directory. In the output directory, it consists of all intermediate files and results, including a list of predicted miRNA candidates.

-Line 150: What is the extension of the directory and file? (-15-0-10)
We thank the Reviewer for suggesting us to make this clearer. 15, 0 and 10 correspond to three parameters of miRDP2, －L (location limit), -M (mismatches allowed), and –R (threshold of RPM). In this case, location limit was 15, no mismatches were allowed and threshold of RPM was 10.

-Line 165: Does the tarball contains a script to perform the removal of redundant reads and their proper renaming?
There is an embedded script collapsing and renaming identical reads in miRDP2. We updated miRDP2 to a new version (version 1.1.3) which includes one more new option of transforming input reads formats.

-Line 194: Explain all options of the bash script (or mention at least its help option).
We thank the Reviewer raising this question. We added more comments and explanations on all of commands. Meanwhile, all options of this bash script were explained in its help option.

-Line 208: What are the exact criteria to give a positive/negative prediction?
The plant miRNA criteria are cited from Axtell and Meyer’s 2018 plant cell paper. We added a supplementary file 3 which includes all details.

-Line 238: What is a "signature" file?
A “signature” file includes the information of position and numbers of sRNA reads

mapped to potential precursors. An example of “signature” file is in TestResult of miRDP2 TestData folder.

-Fig. 1B,C,D: What are the two dots for each program?
The 2 dots indicate two tests were taken in each program. The explanation has been added into Figure legend.

-Fig. 1B: Move the `(h)' behind the y-label `Runtime'
We have changed this in the Figure.

**Minor Concerns:**

-Line 185: "need to re-indexed." => "need to be re-indexed." Why should a further indexing be necessary?
We thank the Reviewer for suggesting us to make this clearer. In fact, this is a step not for re-indexing because the TestData we provided only includes un-indexed sequences of Arabidopsis genome and ncRNA of RFAM considering the size of indexed files is much larger than that of original sequence files. In addition, we added more comments and explanation on these commands.

-Line 219: "The input files are FASTA format of original reads files" = > "The input files are original reads in FASTA format"
Is this the file with the re-formatted reads (as described in section 2.1) or the reads just after adapter removal?
We thank the Reviewer for raising the question. As suggested by you and other Reviewers, we have added option of miRDP2 which can handle the transformation between different formats (FASTA and FASTQ) and generate an input file in required format. The new version of miRDP2 has been updated in its Sourceforge webpage.

-Line 266: What is meant by "loose" result? (less stringent/specific/sensitive?)
We thank the Reviewer for catching this. "loose" result includes less stringent candidates. We have changed loose to less stringent in the revised manuscript.

-Line 90: "Taylor et al[21]. found" => "Taylor et al[21] found"
-Line 144: Is a call via bash necessary? By the way, the formatting of this line is very ugly; what about a continuation line (\)? A similar modification would help with lines 196--197.
-Line 145: "-i ./TestData/TAIR10.genome" => "-i ./TestData/TAIR10_genome"
-Line 193: "miRDP2 to detecting new" => "miRDP2 to detect new"
-Line 225: Give nominator and denominator in normal font (non-italics).
-Line 252: "Reads map to too many sites" => "Reads mapping to too many sites"
-Line 260: "Only reads exceeded a certain RPM threshold" => "Only reads exceeding a certain RPM threshold"

-Line 277: "gradually increased genome," => "gradually increased genome length,"
-Line 279: "Overall, to each species" => "Overall, for each species"
-Line 281: "are proceeded as" => "are processed as"
-Line 281f: "5 miRNA computational prediction tools including the old version miRDeep-P and other 4 tools, miRPlant, miR-PREFeR, miRA, miReNA, were selected to make the comparison." => "Five miRNA computational prediction tools (miRPlant, miR-PREFeR, miRA, miReNA, miRDeep-P) were selected to make the comparison."
-Line 289: "programs are run" => ``programs were run"
-Line 290: "We instead tested" => "Instead we tested"; "and is not able" => "and was not able"
-Line 295: ``programs could run" => ``programs ran"
-Line 299: "miRPlant has consumed" => "miRPlant consumed"; "result, it could not be able" => "result was not able"
-Line 301: ``Instead, miRDP2 could finish" => ``miRDP 2 finished"
-Line 309: formula => formulas
-Line 318: "to mature & star" = > "to mature and star"
-Line 319: "As Fig. 1C&D displayed, miRDP2 ..." => "miRDP2 ... (Figs 1C,D)"
-Line 328: aes-tivum => aestivum
-Line 332: "Resources of testing miRDP2 and other tools." => "Genome and sRNA library sizes used for testing."
-Line 353: "relatively uniform animal miRNA precursors" => "relatively uniform length of animal miRNA precursors"
-Line 397: "it is likely even more" => "it is likely that even more"
-Fig. 1: In print the dot's colors of miRDeep-P2 and miReNA are very close; choose one different color, please.
We thank the Reviewer for suggesting these above improvements. We have made these changes in the revised manuscript.

**Reviewer #2:**

miRDeep-P2 is a plant miRNA prediction tool and it is an update to the original miRDeep-P. Much of the core of miRDeep-P remains, but the scoring parameters have been updated to accommodate a new set of plant miRNA annotation criteria published a year ago. Overall, the paper is relatively clear, so I've focused my comments on the use of miRDeep -P2.
We really appreciate all the effort and comments the Reviewer made.

Installing and running miRDeep-P2 was relatively simple to do, though I did find the format of the sRNA files a bit annoying to create from any library files I chose to feed it. I wrote a quick python script to do this, but others may not find this as easy. I think it would be better if there was a more standard file format that could be utilized for this.
We thank the Reviewer for suggesting this improvement. miRDP2 has been updated to a new version 1.1.3, where we embedded an additional option by which the users can

choose the format of sRNA files (Fastq and Fasta, two of the most common formats of sRNA files). This change has greatly simplified the format transformation.

During a read through of the miRDeep-P2's methods, I noticed the statement: "reads conserved with annotated plant mature or star miRNAs in miRBase are separately processed" which made me curious to see what that meant. After looking through their code, it appears that early on in the script, some reads are filtered if they map to tRNA, rRNA, or snRNA as well as if the reads have too low of an RPM (10 is the default). However, reads will bypass this RPM filter if the sequence matches with up to 1 mismatch of any miRBase miRNA. These are still fed these through the rest of their filters, but there is preferential treatment given to miRBase miRNAs from the start by allowing that first filter be skipped. While utilizing miRBase annotations is surely a viable strategy for predicting miRNAs, similar to the method that the Axtell lab used when discussing accuracy and sensitivity of ShortStack, these authors use miRBase miRNAs as true positives. Thus, miRDeep-P2 has access to its true positive test dataset, and it actively allows each of these true positives (and those with 1 mismatch) to bypass its high RPM filter which was set to prevent long runtimes and prevent false positives. We thank the reviewer for pointing out these details. There are several changes we made when updating miRDeep-P to miRDP2. One of them is the mapping strategy as the Reviewer mentioned. In fact, we found that the processing time became horrible when the size of species genome and sRNA-seq data increasing. For instance, the processing time became tens of hours to days when employing the same strategy in Arabidopsis to maize or even other species with larger size genome (like we observed in Figure 1B). In addition, the major concern of miRNA annotation currently is how to control the false positives (As stated in Axtell and Meyer 2018 Plant Cell paper). Taken these two factors into consideration, we tested the above strategy, separating reads conserved with annotated plant mature or star miRNAs and others. In fact, as the result displayed in Figure 1C,D, the strategy was successful at both sensitivity and accuracy beside the computing time was dramatically cut. We agree with the Reviewer that the double standard might filter out some lowly expressed miRNA candidates, but it looks like that it is more effective to minimize the large number of false positives.

I also have concerns with regards to the provided definition of accuracy as: 'Known miRNAs number/predicted miRNAs number'. The definition of accuracy is (TP+TN)/(TP+FP+TN+FN). Because we do not know anything about the negative set, the numerator should only be the number of miRBase miRNAs that miRDeep-P2 predicted and the denominator should be the total number of miRNAs that it predicted. The accuracy should not take into consideration the total number of known miRBase miRNAs on its own, as this number and the total number of miRNAs predicted by miRDeep-P2 does nothing to distinguish how many of those predicted miRNAs are actually present in miRBase (i.e. what in this set of data is a false positive). This appears to be an issue and I would surmise that this is the cause for the near perfect accuracy when my tests seemed to suggest a non-trivial number of non-miRBase miRNAs are predicted when running miRDeep-P2. In other words, miRDeep-P2 has nowhere near

perfect accuracy because it is trusting miRBase to be free of poorly annotated miRNAs, which it is not. This problem is actually mentioned by the authors on lines 90-91, but they don't seem to have addressed it in the code. With all this said, I was able to remove this bypass and run miRDeep-P2 without the RPM bypass for miRBase miRNAs and predict miRNAs using several test libraries from Arabidopsis and maize. miRDeep-P2 ran exceptionally well with fewer false positives than many other tools while still identifying a large number of miRBase miRNAs. My primary suggestions to the authors are as follows:
We greatly appreciate all the effort and the comments from the Reviewer.

*allow a standard file format as an input for sRNA data files
As we answered the Reviewer's first question, we have embedded an additional option in the new version 1.1.3 of miRDP2, by which users could choose either Fasta or Fastq format of sRNA data files.

*allow a batch running mode so that users who wish to process multiple libraries do not need to make several individual calls to the primary bash file
We thank the Reviewer for suggesting this improvement. We have added a batch running mode in the new updated version 1.1.3 of miRDP2.

*Take a look at the sensitivity and accuracy as something seems off with these numbers. They appear to be way too high.
We thank the Reviewer for raising this great question. We would like to point out that in fact, there is not a widely accepted standard to define the sensitivity and accuracy of miRNA prediction with several reasons. First of all, even to the collection of miRNAs in Arabidopsis, the most intensively studied species, no one can guarantee all of the miRNA candidates are authentic miRNAs. It is a compromise that miRNA research community employ the collection of miRNAs in Arabidopsis as a testing dataset. Second, the definition of the sensitivity and accuracy we used is the one relatively accepted by other scientists (Axtell and Meyers, 2018, plant cell; Lei and Sun, 2014, bioinformatics). Third, one of the reason why the sensitivity and accuracy is very high is that we employed the new miRNA annotation criteria (Axtell and Meyers, 2018, plant cell), mainly based on the knowledge of the collection of Arabidopsis miRNAs.

**Reviewer #3:**

Identifying miRNAs in plant species is still an active research field. Thus, a tool that can provide more accurate and efficient annotation is still in need. The authors made some contributions towards this goal.
We greatly appreciate the comments from the Reviewer.

Major Concerns:

As the authors emphasized two challenges for miRNA annotation in plants, I would expect some explanations about how this work addressed those challenges. But I cannot find any. In particular, why is this version faster than others including the first version? Is it because of a better implementation? Not clear.
We thank the Reviewer for raising this question. The reason why miRDP2 is much faster than the old version and others is mainly caused the new strategy we employed. Briefly, we filtered out reads before mapping by RPM and subsequently, the time of precursor generation and secondary structure prediction is greatly shortened. More details are in manual of miRDP2. Meanwhile, we briefly explained this in discussion.

Also, why is it more accurate than others?
The mainly reason is that we employed the new plant miRNA annotation criteria (Axtell and Meyers, 2018, the Plant Cell), and added it as a filter. More details are included in a new supplementary file 3.

How did it handle the challenge of mixing miRNAs and siRNAs in plants? Including these will give the users a peaceful of mind when adopting this tool.
A related concern is that the results did not include any discussion about distinguish siRNAs from miRNAs. The motivation in the Intro, the methods, and the results should be synchronized to some extent.
The new plant miRNA criteria (Axtell and Meyers, 2018, the Plant Cell) emphasized much on how to remove the siRNA noise when predicting miRNAs. We added a new supplementary file 3 to illustrate this. In particular, we added a new part on how to separate 23-24 nt siRNA from miRNAs. Additionally, as the Reviewer suggested, we briefly included the reason in discussion section of main text.

Considering that Bowtie 2 is faster, the users should use bowtie 2 as the default read mapping tool. I don't know why Bowtie is still the default. After all, you need faster indexing and read mapping for large plant genomes right?
We thank the Reviewer for pointing this out. We have updated Bowtie and Bowtie2 as both options for mapping tool, which is updated in the new version of miRDP2 (version 1.1.3).

Please add more detailed information about the running time evaluation. Does it include read mapping stage? Bowtie's indexing stage? Not clear. Same for other tools, which might include several scripts/steps. Please ensure a fair comparison.
We thank the Reviewer for suggesting to provide more details on running time evaluation. The running time does not include Bowtie’s indexing stage, but includes all mapping stage. We employed the same criteria to all tools. Thus, our comparison is fair.

What is the theoretical running time complexity of miRDeep P2? It looks better than linear according to the figure1. Again, it is the running time of whole package (including read mapping, indexing...)?
We thank the Reviewer for suggesting to explain this clearer. The bottle neck or the

most time-consuming step of miRDP2 is to excise potential miRNA precursors based on the mapping result of sRNA reads. Mapped reads number and size of genome will result in the growth of running time exponentially. That is the reason why we change the mapping strategy and stringently control the number of reads accessing to mapping process. As described in manuscript, by a pretreatment, we only keep sRNA reads, 1) conserved with known mature or star miRNAs and 2)non-conserved but with high RPM value (10 default). Through this filter, less than 5% reads of total sRNA dataset were generally processed into the following mapping process. Thus, the running time is related to how many reads through the filter. For each species, the reads number is not linearly increased, which is also the reason why the running time in Figure 1 is not linear.

The definition of the sensitivity is confusing. Suppose that there are X expressed known miRNAs, out of X, Y are successfully detected by a tool, the sensitivity is Y/X. To me this is the standard way to compute the sensitivity using *known* miRNAs. I am not clear whether the given definition is the same as this. If not, please explain. The accuracy is what I expect. According to the accuracy equation, should sensitivity be: predicted known miRNAs /expressed known miRNAs?
We thank the Reviewer for catching this. In fact, we used the same strategy to calculate accuracy. We have made changes as the Reviewer suggested.

Line 149. Instead of saying "taking a while", better give real running time, which should be available as you already did the experiments.
We thank the Reviewer for pointing out this detail. Indexing Arabidopsis genome took around ten minutes in our computing cluster (hardware details in Supplementary File 1), and we have added this piece of information to the revised manuscript.

2.1 Formatting reads, better provide script for users as well. Will make this tool easier to use.
We thank the Reviewer for suggesting this improvement. We have updated miRDP2 to version 1.1.3, including the option of reads format. Via this option, users can choose the most common used two formats, Fastq or Fasta.

**Supplementary File 1. Comparison of runtime, sensitivity and accuracy of miRDP2 and other five tools.**

To compare runtime and performance of miRDP2 and other five tools, miRDeep-P (Yang and Li, 2011), miRPlant (An, et al., 2014), miR-PREFeR (Lei and Sun, 2014), miRA (Evers, et al., 2015), miReNA (Mathelier and Carbone, 2010), we installed all six tools in cluster server with Cent OS release 6.5 system. These programs are run with same input sequencing files and genomes with 2x Intel Xeon Processor E5-2670 v2 10C 2.5GHz. All programs are run using 1 thread, and 40Gb memory in computing node of our server. Specially, miRPlant is controlled from GUI written in Java and is not able to run on the server. We instead test miRPlant on a PC with Windows 10 system, Intel Core i7-4720HQ 2.6GHz and 16Gb memory. We have also tested miRDP2 and miRDeep-P on this PC. There are no significant difference on time consumption between programs running on PC and on server.

For small genomes as *Arabidopsis thaliana*, *Oryza sativa*, and *Solanum lycopersium*, all the programs could run properly. However, for large genomes of *Zea mays* and *Triticum aestivum* (including *Solanum lycopersium* for miRA), some of the programs depleted all computing resource and break down halfway. miReNA, miRA, and miR-PREFeR have failed to generate results for some or all of the input sequencing files, probably due to memory deficiency while dealing with .sam files or intermediate files. miRPlant has consumed too much space in C:/, possible temporary files, that are not able to run on our PC.

The commands (including preprocessing steps) and parameters for miRA, MIReNA, miR-PREFeR, and miRPlant are listed as the following.

**miReNA:**
Formatting of reads file using custom perl script.
Running MIReNA.sh with `-D' option.

**miR-PREFeR:**
Formatting of reads file using custom perl script.
Running miR_PREFeR.py using `-L pipeline' option.

Parameters:
*PRECURSOR_LEN = 300*
*READS_DEPTH_CUTOFF = 20*
*NUM_OF_CORE = 1*
*MAX_GAP = 100*
*MIN_MATURE_LEN = 18*
*MAX_MATURE_LEN = 24*
*ALLOW_NO_STAR_EXPRESSION = Y*

*ALLOW_3NT_OVERHANG = N*
*CHECKPOINT_SIZE = 3000*

**miRPlant:**
Formatting of reads file using custom perl script.

Parameters:
*Adapter =*
*precursor length = 200*
*min loop length = 20*
*flank length = 10*
*max inconRead ratio = 0.1*
*miR Lnegth = 18 to 23*
*min phred = 20*
*max multimap = 101*
*min reads = 5*
*min score = -10*

**miRA:**
Formatting of reads file using custom perl script.
Mapping reads using bowtie, with option -a -v 0 -S.
Running miRA pipeline with `full' option.

Parameters:
*log_level = 2*
*openmp_thread_count = 1*
*cluster_gap_size = 10*
*cluster_min_reads = 10*
*cluster_flank_size = 200*
*cluster_max_length = 2000*
*min_precursor_length= 50*
*max_precursor_length= 0*
*max_mfe_per_nt = -0.2*
*max_hairpin_count = 4*
*min_double_strand_length = 18*
*permutation_count = 100*
*max_pvalue = 0.01*
*min_coverage = 0.01*
*min_paired_fraction = 0.55*
*min_duplex_length = 18*
*max_duplex_length = 30*
*allow_three_mismatches = 1*
*allow_two_terminal_mismatches= 1*

```
create_coverage_plots = 1
create_structure_plots = 1
create_structure_coverage_plots = 1
cleanup_auxiliary_files = 1
```

## References

An, J., *et al.* miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. *BMC Bioinformatics* 2014;15:275.

Evers, M., *et al.* miRA: adaptable novel miRNA identification in plants using small RNA sequencing data. *BMC Bioinformatics* 2015;16:370.

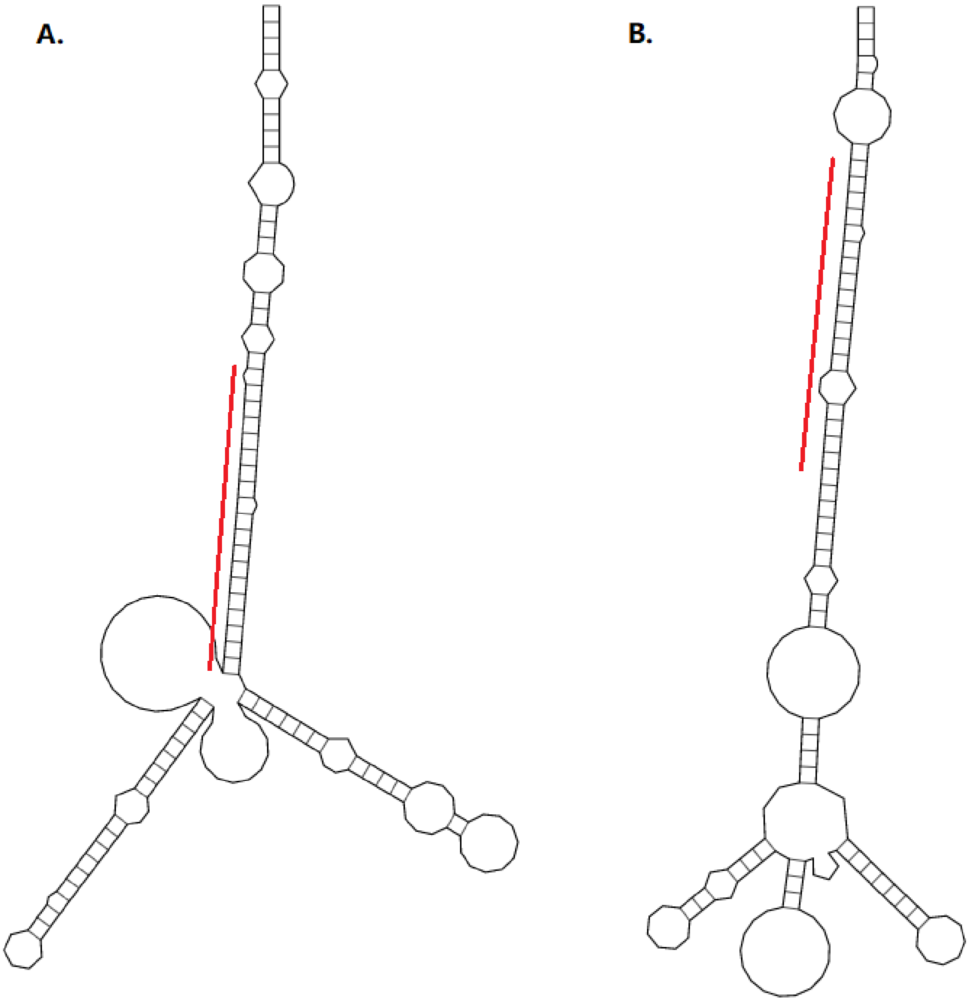
Lei, J. and Sun, Y. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics* 2014;30(19):2837-2839.

Mathelier, A. and Carbone, A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* 2010;26(18):2226-2234.

Yang, X. and Li, L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 2011;27(18):2614-2615.

**Supplementary File 2. Examples of authentic miRNAs with bifurcate structure in loops.**

Predicted secondary structure of two authentic miRNAs in Arabidopsis that failed to be detected by miRDeep-P, but could be retrieved by miRDeep-P2 (miRDP2). The red lines indicate locations of mature sequences. A. The secondary structure of *Ath-MIR157c*. B. The secondary structure of *Ath-MIR858*.



**Supplementary File 3. Updated criteria for plant miRNA annotation and criteria for 23-nt and 24-nt miRNAs.**

Criteria for normal miRNAs (Axtell and Meyers, 2018):

1. One or more miRNA/miRNA* duplexes with two-nucleotide 3’ overhangs, excluding secondary stems or large loops in the miRNA/miRNA* duplex and limiting precursor length to 300 nucleotides.
2. Confirmation of both the mature miRNA and its miRNA* only by sRNA-seq.
3. miRNA/miRNA* duplex contains ≤ 5 mismatched bases, and has at most one asymmetric bulge containing at most 3 bulged nucleotides.
4. ≥75% of reads from exact miRNA or miRNA*, including one-nucleotide positional variants of miRNA and miRNA* when calculating precision.
5. Novel annotations should meet all criteria in at least two sRNA-seq libraries (biological replicates).
6. Homology-based annotations should be noted as provisional, pending actual fulfillment of all criteria by sRNA-seq.
7. No RNAs <20 nucleotide or >24 nucleotides should be annotated as miRNAs. Annotations of 23- or 24-nucleotide miRNAs require extremely strong evidence.

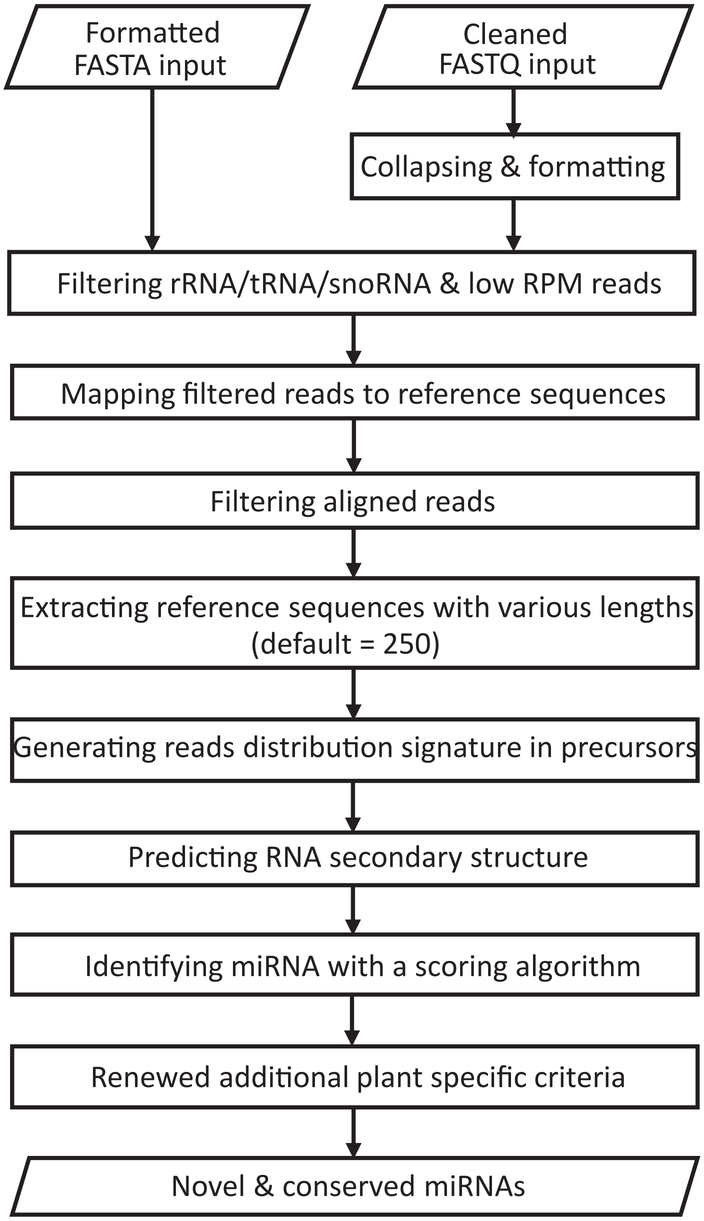
Criteria for 23- and 24-nt miRNAs: Beside all criteria above, the following 2 requirements added.

1. Reads corresponding to mature miRNA with RPM (reads per million ) ≥ 20.
2. miRNA/miRNA* duplex contains at least 1 mismatched bases or bulge.
3. The miRNA* must have corresponding reads.

References

Axtell, M.J. and Meyers, B.C. Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. *Plant Cell* 2018;30(2):272-284.

# Supplementary File 4. Diagram of the workflow of miRDP2


Formatted FASTA input
Cleaned FASTQ input
Collapsing & formatting
Filtering rRNA/tRNA/snoRNA & low RPM reads
Mapping filtered reads to reference sequences
Filtering aligned reads
Extracting reference sequences with various lengths (default = 250)
Generating reads distribution signature in precursors
Predicting RNA secondary structure
Identifying miRNA with a scoring algorithm
Renewed additional plant specific criteria
Novel & conserved miRNAs

# OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Feb 13, 2019

This Agreement between Xiaozeng Yang ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

| | |
|---|---|
| License Number | 4527340933846 |
| License date | Feb 13, 2019 |
| Licensed Content Publisher | Oxford University Press |
| Licensed Content Publication | Bioinformatics |
| Licensed Content Title | miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants |
| Licensed Content Author | Kuang, Zheng; Wang, Ying |
| Licensed Content Date | Dec 6, 2018 |
| Type of Use | Journal |
| Requestor type | Author of this OUP content |
| Pharmaceutical support or sponsorship for this project | No |
| Format | Electronic |
| Portion | Figure/table |
| Number of figures/tables | 5 |
| Will you be translating? | No |
| Circulation/distribution | 1000 |
| Title of new article | A Bioinformatics Pipeline to Accurately and Efficiently Analyze the microRNA Transcriptome in Plants |
| Lead author | Ying Wang |
| Title of targeted journal | Journal of Visualized Experiments |
| Publisher | MyJove Corp. |
| Expected publication date | Mar 2019 |
| Portions | figure 1， supplementary material 1， 3 4 and 5 |
| Requestor Location | Xiaozeng Yang<br>shu guang hua yuan zhong lu 11 hao<br><br>Beijing, other<br>China<br>Attn: |
| Publisher Tax ID | GB125506730 |
| Billing Type | Invoice |
| Billing Address | Xiaozeng Yang<br>shu guang hua yuan zhong lu 11 hao<br><br>Beijing, China<br>Attn: Xiaozeng Yang |

| | |
|---|---|
| Total | 0.00 USD |

Terms and Conditions

## STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all

claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
12. Other Terms and Conditions:
v1.4

**Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.**