A user's manual for miRDeep-P

Last updated: 23-June-2011
Current as of miRDP version 1.3

Xiaozeng Yang
University of Virginia
Contact: xzyang@virginia.edu

Contents

1. OVERVIEW

1.1 BACKGROUND

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 miRNA, the final gene product, and ignorance of sequence information associated with other parts of the miRNA genes. New strategies and/or tools are thus highly desirable to analyze the increasingly available sequencing data to gain insights into the miRNA transcriptomes. The development of miRDeep-P, miRDP for short, was motivated by this need.

1.2 SUMMARY OF MIRDP FUNCTION

Based on ultra deep sampling of small RNA libraries by next generation sequencing, miRDP has two main purposes. First, miRDP can be used to identify miRNA genes in plant species, even for those without detailed annotation. Second, miRDP is designed to assign expression status to individual miRNA genes, which is critical as more miRNAs in plants belong to paralogous families with multiple members encoding identical or near-identical miRNAs.

1.3 IMPLEMENTATION AND ALGORITHM

MiRDP is documented by Perl (Perl 5.8 or later versions) and makes use of fundamental packages from Perl library. All the scripts have been tested on two Linux platforms, SUSE 10 and Fedora 14, and should work on similar systems that support Perl.

The core algorithm of miRDP was developed by modifying miRDeep (Friedlander et al., 2008), which is based on a probabilistic model of miRNA biogenesis in animals, with a plant-specific scoring system and filtering criteria.

1.4 LICENSE AND AVAILABILITY

MiRDP is freely available under a GNU Public License (Version 3) at:

http://faculty.virginia.edu/lilab/miRDP/index.html and http://sourceforge.net/projects/mirdp/

The miRDeep-P scripts, demos and user manual can be obtained from both web sites.

1.5 CONVENTIONS AND RECOMMENDATIONS

All command lines, filenames and directory names are in The command lines are separated by a blank line. The line starting with # is an interpretation of the following command line.

Two attached demo packages, (to explore expression patterns of known miRNA genes in Arabidopsis (Yang et al., 2011)) and (to detect new miRNA genes in Arabidopsis), which include the files by which users can reproduce every step and gain familiarity with miRDP. Please note that some of the intermediate files of the (e.g. the bowtie aligned result) are not included due to the size of these files. The users are recommended to generate these files by themselves.

2. INSTALLATION

Several dependencies are required to run miRDP. First, the Bowtie package can be downloaded from the site: http://bowtie-bio.sourceforge.net/index.shtml. Second, the Vienna package should be downloaded from the site: http://www.tbi.univie.ac.at/~ivo/RNA/. Third, you should

set paths to include the location of the downloaded miRDP scripts, as well as the directories where you have the bowtie, and Vienna executables. This is done by adding the lines:

to the or equivalent file ('location' designates the desired location in the file system).

3. PREPROCESSING THE READS

Before reads are mapped to the genome, they must be preprocessed. First, the deep sequencing reads should have the adapters removed from 5' and 3' ends (if present). Second, deep sequencing reads shorter than 15 nt should be discarded, since they will otherwise flood the mapping output. Third, the deep sequencing reads must be parsed into FASTA format. Forth, redundancy should be removed such that reads with identical sequence are represented with a single FASTA entry. Therefore, each sequence identifier must end with a '_x' and an integer, with the integer indicating the number of times the exact sequence was retrieved in the deep sequencing dataset. Finally, all of the FASTA ids should be unique. One way to ensure this is to include a running number in the id. For reference, see the file, , in the package. The following are several examples:

>ShootPi1000000_x3
AGAGAGTTCTTACATAAGATCTA
>ShootPi1000001_x1
TCCTTGATTTGATGCAACTAAAT
>ShootPi1000002_x3
TCTTGATTCTATGGGTGGTGGTG
>ShootPi1000003_x1
TAAGATCTCAAAGGAATTAGCAT

4. RETRIEVING EXPRESSION PATTERNS OF KNOWN MIRNA GENES AND OBTAINING THE OPTIMAL LENGTH OF PRECURSORS

One main function of miRDP is designed to assign expression status to individual miRNA genes. The following steps show how to get the expression patterns of known miRNA genes in Arabidopsis.

First, the annotated precursors of miRNAs are extended based on genomic sequences and general feature information by the script, .

For Arabidopsis, we add 500 nt at both sides of annotated precursors of miRNAs. Note that the sequence ids in must be identical to those in . downloaded from miRBase (release15) in the GFF (General Feature Format) format, including annotated miRNA general information.

Second, retrieve annotated miRNA precursors at a specific length by the following commands.

#indexing for a bowtie mapping search (please check bowtie manual for how to use bowtie-build).

#mapping reads to (only keep the perfect alignments (full length, 100% identity) and all the valid alignments. Please check bowtie manual for how to use bowtie.

#converting bowtie format to blastparsed format (a miRDeep format).

#excising candidate precursors at a specific length (the third option of the script, , is the length of candidate precursors, here we choose 250 as an example)

#indexing the candidate precursors and predicting their secondary structure (the secondary structures of the potential precursors are predicted using RNAfold and -noPS means that no graphical output is produced).

#aligning reads to candidate precursors and achieving signatures.

note that it is necessary for the prediction to do this sorting.

#retrieving annotated miRNAs

At different lengths of candidate precursors, the retrieved number of miRNAs is different. In Arabidopsis, our result shows at 250 bp most miRNA genes could be retrieved. Thus, 250 is the optimal length for detecting miRNAs in Arabidopsis (Yang et al., 2011). As an option, different lengths, e.g. 100, 150, 200, 250, 300, 350 and 400 bp, can be tested for retrieval of annotated

miRNAs in other plant species with the same commands as above. Meanwhile, if multiple libraries prepared from different biological samples are employed, expression profiling of individual miRNA genes can be achieved.

5. DETECTING NEW MIRNAS

5-1. ALIGNING THE READS TO THE REFERENCE GENOME

This step is to align the deep sequencing reads to the genomic (or transcriptomic if preferred) sequences. It is computationally demanding when the reference genome or small RNA library is large. The indexed TAIR genomic file can be downloaded from the bowtie website (http://bowtie-bio.sourceforge.net/index.shtml) or indexed by the users.

Then, align the deep sequencing reads to the indexed genome. Only keep the perfect alignments and all the valid alignments.

After that, convert it into blast format.

5-2. FILTERING UBIQUITOUS ALIGNMENTS

The reads are now filtered such that only perfect alignments (full length, 100% identity) are retained. Then, filter the reads which are mapped to multiple positions in the genome. For Arabidopsis, 15 is set as the cutoff since the largest miRNA family, miR169, has 14 members.

For other species, different cutoffs, based on the known family sizes or other empirical considerations such as genome sizes, might be selected.

5-3. FILTERING ALIGNMENTS BY ANNOTATION

The reads which mapped to exons or other non-coding RNAs, including rRNA, snRNA, snoRNA, and tRNA, are filtered by:

Only alignments where the read ids are not included in the file are retained. –g designates that lines where the query read ids are included in the file are discarded:

Reads can also be filtered such that only reads that have one or more remaining alignments are kept. -b designates that the output should be FASTA entries and not alignments:

5-4. PREDICTING MICRORNAS

Using the remaining alignments as guidelines, the potential precursor sequences are excised from the genome. This step is time-consuming, especially when the reference genome is large.

The secondary structures of the potential precursors are predicted using RNAfold. -noPS means that no graphical output is produced.

The signatures are generated by aligning the remaining reads to the potential precursors.

#Note that it is necessary for the prediction to do this sorting.

Predictions are made:

Note that there are several parameters of miRDeep that can be custom adjusted. See the miRDeep reference.

5-5. REMOVING REDUDANT PREDICTED MICRORNAS AND FILTERING PREDICTED ONES BY PLANT CRITERIA

One issue that miRDeep has not dealt with is that in some cases, there are redundant predicted items if the precursors are extracted by mapped reads that are closely located at the same

chromosome loci. Meanwhile, recently updated criteria of plant miRNAs are considered critical in identifying new species- or tissue-specific miRNAs (Meyers et al. 2008). The script, can remove redundant items and filter the items out that do not meet the criteria of plant miRNAs.

The file, , includes the information on chromosome length. The file, , is the output file which contains non-redundant predicted miRNA information. The file, , is also the output file, which contains predicted miRNAs that meet the criteria of plant miRNAs. For the details of the format of the two output files, please see Section 6 of this manual.

6. THE MIRDP SOFTWARE PACKAGE

The miRDP package consists of nine documented Perl scripts that should be run sequentially by the user. Of the nine scripts, three, and are inherited from miRDeep (Friedlander et al., 2008). The other scripts are either novel or have been modified from the original miRDeep version. Functions of the nine scripts are described in the following:

a. fetches the extended precursors from reference sequences based on the location information of annotated miRNAs. The gff file could be downloaded from miRBase (http://www.miRBase.org).

b. changes the bowtie format into blastparsed format. Blastparsed format is a custom tabular separated format derived from standard NCBI blast output format.

c. changes the SAM format into blastparsed format. Note that when use aligners which could generate SAM format file, all perfect valid alignments should be kept. When using bowtie, for instance, the option –a and –v 0 indicate reporting all perfect valid alignment.

d. 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 blastparsed format.

e. can be used (user-specified) to remove reads that align to the genome in positions that overlap with selected annotation tracks provided by the user (for example known rRNAs, tRNAs, etc). The basic input is a file in blastparsed format and an annotation file in standard gff format. In fact, just some aspects of the GFF format, including seqname, start, end, and strand, are used. (For details on GFF format, please check http://genome.ucsc.edu/FAQ/FAQformat.html#format3)

f. cleans the ids overlapping with ncRNA or exons.

g. cuts out potential precursor sequences from a reference sequence using aligned reads as guidelines. The basic input is a file in blastparsed format and a FASTA file. The basic output is also in FASTA format.

h. needs two input files, signature file and structure file, which is modified from the core miRDeep algorithm by changing the scoring system with plant specific parameters.

i. exactly picks out the predicted miRNAs which are the annotated ones.

j. needs three input files: chromosome_length, precursors and original_prediction generated by . It generates two output files, non-redundant

predicted file and predicted file filtered by plant criteria. The tab-delimited files contain columns that indicate chromosome id, strand direction, reads id, precursor id, mature miRNA location, precursor location, mature sequences, and precursor sequences.

7. ISSUES USING MIRDP

7.1 REDUNDANCY AND MIRNA*

In some cases, the output miRNAs from miRDP 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. However, if users desire to exactly extract the annotated miRNAs, the optional script
could be used to assist for this purpose.

is a file including miRNA precursor and mature miRNA location information from the annotation.

7.2 COMPUTATIONAL TIME

There are three steps that might be time-consuming when use miRDP, especially when the size of small RNA library and the genome are both large.

a. the process of mapping reads to genome sequences (command line likes:

). To save time, you may want to download bowtie index files from the bowtie website (http://bowtie-bio.sourceforge.net/index.shtml) if the genome sequences of the species you are working with have been indexed. Otherwise, you should index reference sequences by yourself. Please keep the index file for a while till you have finish your project since you might need to re-index your genome.

b. the process of obtaining candidate precursors(command line likes:

). This step might take hours of computational time on a modest-sized cluster. A good strategy is to divide the xxx.bst file into smaller sub-groups based on chromosomes or contigs. Running the divided sub-groups simultaneously could significantly shorten the computational time.

c. the process of exploring secondary structure of candidate precursors(command line likes:

). If the number of candidate precursors is huge, the strategy of dividing them into smaller sub-groups and running the sub-groups at the same time can be used.

8. REFERENCES

Bartel, D.P. (2009) MicroRNAs: Target Recognition and Regulatory Functions, , **136**, 215-233.

Fahlgren, N., et al. (2007) High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes, , **2**, e219.

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

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

Wark, A.W., Lee, H.J. and Corn, R.M. (2008) Multiplexed detection methods for profiling microRNA expression in biological samples, , **47**, 644-652.

Yang, X., Zhang, H. and Li, L. (2011) Global analysis of gene-level microRNA expression in Arabidopsis using deep sequencing data, , Epub ahead of print.

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