BIOINFORMATICS DISCOVERY NOTE

Vol. 29 no. 20 2013, pages 2521–2523
doi:10.1093/bioinformatics/btt424

Sequence analysis

Advance Access publication August 14, 2013

isomiRID: a framework to identify microRNA isoforms

Luiz Felipe Valter de Oliveira[1], Ana Paula Christoff[1] and Rogerio Margis[1,2,*]

[1]Genetics and Molecular Biology Graduation Program and [2]Department of Biophysics, Center of Biotechnology, Universidade Federal do Rio Grande do Sul – UFRGS, Avenida Bento Goncalves 9500, Predio 43431, Sala 213, Porto Alegre, Brasil, CEP 91501-970

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: MicroRNAs (miRNAs) have been extensively studied owing to their important regulatory roles in genic expression. An increasingly number of reports are performing extensive data mining in small RNA sequencing libraries to detect miRNAs isoforms and also 5′ and 3′ post-transcriptional nucleotide additions, as well as edited miRNAs sequences. A ready to use pipeline, isomiRID, was developed to standardize and automatize the search for miRNAs isoforms in high-throughput small RNA sequencing libraries.

Availability: isomiRID is a command line Python script available at http://www.ufrgs.br/RNAi/isomiRID/.

Contact: rogerio.margis@ufrgs.br

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

Received on January 31, 2013; revised on July 16, 2013; accepted on July 17, 2013

1 INTRODUCTION

Several high-throughput small RNA (sRNA) sequencing projects have increased the discovery and number of small non-coding RNAs related to gene expression regulation. MicroRNAs (miRNAs) are endogenous sRNAs with 19–24 nt in length (Bartel, 2004). To regulate protein-coding genes, mature miRNA binds in the mRNA target sites leading to their degradation or repression of translation (Pasquinelli, 2012; Voinnet, 2009).

Recently, reports have demonstrated the existence of isomiRs (Cloonan *et al.*, 2011; Guo and Lu, 2010; Körbes *et al.*, 2012). The isomiR classification relies in three major categories: 5′, 3′ and polymorphic isomiRs, with the sub classification of 5′ and 3′ isomiRs into templated or non-templated modifications, according to the miRNA precursor sequence (Neilsen *et al.*, 2012). Furthermore, efforts have been made to identify and understand the biological processes where non-templated nucleotides are added in the 3′ end of mature miRNAs, altering the miRNA stability and efficiency (Burroughs *et al.*, 2010; Ebhardt *et al.*, 2009; Lu *et al.*, 2009; Wyman *et al.*, 2011). Nucleotide additions at the 5′ end have also been reported altering the miRNA seed region and consequently its functionality (Bizuayehu *et al.*, 2012; Ebhardt *et al.*, 2010). Several isomiR variants have been discovered with deep sequencing technologies. Some of these variations may have biological origin and functions, once the modifications were seen repeatedly at the same sites with higher frequencies than random errors (Ebhardt *et al.*, 2009).

*To whom correspondence should be addressed.

To improve and automate the search for isomiRs in sRNA sequenced libraries, we developed a simplified workflow. isomiRID allows the identification of 5′, 3′ and polymorphic isomiRs based on the canonical miRNA known sequence and also from other regions on the same miRNA precursor. Additionally, the program can also identify non-templated 5′ or 3′ end variations by mapping the sRNAs in the known pre-miRNAs. This framework is a simple method to identify and compare isomiRs abundance in different sequencing libraries, providing an initial view for biological relevant isomiRs. We also highlight that this pipeline could be applied to study isomiRs from plants, animals or any other living organisms.

2 METHODS

The isomiRID workflow (Fig. 1) has four main steps:

The first step generates the file of sRNAs with perfect matches on the pre-miRNA (R0 = round 0) and create the dataset for the subsequent analysis with the pre-miRNA unmapped reads. Optionally, it is possible to use a reference genome (or transcriptome), to filter the unmapped sRNAs from other genomic regions.

On the second step, sRNAs mapping is performed, allowing a single mismatch in the sequence. This first part generates the R1 file, containing sequences with one mismatch in the 5′ end (5′MM), in the middle of sequence (MM) and in the 3′ end (3′MM). Posteriorly, the small reads, which still do not match with the pre-miRNA sequence, are submitted to trimming rounds in their 5′ and 3′ ends. Each trim is performed once in a round, where a single 5′ or 3′ nt is removed for subsequent sRNA mapping on the pre-miRNA reference file. This procedure can be performed N times, according to the number of nucleotide trimmings determined by the user. Reads that match to the pre-miRNA reference after the trimming will be separated in R2, R3…RN files. These files contain a raw table with the mapped sequences, the name of miRNA precursor from which it potentially originated, their lengths, the variant nucleotide(s) and the abundance of such variations in each sequence on the analyzed library(ies) (Supplementary File 1).

Additionally, in a third step, the files with mapped reads can undergo an abundance cutoff filter, to restrict the read values, specified in the configuration file. Finally, the fourth step relies on concatenation of the isomiRs mapping results in a sub folder (MapResults). A graphical output is built based on the mappings with or without abundance cutoff. In graphical outputs, the pre-miRNA sequence is aligned with the miRNA and its isomiRs sequences, also detailing the individual sRNA variation and the abundance of each sequence in different libraries that were analyzed.

3 IMPLEMENTATION

The isomiRID was implemented with python 2.7 and use Bowtie for mapping the reads (Langmead *et al.*, 2009). This framework

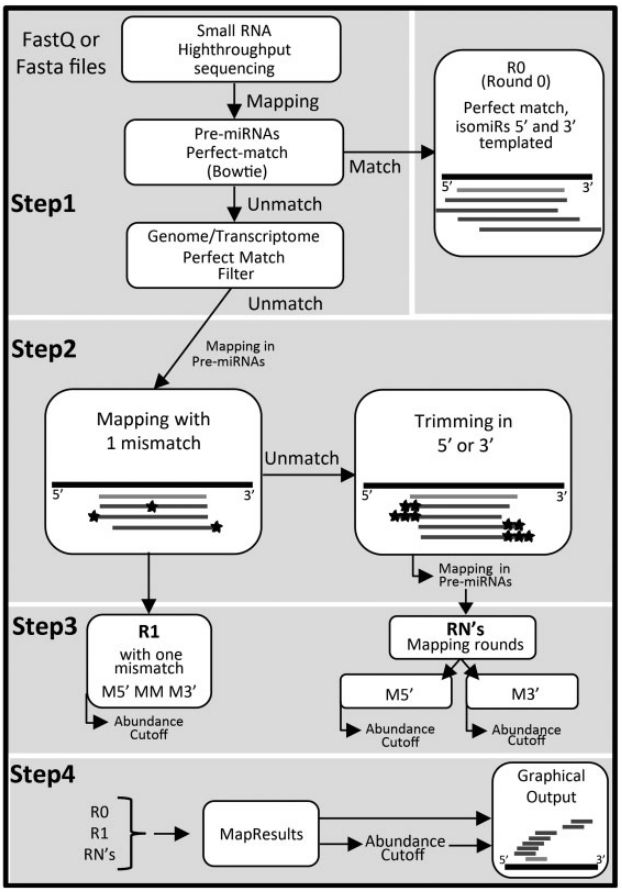

Fig. 1. Graphical representation of isomiRID workflow. (thick black line) pre-miRNA. (grey line) canonical miRNA. (thin black line) isomiRs. nucleotides mismatch. M5′ mismatch in 5′ end. MM mismatch in the middle of the sequence (asterisk). M3′ mismatch in the 3′ end (see text and supplementary material for detailed information)

was tested on Linux and Mac OSX environment and settled in a simple and standardized text file (see Supplementary File 1). With the use of this configuration file, it is possible to set more than one input sRNA sequencing file to be analyzed and compared. Sequence files can be in fastQ or fasta format. The filtering process using reference genome or transcriptome is optional, as well as the abundance cutoff, which can be applied for more than one value (e.g. 10, 50 and 100) corresponding to the reads sum in all libraries analyzed.

4 APPLICATIONS

The main goal of this workflow is to provide a tool to automate, standardize and simplify the searches for isomiRs and non-templated nucleotides in 3′ and 5′ miRNA ends, using as input single or multiple sequencing data files in a comparative way. The output files produced from this workflow are understandable tab delimited.txt files, which can be easily imported for spreadsheet or statistical suits, as R package (Dimitriadou *et al.*, 2010). Also, the aligned output provides a broad view of particular miRNAs and the abundance of their isomiRs sequences in different sequencing datasets.

To validate our workflow, we applied the isomiRID to a sRNA sequencing library previously reported (Wyman *et al.*, 2011) (GEO:GSM740469).This library from normal human prostate tissue encompasses several cases of 3′ miRNA additions, reported by the previous authors. isomiRID was also tested with *Arabidopsis thaliana* immunoprecipitated AGO1 sRNAs libraries, from two tissues (GSM707682 – flower and GSM707683 – leaf) (Wang *et al.*, 2011). These analyses were performed with miRNAs and pre-miRNA sequences retrieved from miRBase release 19 (Kozomara *et al.*, 2011), for each species, and filtered against the human genome release 11, available on Metazome (http://www.metazome.net) and arabidopsis genome v.9 (http://www.phytozome.net). The isomiRID outputs for prostate data can be visualized in Supplementary Table S1 and Supplementary Figure S1. Arabidopsis results can be achieved in Supplementary Table S2 and Supplementary Figure S2. In these result examples, the.xt files generated by the mappings (r0, r1, M5 and M3), with the cutoff of 50, were summarized in four different tabs in the Supplementary Tables S1 and S2. The Supplementary Figures S1 and S2 show a few examples of the graphical output for some miRNAs aligned with their respective pre-miRNA and isomiRs. More details about the output data interpretation can be found in Supplementary File 1, also supplied as a quick start guide in the software download page.

With the isomiRID, we are able to detect miRNA variations, known as isomiRs. Regarding nucleotide variations detected as a middle mismatch, they should be considered as individual biological variation (SNP) or the result of enzymatic modification like A-I editing (Bahn *et al.*, 2012). The 5′ and 3′ modifications that the isomiRID detects could be also originated by the two mechanisms cited earlier in the text, or they can also be the result from biological events of nucleotides additions by nucleotidyl transferases in the extremities of miRNAs. The use of more than one round of trimming allows higher accuracy in identification of 5′ and 3′ addition events for two or more nucleotides.

Funding: Support was provided by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, for the Ph.D fellowship to L.F.V.O. and A.P.C.; Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq for financial resources and CESUP/UFRGS for computational infrastructure.

Conflict of Interest: none declared.

REFERENCES

Bahn,J.H. (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.*, **22**, 142–150.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bizuayehu,T.T. *et al.* (2012) Differential expression patterns of conserved miRNAs and isomiRs during Atlantic halibut development. *BMC Genomics*, **13**, 11.

Burroughs,A.M. *et al.* (2010) A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res.*, **20**, 1398–1410.

Cloonan,N. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, R126.

Dimitriadou,E. *et al.* (2010) e1071: misc functions of the Department of Statistics (e1071), TUWien. R package version 1.5-24.

Ebhardt,H.A. *et al.* (2010) Naturally occurring variations in sequence length creates microRNA isoforms that differ in argonaute effector complex specificity. *Silence*, **1**, 12.

Ebhardt,H.A. *et al.* (2009) Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.*, **37**, 2461–2470.

Guo,L. and Lu,Z. (2010) Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data. *Comput. Biol. Chem.*, **34**, 165–171.

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

Körbes,A.P. *et al.* (2012) Identifying conserved and novel microRNAs in developing seeds of brassica napus using deep sequencing. *PLoS One*, **7**, e50663.

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

Lu,S. *et al.* (2009) Adenylation of plant miRNAs. *Nucleic Acids Res.*, **37**, 1878–1885.

Neilsen,C.T. *et al.* (2012) IsomiRs – the overlooked repertoire in the dynamic microRNAome. *Trends Genet.*, **28**, 544–549.

Pasquinelli,A.E. (2012) MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat. Rev. Genet.*, **13**, 271–282.

Voinnet,O. (2009) Origin, biogenesis, and activity of plant MicroRNAs. *Cell*, **136**, 669–687.

Wang,H. *et al.* (2011) Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. *Plant J.*, **67**, 292–304.

Wyman,S.K. *et al.* (2011) Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res.*, **21**, 1450–1456.