BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 18 2011, pages 2614–2615
doi:10.1093/bioinformatics/btr430

Data and text mining

Advance Access publication July 19, 2011

miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants

Xiaozeng Yang and Lei Li*

Department of Biology, University of Virginia, Charlottesville, VA 22904, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Ultra-deep sampling of small RNA libraries by next-generation sequencing has provided rich information on the microRNA (miRNA) transcriptome of various plant species. However, few computational tools have been developed to effectively deconvolute the complex information.

Results: We sought to employ the signature distribution of small RNA reads along the miRNA precursor as a model in plants to profile expression of known miRNA genes and to identify novel ones. A freely available package, miRDeep-P, was developed by modifying miRDeep, which is based on a probabilistic model of miRNA biogenesis in animals, with a plant-specific scoring system and filtering criteria. We have tested miRDeep-P on eight small RNA libraries derived from three plants. Our results demonstrate miRDeep-P as an effective and easy-to-use tool for characterizing the miRNA transcriptome in plants.

Availability: http://faculty.virginia.edu/lilab/miRDP/

Contact: ll4jn@virginia.edu

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

Received on May 12, 2011; revised on June 23, 2011; accepted on July 11, 2011

1 INTRODUCTION

miRNAs are an important class of endogenous small RNAs that regulate gene expression at the post-transcription level (Bartel, 2009). There has been a surge of interest in the past decade in identifying miRNAs and profiling their expression pattern using various experimental approaches (Wark *et al.*, 2008). Most recently, deep sequencing of specifically prepared low-molecular weight RNA libraries has been used for both purposes in diverse plant species (Fahlgren *et al.*, 2007; Zhu *et al.*, 2008). A major drawback of these efforts is the exclusive focus on mature miRNAs, the final gene product and ignorance of sequence information associated with other parts of the miRNA genes. New strategies and tools are thus highly desirable to analyze the increasingly available sequencing data to gain insights into the miRNA transcriptomes.

Although miRNAs are only 20–24 nt long, they are processed from longer, stem–loop structured precursors called pre-miRNAs (Bartel, 2009). Maturation of miRNAs releases small RNAs derived from different parts of the stem–loop structure with asymmetric abundance. The program miRDeep employs a probabilistic model of miRNA biogenesis in animals to score compatibility of the nucleotide position and frequency of sequenced small RNA reads with the secondary structure of pre-miRNAs (Friedlander *et al.*, 2008). However, two significant differences in miRNA precursors between animals and plants prevent straightforward adaptation of miRDeep to the plant systems. First, plant pre-miRNAs are much longer with more variable lengths. Second, more miRNAs in plants belong to paralogous families with multiple members encoding identical or near-identical miRNAs (Supplementary Materials 1 and 2). We have demonstrated that miRDeep modified with plant-specific parameters is useful in analyzing the miRNA transcriptome in the model plant *Arabidopsis* (Yang *et al.*, 2011). Here we describe the improved package, called miRDeep-P, and its applications in plants.

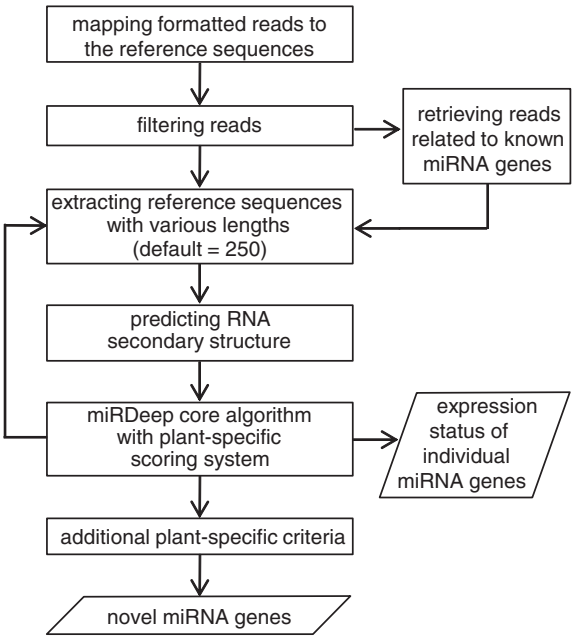

Fig. 1. Diagram of the workflow of miRDeep-P.

2 APPLICATION DESCRIPTION

2.1 Workflow of miRDeep-P

Based on ultra-deep sampling of small RNA libraries by next-generation sequencing, miRDeep-P enables users to explore expression patterns of annotated miRNA genes and discover novel ones. Figure 1 illustrates the workflow of miRDeep-P. To run

*To whom correspondence should be addressed.

this application, the reads should be preprocessed by removing adapters, discarding reads <15 nt and parsing them into FASTA format with their copy number recorded. With correctly formatted input files, miRDeep-P maps the reads to the reference (either genomic or transcriptomic) sequences using Bowtie (Langmead *et al.*, 2009). For a given mapped read, the optimal size of the window from which to extract reference sequences for predicting RNA secondary structure has been shown to be 250 bp (Yang *et al.*, 2011). However, miRDeep-P contains a module for users to empirically determine what window sizes to use in case a set of validated miRNA genes is available (Fig. 1). The secondary structures of the extracted reference sequences along with all reads mapped to such sequences are processed by the miRDeep core algorithm (Friedlander *et al.*, 2008) with a plant-specific scoring system (Supplementary Material 3). The output from the core algorithm is then filtered with additional plant-specific criteria based on known characteristics of plant miRNA genes (Meyers *et al.*, 2008). The overall process quantifies the signature distribution of small RNA reads and thereby provides reliable information on the transcription and processing of the pre-miRNAs. miRDeep-P uses such information to effectively profile the miRNA transcriptome (Fig. 1).

2.2 Identification of new miRNA genes

A major utility of miRDeep-P is to identify miRNA genes in plant species without detailed annotation. As long as there is sufficient read coverage, quantification of the signature small RNA distribution along reference sequences will be effective in revealing expressed pre-miRNAs from deeply sampled small RNA libraries. An advantage of miRDeep-P is that it outputs not only sequences of the putative miRNAs but also the stem–looped precursors and their location in reference sequences, which can be used to distinguish individual miRNA genes. Another advantage of miRDeep-P is that it does not require *a priori* information on sequence homology to known miRNA genes. This feature should be especially helpful to study the large complements of species-specific miRNA genes in plants (Fahlgren *et al.*, 2007).

2.3 Determination of the expression status of individual miRNA genes

A novel application of miRDeep-P is to assign expression status to individual miRNA genes. Although normalized frequency of the miRNA-matching reads can be used to estimate the expression level of miRNA genes (Fahlgren *et al*., 2007), the short length of the reads would mean cross-contamination among paralogous genes due to sequence similarity is potentially an issue. In miRDeep-P, this issue is overcome by quantifying the signature distribution of reads along the entire length of the miRNA precursors. This feature is especially useful to determine the expression status of paralogous miRNA genes that encode identical mature miRNAs. Meanwhile, if multiple libraries prepared from different biological samples (e.g. leaf, root, etc.) are employed, expression profiling of individual miRNA genes can be achieved as well.

3 IMPLEMENTATION AND RESULTS

The miRDeep-P package was developed in Perl by combining the core algorithm of miRDeep (Friedlander *et al.*, 2008), the mapping tool Bowtie (Langmead *et al.*, 2009) and the Vienna RNA package for predicting RNA secondary structure (Hofacker, 2003). Current version of miRDeep-P includes nine Perl scripts, which can be executed sequentially in a command line environment. All scripts have been tested on two Linux platforms, SUSE 10 and Fedora 14, and should work on similar systems that support Perl. The miRDeep-P scripts and user manual can be obtained from http://faculty.virginia.edu/lilab/miRDP/index.html as well as http://sourceforge.net/projects/mirdp/.

miRDeep-P has been tested using eight small RNA libraries from three plant species, *Arabidopsis*, rice and papaya. Both *Arabidopsis* and rice are well annotated for miRNA genes while there is no annotation in papaya (Griffiths-Jones, 2010). Based on these tests, it has been shown the optimal window size for extract precursor reference sequences is 250 bp for both dicot and monocot plants (Yang *et al.*, 2011). From the three *Arabidopsis* libraries, a total of 108 expressed (90 annotated and 18 novel) miRNA genes were detected. The two rice libraries yielded 158 annotated and 51 novel miRNA genes. Results from *Arabidopsis* have been successfully validated using other experimental approaches (Yang *et al.*, 2011), demonstrating the reliability of miRDeep-P. From the three papaya libraries, we detected 104 putative expressed miRNA genes of which 56 are conserved in other plant species and 48 are novel, further indicating that miRDeep-P is of broad use in plants.

Funding: This work was supported by a grant from National Science Foundation (DBI-0922526 to L.L.).

Conflict of Interest: none declared.

REFERENCES

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

Fahlgren,N. *et al.* (2007) High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS One*, **2**, e219.

Friedlander,M.R. *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.

Griffiths-Jones,S. (2010) miRBase: microRNA sequences and annotation. *Curr. Protoc. Bioinformatics*, **Chapter 12**, Unit 12 19 11–10.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

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

Wark,A.W. *et al.* (2008) Multiplexed detection methods for profiling microRNA expression in biological samples. *Angew. Chem. Int. Ed. Engl.*, **47**, 644–652.

Yang,X. *et al.* (2011) Global analysis of gene-level microRNA expression in Arabidopsis using deep sequencing data. *Genomics*, **98**, 40–46.

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