D74–D77 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 15 November 2013*
doi:10.1093/nar/gkt1156

miRNEST 2.0: a database of plant and animal microRNAs

Michał W. Szcześniak* and Izabela Makałowska*

Laboratory of Bioinformatics, Adam Mickiewicz University in Poznań, Poznań, Poland

Received September 16, 2013; Revised October 21, 2013; Accepted October 26, 2013

ABSTRACT

Ever growing interest in microRNAs has immensely populated the number of resources and research papers devoted to the field and, as a result, it becomes more and more demanding to find miRNA data of interest. To mitigate this problem, we created miRNEST database (http://mirnest.amu.edu.pl), an integrative microRNAs resource. In its updated version, named miRNEST 2.0, the database is complemented with our extensive miRNA predictions from deep sequencing libraries, data from plant degradome analyses, results of pre-miRNA classification with HuntMi and miRNA splice sites information. We also added download and upload options and improved the user interface to make it easier to browse through miRNA records.

INTRODUCTION

microRNAs (miRNAs) are a class of negative regulators of gene expression, widely identified in animals and plants. In plants, miRNAs participate in different aspects of growth and developmental processes, including lateral root formation or transition from juvenile to adult vegetative phase (1). They are also key players in response to stress conditions, like drought, low temperatures or nitrogen deficiency (2). Animal miRNAs are believed to regulate more than half of protein-coding genes and, like in plants, are implicated in a number of biological processes (3). Notably, multiple miRNAs have been associated with diseases, like cancers or rheumatoid arthritis (4).

The fact that miRNAs are key regulators of molecular processes in a cell and that they could find multiple applications in biotechnology, molecular biology or medicine, motivated extensive development of methods for their identification and study. The growing number of miRNA studies allowed better understanding of their biology and, consequently, led to accumulation of miRNA databases. However, many of them are limited to species of high interest, selected taxa or miRNAs involved in some specific processes. For instance, miRNeye (5) collects data about miRNA expression in mouse eye, whereas GrapeMiRNA stores sequences from *V. vinifera* (6). miRBase (7), on the other hand, although accommodates data from a wide range of species, contains only already published results. As a result, a single universal repository is required so that there was no necessity to browse through a number of dispersed data sets to collect information related to specific species or miRNA type.

Previously, we took up this challenge and we developed miRNEST, a comprehensive online resource for plant, animal and virus miRNAs. Using a comparative approach, we identified 10 004 miRNA candidates in 221 animal and 199 plant species. As our goal was not only to identify new miRNAs but also to develop a resource that would integrate miRNA data scattered across literature and databases, we also incorporated miRNA sequences from three other databases and two publications. Additionally, based on availability, we used data from 12 resources providing further annotation for miRNAs from selected species. Here we present miRNEST 2.0, an updated version of the database. In addition to 39 122 miRNAs from miRNEST 1.0 (10 004 from our EST analysis and 29 118 from other resources), we predicted 18 043 pre-miRNAs using small RNA deep sequencing data from 21 species. For miRNAs in 10 species, we provided targets inferred from degradome libraries. We also added miRNA splice sites information, HuntMi (8) predictions and some database functionalities, including download option. Taken together, miRNEST 2.0 is a large and comprehensive resource of miRNA data that bears distinct improvements over its previous version.

MATERIALS AND METHODS

miRNA prediction from sRNA deep sequencing data

For miRNA predictions we downloaded, from GEO database (9), 171 small RNA deep sequencing libraries from 8 plant and 13 animal species (Figure 1,

*To whom correspondence should be addressed. Tel: +48 61 829 5836; Fax: +48 61 829 5949; Email: miszcz@amu.edu.pl
Correspondence may also be addressed to Izabela Makałowska. Tel: +48 61 829 5835; Fax: +48 61 829 5949; Email: izabel@amu.edu.pl

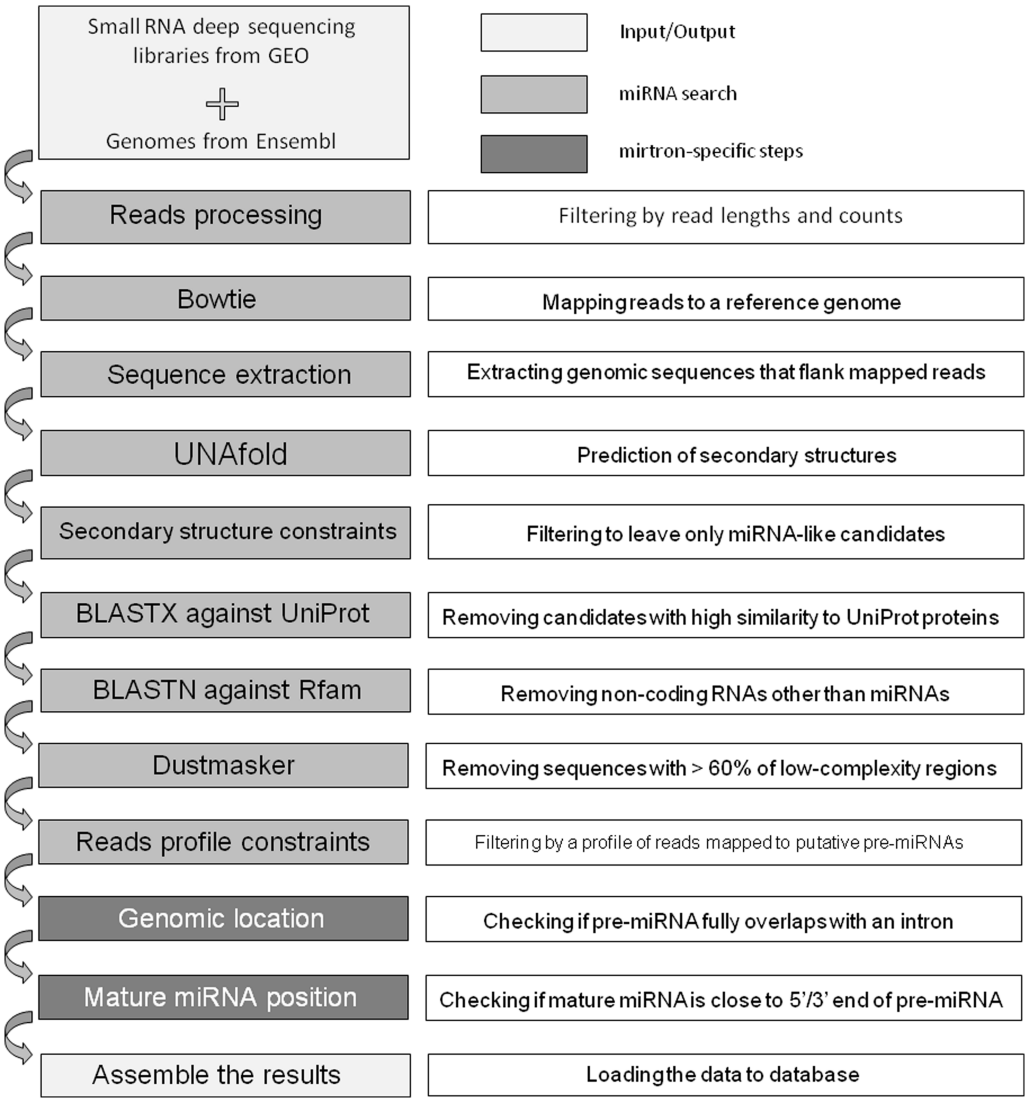

Figure 1. The pipeline used for large-scale miRNA discovery from sRNA deep-sequencing data.

Supplementary Table S1). Reads 19–26 bases long were kept and we mapped them to corresponding plant or animal genomes using Bowtie (10). In the mapping step, no mismatches were allowed and reads mapping to >20 distinct locations were discarded. Mapped reads that were 19–22-nt long and with count ≥ 5 were considered 'potential mature miRNAs'. We retrieved their sequences from genomes along with flanking genomic sequences of 150 bases in animals and 250 bases in plants, and then we predicted secondary structures using hybrid-ss-min from UNAFold package (11). We kept only sequences with miRNA-like secondary structures: a stem loop-structure with 'potential mature miRNA' located in a single hairpin arm; no more than six mismatches and three bulges (animals) or five mismatches and two bulges (plants) between mature miRNA and the opposite hairpin arm. If a stem-loop structure was surrounded by additional nucleotides, the flanking regions were cutoff. Subsequently, we checked similarity to non-coding RNAs from RFAM (12) and proteins from UniProt (UniProtKB/Swiss-Prot protein data set) (13) using BLAST (14). Sequences showing similarity to RFAM non-miRNAs with $E < 1e\text{-}10$ or UniProt proteins with $E < 1e\text{-}20$ were discarded. After that we searched for low-complexity regions using Dustmasker (14); sequences bearing >60% of low-complexity regions were removed. Finally, we made sure that there is a miRNA-like profile of reads mapped to the hairpin. To achieve this we kept only the hairpins where (i) 'potential mature miRNA' corresponded to the most abundant read in at least one library, (ii) abundance of 'potential mature miRNA' constituted minimal 20% of total read counts in at least one library and (iii) the total count of reads starting at 5′ position of 'potential mature miRNA' was the maximal one in at least one library.

Newly identified miRNA candidates were checked against intronic sequences in corresponding species and sequences that fully overlapped with introns, with 'potential mature miRNA' located no more than four bases away from 5′ or 3′ intron end became mirtron candidates.

We supplemented these candidates with already published predictions in mouse and human (15).

Degradome analysis

We downloaded 18 degradome libraries from GEO (9) that corresponded to 10 plant species: *Arabidopsis thaliana, Glycine max, Hordeum vulgare, Malus domestica, Medicago truncatula, Physcomitrella patens, Prunus persica, Solanum lycopersicum, Triticum aestivum* and *V. vinifera* (Supplementary Table S2). Transcript sequences (cDNAs) were downloaded from Ensembl Plants (16), and mature miRNA sequences were retrieved from miRNEST (17). Using PAREsnip (18), we searched for miRNA targets evidenced by degradome reads. We adjusted the program settings to look only for category 0, 1 and 2 targets, i.e. only high confidence candidates. For obtained candidates, we prepared degradome reads alignment files and corresponding plots for graphical representation of read mapping.

HuntMi predictions

HuntMi (8) is a machine learning tool for discrimination between true and false pre-miRNAs in plants, animals and viruses based on properties of pre-miRNA sequence and its secondary structure. We used this tool with default settings to better annotate pre-miRNAs stored in miRNEST. For animal, plant and virus sequences, different taxon-specific classifiers were used.

miRNA splice sites prediction

To infer miRNA splicing events from EST sequences, we applied a strategy previously used in ERISdb (19). In the first step, pre-miRNAs were searched against dbEST (20) using Megablast (14). It was required that the identity was 97% or higher and that the EST sequence contained at least 90% of known pre-miRNA sequence. The selected ESTs were subsequently mapped to the corresponding genome using Splign (21) with default settings. The alignments were finally checked manually to remove cases where ESTs came from the antisense strand and to improve the alignment in every case when splice site was broken because of imperfection of EST alignment software. Additionally, gene structures for 45 plant miRNAs were downloaded from ERISdb (19). We also obtained gene structures from RACE experiments in *Populus trichocarpa* (22), and RNA-Seq-evidenced splice sites in *V. vinifera* (23).

RESULTS

In current version, miRNEST has been extensively enlarged by results of small RNA deep sequencing analyses. First of all, we predicted 18 043 pre-miRNAs in 21 plant and animal species, and because miRNAs were often found independently in different sRNA libraries, this corresponds to as many as 36 468 new records in the database. In the search pipeline, we applied a number of strict criteria from the literature (17,24,25). In all, 38.1% of new sequences overlap with miRNAs already stored in miRNEST 1.0, thus providing experimental support for them (Supplementary Table S3). Moreover, as the database encompasses multiple libraries per species, it is possible to investigate isomiRs and changes in small RNA counts in different tissues and conditions. Although a similar functionality is available at miRBase (7), the analyzed species and selected deep sequencing libraries overlap only partly. Furthermore, for all miRNAs stored in miRNEST, including new predictions, we run classification analysis using HuntMi, which helped in much better annotation. Altogether, 91.16% of miRNEST sequences were considered true miRNAs, including miRNEST EST predictions (77.85%), miRNEST deep sequencing predictions (71.9%) and miRNAs from external databases (96.91%). Relatively high fraction of sequences recognized as true miRNAs in case of external databases [miRBase (7), PMRD (26), microPC (27)] might be due to the fact that this data set largely overlaps with miRNAs used to train HuntMi. Another aspect of deep sequencing analysis was identification of degradome-evidenced miRNA targets in 10 plant species. As we wanted to achieve highest quality results, only category 0, 1 and 2 candidates, as returned by PAREsnip, were considered. This allowed us to identify 2041 miRNA-target associations (Supplementary Table S4).

Splicing in miRNA genes is an underestimated aspect of miRNA biology. So far, there is only one repository that stores miRNA splice sites information (19). We incorporated that data into miRNEST 2.0 and additionally performed splice site search in several species, which allowed us to find 17 miRNAs with introns in 5 plant species. We also complemented that data with miRNA gene structures from the literature (*P. trichocarpa*, *V. vinifera*).

CONCLUSIONS

The current version of the miRNEST database contains twice as many miRNA records as the version 1.0. Thanks to the small RNA deep sequencing data analysis, almost 40% of previously predicted miRNAs is now validated by the experimental data. Moreover, target predictions for miRNAs from 10 species are supported by degradome data. miRNEST 2.0 has also an updated user interface and works faster than its predecessor. We added both bulk data download and download available from 'Browse' page (for user-selected miRNAs). As we want miRNEST to grow and be a truly comprehensive miRNA resource, we also enabled upload option for miRNA-associated data.

AVAILABILITY AND REQUIREMENTS

miRNEST is freely available at http://mirnest.amu.edu.pl. Its previous version, miRNEST 1.0, can still be accessed at http://lemur.amu.edu.pl/share/php/mirnest_1.0. The database was constructed using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), PHP 5.2.11 (http://www.php.net/) and MySQL 4.0.31 (http://www.mysql.com/). pre-miRNA secondary structures are drawn using Java lightweight applet VARNA (28), which requires installation of Java plugin.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: National Science Centre grant [2011/01/N/NZ2/01653 to M.W.S].

Conflict of interest statement. None declared.

REFERENCES

1. Mallory,A.C. and Vaucheret,H. (2006) Functions of microRNAs and related small RNAs in plants. *Nat. Genet.*, **38**, S31–S36.
2. Sunkar,R., Chinnusamy,V., Zhu,J. and Zhu,J.K. (2007) Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. *Trends Plant Sci.*, **12**, 301–309.
3. Siomi,H. and Siomi,M.C. (2010) Posttranscriptional regulation of microRNA biogenesis in animals. *Mol. Cell*, **38**, 323–332.
4. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
5. Karali,M., Peluso,I., Gennarino,V.A., Bilio,M., Verde,R., Lago,G., Dollé,P. and Banfi,S. (2010) miRNeye: a microRNA expression atlas of the mouse eye. *BMC Genomics*, **11**, 715.
6. Lazzari,B., Caprera,A., Cestaro,A., Merelli,I., Del Corvo,M., Fontana,P., Milanesi,L., Velasco,R. and Stella,A. (2009) Ontology-oriented retrieval of putative microRNAs in Vitis vinifera via GrapeMiRNA: a web database of *de novo* predicted grape microRNAs. *BMC Plant Biol.*, **9**, 82.
7. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
8. Gudyś,A., Szcześniak,M.W., Sikora,M. and Makałowska,I. (2013) HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics*, **14**, 83.
9. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
10. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
11. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.
12. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
13. UniProt Consortium. (2013) Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
14. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
15. Ladewig,E., Okamura,K., Flynt,A.S., Westholm,J.O. and Lai,E.C. (2012) Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Res.*, **22**, 1634–1645.
16. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
17. Szcześniak,M.W., Deorowicz,S., Gapski,J., Kaczyński,Ł. and Makalowska,I. (2012) miRNEST database: an integrative approach in microRNA search and annotation. *Nucleic Acids Res.*, **40**, D198–D204.
18. Folkes,L., Moxon,S., Woolfenden,H.C., Stocks,M.B., Szittya,G., Dalmay,T. and Moulton,V. (2012) PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic Acids Res.*, **40**, e103.
19. Szcześniak,M.W., Kabza,M., Pokrzywa,R., Gudyś,A. and Makałowska,I. (2013) ERISdb: a database of plant splice sites and splicing signals. *Plant Cell Physiol.*, **54**, e10.
20. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
21. Kapustin,Y., Souvorov,A., Tatusova,T. and Lipman,D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct*, **3**, 20.
22. Kruszka,K., Pacak,A., Swida-Barteczka,A., Stefaniak,A.K., Kaja,E., Sierocka,I., Karlowski,W., Jarmolowski,A. and Szweykowska-Kulinska,Z. (2013) Developmentally regulated expression and complex processing of barley pri-microRNAs. *BMC Genomics*, **14**, 34.
23. Mica,E., Piccolo,V., Delledonne,M., Ferrarini,A., Pezzotti,M., Casati,C., Del Fabbro,C., Valle,G., Policriti,A., Morgante,M. *et al.* (2009) High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera. *BMC Genomics*, **10**, 58.
24. Meyers,B.C., Axtell,M.J., Bartel,B., Bartel,D.P., Baulcombe,D., Bowman,J.L., Cao,X., Carrington,J.C., Chen,X., Green,P.J. *et al.* (2008) Criteria for annotation of plant MicroRNAs. *Plant Cell*, **20**, 3186–3190.
25. Chung,W.J., Agius,P., Westholm,J.O., Chen,M., Okamura,K., Robine,N., Leslie,C.S. and Lai,E.C. (2011) Computational and experimental identification of mirtrons in *Drosophila melanogaster* and *Caenorhabditis elegans*. *Genome Res.*, **21**, 286–300.
26. Zhang,Z., Yu,J., Li,D., Zhang,Z., Liu,F., Zhou,X., Wang,T., Ling,Y. and Su,Z. (2010) PMRD: plant microRNA database. *Nucleic Acids Res.*, **38**, D806–D813.
27. Mhuantong,W. and Wichadakul,D. (2009) MicroPC (microPC): a comprehensive resource for predicting and comparing plant microRNAs. *BMC Genomics*, **10**, 366.
28. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.