miRAuto: An Automated User-Friendly MicroRNA Prediction Tool Utilizing Plant Small RNA Sequencing Data

Jeongsoo Lee^{1,4}, Dong-in Kim^{1,4}, June Hyun Park^{2,4}, Ik-Young Choi^{1,*}, and Chanseok Shin^{2,3,*}

MicroRNAs (miRNAs) are a class of small RNAs that post-transcriptionally regulate gene expression in animals and plants. The recent rapid advancement in miRNA biology, including high-throughput sequencing of small RNA libraries, inspired the development of a bioinformatics software, miRAuto, which predicts putative miRNAs in model plant genomes computationally. Furthermore, miRAuto enables users to identify miRNAs in non-model plant species whose genomes have yet to be fully sequenced. miRAuto analyzes the expression of the 5'-end position of mapped small RNAs in reference sequences to prevent the possibility of mRNA fragments being included as candidate miRNAs. We validated the utility of miRAuto on a small RNA dataset, and the results were compared to other publicly available miRNA prediction programs. In conclusion, miRAuto is a fully automated user-friendly tool for predicting miRNAs from small RNA sequencing data in both model and non-model plant species. miRAuto is available at <http://nature.snu.ac.kr/software/miRAuto.htm>.

INTRODUCTION

Several classes of small RNAs have been discovered in the past decade. One such class of non-coding RNAs (ncRNAs) is the microRNA (miRNA), which is found in many organisms and implicated in a wide range of biological processes, including differentiation during development and metabolism (Bartel, 2009). Lately, small RNA deep sequencing technology has shown significant promise for small RNA identification and has become commonly available and affordable (Grimson et al., 2008; Ruby et al., 2006). According to the miRBase database (Release 18, Nov. 2011), 18,226 hairpin precursor miRNAs (pre-miRNAs) have been discovered in 168 organisms. In the Plant miRNA Database (PMRD), approximately 9,200 pre-miRNAs have been cataloged in 123 plant species, including *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, and *Glycine max*.

Characteristics, such as a miRNA's stem-loop structure, the free-folding energy of its precursor, and the conservation of miRNA sequences, are widely used by miRNA detection programs to effectively predict miRNAs (Berezikov et al., 2005; Ritchie et al., 2007; Xue et al., 2005). Using these characteristics, many studies about miRNAs from non-model plant species have been reported as well (Donaire et al., 2011; Kim et al., 2012; Mica et al., 2009).

Here, we developed the miRNA auto-detection ("miRAuto") software, which enables users to predict conserved and novel miRNA candidates in both model and non-model plant genomes. miRAuto predicts candidate miRNAs based on expression analysis of the 5'-end position of mapped small RNAs in reference sequences (i.e., genomic sequences, contigs, or EST databases), thereby reducing the probability that mRNA fragments are included as candidate miRNAs. miRAuto is different from the existing miRNA prediction software because of its user-friendly interface and integrated analysis. To predict miRNAs, miRAuto uses database information and predicted/statistical approaches, which provide reliable results in both model and non-model plant species for conserved and novel miRNAs. Furthermore, we tested miRAuto on a small RNA dataset to illustrate the usage of the program. These results were compared to existing miRNA prediction software to evaluate the performance of miRAuto. Furthermore, some miRNA candidates predicted by miRAuto were validated experimentally using Northern blot analysis. miRAuto provides an efficient, user-friendly solution for miRNA prediction from small RNA deep sequencing data, thereby providing a convenient tool for miRNA research in plants.

MATERIALS AND METHODS

Work flow and data processing

Having consulted with the previously developed programs, such as miRDeep (Friedlander et al., 2008) and PIPmiR (Breakfield et al., 2012), the following miRNA prediction pipeline was constructed. The automated pipeline of miRAuto consists of

¹National Instrumentation Center for Environmental Management, College of Agriculture and Life Sciences, ²Department of Agricultural Biotechnology, ³Plant Genomics and Breeding Institute, Seoul National University, Seoul 151-921, Korea, ⁴These authors contributed equally to this work.
*Correspondence: choii@snu.ac.kr (IYC); cshin@snu.ac.kr (CS)

Fig. 1. miRAuto architecture using the use-case model. The system architecture of miRAuto using the use-case model, which enables users to work step-by-step through an automated pipeline.

four modules: (1) data pre-processing, (2) classification of other small RNAs, (3) identification of conserved and novel miRNA candidates, and (4) prediction of their target genes (Fig. 1). The data pre-processing module is designed to remove the 3' adaptor sequence of deep sequencing reads and filter low quality sequencing reads using fastx_toolkit-0.0.13.2 (Blankenberg et al., 2010). After this process, the distinct reads are collapsed into a unique read and counted. Read types exhibiting the following properties were discarded during this step: (1) less than 15 nucleotide (nt) in length, (2) ambiguous nucleotides, and (3) no adaptor sequence. To classify other small RNA molecules, these refined reads are queried against ncRNAs (rRNA, tRNA, snRNA, and snoRNA) from known databases (rfam, ribosomal Database Project - release 10: <http://rdp.cme.msu.edu>, snoRNA database: <http://lowelab.ucsc.edu/snoRNAdb>, tRNAscan, and genomic tRNA database: <http://lowelab.ucsc.edu/GtRNAdb>) using the Burrows-Wheeler Aligner (BWA) search. Matched reads are excluded from further analysis. This feature is depicted in Fig. 1.

To predict conserved miRNAs in model plant species, refined data were compared to mature sequences of known miRNAs cataloged in a database (miRBase and/or PMRD; Fig. 2). Matched data were considered conserved miRNA candidates, whereas unmatched data were presumed to be novel miRNA candidates. For novel miRNA candidates, expression of the 5'-end position of mapped small RNAs (the minimum default depth was set to 100) was analyzed using genomic sequences (Fig. 2). Based on this expression analysis, flanking sequences were obtained from highly expressed small RNA sequences. Using sliding window methods (i.e., window size: 200 nt + the length of small RNA + 5 nt, moving size: 30 nt), sub-sequences were extracted from the flanking sequences, which were further subjected to secondary structure analysis using the RNAfold program (Hofacker et al., 1994) (<http://www.tbi.univie.ac.at/~ivo/RNA/vienna> package). Stem-loop structures exhibiting the lowest acceptable minimum free energy of folding (MFE) were considered best and were therefore selected. The default parameters for annotating the stem-loop structure were: mismatches ≤ 6 , bulge size ≤ 2 , and bulge frequency ≤ 2 .

To predict miRNAs in non-model plant species, expression of the 5'-end position of a miRNA was primarily analyzed from the contigs of a reference genome or an EST database as a reference sequence (Fig. 3). The method used to predict the flanking sequences and secondary structures of miRNAs from non-model plants was identical to that described above (Supplementary Fig. S2). miRNA target genes were predicted using two different approaches: (1) Plant Small RNA target Analysis Server with the default settings, and (2) BWA antisense matching allowing G-U pairs with up to four mismatches.

Construction of a small RNA library

A small RNA sequencing library was constructed using the Small RNA Sample Prep Kit (Illumina, USA) according to the manufacturer's instructions. Briefly, five micrograms of total RNA from hot pepper fruit samples [*Capsicum annuum*; CM334; 25 days after pollination (25 DAP)] were ligated to 3' and 5' adaptors sequentially and converted to cDNA by reverse transcription-polymerase chain reaction (RT-PCR). The resulting cDNAs were amplified by PCR, gel-purified, and submitted for Illumina/Solexa sequencing. The GEO accession number for this series is GSE42219.

Northern blot analysis

Total RNA was extracted from three different hot pepper tissues using TriReagent (Ambion). Twenty micrograms of each total RNA sample from the fruit (25 DAP), leaf, and flower were individually separated on a 15% UREA-polyacrylamide gel, transferred to Hybond-NX membrane (GE Healthcare), and chemically cross-linked via 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) (Pall and Hamilton, 2008). For labeling reaction of probes for Northern blotting, 2 μ l of 10 μ M oligonucleotide, 2 μ l 10X T4 PNK buffer (Takara, Japan), 2.5 μ l of [γ - 32 P] ATP [$> 7,000$ Ci/mmol (~ 150 μ Ci/ μ l)], 12.5 μ l of dH₂O, and 1 μ l of T4 Poly Nucleotide Kinase (Takara, Japan) were combined and incubated for 1 h at 37°C. The labeled probes were further purified using PERFORMA Spin Columns (Edge Bio) according to the manufacturer's instructions. Probe sequences used for Northern blot analysis are; miR167a: 5'-TAGATCATGCTGGC-

Fig. 2. Module flow of miRAuto on model plant genomes. The pre-processed small RNA reads are classified into two major groups (conserved and novel miRNA candidates) using the similarity search with miRBase and/or PMRD databases. The putative conserved miRNA sequences are mapped to a reference genome to extract their flanking sequences, whose lengths are dependent on the length of the stem-loop precursor. For novel miRNA candidates, the expression of 5'-end position of mapped small RNAs is investigated to find the genomic distribution. The highly expressed small RNA peaks (e.g., default: 500) are selected, and the corresponding flanking sequences are extracted from nearby positions of these peaks. These flanking sequences are subjected to secondary structure analysis. Stem-loop precursors with a minimum value of folding free energy are ultimately chosen to be the best miRNA precursor candidates.

Fig. 3. Expression analysis of the 5'-end position of mapped small RNAs on a reference sequence. The expression analysis of 5'-end positions of mapped small RNA reads is performed to prevent the possibility of mRNA fragments being included in the group of candidate miRNAs.

AGCTTCA-3', sRNA_538_21_2250: 5'-AACAACTGTGCTCGATCTTA-3', sRNA_2159_21_644: 5'-CGTGTGAACCTTATCTCATAA-3', and sRNA_392_22_3072: 5'-ACCTGAAGCATGCTTCTCTA-3'. Hybridization and washing procedures were performed as described previously (Sunkar and Zhu, 2004). The membranes were exposed to a phosphorimager, and signals were analyzed using BAS-2500 (Fuji).

RESULTS AND DISCUSSION

Graphical user interface (GUI) overview

Here, we developed an integrated automated miRNA prediction program, miRAuto, based on a GUI system. The featured characteristics of miRAuto are described in Table 1. In general,

GUI system is not flexible because most options are fixed to default values. On the other hand, users can control the analysis options in a command line interface (CLI) system. miRAuto's automated pipeline is as flexible as a CLI system. For example, contrary to other GUI systems, miRAuto is capable of adjusting the number of central processing units (CPUs), therefore providing a faster speed for analysis. This system can handle two modes of miRNA analysis, namely miRNA prediction from model and non-model plant genomes. Such GUI systems are depicted in Fig. 4. miRAuto performs miRNA prediction differently, depending on which genome is chosen. The system is a user-friendly GUI, and users can select from blastall 2.2.25, blastn 2.2.26+, and BWA 0.6.1-r104 for sequence alignment. Users can also adjust the options for read

Table 1. The featured characteristics of miRAuto

miRAuto
Available for novel miRNA prediction
Available for visualization of hairpin structures
Web service mode and stand-alone service mode
Graphical user interface (GUI)
Available for homology search with mature miRNA sequences
Automation for data pre-processing
Available for analyzing both model and non-model plant genomes
Input data files
Small RNA sequencing data
Reference sequences (genomic sequences, contigs, or EST databases)
Annotation file for reference sequences (.gff)
Optional : data for classification of other ncRNAs (rRNA, tRNA, snRNA and snoRNA)

mapping and secondary structure prediction. After a full analysis, the following results are provided: (i) Read statistics of miRNAs: the 5'-terminal nucleotide information (A, U, G, or C) of refined miRNA candidates, along with their length distribution. (ii) Precursor, mature, and star sequences for conserved and novel miRNA candidates. (iii) A list of conserved and novel miRNA candidates; the conserved miRNAs are extracted from miRBase and PMRD. (iv) The MFE, GC contents, and stem-loop structures of miRNA candidates. Secondary structure information of miRNA precursors as well as the mature miRNA sequences (red-colored). (v) Homology search for analyzing the conservation of miRNA sequences and their secondary structures among species.

Simulation of prediction of miRNA candidates

To demonstrate that miRAuto can predict conserved and novel miRNA candidates, we implemented the miRAuto on a small RNA dataset isolated from the fruit tissue of hot pepper (25 DAP), a non-model plant species. The complete results from

Fig. 4. Options available for selection on the miRAuto GUI. miRAuto allows users to select the genome type (model or non-model plant species) to be analyzed and adjust the number of CPUs for faster analysis.

Fig. 5. Venn diagram comparing miRAuto with other miRNA prediction programs. Venn diagrams showing comparative results of the number of miRNA candidates predicted from each of the three miRNA prediction programs, miRAuto, miRDeep-P, and miREvo, using small RNA dataset. Small RNAs with read count of 500 or higher and 18-25 nt in length were selectively compared.

Fig. 6. Northern blot analysis of miRNA candidates predicted from miRAuto. Total RNA was extracted from three different hot pepper tissues, including fruit (25 DAP), leaf and flower. U6 snRNA was used as a loading control.

this analysis, including a list of miRNA candidates and secondary structures of the precursor miRNAs, are provided in Supplementary Data S1-S3.

A total of 47,728,864 reads were sequenced. After removing adaptor sequences and filtering out low-quality reads (Phred score ≥ 20), 42,961,532 reads, representing 17,669,084 non-redundant reads, were obtained. Among them, 1,517,010 (3,383 unique reads) were considered to be putative miRNAs that are highly expressed (depth ≥ 50), more than 15 nt in length with typical stem-loop structures. We next investigated the 5'-terminal nucleotide sequence and the length distribution of miRNA candidates (Supplementary Fig. S1 and Data S1). Based on this analysis, we found that the vast majority of small RNAs were 21 nt in length with a 5'-terminal U, the characteristics of canonical plant miRNAs (Voinnet, 2009). These results are generally consistent with previous studies investigating small RNA distribution (Fahlgren et al., 2007; Rajagopalan et al., 2006).

Prediction of conserved miRNAs

Next, we sought to demonstrate that miRAuto can effectively predict conserved miRNAs in the hot pepper fruit. Our results showed that draft genome-aligned small RNA sequences, with a total frequency of 50 or higher that were BWA searched against currently known miRNAs in miRBase (Release 18) and PMRD, had perfect matches between sequences. In total, 42 conserved miRNA families were predicted from miRBase and PMRD (Supplementary Tables S1 and S2). Twenty-four families were detected in both miRBase and PMRD, while eleven and seven families were uniquely detected by miRBase and PMRD, respectively. The stem-loop structures of miRNA precursors were successfully identified using RNAfold program. Examples of representative stem-loop structures of conserved miRNA candidates predicted from miRBase and PMRD are shown in Supplementary Data S2 and S3 ('miRBase conserved highest Depth' tab and 'PMRD conserved highest Depth' tab). The homology search revealed that 42 miRNA families were conserved across a variety of plant species, including *Arabidopsis thaliana*, *Oryza sativa*, *Glycine max*, *Solanum lycopersicum*, and *Zea mays*. A full list of this data is provided in Sup-

plementary Data S2 and S3 ('mature homology search' tab).

Prediction of novel miRNAs

By the homology search in miRBase and PMRD, conserved miRNAs were excluded to distinguish them from novel miRNA candidates. From this, we predicted 3,318 and 3,331 putative novel miRNAs from miRBase and PMRD, respectively. Among these, 3,182 (miRBase) and 3,195 (PMRD) corresponded to miRNA candidates that were 18-25 nt in length and exhibited a total frequency of 50 or higher. A full list of novel miRNA candidates, including their mature sequences and the stem-loop structures of their precursors, is provided in Supplementary Data S1 and S2 ('miRBase novel Highest Depth' tab and 'PMRD novel Highest Depth' tab).

Performance evaluation

To assess how miRAuto performs relative to existing programs, we conducted a comparison test with other publicly available programs, namely miRDeep-P and miREvo, using small RNA sequencing data from the hot pepper fruit tissue (25 DAP). These programs were selected because miRDeep-P is the most widely known/used tool in plants and miREvo is a GUI-supported program. In the present comparison test, small RNAs with a frequency of 500 or higher and 18-25 nt in length were selectively chosen to determine the number of miRNA candidates predicted by all three programs (Fig. 5 and Supplementary Data S4). For conserved miRNA candidates, a total of 38 and 35 miRNAs aligned from the miRBase and PMRD databases, respectively, were predicted by all three programs. Among these, 20 (52.6%, miRBase) and 18 (51.4%, PMRD) miRNAs were predicted by all three programs, suggesting that miRAuto performs comparably to the other two programs. For novel miRNA candidates, a total of 206 miRNAs were predicted by all three programs. Of these, 5 (2.4%) miRNAs were commonly predicted by all three programs, and 23 (11.2%) and 9 (4.4%) miRNAs from miRAuto/miREvo and miRAuto/miRDeep-P, respectively. Interestingly, miRAuto predicted a higher number of novel miRNA candidates.

Validation of miRNA prediction using small RNA Northern blot analysis

We performed small RNA Northern blot analysis for some representative miRNA candidates to confirm the reliability of miRAuto. As a result, miR167a, one of the most abundantly expressed conserved miRNAs in plants (Wu et al., 2006), was readily detected, as expected. In addition, we were able to validate the expression of the novel miRNA candidates (i.e., sRNA_392_22_3072, sRNA_538_21_2250, and sRNA_2159_21_644), all of which were predicted by miRAuto (Fig. 6).

Taken together, these results suggest that miRAuto provides reliable prediction of both conserved and novel miRNAs in non-model plant species. Nonetheless, results produced by miRAuto require careful interpretation and should be validated experimentally using methods such as Northern blot analysis and/or 5' rapid amplification of cDNA ends (RACE). Overall, our results demonstrate that miRAuto is a useful and easy-to-use tool for predicting miRNAs in plants.

Availability and system requirements

Application name: miRAuto

Application home page: <http://nature.snu.ac.kr/software/miRAuto.htm>

Operating System(s): Unix/Linux (64-bit based only)

Programming language: Bash, Perl, Java

The following open source packages are required before starting the application

- 1) Java - <http://www.oracle.com/technetwork/java/javase/downloads/index.html> over version "1.6.0_14"
- 2) Perl - <http://www.perl.org/get.html> over version "5.10.1"

The built-in programs are listed as follows (3~9)

- 3) BioJava - <http://biojava.org/wiki/BioJava:Download>
- 4) iText (java PDF library) - <http://itextpdf.com/download.php>
- 5) Vienna RNA package - <http://www.tbi.univie.ac.at/~ivo/RNA/>
- 6) BWA (Burrows-Wheeler Aligner) - <http://bio-bwa.sourceforge.net/>
- 7) BLAST - http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
- 8) FASTX_Toolkit - http://hannonlab.cshl.edu/fastx_toolkit/
- 9) mothur - http://www.mothur.org/wiki/Download_mothur

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

We would like to thank Daehyun Baek and Jae Yun Lim for advising on program development, Gregory Reeves for critically reading the manuscript, and Doil Choi's laboratory for kindly providing the hot pepper fruit samples. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST; No. 2012R1A2A2A01045528) and the Next-Generation BioGreen 21 Program (No.PJ008011), Rural Development Administration, Republic of Korea.

REFERENCES

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

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 120, 21-24.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., and Nekrutenko, A. (2010). Manipulation of FASTQ data with galaxy. *Bioinformatics* 26, 1783-1785.

Breakfield, N.W., Corcoran, D.L., Petricka, J.J., Shen, J., Sae-Seaw, J., Rubio-Somoza, I., Weigel, D., Ohler, U., and Benfey, P.N. (2012). High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis. *Genome Res.* 22, 163-176.

Donaire, L., Pedrolà, L., Rosa Rde, L., and Llave, C. (2011). High-throughput sequencing of RNA silencing-associated small RNAs in olive (*Olea europaea* L.). *PLoS One* 6, e27916.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangel, J.L., et al. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of miRNA genes. *PLoS One* 2, e219.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26, 407-415.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455, 1193-1197.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh Chem.* 125, 167-188.

Kim, J., Park, J.H., Lim, C.J., Lim, J.Y., Ryu, J.Y., Lee, B.W., Choi, J.P., Kim, W.B., Lee, H.Y., Choi, Y., et al. (2012). Small RNA and transcriptome deep sequencing proffers insight into floral gene regulation in *Rosa* cultivars. *BMC Genomics* 13, 657.

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, 558.

Pall, G.S., and Hamilton, A.J. (2008). Improved Northern blot method for enhanced detection of small RNA. *Nat. Protoc.* 3, 1077-1084.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* 20, 3407-3425.

Ritchie, W., Legendre, M., and Gautheret, D. (2007). RNA stem-loops: to be or not to be cleaved by RNase III. *RNA* 13, 457-462.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127, 1193-1207.

Sunkar, R., and Zhu, J.K. (2004). Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell* 16, 2001-2019.

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* 136, 669-687.

Wu, M.F., Tian, Q., and Reed, J.W. (2006). Arabidopsis microRNA 167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. *Development* 133, 4211-4218.

Xue, C., Li, F., He, T., Liu, G.P., Li, Y., and Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 6, 310.