ABSTRACT

Summary: Integrative Short Reads NAVigator (ISRNA) is an online toolkit for analyzing high-throughput small RNA sequencing data. Besides the high-speed genome mapping function, ISRNA provides statistics for genomic location, length distribution and nucleotide composition analysis of sequence reads. Number of reads mapped to known microRNAs and other classes of short non-coding RNAs, coverage of short reads on genes, expression abundance of sequence reads as well as some other analysis functions are also supported. The versatile search functions enable users to select sequence reads according to their sub-sequences, expression abundance, genomic location, relationship to genes, etc. A specialized genome browser is integrated to visualize the genomic distribution of short reads. ISRNA also supports management and comparison among multiple datasets.

Availability: ISRNA is implemented in Java/C++/Perl/MySQL and can be freely accessed at http://omicslab.genetics.ac.cn/ISRNA/.

Contact: xjwang@genetics.ac.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on July 25, 2013; revised on November 3, 2013; accepted on November 18, 2013

1 INTRODUCTION

The wide application of high-throughput sequencing technology has given rise to significant needs for powerful bioinformatics tools to analyze the resulting data and retrieve useful information. Although there have been some web-based and standalone tools for high-throughput small RNA analysis available, many of them mainly focus on microRNA (miRNA) identification (An et al., 2013; Hackenberg et al., 2011; Ronen et al., 2010; Sablok et al., 2013; Zhao et al., 2011; Zhu et al., 2010) or small RNA characterization (Chen et al., 2012; Fasold et al., 2011; Gupta et al., 2012; Huang et al., 2010; Pantano et al., 2011; Stocks et al., 2012; Zhang et al., 2012), few provide comprehensive feature analysis, mining as well as visualization and cross-sample comparison functions (see Supplementary Table S1 for detailed comparison).

Here we introduce Integrative Short Reads NAVigator (ISRNA), an online toolkit for searching, analyzing, visualizing and comparing short sequence reads. ISRNA integrates several useful tools in a user-friendly way. Besides mapping short sequence reads to a given genome and allowing a genome browser to visualize the position and distribution of sequence reads, ISRNA also calculates nucleotide composition, identifies known miRNAs and their homologous sequences, predicts the secondary structure for the surrounding genomic region of a short sequence and reveals genomic small RNA clusters and genes producing small RNAs. It also provides versatile searching functions and allows cross-dataset comparisons. All these functions are easy to use with user-friendly outputs, making ISRNA suitable for researchers with limited bioinformatics skills.

2 DESIGN AND KEY FEATURES

2.1 Data processing and genome mapping

ISRNA takes raw FASTQ files or text files with sequence and read count information, which can be generated by the Perl script provided by ISRNA as inputs. For details, please refer to the online user manual. The uploaded sequences will be mapped to the corresponding genome by Bowtie (Langmead et al., 2009) with user-defined parameters and stored in a MySQL database. Sequences with a number of mapped loci exceeding the threshold value will be discarded.

2.2 Sequence analysis

The analysis functions of ISRNA include:

2.2.1 Data overview This module analyzes some basic features of the uploaded dataset to provide users the general information, including sequencing depth, genome mapping results, sequence length distribution, etc.

2.2.2 Nucleotide preference analysis This module enables users to calculate the nucleotide composition at any defined position among sequence reads with genomic matches, in both total and non-redundant datasets.

2.2.3 Sequence annotation ISRNA uses the BLAST+ program (Camacho et al., 2009) and Rfam (Burge et al., 2013) database to annotate short sequence reads. The classification results of sequences will be provided in a pie chart and be downloaded via the ‘Export’ function.

2.2.4 Known miRNA identification ISRNA identifies known miRNAs as well as isomiRNAs by mapping reads to the
mature and precursor sequences of known miRNAs collected in the miRBase. The abundance of miRNAs or isoforms of miRNAs (isomiRNAs) is estimated by the reads per million values.

2.2.5 Secondary structure prediction ISRNA extracts the surrounding genomic regions of input sequences and predicts their secondary structures using the CentroidFold program (Sato et al., 2009), which allows users to identify new miRNAs or other classes of small RNAs with precursors of defined structural features.

2.3 Search functions ISRNA provides powerful search functions that enable users to analyze and curate their data with different needs (Fig. 1a), including search by sequences (allowing mismatches, Fig. 1b), search by miRNA, search by short read count, search by genomic location, search by related genes/gene families and search by read coverage on genes. A detailed description of the search functions can be found in the Supplementary File.

2.4 Cross-dataset comparison ISRNA supports comparison among multiple datasets within a project. Differentially expressed sequence reads among datasets can be identified by the implemented edgeR package (Robinson et al., 2010), with user-adjustable P-value and false discovery rate thresholds. Genome distribution of sequence reads in multiple datasets can be examined via the cross-dataset comparison function (Fig. 1c), which assists users in identifying differentially expressed genomic regions.

2.5 Genomic cluster analysis This function permits users to identify clustered short sequences on the genome, such as piwi-interacting RNAs or phased small RNAs. Comparing the genomic cluster distribution of sequence reads can also help to identify highly expressed sequence loci and sequence read production hotspots.

2.6 Sequence read browser In all search result display pages, a specialized genome browser is integrated to show the genomic mapping results of short sequence reads and genes. Users can easily identify the positional relationship among short sequence reads, as well as between short sequence reads and annotated genes. The ID, genomic position and annotation of genes, together with the information of sequence reads, will be displayed via mousing over the corresponding gene or sequence read. Short reads are displayed by cumulated short segments with color gradients representing the sequence abundance (Fig. 1d). The genome browser also enables users to examine the overall sequence distribution on each chromosome, and therefore to identify regions enriched or deprived of sequence reads.

3 CONCLUSIONS We introduce ISRNA, a piece of online software designed for storage, visualization and analysis of small RNA sequencing data produced by high-throughput sequencing technologies. The comprehensive data analysis, search and cross-dataset comparison functions, as well as the user-friendly output of ISRNA, make it a practical and easy-to-use tool for researchers with limited bioinformatics analysis skills.

Funding: This work was supported by the grants from the Ministry of Science and Technology of China (number 2011CBA01101), the 863 program of China (2011AA020108) and from the Chinese Academy of Sciences (numbers XDA01020105 and KSCX2-EW-R-01-03) to X.-J.W.

Conflict of Interest: none declared.

REFERENCES


