Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Sequence analysis

TarHunter, a tool for predicting conserved microRNA targets and target mimics in plants

Xuan Ma[1,2], Chunyan Liu[2], Lianfeng Gu[3,4], Beixin Mo[1], Xiaofeng Cao[2], and Xuemei Chen[1,4,5,*]

[1] Guangdong Provincial Key Laboratory for Plant Epigenetics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518060, China

[2] State Key Laboratory of Plant Genomics and National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, West Beichen Road, Chaoyang District, Beijing 100101, China

[3] Haixia Institute of Science and Technology (HIST), Fujian Agriculture and Forestry University, Fuzhou 350002, China

[4] Department of Botany and Plant Sciences, Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA

[5] Howard Hughes Medical Institute, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Abstract

Summary: In plants, the targets of deeply conserved microRNAs (miRNAs) were comprehensively studied. Evidence is emerging that targets of less conserved miRNAs, endogenous target mimics (eTM) and non-canonical targets play functional roles. Existing plant miRNA prediction tools lack a cross-species conservation filter and eTM prediction function. We developed a tool named TarHunter that features a strict cross-species conservation filter and capability of predicting eTMs. TarHunter has higher recall or precision rate as compared with other tools, and the conservation filter effectively increases prediction precision. TarHunter prediction combined with degradome analysis uncovered previously neglected miRNA targets including non-canonical target sites from various plant species, which are available at the TarHunter website (http://tarhunter.genetics.ac.cn/).

Availability: The code of TarHunter is available on Github (https://github.com/XMaBio).

Contact: xuemei.chen@ucr.edu

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

1 Introduction

Plant microRNAs (miRNAs) are endogenous ~21 nucleotide RNAs that pair with their target RNAs in a near-complementary manner (Rhoades, et al., 2002). Various bioinformatics tools such as Targetfinder (Fahlgren and Carrington, 2010), psRNATarget (Dai and Zhao, 2011), comTAR (Chorostecki and Palatnik, 2014), psRobot (Wu, et al., 2012), CleaveLand (Addo-Quaye, et al., 2009) and sPARTA (Kakrana, et al., 2014) have been developed to predict miRNA targets in plants. However, these tools have limited ability to predict cross-species conserved miRNA targets. psRobot's conservation analysis is only limited to eight species. comTAR only focuses on 22 conserved miRNAs and neglects less conserved miRNAs. Currently, there is no standalone pipeline available for conservation analysis of miRNA targets in plants.

Additionally, only a handful of miRNA target mimics (eTM) have been experimentally identified in plants (Franco-Zorrilla, et al., 2007; Li, et al., 2015; Wu, et al., 2013). It was recently shown that central mismatches rather than bulges confer effective target mimicry in plants (Liu, et al., 2014); thus, the currently available eTM prediction algorithm (Wu, et al., 2013; Karakülah et al., 2016) needs to be improved.

To facilitate the identification of conserved miRNA targets and eTMs in plants, we developed a local pipeline TarHunter, as well as a website that collects TarHunter prediction and degradome analysis results from various plant species.

2 Method

2.1 TarHunter workflow

The TarHunter pipeline includes miRNA target prediction in coding sequences (CDS) and noncoding regions, as well as eTM prediction (Fig. 1). CDS target prediction based on the ortho_mode (blue lines in Fig. 1) is performed in the following four steps. First, TarHunter implements CD-HIT (Li and Godzik, 2006) for orthologous miRNA clustering from desired species. Second, TarHunter adopts the Reciprocal Best Hits method to group orthologous genes; it implements UBLAST (Edgar, 2010) to find orthologous genes from diverse species, and clusters these genes using MCL (Enright, et al., 2002) followed by orthologous sequence alignments with MUSCLE (Edgar, 2004). Third, TarHunter implements FASTA (or RNAhybrid) for searching miRNA targets. Finally, a cross-species conservation filter requires: (i) orthologous miRNAs target orthologous genes; (ii) targeting occurs at the corresponding position in multiple sequence alignments of orthologous proteins.

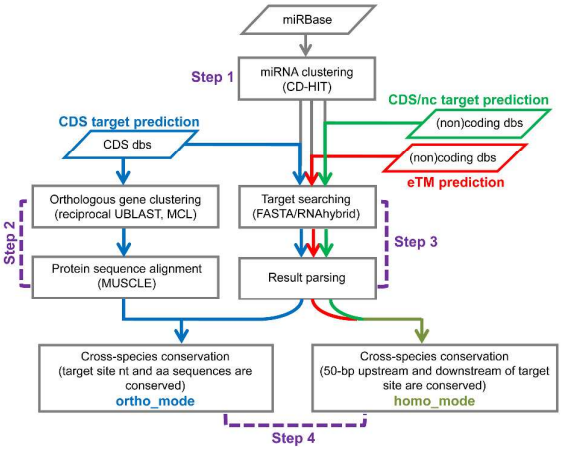

Fig 1. A schematic overview of TarHunter.

CDS target prediction can also be performed by the homo_mode (green lines in Fig. 1). Homo_mode differs from ortho_mode in that it requires a sequence alignment identify of >70% for the sequences surrounding the target site. It identifies homologous targets instead of orthologous targets. It is useful when genome annotations for some species are lacking and only EST databases are available.

Noncoding target prediction (green lines in Fig. 1) and eTM prediction (red lines in Fig. 1) consist of three steps: orthologous miRNA clustering, target searching and conservation filtering (homo_mode). The conservation analysis implements USEARCH, and requires a sequence alignment identity of >70% for the sequences surrounding the target site. The criterion for eTM prediction is that the central mismatches/bulge at guide miRNA positions 9-11 should be flanked by two highly complementary regions.

TarHunter outputs two files including all predicted miRNA targets in queried species and conserved targets. To further support our target prediction, we analyzed 43 degradome datasets from 13 plant species by CleaveLand4 and placed these results on the TarHunter website.

2.2 Sequencing data analysis

43 published degradome sequencing data (http://tarhunter.genetics.ac.cn/deg_data.htm) were analyzed by CleaveLand4 (Addo-Quaye, et al., 2009). High-confidence cleavage sites were identified using the following criteria: (i) category 0 or 1; (ii) penalty score <5. To identify non-canonical slicing sites, we used CleaveLand4 to generate a degradome density file and a GSTAr alignment file, which were parsed by an in-house perl script.

3 Results

To evaluate the performance of TarHunter, we first compared TarHunter without the conservation filter with four other plant miRNA target prediction algorithms, Targetfinder, psRNATarget, CleaveLand4, and sPARTA. We applied these tools on a self-compiled, experimentally validated CDS target dataset (Supplementary Table S1), and measured the recall and precision values at various penalty scores. The precision-recall plot clearly shows that TarHunter has higher precision/recall rate than the other tools tested (Supplementary Fig. S1A, Supplementary Table S2). By quantifying the Area Under the precision-recall Curve (AUC), we show that TarHunter performs best among these tools (Supplementary Fig. S1B). Note that Targetfinder performs similarly as TarHunter, but it lacks conservation-based or eTM prediction functions. Additionally, Targetfinder allows users to input only one miRNA each time, which is inconvenient when analyzing multiple miRNAs.

Compared with other tools, TarHunter's conservation filter is novel. TarHunter with either the 'ortho_mode' or 'homo_mode' conservation filter produced strikingly higher prediction precision at scores >3 (Supplementary Fig. S1C). Therefore, TarHunter's conservation filter is particularly useful when high confidence miRNA targets are being sought for. Note that the conservation filter can be used not only for deeply conserved miRNAs but also lineage-specific miRNAs.

TarHunter prediction combined with degradome analysis has identified novel miRNA targets from various species. These *in silico* predicted and sequencing data supported targets are available on the TarHunter website. Of particular interest, these targets include previously unreported non-canonical miRNA targets, e.g., Supplementary Fig. S2 illustrates four unreported *Arabidopsis* miRNA targets that have central mismatches/bulge.

Acknowledgements

The authors would like to thank Drs. Shaofang Li, Chenjiang You, Lei Gao and Lin Liu for comments on the manuscript.

Funding

This work was supported by grants from National Science Foundation of China [31210103901, 91440105, 31571332]; National Institute of Health [GM061146]; and Guangdong Innovation Research Team Program [2014ZT05S078].

Conflict of interests: none declared.

References

Addo-Quaye, C., Miller, W. and Axtell, M.J. (2009) CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets, *Bioinformatics*, **25**, 130-131.

Chorostecki, U. and Palatnik, J.F. (2014) comTAR: a web tool for the prediction and characterization of conserved microRNA targets in plants, *Bioinformatics*, **30**, 2066-2067.

Dai, X. and Zhao, P.X. (2011) psRNATarget: a plant small RNA target analysis server, *Nucleic Acids Res*, **39**, W155-159.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*, **32**, 1792-1797.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, **26**, 2460-2461.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res*, **30**, 1575-1584.

Fahlgren, N. and Carrington, J.C. (2010) miRNA Target Prediction in Plants, *Methods in Molecular Biology*, **592**, 51-57.

Franco-Zorrilla, J.M., *et al.* (2007) Target mimicry provides a new mechanism for regulation of microRNA activity, *Nat Genet*, **39**, 1033-1037.

Kakrana, A., *et al.* (2014) sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software, *Nucleic Acids Res*, **42**, e139.

Karakülah G., *et al.* (2016) PeTMbase: A Database of Plant Endogenous Target Mimics (eTMs), PLoS One, Dec 9;11(12):e0167698

Li, F., *et al.* (2015) Regulation of Nicotine Biosynthesis by an Endogenous Target Mimicry of MicroRNA in Tobacco, *Plant Physiology*, **169**, 1062-1071.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.

Liu, Q., Wang, F. and Axtell, M.J. (2014) Analysis of complementarity requirements for plant microRNA targeting using a Nicotiana benthamiana quantitative transient assay, *The Plant Cell*, **26**, 741-753.

Rhoades, M.W., *et al.* (2002) Prediction of plant microRNA targets, *Cell*, **110**, 513-520.

Wu, H.J., *et al.* (2012) PsRobot: a web-based plant small RNA meta-analysis toolbox, *Nucleic Acids Res*, **40**, W22-28.

Wu, H.J., *et al.* (2013) Widespread Long Noncoding RNAs as Endogenous Target Mimics for MicroRNAs in Plants, *Plant Physiology*, **161**, 1875-1884.