Vol. 29 no. 24 2013, pages 3225–3226
doi:10.1093/bioinformatics/btt545

Advance Access publication September 18, 2013

miREval 2.0: a web tool for simple microRNA prediction in genome sequences

Dadi Gao[1,2], Robert Middleton[1], John E. J. Rasko[2,3] and William Ritchie[1,2,*]

[1]Bioinformatics Laboratory, Centenary Institute, [2]Gene and Stem Cell Therapy Program, Centenary Institute, University of Sydney, Sydney, New South Wales, Australia and [3]Cell and Molecular Therapies, Royal Prince Alfred Hospital, Camperdown, New South Wales 2050, Australia

Associate Editor: Alfonso Valencia

ABSTRACT

Result: We have developed miREval 2.0, an online tool that can simultaneously search up to 100 sequences for novel microRNAs (miRNAs) in multiple organisms. miREval 2.0 uses multiple published *in silico* approaches to detect miRNAs in sequences of interest. This tool can be used to discover miRNAs from DNA sequences or to validate candidates from sequencing data.

Availability: http://mimirna.centenary.org.au/mireval/.

Contact: w.ritchie@centenary.org.au

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

Received on July 1, 2013; revised on August 26, 2013; accepted on September 13, 2013

1 INTRODUCTION

microRNAs (miRNAs) are ~22-nt non-coding RNAs (ncRNAs) that post-transcriptionally regulate the expression of target mRNAs. Although recent advances in sequencing technology allow a comprehensive analysis of expressed small RNAs, including miRNAs, distinguishing miRNAs from other RNA molecules within sequencing data remains difficult (Ritchie *et al.*, 2012). Moreover, sequencing data will not detect miRNAs specific to other tissue types and may not be available for partially sequenced genomes.

miREval 2.0 detects novel miRNAs from input DNA or RNA sequences using multiple bioinformatics approaches such as predicted secondary structure, phylogenetic conservation and shadowing and positional clustering. It also allows users to visualize RNA molecules in the sequence of interest that may be mistaken for miRNAs. miREval 2.0 also displays transcription factor (TF) binding sites that may regulate miRNAs in the input sequence and overlapping mRNA transcripts.

Improvements over our previous version of miREval (Ritchie *et al.*, 2008) include a bigger genome database (31 species instead of 10), multiple input sequences per enquiry, improved performance of the secondary structure prediction and parallelization to increase the speed of each analysis by over 10-fold. Results are now displayed in a circle graph to allow for a more compact visualization of long input sequences.

New features include phylogenetic footprinting to discover miRNAs that are specific to a given clade, the display of RNA molecules from rFam that may be mistaken for miRNAs and of TF binding sites that may regulate miRNAs in the input sequence. miREval 2.0 offers more features than most current prediction tools and displays high performance scores (Supplementary Table S1).

2 METHODS

miREval accepts up to 100 sequences of DNA or RNA in FASTA format. Comprehensive analysis is available for 31 species; partial analysis, including secondary structure prediction and alignment to known miRNAs, is still available for other species.

2.1 miRNA prediction

SVM prediction of miRNAs: Support vector machines (SVM) are an efficient machine learning technique used to predict miRNAs (Xue *et al.*, 2005). Our SVM is trained on 57 features including secondary structure, free energy and sequence composition (Supplementary Table S2). The positive dataset used to train the SVM is a collection of miRNA precursors from miRBase v20 across 15 vertebrates. The negative training set contains sequences from exonic regions of our 31 available genomes and ncRNAs randomly selected from rFam. Training our SVM on rFam ncRNAs allows miREval 2.0 to distinguish miRNA hairpins from hairpins in other ncRNAs. Tests run on 715 known miRNA precursors and 715 exonic fragments from mouse found the following performance rates: precision 0.92, accuracy 0.91, specificity 0.92 and sensitivity 0.89. For sequences longer than average miRNA precursors, miREval 2.0 uses a sliding window with a 10-bp step and size equal to the average length of miRNAs for the input species [69 (Wallaby)-148 bp (Tasmanian Devil)].

Phylogenetic conservation: miRNAs often show evolutionarily conserved patterns in their precursor regions. The conservation of a typical miRNA is higher at the two stems of the hairpin and drops to a low level toward the loop region (Berezikov *et al.*, 2005). miREval 2.0 uses conservation scores for alignments of 45 vertebrate genomes (10 genomes in the previous version) with the enquiry species from UCSC. These scores are calculated by PHAST (Siepel *et al.*, 2006) to give conservation at base resolution.

Phylogenetic shadowing (new): Phylogenetic shadowing uses multiple alignments of closely related species to discover functional elements (Blanchette and Tompa, 2002). Phylogenetic shadowing can pick up clade-specific miRNAs that would be overlooked using phylogenetic conservation. Phylogenetic shadowing scores are calculated in the same manner as conservation scores except that multiple (>5) related species are used instead of distant species.

miRNA clustering (improved): Sewer *et al.* and Marco found that miRNAs often clustered together (Marco *et al.*, 2013; Sewer *et al.*, 2005). This property can be used to enhance the accuracy of miRNA

*To whom correspondence should be addressed.

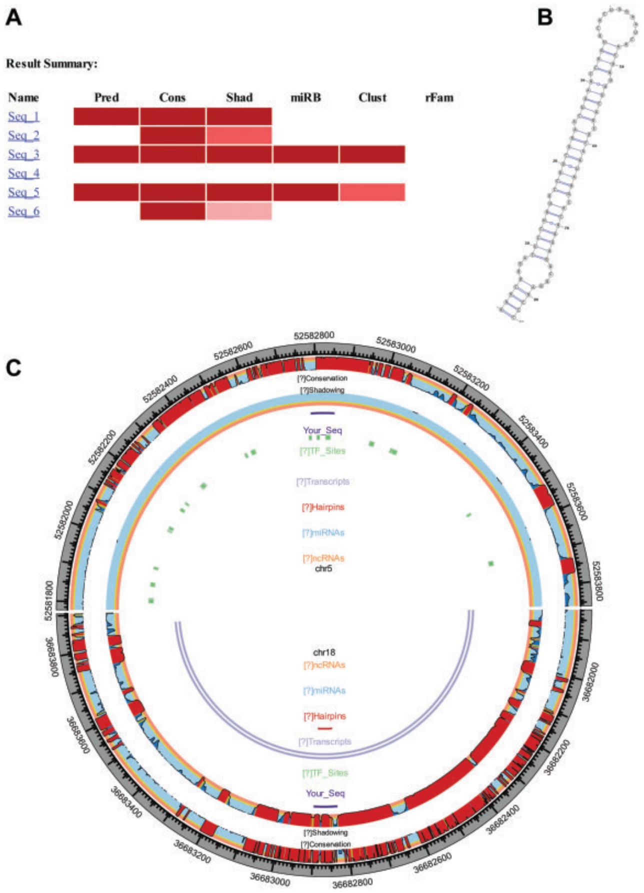

Fig. 1. Summary and final results page of miREval. (**A**) The summary page displays features of interest for each sequence as a heatmap. (**B**) In the final results page, secondary structure is visualized by Varna GUI and (**C**) genomic information is displayed as a circle graph

prediction. miREval searches 1000 bp around the enquiry sequence for miRNAs (miRBase v20) that cluster with the enquiry sequence.

2.2 Genomic information

If the user selects one of the 31 available species, miREval provides additional genomic information.

TF binding sites and transcripts (new): miRNAs are directly transcribed from DNA by RNA polymerase II or III (Lee *et al.*, 2004) or co-transcribed with mRNA transcripts (Marco *et al.*, 2013). Identification of nearby TF binding sites or known transcripts that overlap with the sequence of interest may aid miRNA prediction (Marco *et al.*, 2013) and provide insight into miRNA regulation. Transcript lists were downloaded from UCSC, and score matrices of TF binding sites were built from the JASPAR database (Portales-Casamar *et al.*, 2010). miREval 2.0 compares the coordinates of enquiry sequences with transcript list and uses the FIMO tool (Grant *et al.*, 2011) to find TF binding sites 1000 bp flanking the enquiry or flanking any mRNA transcript overlapping the enquiry.

Overlap with ncRNAs (new): Some ncRNAs have hairpin-like structures that may confound secondary structure prediction. miREval 2.0 displays transcripts from rFam that overlap the enquiry sequence and displays possible miRNA-like hairpins within them.

3 RESULTS

miREval 2.0 now supports multiple sequences per enquiry. Before displaying the final result page, it displays a heatmap of all of its features for each input sequence (Fig. 1A). This enables the user to visualize and select sequences with the most interesting features. Predicted hairpins are visualized using Varna v3.9 (http://varna-gui.software.informer.com/) (Fig. 1B). miREval 2.0 displays the final output in a circle graph using Circos 0.64 (Krzywinski *et al.*, 2009) (Fig. 1C). Users may hover their mouse over the graph to display more detailed information for each feature. miREval is fully compatible with Chrome, Firefox and IE after version 8.0.

ACKNOWLEDGEMENTS

The authors thank their colleagues at the Gene and Stem Cell Therapy Lab of Centenary Institute for useful comments on Web site display.

Funding: Australian National Health and Medical Research Council [Training Fellowship 571156], Cancer Institute NSW [career Development Fellowship 12/CDF/2-28], Cancer Council NSW [Project Grant 1006260], the Rebecca L. Cooper Foundation and the Cell and Gene Trust [Cure The Future].

Conflict of Interest: none declared.

REFERENCES

Berezikov,E. *et al.* (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.

Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.

Grant,C.E. *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Lee,Y. *et al.* (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051–4060.

Marco,A. *et al.* (2013) Clusters of microRNAs emerge by new hairpins in existing transcripts. *Nucleic Acids Res.*, **41**, 7745–7752.

Portales-Casamar,E. *et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.

Ritchie,W. *et al.* (2012) Defining and providing robust controls for microRNA prediction. *Bioinformatics*, **28**, 1058–1061.

Ritchie,W. *et al.* (2008) miREval: a web tool for simple microRNA prediction in genome sequences. *Bioinformatics*, **24**, 1394–1396.

Sewer,A. *et al.* (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267.

Siepel,A. *et al.* (2006) New methods for detecting lineage-specific selection. In: *Proceedings of Research in Computational Molecular Biology*. Vol. 3909, Springer Berlin-Heidelberg, Berlin, Germany, pp. 190–205.

Xue,C.H. *et al.* (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.