BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 19 2014, pages 2837–2839
doi:10.1093/bioinformatics/btu380

Data and text mining

Advance Access publication June 14, 2014

miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data

Jikai Lei and Yanni Sun*

Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

Associate Editor: Igor Jurisica

ABSTRACT

Summary: Plant microRNA prediction tools that use small RNA-sequencing data are emerging quickly. These existing tools have at least one of the following problems: (i) high false-positive rate; (ii) long running time; (iii) work only for genomes in their databases; (iv) hard to install or use. We developed miR-PREFeR (miRNA PREdiction From small RNA-Seq data), which uses expression patterns of miRNA and follows the criteria for plant microRNA annotation to accurately predict plant miRNAs from one or more small RNA-Seq data samples of the same species. We tested miR-PREFeR on several plant species. The results show that miR-PREFeR is sensitive, accurate, fast and has low-memory footprint.

Availability and implementation: https://github.com/hangelwen/miR-PREFeR

Contact: yannisun@msu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 29, 2013; revised on May 18, 2014; accepted on June 2, 2014

1 INTRODUCTION

Plant microRNAs (miRNAs) are ~21-nt-long non-coding RNAs that play important roles in transcriptional and post-transcriptional regulation of gene expression (Zhang *et al.*, 2006). Recently developed genome-wide miRNA annotation tools all use small RNA-Seq data to quantify the expression of annotated miRNAs and to predict novel ones (Wang *et al.*, 2009; Yang and Li, 2011; Hackenberg *et al.*, 2011; Axtell, 2013). These tools suffer from several of the following problems. First, these tools usually have variable sensitivity and high false-positive (FP) rate when applied to different species. Second, most of existing NGS-based tools are slow. Third, most existing command-line-based tools are not user friendly. Web-server tools, such as miRanalyzer (Hackenberg *et al.*, 2011), are easy to use. But they usually only work for genomes in their databases, which inhibit users to analyze new genomes. In addition, most of web-server tools also have other problems listed here. Thus, there is a need for a plant miRNA prediction tool that has good performance (high sensitivity, low FP rate and accurate), works for all plant genomes, runs fast, has a small memory footprint and is easy to use.

miR-PFEFeR uses expression patterns of miRNAs and follows the criteria for plant microRNA annotation (Meyers *et al.*, 2008) to accurately predict plant miRNAs from one or more small RNA-Seq data samples. It has high sensitivity and low FP rate. miR-PREFeR is much faster and uses less memory than existing tools. Using miR-PREFeR requires minimum informatics expertize: it has low dependency on other programs; there is no need to compile or install the pipeline; it provides a checkpoint feature, which makes it easy to continue an unfinished job from where it was stopped; and the documentation is publicly available. The miR-PREFeR pipeline is in the process of being incorporated into the MAKER-P genome annotation engine (Campbell *et al.*, 2014).

2 DESCRIPTION OF THE PIPELINE

The miR-PREFeR pipeline takes a FASTA format genome file of a species and one or multiple SAM format small RNA-Seq read alignment files of the same species as input. Users can input an optional annotation file in GFF3 format to specify regions that are already annotated. The output includes the following: (i) An HTML table that summarizes the results of the predicted miRNAs. It contains the number of predictions, the length distribution of the predicted mature miRNA sequences, the list of all pre-miRNA sequences and their secondary structures and the detailed read mapping profiles against the pre-miRNA sequences. In addition, this table provides search links to mirBase (Kozomara and Griffiths-Jones, 2011). (ii) Separate files for easy downstream analysis. Details about the files can be found in the Supplementary File.

The pipeline first generates candidate regions and candidate mature sequences of each candidate region based on the alignment depth. In the next step, these regions are folded using RNALfold (Lorenz *et al.*, 2011). Regions with qualified stem-loop structures are then examined using published plant miRNA annotation criteria (Meyers *et al.*, 2008). The primary criteria is that the small RNA-Seq data should provide evidence of precise miRNA/miRNA* excision. Other criteria are mainly related to structure characteristics of the miRNA/miRNA* duplex. In addition to the annotation criteria, expression information from multiple small RNA-Seq data samples (if the input contains more than one sample) is also used to improve the accuracy of the prediction. Details of the method can be found in the Supplementary File.

By default, the pipeline makes a checkpoint after each major step of a job, and makes checkpoints periodically within the time-consuming folding stage. Users can easily set the checkpoint granularity in the configuration file. This provides users an easy way to restart unfinished jobs. For example, it is possible that a

*To whom correspondence should be addressed.

Table 1. Performance comparison of benchmarked tools

Dataset	miR-PREFeR	ShortStack	mirDeep-P	miRanalyzer	mirDeep2	mirDeep*	MIReNA
Software version	v0.09	0-5-0	1.3	unversioned	0.0.5	v31	2.0
Athl-2 (two samples. Number of known miRNAs expressed[a]: 240)							
Number of predicted miRs	155	113	1263	2182	182	2018	152
Number of expressed miRs predicted	127	107	86	201	64	10	35
Number of novel predictions	28	6	1177	1981	118	2008	117
Sensitivity	0.53	0.45	0.36	0.84	0.27	0.04	0.15
Athl-6 (six samples. Number of known miRNAs expressed[a]: 243)							
Number of predicted miRs	185	136	3021	13 114	291	1472	411
Number of expressed miRs predicted	136	125	128	209	79	7	44
Number of novel predictions	49	11	2893	12 306	212	1465	367
Sensitivity	0.56	0.51	0.53	0.86	0.33	0.03	0.18

[a]Known miRNA annotations are from miRBase v20. An miRNA is expressed if at least 20 reads were mapped to the miRNA precursor region in the dataset.

job running on a high-performance computing system is killed because of resource limits, or the laptop that the user works on runs out of battery. By restarting a job from the latest checkpoint other than starting it from the beginning, a lot of time/resources can be saved for long-running jobs on large plant genomes. As far as we know, miR-PREFeR is the first microRNA prediction tool that provides such a useful feature.

3 IMPLEMENTATION AND RESULTS

miR-PREFeR is implemented using Python. There is no need for the users to install it, and it works under Linux/MacOS/Windows. The pipeline is able to use multiple CPUs/cores automatically. This makes the pipeline much faster than existing programs and is easier for the users to use their computational resources.

We benchmarked miR-PREFeR and several popular or newly developed NGS-based miRNA prediction tools including ShortStack (Axtell, 2013), miRDeep-P (Yang and Li, 2011), miRanalyzer (Hackenberg *et al.*, 2011), miRDeep2 (Friedlander *et al.*, 2012), miRDeep* (An *et al.*, 2013) and MIReNA (Mathelier and Carbone, 2010). The features of the tools can be found in Supplementary Table S1. We used two datasets from *Arabidopsis thaliana* to evaluate the performance of the tools. The first dataset, Athl-2, which contains two samples, is the dataset used in the ShortStack paper (Axtell, 2013). The second dataset, Athl-6, contains six published samples from different tissues/conditions. The details of the datasets can be found in the Supplementary Table S2. Table 1 shows the performance of each tool on the two datasets. miRanalyzer has the best sensitivity, but the numbers of predictions are large on both datasets. It is likely that most of these predictions are FP predictions. It should be noted that miRanalyzer tries to detect annotated miRNAs that are saved in its own database. This is why its sensitivity is higher than all other tools. miRDeep2, miRDeep* and MIReNA have low sensitivity on both datasets, which indicates that they should not be the first choice for annotating plant miRNA. ShortStack is designed to be specific; thus, it has the lowest FP rate and reasonable sensitivity. miR-PREFeR has the second highest sensitivity and low FP rate. The predictions from miR-PREFeR have large intersections with predictions from ShortStack (See Supplementary Fig. S4). Details about the datasets and experiments from more species (maize, Medicago, papaya and peach) can be found in the Supplementary File.

Most of the predictions from miR-PREFeR correspond to previously annotated miRNAs. In all, 77.8% of the miR-PREFeR-predicted miRNAs have the same start positions as their annotations. In all, 81% of the predictions have the same lengths as the annotations. The high consistency was also observed in a large-scale study about miRNA isoforms in *A.thaliana* (Jeong *et al.*, 2013). On the other hand, other tools show much lower consistency with previously annotated miRNAs. More detailed analysis can be found in section 3.1 of the Supplementary File.

miR-PREFeR achieves $4\times \sim 37\times$ speedup compared with ShortStack, which can annotate other types of small RNAs and is the second fastest tool (see Supplementary Tables S4 and S5). miRDeep-P and MIReNA have long running time on both datasets. Without manually parallelizing the jobs, it is even difficult to run the two tools on small genomes on a personal computer. miR-PREFeR uses less memory than the other tools on both datasets.

Funding: This work was supported by NSF CAREER Grant DBI-0953738 and NSF IOS-1126998.

Conflict of Interest: none declared.

REFERENCES

An,J. *et al.* (2013) miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res.*, **41**, 727–737.

Axtell,M.J. (2013) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, **19**, 740–751.

Campbell,M. *et al.* (2014) MAKER-P: a tool-kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.*, **164**, 513–524.

Friedlander,M.R. *et al.* (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.

Hackenberg,M. *et al.* (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.

Jeong,D.H. *et al.* (2013) Comprehensive investigation of microRNAs enhanced by analysis of sequence variants, expression patterns, ARGONAUTE loading, and target cleavage. *Plant Physiol.*, **162**, 1225–1245.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

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

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

Wang,W.C. *et al.* (2009) miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics*, **10**, 328.

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

Zhang,B. *et al.* (2006) Plant microRNA: a small regulatory molecule with big impact. *Dev. Biol.*, **289**, 3–16.