bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted April 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SeedVicious: analysis of microRNA target and near-target sites

Antonio Marco*

School of Biological Sciences, University of Essex, Colchester, United Kingdom

Running head: SeedVicious, versatile microRNA target prediction

* To whom correspondence should be addressed

School of Biological Sciences

University of Essex

Wivenhoe Park, Colchester CO4 3SQ

United Kingdom

Email: amarco.bio@gmail.com

Telephone: +44 (0) 120 687 3339

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted April 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

Summary: Here I describe seedVicious, a versatile microRNA target site prediction software that can be easily fitted into annotation pipelines and run over custom datasets. SeedVicious finds microRNA canonical sites plus other, less efficient, target sites. The program also detects near-target sites, which have one nucleotide different from a canonical site. Near-target sites are important to study population variation in microRNA regulation. Among other features, it can also predict targets on alignments and compute evolutionary gains/losses of target sites using maximum parsimony.

Availability and implementation: The program is written in Perl and runs on 64-bit Unix computers. Users can also try the program in a dedicated web-server by uploading custom data, or browsing pre-computed predictions. SeedVicious and its associated web-server and database (SeedBank) are distributed under the GPL/GNU license, and available at http://seedvicious.essex.ac.uk/

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

Contact: amarco.bio@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted April 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

Animal microRNAs target gene transcripts by partial sequence complementarity (Bartel, 2009). There are multiple microRNA target prediction tools that use different strategies. Most prediction programs look at the existence of seed sequences, six-nucleotide-long sequences in the transcripts that are complementary to nucleotides 2 to 7 in the microRNA (Bartel, 2009). Depending on additional nucleotide pairings these sites can be canonical or marginal. Many programs exploit additional features, mainly evolutionary conservation and RNA folding features (Agarwal *et al.*, 2015; Maragkakis *et al.*, 2009). Although these programs have an increased accuracy, they lose power as many real targets remain undetected. Often, microRNA target predictions are available only on selected datasets, and stand-alone programs are not available for use on custom data. Here I describe a new microRNA target prediction software that does not aim to replace but to complement the existing tool-kit, and allow the high-throughput analysis of custom microRNA/transcript data as well as the exploration of additional features not covered by other programs.

The first population genetics studies on microRNA targets sites already described selective pressures on seed sequences (Chen and Rajewsky, 2006; Saunders *et al.*, 2007). In a recent study I showed that, in *Drosophila* populations, there is selection against microRNA target sites (Marco, 2015). To study selection on non-target sites that may become targets, I defined 'one-mutant neighbors' as six-nucleotide-long sequences that have one nucleotide different to a seed sequence. Here I define a broader term: near-target sites, which have one nucleotide different to a putative microRNA target site, and are not targets themselves. Detection of near-target sites are important for population genetics and evolutionary studies, and will allow us to explore the selective pressures on gene transcripts. Additionally, reference genome sequences, in which microRNA target sites are usually predicted, do not capture the true diversity of a species and may miss *bona fide* target sites. Scanning for near-target sites may reveal targets that would be otherwise ignored. SeedVicious allows the exploration of near-target sites.

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted April 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

METHODS

Prediction of microRNA sites

SeedVicious initially scans custom sequences to detect canonical sites as described in Bartel (2009). Optionally, marginal sites can be reported. Other options include merging input microRNAs into microRNA seed families (microRNAs with the same potential targets), scanning only the longest 3'UTR of multiple transcripts for the same genes, and computing the free energy of the RNA:RNA duplex (Hofacker, 2009). The program can read aligned sequences.

Gains and losses of microRNA target sites

SeedVicious permits the inference of gains and losses of microRNA target sites by first predicting individual target sites at all transcripts in a given alignment, and then fitting a maximum parsimony (MP) model to a tree provided by the user. The MP reconstruction of ancestral states is computed with the 'dollop' program from the Phylip package (Felsenstein, 2005). We first used this method to study the evolution of post-transcriptional regulation in a gene family in *Drosophila* (Clifton *et al.*, 2017).

Additional analysis

A number of additional analyses can be performed with seedVicious. It can compare different 3'UTRs and detect either common microRNAs targeting pairs of transcripts or common target sites in an alignment. It can also report the minimum distance between pairs of target sites for the same microRNA. This distance can be used as a filtering criteria for potential targets, based on a recent work that suggests that neighboring sites cooperate during lateral diffusion of the Ago-miRNA complex during target recognition (Chandradoss *et al.*, 2015). A major feature of seedVicious is the detection of near-target sites which are detected for both canonical and marginal target sites.

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted April 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

IMPLEMENTATION

The main program can be run from the command line using Perl 5 or above, and the required modules and external binary files (compiled in a 64-bit Unix computer) are included in the distributed version. Input sequence files should be in FASTA format, and tree files in NEWICK format. Input files can be compressed in Gzip format. A fully referenced User Guide is available form the package or as Supplementary Information to this paper, and it includes different protocols. The program is available for download at http://seedvicious.essex.ac.uk/download. A web version is available, which allows users to run the program using our server. Precomputed targets for selected species can also be browsed from our SeedBank database. The web server and the database are accessed via http://seedvicious.essex.ac.uk

EXAMPLE

The transcript from the gene *lin-4* in *Caenorhabditis elegans* has seven targets sites for the microRNA *lin-4* (He and Hannon, 2004). However, other prediction programs detect only three (TargetScan) or none, probably due to the stringent criteria. Using seedVicious I scanned the lin-4 3' UTR (Jan *et al.*, 2010) for canonical target and canonical near-target sites. Three canonical sites were detected, the same that are reported in TargetScan. Additionally, five near-target sites were reported, four of them corresponding to the other four sites originally described. One of the near-target sites was not previously described and may be a good candidate to further explore. The detailed analysis as well as the input files are available with the package and fully described in the User Guide (Supplementary File 1). This example illustrates the potential of studying near-target sites, not only in evolutionary studies, but in the detection of potential functional targets. Further examples are described in the software manual.

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted April 5, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FUNDING INFORMATION

This work was supported by the Wellcome Trust [grant number 200585/Z/16/Z].

ACKNOWLEDGEMENTS

This software would not have been made publicly available without the encouragement and support of my colleagues, in particular Mohab Helmy and Sam Griffiths-Jones. I am also grateful to M. Helmy and Andrea Hatlen for discussion, and to Stuart Newman for his invaluable help setting up the web-server.

REFERENCES

Agarwal,V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Chandradoss,S.D. *et al.* (2015) A Dynamic Search Process Underlies MicroRNA Targeting. *Cell*, **162**, 96–107.

Chen,K. and Rajewsky,N. (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nat. Genet.*, **38**, 1452–1456.

Clifton,B.D. *et al.* (2017) Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in Drosophila. *Mol. Biol. Evol.*, **34**, 51–65.

Felsenstein,J. (2003) Inferring Phylogenies Sinauer Associates.

Felsenstein,J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.

Hofacker,I.L. (2009) RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al*, **Chapter 12**, Unit12.2.

Jan,C.H. *et al.* (2010) Formation, regulation and evolution of Caenorhabditis elegans 3[prime]UTRs. *Nature*, **469**, 97-101.

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493-496.

Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68-73.

Maragkakis,M. *et al.* (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.

Marco,A. (2015) Selection Against Maternal microRNA Target Sites in Maternal Transcripts. *G3 GenesGenomesGenetics*, **5**, 2199–2207.

Saunders,M.A. *et al.* (2007) Human polymorphism at microRNAs and microRNA target sites. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 3300–3305.