Molecular and Cellular Probes 26 (2012) 208–211

Contents lists available at SciVerse ScienceDirect

Molecular and Cellular Probes

journal homepage: www.elsevier.com/locate/ymcpr

miRSeqNovel: An R based workflow for analyzing miRNA sequencing data

Kui Qian [a,*], Eeva Auvinen [b,c], Dario Greco [d], Petri Auvinen [a]

[a] DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki, Helsinki, Finland
[b] Haartman Institute, Department of Virology, Helsinki, Finland
[c] Helsinki University Hospital Laboratory, Department of Virology and Immunology, Helsinki, Finland
[d] Department of Bioscience and Nutrition, Karolinska Institutet, Sweden

ARTICLE INFO

Article history:
Received 28 February 2012
Received in revised form
4 May 2012
Accepted 4 May 2012
Available online 17 May 2012

Keywords:
miRNA sequencing
Differentially expressed miRNAs
Novel miRNA prediction

ABSTRACT

We present miRSeqNovel, an R based workflow for miRNA sequencing data analysis. miRSeqNovel can process both colorspace (SOLiD) and basespace (Illumina/Solexa) data by different mapping algorithms. It finds differentially expressed miRNAs and gives conservative prediction of novel miRNA candidates with customized parameters.

miRSeqNovel is freely available at http://sourceforge.net/projects/mirseq/files.

With the advantages of next generation sequencing (NGS) platforms, new opportunities have arisen to quantitate the expression of known miRNAs as well as to predict novel miRNAs. A number of tools have been developed for the analysis of miRNA sequencing data, such as stand-alone software miRDeep2 [1], or web based tools like miRanalyzer [2] and UEA sRNA toolkit [3] (the latter two also provide stand-alone versions). However, there are limitations in these methods. First, some of them are designed to support basespace format only, such as miRDeep2. Second, methods such as miRDeep2 and miRanalyzer have their mandatory mapping method (i.e. Bowtie [4]) and they do not allow utilization of custom mapping methods or mapping parameters. Third, web versions of miRanalyzer and UEA sRNA toolkit only support a limited number of reference genomes, and have limitations on the number of input reads. Fourth, miRanalyzer and UEA sRNA toolkit depend on their inherent outdated miRNA annotations for novel miRNA prediction. miRDeep2 requires miRNA annotations of related species for novel miRNA discovery. Fifth, those methods have limited options for adjusting prediction parameters, which are needed to obtain reliable miRNA predictions because of varying lengths of pre-miRNA sequences and different complementarities between mature and star miRNAs in different species. An example is the length of gaps between star and mature sequences: long (~400 nt) in plants and short (~40 nt) in mammals. This option, for instance, is lacking in miRDeep2. Moreover, those methods do not have the flexibility of choosing statistical methods to find differentially expressed miRNAs, but they reply on a limited number of pre-defined methods. Based on the above facts, we find the existing methods not flexible enough for processing miRNA sequencing data.

We have developed a new workflow with improved flexibility and prediction accuracy for miRNA sequencing data processing, which we have named miRSeqNovel. miRSeqNovel can use output from popular mapping software, e.g. RNA2MAP [5], Bowtie [4] or BWA [6], which can map sequencing data from multiple platforms, including SOLiD ("csfasta" format) and Illumina/Solexa ("fastq" format) platform, to any reference genomes. The genome mapping output is combined with the known miRNA information from miRBase [7] (Fig. 1. A) to produce a table of read counts for known miRNAs. miRSeqNovel applies widely-used statistical methods, such as functions implemented in popular Bioconductor packages edgeR [8] and DESeq [9], or other user-specified methods, to discover differentially expressed miRNAs (Fig. 1. B). Next, reads mapped to known non-coding RNAs and exon regions are filtered according to Ensembl annotation [10] (optional step). Finally, the remaining reads will be used as an input for novel miRNA prediction (Fig. 1. C). miRSeqNovel uses mapped reads information to find candidate miRNA precursor sequences by screening their secondary structures. By assigning different sets of predicting parameters optimized for animal and plant genomes, we demonstrated that miRSeqNovel can successfully predict most known miRNAs and find conservative novel candidates.

* Corresponding author.
E-mail address: kui.qian@helsinki.fi (K. Qian).

0890-8508/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mcp.2012.05.002

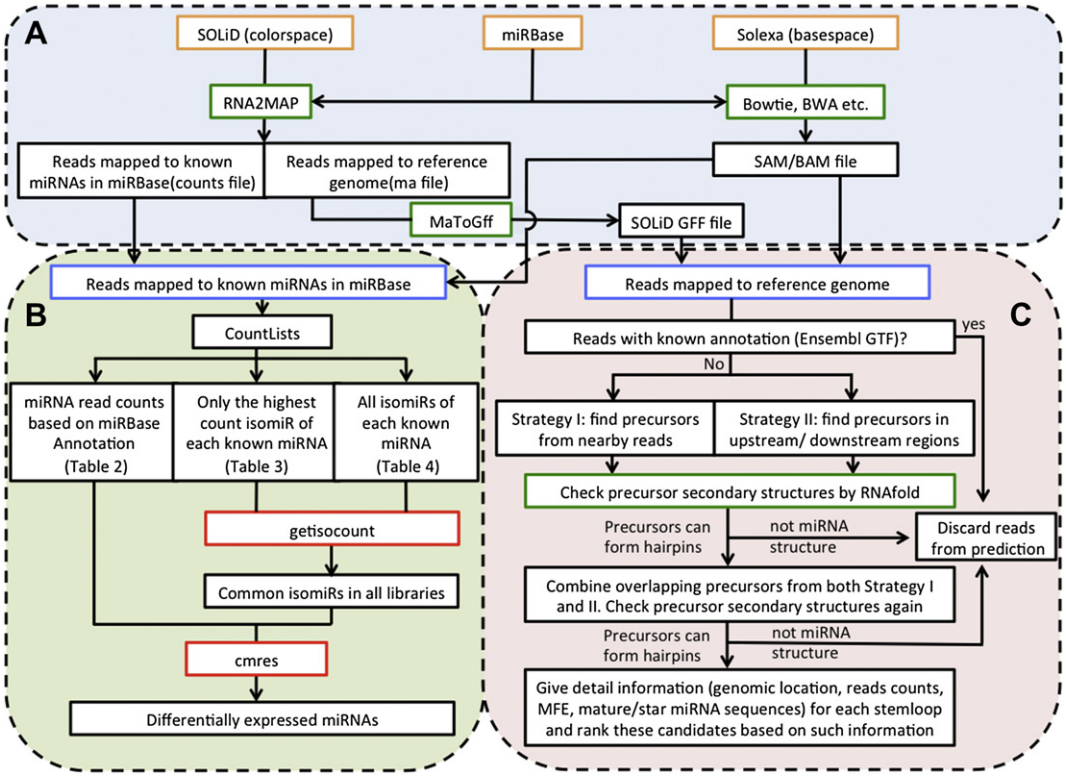

Fig. 1. miRSeqNovel analyzing workflow the miRNA sequencing data are mapped to reference genome (A) to identify differentially expressed miRNAs (B) and predict novel miRNA candidates (C). A) Raw data are mapped to the reference genome of interest by other suitable software. miRNA annotations from miRBase are used to measure expression levels of known miRNA. B) Common isomiRs of known miRNAs in CountLists are found by "getisocount" and differentially expressed miRNAs are tested by "cmres". C) Reads mapped in genome are used for two prediction strategies to discover novel miRNA candidates. Good candidates with miRNA-like hairpin structures are checked by RNAfold and ranked based on expression and structure information. Input data are in orange frames. Mapping software and other needed software are in green frames. Inputs for miRSeqNovel are in blue frames. Functions in miRSeqNovel are in red frames. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Below we describe the detailed processing schemes of the workflow (Fig. 1). SOLiD reads mapped onto the reference genome are saved as .ma files by RNA2MAP [5]. This can be converted into SOLiD GFF format by MaToGff [11]. SAM/BAM formatted output from Illumina/Solexa data can be loaded into R by ShortRead [12], rtracklayer [13] or Rsamtools [14] (Fig. 1. A). For known miRNAs, their reads will be exacted from total mapped reads and used as a GRange object. For the remaining reads, a GenomicRange object is created for each strand of every chromosome, which will be used individually in the novel miRNA prediction step. In the GenomicRange object, reads with the same location are combined into one record reporting its count number information.

The known miRNA annotations from miRBase are required to find differentially expressed miRNAs. It is known that isomiRs are common for miRNAs and the functional isomiRs may not be annotated in miRBase [15,16]. miRSeqNovel presents options for the user to decide whether or not to consider isomiRs (Fig. 1.B). For each sequencing library, the GRange object from pre-processing step will be converted to a CountList, which is an R list consisting of four different tables: Table 1, count number of reads whose genomic location matches exactly with each mature miRNA in miRBase; Table 2, count number of reads overlapping with each mature miRNA; Table 3, count number of the most abundant isomiR of each mature miRNA; Table 4, count numbers of all isomiRs that give certain contribution, for example, bigger than 10% of total counts, of each mature miRNA. Then for each table in the CountList, common isomiRs in each sample library are combined ("getisocount") to build a table with isomiR names in rows and sample names in columns for statistical testing. The new table can be directly used to find differentially expressed miRNAs by Bioconductor package tools such as edgeR [8] and DESeq [9] or other statistical methods the user prefers. In addition, miRSeqNovel provides a "cmres" function to integrate functions from edgeR and DESeq for the ease of usage. Table 1 is not used here, because some miRNAs have no reads in annotated locations but their isomiRs have many mapped reads. We assume that such isomiRs are the functional ones, and thus the count numbers of annotated sequences do not represent the expression of these miRNAs [15,16].

Prediction algorithm of novel miRNA using miRSeqNovel follows two strategies (Fig. 1. C), which are popularly used in other software, such as miRDeep2 and miRanalyzer. First, pairs of locations mapped within a certain range of each other onto the same DNA strand are considered as the mature and the star sequences of a putative pre-miRNA. Second, for a mapped read whose distance from its closest mapped read is longer than the defined gap, the genomic regions upstream and downstream of the read are searched for possible sequences forming a hairpin secondary structure. A similar strategy has also been used in miRanalyzer. Compared to miRDeep2 and miRanalyzer, miRSeqNovel would automatically adjust this flanking region by keeping the core hairpin sequence and trimming the unpaired sequence in both ends of the flanking region. This would identify better novel candidates because the unpaired sequence affects the secondary structure prediction. We applied the RNAfold program in ViennaRNA [17], to predict minimum energy secondary structures and base-pair probabilities.

We tested miRSeqNovel on three SOLiD/Solexa miRNA sequencing datasets of human/Arabidopsis samples from NCBI