bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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 August 4, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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

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. Here I show that near-target sites can also be functional sites. Among other features, seedVicious can also compute evolutionary gains/losses of target sites using maximum parsimony. SeedVicious does not aim to outperform but to complement existing microRNA prediction tools. For instance, the precision of TargetScan is doubled (from 11% to ~22%) when we filter predictions by the distance between target sites using our program. The software is written in Perl and runs on 64-bit Unix computers (Linux and MacOS X). 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.

Contact: amarco.bio@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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 (1). 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 (1). Depending on additional nucleotide pairings these sites can be canonical or marginal. Many programs exploit additional features, mainly evolutionary conservation and RNA folding features (2, 3). 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 (4, 5). In a recent study I showed that, in *Drosophila* populations, there is selection against microRNA target sites (6). 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 (Figure 1), 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 August 4, 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 AND IMPLEMENTATION

SeedVicious initially scans custom sequences to detect canonical microRNA target sites as described in (1). Optionally, marginal sites can be reported (Figure 1). This strategy produces a large number of false positives, and other prediction tools use evolutionary conservation to narrow down the number of potential target sites (2, 7). However, evolutionary conservation is not necessarily associated with functionality of the target site (8). SeedVicious allows an alternative (and complementary) way of finding putative target sites. The complete list of canonical (and marginal) sites can be ranked according to selected features. First, seedVicious computes the free energy of the microRNA:transcript duplex using the RNAeval program in the Vienna Package (9). The lower the energy, the more stable is the RNA:RNA pair, which can be interpreted as a more efficient target site. Second, seedVicious can also report the minimum distance between pairs of target sites for the same microRNA.

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, following Dollo's criteria (10). The MP reconstruction of ancestral states is computed with the 'dollop' program from the Phylip package (11). We first used this method to study the evolution of post-transcriptional regulation in a gene family in Drosophila (12). 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. A major feature of seedVicious is the detection of near-target sites which are detected for both canonical and marginal target sites.

The main program can be run from the command line using Perl 5 or above, and the required

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

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.

For the analysis presented below, all microRNAs were retrieved from miRBase release 21 (13), and all transcript sequences from ENSEMBL Genes 89 via Biomart (14). TargetScan targets were retrieved from targetscan.org, release 7.1 (2). Predictions were compared to the experimentally validated targets in miRTarBase 6.0 (15), for both weak and strong interactions. Precision was calculated as the ratio TP/(TP+FP), where TP (True Positives) was the number of predicted targets that were experimentally validated, and FP (False Positives) was the number of predicted targets not validated in miRTarBase.

RESULTS AND DISCUSSION

I describe in this section a few examples that illustrate different ways of analyzing target sites with seedVicious. First, I computed with seedVicious the minimum distance between canonical target sites for all human microRNAs in 3'UTRs (see Methods and Implementation). 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 (16). Overall, the minimum distance between target sites for microRNA/transcript pairs with experimentally validated targets is smaller than for non-validated targets (Figure 2A). This indicates that the closest two target sites are in a transcript, the more likely that the microRNA will regulate the targeted transcript. As a matter of fact, two contiguous target sites (6 nt distance) significantly increases the chances of work as a regulatory sequence (Figure 2B).

The precision of a target prediction is measured as the fraction of true positives with respect to the

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

total number of predictions. By using a compilation of experimentally validated microRNA/transcript interactions (see Methods and Implementation) I estimated the precision of target prediction based on the minimum distance between canonical sites. If we predict as microRNA targets transcripts with at least 2 canonical sites, the precision of the prediction is 0.041. This value significantly increases (to 0.087) if we only consider pairs of canonical sites next to each other (distance of 6 nucleotides). Precision values in microRNA target prediction are generally low, in particular when one uses a small dataset of validated targets (as in this case), but it is useful to compare the predictive power of different strategies. Using this same dataset, the precision of TargetScan (using conserved targets only) is 0.11. Strikingly, if we combine TargetScan and pairs of targets at 6 nucleotides in distance, the precision goes up to 0.21. In conclusion, combining some of the features of seedVicious to other programs can be use to increase the precision in microRNA target prediction.

In previous studies, the occurrence of two or more seed sequences in transcripts has been associated to a higher chance of predicting a *bona fide* targeting interaction (17). Here I compared the proportion of predicted targets that are validated by experiments (precision) and the number of canonical target sites. The graph in Figure 2C shows that, indeed, as the number of target sites increase, the proportion of validated targets also increases, although there is a global maximum around 50. It will be interesting to explore whether distinct microRNAs have a different optimum number of target sites for an efficient repression of the target transcript.

The analysis of targets and near-targets is of particular interest to analyse populations with segregating alleles in target sites (4, 6). For instance, the longest transcript from tumour-suppressor gene *p53* have no target sites (neither canonical nor marginal) for miR-7-5p, a well known cancer-related microRNA (18). However, it contains three near-target sites. That is, there are three potential nucleotide substitutions that will create a novel target site for this particular microRNA. Indeed,

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

miR-7 promotes tumour growth via a mechanisms that includes p53 (19).Somatic mutations in the 3'UTR of p53 leading to new miR-7-5p target sites in cancer patients is then plausible. Current work in our lab reveals that selective pressures against this type of substitutions in cancer-associated genes is prevalent (Hatlen and Marco, unpublished results).

The transcript from the gene *lin-14* in *Caenorhabditis elegans* has seven targets sites for the microRNA *lin-4* (20). However, other prediction programs detect only three (TargetScan) or none, probably due to the stringent criteria. Using seedVicious I scanned the lin-14 3' UTR (21) 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.

The canonical target sites for lin-4-5p in lin-14 are highly conserved. Actually, in five species of worm studies all sites remain the same. However, let-7-5p, which also targets lin-14, has a more dynamic evolution. Using seedVicious I predicted the canonical target sites in five worm species and to reconstruct the gains and losses using maximum parsimony. Results are shown in Figure 2C, where we observe that there have been 4 target sites gained among the studied species whilst *Caenorhabditis briggsae* lost its three ancestral target sites.

In conclusion, seedVicious is a valuable addition to the toolkit of microRNA biologists. It works on custom datasets and can be easily fitted into annotation pipelines. It provides a range of analysis

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

that can be combined with other programs to generate robust microRNA target predictions. The analysis of near-target sites is also useful to study population dynamics at target sites, and the inbuilt maximum parsimony functionality is valuable to study the evolution of microRNA-mediated regulation at a large scale.

AVAILABILITY

The program is available for download at http://seedvicious.essex.ac.uk/download.html. 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

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, to Roman Cheplyaka for feedback on the program, and to Stuart Newman for his invaluable help setting up the web-server.

FUNDING

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

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

REFERENCES

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

2. Agarwal,V., Bell,G.W., Nam,J.-W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**.

3. Maragkakis,M., Alexiou,P., Papadopoulos,G., Reczko,M., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K., Simossis,V., *et al.* (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.

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

5. Saunders,M.A., Liang,H. and Li,W.-H. (2007) Human polymorphism at microRNAs and microRNA target sites. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 3300–3305.

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

7. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Vlachos,I.S., Vergoulis,T., Reczko,M., Filippidis,C., Dalamagas,T. and Hatzigeorgiou,A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, **41**, W169–W173.

8. Pinzón,N., Li,B., Martinez,L., Sergeeva,A., Presumey,J., Apparailly,F. and Seitz,H. (2016) The number of biologically relevant microRNA targets has been largely overestimated. *Genome Res.*, 10.1101/gr.205146.116.

9. 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.

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

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

12. Clifton,B.D., Librado,P., Yeh,S.-D., Solares,E.S., Real,D.A., Jayasekera,S.U., Zhang,W., Shi,M., Park,R.V., Magie,R.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.

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

14. Kinsella,R.J., Kähäri,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A., *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database J. Biol. Databases Curation*, **2011**, bar030.

15. Chou,C.-H., Chang,N.-W., Shrestha,S., Hsu,S.-D., Lin,Y.-L., Lee,W.-H., Yang,C.-D., Hong,H.-C., Wei,T.-Y., Tu,S.-J., *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

16. Chandradoss,S.D., Schirle,N.T., Szczepaniak,M., MacRae,I.J. and Joo,C. (2015) A Dynamic Search Process Underlies MicroRNA Targeting. *Cell*, **162**, 96–107.

17. Alexiou,P., Maragkakis,M., Papadopoulos,G.L., Reczko,M. and Hatzigeorgiou,A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.

18. Zhao,J., Tao,Y., Zhou,Y., Qin,N., Chen,C., Tian,D. and Xu,L. (2015) MicroRNA-7: a promising new target in cancer therapy. *Cancer Cell Int.*, **15**, 103.

19. Hong,C.-F., Lin,S.-Y., Chou,Y.-T. and Wu,C.-W. (2016) MicroRNA-7 Compromises p53 Protein-dependent Apoptosis by Controlling the Expression of the Chromatin Remodeling Factor SMARCD1. *J. Biol. Chem.*, **291**, 1877–1889.

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

21. Jan,C.H., Friedman,R.C., Ruby,J.G. and Bartel,D.P. (2010) Formation, regulation and evolution of Caenorhabditis elegans 3[prime]UTRs. *Nature*, **advance online publication**.

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

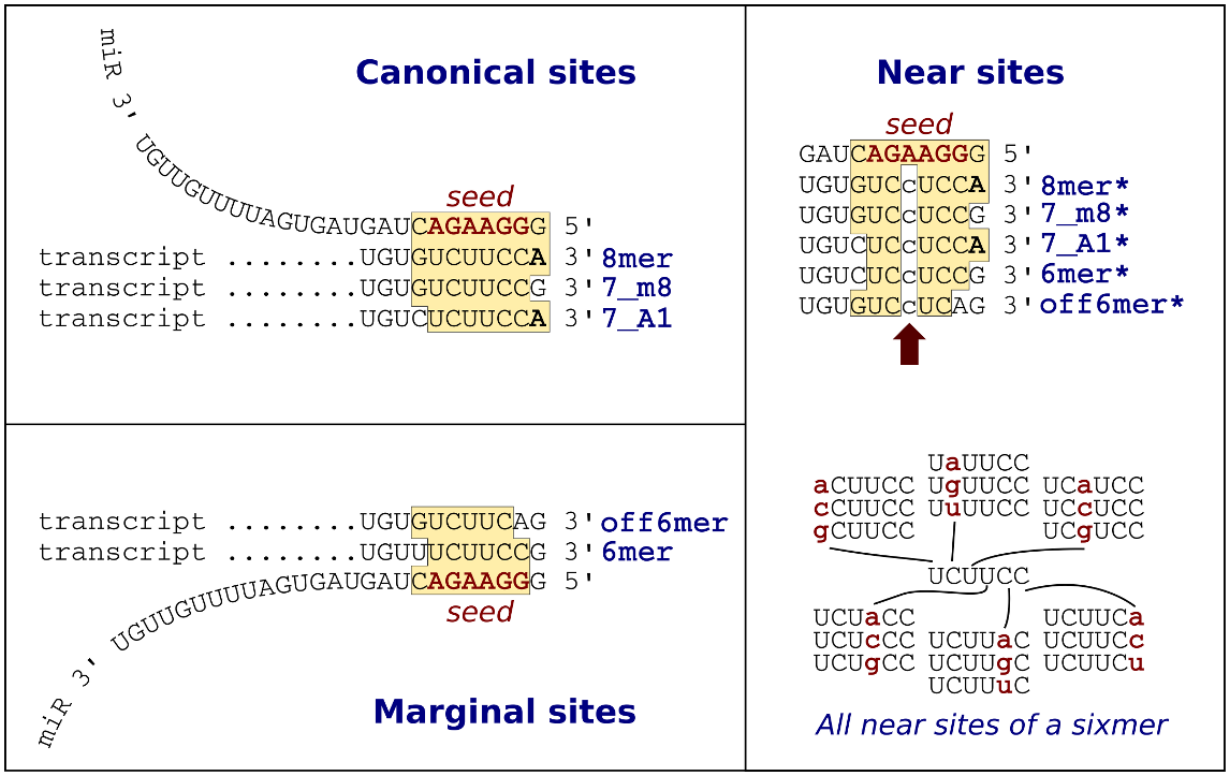
FIGURE LEGENDS

Figure 1. MicroRNA target and near-target sites. Canonical and marginal sites are described in Bartel (2009). Near-target sites can be defined for all type of sites. The figure also shows that each seed (sixmer) 18 possible near-target sites as defined by Marco (2015).

Figure 2. Multiple uses of seedVicious. A) Cummulative distribution of the number of transcripts with at least tow target sites with respect to the minimum distance between pairs of targets. The graph compares the distribution of all transcript (black), transcript/microRNA pairs with experimentally validated interactions (blue) and those with strongly validated interactions (red) according to miRTarBase. B) Precision of microRNA target prediction for different cut-offs of minimum distance between pairs of targets. The peak corresponds to a distance of 6, that is, two contiguous target sites. C) Precision of microRNA target prediction for multiple targets sites for the same microRNA. D) Evolutionary turnover of microRNA canonical target sites for let-7 in lin-14 in roundworm species using maximum parsimony.

bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/124529; this version posted August 4, 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.

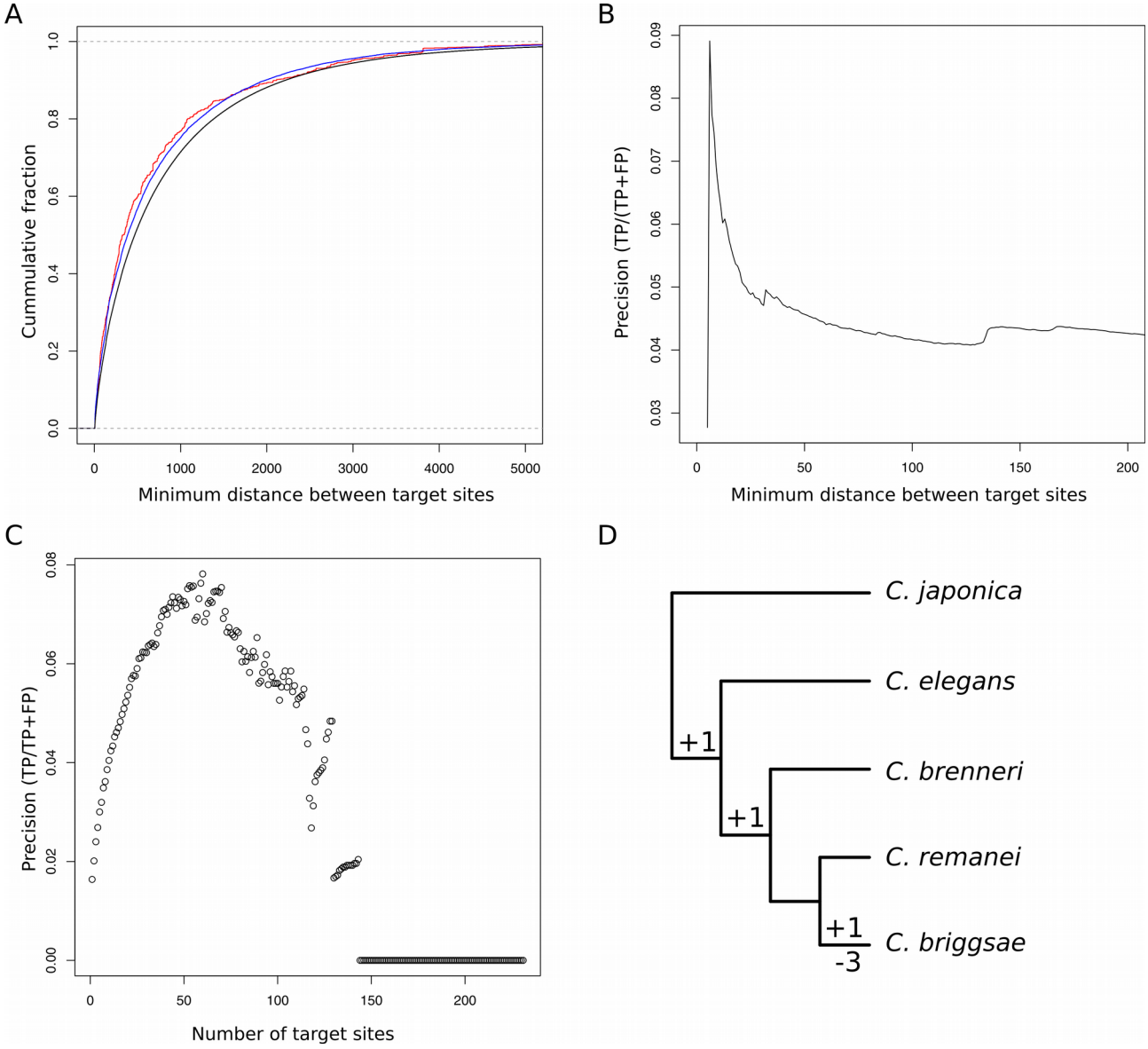



Figure 2