isomiRex: Web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets

Gaurav Sablok a,⇑,1, Ivan Milev b,1, Georgi Minkov b, Ivan Minkov b, Claudio Varotto a, Galina Yahubyan b, Vesselin Baev b,⇑

a Department of Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, Via E Mach 1, 380105 Michele all’Adige, TN, Italy
b Department of Plant Physiology and Molecular Biology, Bioinformatics Group, University of Plovdiv, Tzar Assen 24, Plovdiv, Bulgaria

1 These authors contributed equally to this work.

E-mail addresses: sabloksg@gmail.com (G. Sablok), vebaev@gmail.com (V. Baev).

A R T I C L E   I N F O
Article info
Received 3 April 2013
Revised 22 June 2013
Accepted 25 June 2013
Available online 4 July 2013

Keywords:
isomiRs
MicroRNA
Next generation sequencing
Web platform

A B S T R A C T
We present an open-access web platform isomiRex, to identify isomiRs and on the fly graphical visualization of the differentially expressed miRNAs in control as well as treated library. The open-access web-platform is not restricted only to NGS sequence dataset from animals and potentially analyzes a wider dataset for plants, animals and viral NGS dataset supporting miRBase (version 19 supporting 193 species). The platform can handle the bloated amount of the read counts and reports the annotated microRNAs from plant, animal and viral NGS datasets. isomiRex also provides an estimation of the the isomiRs, of miRNAs with higher copy number relative to their mature reference sequences indexed in miRBase (version 19 supporting 193 species). Visually enhanced graphs potentially display differentially expressed isomiRs, which will help the user to demonstrate and correlate the abundance of the isomiRs as a signature event to the specific condition. An additional module for estimating the differential expression has been implemented allowing the users to postulate the differential expression across the user input samples. The developed web-platform can be accessed at http://bioinfo1.uni-plovdiv.bg/isomiRex/.

© 2013 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

1. Introduction

MicroRNAs (miRNAs) are single-stranded small RNAs, which are endogenously generated and play varying levels of regulatory roles in modulating the transcriptional landscape well-characterized in several plant and animal model systems as well as non-sequenced organisms [1–4]. The mature miRNAs are short 20–24 nt RNAs that, together with Argonaute proteins, trigger direct post-transcriptional regulation [1]. miRNAs plays a important role as functional regulators in various plant and mammalian development pathways and regulate genes involved in cell division, cell differentiation, metabolism, stress response and apoptosis [1,5]. Several studies have widely reported the demonstrated effect of the regulation of the oncogenic pathways by miRNAs in human oncogenes, tumor-suppressors and a number of cancer-related genes, suggesting that a de-regulation step can pre-dispose to disease and malignancies [6,7]. Previous observation critically evaluated the effect of the fluctuations in the timing, and developmental stages of cancer explaining unique miRNA expression repertoire, and suggests the role of the miRNA genes as novel biomarkers to understand the disease introgression [6,8,9].

NGS widely demonstrated that miRNAs frequently exhibit differences from their corresponding “reference” mature sequences, generating multiple variants specifically abbreviated as “isomiRs”. isomiRs, sequence variants of miRNAs, differentially regulate the target selection, regulating the preferential loading of the RNA-induced silencing complex (RISC) [10]. Next-generation sequencing (NGS) of the small RNAs provides an efficient platform for the explanation of the ubiquitous and differentially expressed behavior of the miRNAs and corresponding isomiRs abundance. Recently small RNA sequencing of human tissues and cell lines suggests that 5'-isomiR-101 is a functional variant and potentially regulates the expression levels of the mir-101 targets by interacting with the different components of the RISC complex [11]. Furthermore, using a deep-sequencing target study, it has been demonstrated that miRNAs and isomiRs act as cooperative dependent manner and an abundant class of isomiRs were derived from the AGO2 cleavage independent of Dicer [12]. Evolutionary implications of isomiRs suggest a consistent expression as compared to their corresponding homologous miRNAs, which clearly depicts the consistent miRNAs maturation process and can be useful to elucidate the ancestral miRNA genes [13]. However, isomiRs generates in a non-random fashion and the degree of mature sequence heterogeneity is different for each locus in Prunus persica L. [14]. Systematic early investigation of the isomiRs from the Arabidopsis thaliana...
suggest the equipotential capability of the 5p and 3p arm of the miRNA precursor to generate the mature sequence [15].

In our previous study, we have demonstrated that in human uterine leiomyoma miRNAome, most of the identified miRNAs have corresponding isomiRs with higher copy numbers as compared to the referenced mature sequence [16]. On the basis of the observed results of the previous study, we developed an integrated computational web platform for handling and analyzing the NGS datasets from small RNA deep sequencing approaches and moreover to be able to intersect different set in order to identify new miRNA homologs or differentially expressed miRNAs at the same time point of the experimental condition. We propose a web-based platform for identifying conserved miRNA and their corresponding isomiR sequences in plant, animal and viral NGS datasets and to estimate the patterns of differential expression to elucidate the differentially expressed members. The developed web platform is written in Perl (v5.12.4) and R (2.13.1) and can be widely accessed via http://bioinfo1.uni-plovdiv.bg/isomiRex/.

2. Materials and methods

2.1. Reference databases and NGS datasets

For the identification, a reference dataset consisting of all known mature and precursor sequences of miRNAs downloaded from miRBase (version 19, http://www.mirbase.org) [17], covering all the available 193 species has been pre-indexed in FASTA format and is made available to the user as reference sequences using the query module. In total, the present version of the isomiRex supports 21,264 precursor miRNAs and 25,141 mature miRNA as a pre-built index file. A calibration and quality check of the developed pipeline integrated in web platform was performed using the recently published four NGS Illumina sequenced libraries from human uterine leiomyoma small RNA-seq (myometrium control) expression data from known miRNAs produced as reads count will be subsequently used as input in the Bioconductor DESeq package in R (http://www.r-project.org/re) [19]. DESeq uses a negative binomial distribution model to test and identify the differential expression in deep sequencing datasets. The list of differentially expressed miRNAs produced by DESeq was further filtered to remove miRNAs with less than five read support in all studied samples to obtain a homogenous and high coverage. A statistical filter of \( P < 0.05 \) has been applied to filter the differentially expressed miRNAs. Visualization of the differential expression is displayed through a heatmap using gplots package in R (v2.10.1).

2.2. Read mapping and identification of isomiRs

To efficiently map the reads, isomiRex by default uses Bowtie as the read mapper, with no mismatches enabled mode (-v 0) [18]. The present version of the isomiRex allows the users to choose one of the two methods for the identification and the quantification of the miRNAs expression values. Firstly, absolute sequence read counts of the mature miRNA and star sequences (miRbase v19; covering 21,264 precursor miRNAs and 25,141 mature miRNA) from each sample were used as a input data for the statistical software DESeq (v1.2.1), which is widely used for the estimation of the read count data from the NGS and for estimating the differential expression [19]. Secondly, all the sequence variants generated from miRNA precursors (pre-miRNA) along with the referenced mature sequence (isomiRs) will be taken into account and subsequently total number of read counts from all isomiRs will be assigned to the corresponding isomiR. Furthermore, the total counts for each miRNA precursor will be used as an input in DESeq module to evaluate the expression divergence across the user input samples.

2.3. Differential expression and pattern inference

To evaluate and hypothesize the observed differential miRNA expression between various experimental conditions (treated vs control), expression data from known miRNAs produced as reads count will be subsequently used as input in the Bioconductor DESeq package in R (http://www.r-project.org/re) [19]. DESeq uses a negative binomial distribution model to test and identify the differential expression across the user input samples.

2.4. Web-platform availability

The entire pipeline is written in Perl (v5.12.4) and R (2.13.1) can be accessed at http://bioinfo1.uni-plovdiv.bg/isomiRex/.

3. Results and discussion

isomiRs, miRNAs with varying heterogeneous ends play an important role in the regulation of the various development phases and show differential expression pattern across the tissues [20,21]. Recently, isomiRs have been found to be biologically active and are postulated to be the functional cooperative partners of canonical miRNAs [12]. There are few contributions to elucidate the potentiality of the isomiRs using the smallRNA NGS dataset. A stand-alone potential repertoire, which reports the identification of the isomiR variants from human 293T cells NGS dataset (http://hood.systemsbiology.net/cgi-bin/isomir/find.pl) requires a specific miRNA as a query field and their corresponding annotation in miRBase and the widely used Ensembl database [22]. A recently released and updated database platform (YM500) provides wider demonstration of the human and mice smallRNA-seq with integrated gene expression omnibus (GEO) for the visualization of the miRNA, isomiR and novel miRNA predictions [23]. Recent reports also present the variation in the miRNAs and their corresponding isomiRs [24]. In accordance with the recent reports, the development of a platform for the visualization of the miRNAs and their corresponding isomiRs, which can be deployed for several reference organisms and can be potentially used for the user-defined NGS data, is still in its naıve form except for the previous report [25]. In our presented research, we propose an integrated web platform for profiling the

![Fig. 1. Web interface of isomiRex web-platform. The platform require the user to upload the NGS small RNAs datasets for control and treated conditions and then select a method for miRNA identification.](image-url)
miRNAs/isomiRs and corresponding differential expression patterns using next-generation sequencing datasets of small RNAs (Fig. 1), which overcomes the limits of the stand-alone database in two ways. Firstly, isomiRex platform supports the de novo prediction of miRNAs and isomiRs in user defined NGS samples and differential analysis, which is previously limited in the stand-alone databases. Secondly, isomiRex is not organism centered platform and can be widely used for variety of reference organisms including plants, animals, model and non-model organisms.

3.1. Structure and web access of isomiRex

The web-based tool isomiRex is currently hosted on a 64-bit Linux server pre-configured with MySQL (http://www.mysql.com), Apache (http://www.apache.org) and PHP (http://php.net). A flow-chart describing the conceptual layout of the isomiRex is presented in Fig. 2. We have developed isomiRex with the aim to provide a user-friendly and simple interface in order to perform a comprehensive and comparative analysis on small RNA datasets from plant, animal and viruses using NGS sequencing. The web-based platform allows the users to analyze multiple data with a wide number of biological replicates. The current version of the isomiRex supports the comprehensive analysis of the multiple datasets at a given point of experimental timeline. A comparative analysis of the treated vs control conditions can be easily performed using the built-in function of the comparative library reads assessment.

Recently, a study illustrates the application of the microRNA profiling using massively parallel sequencing in human embryonic stem cell and demonstrated that the read count abundance provides a more robust support to the observed isomiR variation [21]. In support of the demonstrated observance of isomiR variation, we implemented the algorithm, which provides two methods for the identification and measuring the expression levels of miRNAs: (1) the first workflow is based on absolute sequence read counts of the mature miRNA and star sequences (miRbase v19); (2) the second workflow approach take into account all sequence variants generated from miRNA precursors (pre-miRNA) by mapping of all small RNAs in the library against the precursor sequence of all known miRNAs from the selected species using Bowtie read mapper [18]. The critical difference between the two workflows lies in the parsing of the read counts to the DESeq module for differential expression. In the first case only the mature read counts are taken into account and thus the differential expression is carried out only at the mature sequence expression levels. Furthermore the second method parsed the expression to DESeq module as a sum from mature miRNA as well as all additional isomiR
originating from the miRNA precursor. This potential algorithm allows the estimation of the differential expression at miRNA precursor level as a transcript. The applicability of this workflow will allow the users on his method of preference to identify all tags (iso-miR sequences) in the sample that are originated from the particular miRNA precursor. It also presents an efficient way to characterize the isomiR sequences that are highly expressed than the referenced mature miRNA.

In particular, the validation of the second approach in the algorithm will allow the user to clearly visualize the isomiR variants, which can produce from a particular hairpin by performing a comparative analysis with the reference set of mature miRNAs in miRBase (version 19, as of 27th March, 2013). It has been earlier demonstrated that the variability in either Dicer1 or Drosha cleavage positions within the pre-miRNA hairpin can be the most plausible reason for the variability in the isomiRs [21]. Since variation in isomiR sequence occurs at the 3′ and 5′-ends, they could potentially bind to a different repertoire of targets relative to their mature reference counterparts. Moreover, in cases where isomiRs are expressed at higher levels than their reference mature miRNA, the target network of that particular miRNA will be altered. As per earlier reports using the hESC cell lines, it has been demonstrated that the relative abundance of the isomiRs can vary across the tissue [21], which urge the need for a visualization of the relative abundance across the tissue. In isomiRex, the embedded visualization feature will help to visualize the higher expression of the isomiRs as compared to the indexed mature miRNA sequence (Fig. 3). This function can potentially help to identify the miRNA variants, which may later become functional during the course of the biogenesis pathway of mature miRNAs.

3.2 Web-visualization and inference of isomiR variation

The web-based isomiRex variant visualization platform generates the outputs in three major forms of data (Fig. 4): (1) identified miRNAs across data samples (with detailed view of their isomiR sequences in case of detection by precursor is selected); (2) differentially expressed miRNAs between control and treated datasets in experimental conditions; and (3) graphical visualization of the differentially expressed miRNAs. The change of the expression for a particular miRNA will be represented in fold change (treated vs. control samples) and Log2 fold change. Statistical likelihood estimation is suggested as a strong parameter for the selection of the isomiR to evaluate the abundance of the isomiRs and the total counts of the isomiRs abundantly expressed in the same class [26]. A statistical approximation test of P value <0.05 has been implemented as a threshold to judge the significance of the observed differences in the miRNA expression across the samples.

To conclude, we provide an effective platform for the identification and the visualization of miRNAs and their cognate isomiRs using NGS datasets spanning 193 species (including plants, animals and viruses using miRBase v.19). isomiRex has been developed with the goal to find, annotate and intersect microRNAs from plant and animal NGS datasets on one platform with a single data submission. The web platform provides an effective way for the identification of the miRNAs in the NGS dataset and simultaneously infers the differential expression of the identified miRNAs by taking into account all the sequence variants generated from miRNA precursors (pre-miRNA) along with indexed reference mature miRNA sequence. Moreover, isomiRex can be potentially exploited for the identification of isomiRs with higher copy number as compared to the selected reference mature miRNA in miRBase.

Acknowledgements

This work was supported by EU FP7 (REGPOT, BIOSUPPORT), the Bulgarian National Science Fund (DOO2-071-2008), Ministry of Education and Science (BG051PO001-3.3.05-0001) and the EU BMBS COST Action BM1006. Gaurav Sablok thanks Department of Biodiversity and Molecular Ecology, Research and Innovation
Centre, Fondazione Edmund Mach, Via E Mach 1, 38010 Michele all’Adige (TN), Italy for providing the computational facility. G.S. acknowledges Prof. Paul Bertone for suggestive revisions.

References


