miR-Island: an ultrafast and memory-efficient tool for plant miRNA annotation and expression analysis

Tiantian Gao
Zhejiang Sci-Tech University
Xin Meng
Zhejiang Sci-Tech University
Wei Zhang
Zhejiang Sci-Tech University
Weibo Jin (jwb@zstu.edu.cn)
Zhejiang Sci-Tech University https://orcid.org/0000-0002-7004-0847

Software

Keywords: plant miRNA, prediction tool, quantification, differential expression analysis, Next generation sequencing

Posted Date: December 20th, 2019

DOI: https://doi.org/10.21203/rs.2.19370/v1

Abstract

Background Next-generation sequencing of small RNAs has yielded an abundance of microRNA (miRNA) profiling data for diverse plant species. Many programs have been developed for plant miRNA annotation based on sequencing data, but these programs typically require computers with powerful hardware configurations. At present, few ultrafast computational tools are available for this type of data analysis on standard personal computers.

Results We present miR-Island, an ultrafast tool for plant miRNA identification from deep sequencing data. Two important strategies contribute to the speed of the miR-Island program: (1) extracting precursor candidates using a pseudogenome and (2) using parallel processing for RNA secondary structure prediction. In our analysis, the pseudogenomic strategy reduced the time required for miRNA precursor extraction from 85 seconds to 19 seconds in Salvia miltiorrhiza , and parallel processing significantly reduced the time for the secondary structure prediction of 3957 S. miltiorrhiza miRNA precursors from 90 seconds to 32 seconds. miR-Island completed miRNA annotation for Arabidopsis in 18 minutes on a standard personal computer, which was less than 50% of the time for ShortStack and 9% of that for miRDeep-P. In terms of accuracy, miR-Island identified 128 miRNAs, which included 68 known miRNAs in miRBase. ShortStack predicted 55 total miRNAs, including 38 known miRNAs in miRBase. Of the 175 total miRNAs predicted by miRDeep-P, only 57 miRNAs were registered in miRBase. Agreement between miR-Island and ShortStack was moderate (kappa = 0.47). For the prediction of miRNAs from three other plant datasets, miR-Island spent approximately <50% of the ShortStack run time and <2.5% of the miRDeep-P run time. When the three programs were run on contig-level S. miltiorrhiza data, miR-Island, miRDeep-P and ShortStack finished the prediction in 17 minutes, 11 hours and 40 minutes, respectively.

Conclusion Unlike other approaches, miR-Island is an ultrafast and memory-efficient tool for plant miRNA annotation and quantification on standard personal computer using the strategies of pseudo genome and multiple threads. In addition, miR-Island is a single command Perl script that is convenient to use for miRNA annotation and quantification. miR-Island was implemented in Linux and is available under the GPL GPU license from https://github.com/janeyurigao/miR-Island.

Background

MicroRNAs (miRNAs) are ~ 21 nucleotide noncoding RNAs excised from stem-loop precursors by Dicer polymerase, and they direct transcriptional or post-transcriptional gene expression regulation in metazoans and plants [1]. In the last decade, miRNA detection and expression profiling have become commonly used experimental strategies [2], reflecting the tremendous advances in next-generation sequencing (NGS) technologies [3]. Although mature miRNAs from small RNA (sRNA) sequencing can be quantified without first aligning reads to a reference genome, this approach leads to a high rate of false positives due to the many other classes of sRNAs present in the datasets, including siRNAs, snoRNAs and piRNAs, that could be falsely identified as miRNAs.

In contrast to animal miRNAs, plant miRNAs are more conserved and have variable precursor lengths [4], therefore necessitating special programs for plant miRNA annotation. The currently available tools for plant miRNA identification can be divided into web-based and local tools. The web-based miRNA annotation tools are generally easy to use and include the following: SoMART [5], PlantMiRNAPred [6], miRAuto [7], SeqBuster [8], DSAP [9], CPSS [10], PsRobot [11], DARIO [12] and miRPlant [13]. However, web-based tools are limited in terms of file upload size and reference genome selection. For this reason, local programs have been more widely adopted than web-based tools. MIREAP was available early on for plant miRNA prediction from sRNA-seq data. miRDeep-P [2] was designed for plant miRNA annotation with a plant-specific scoring system and filtering criteria, and it is the most widely known/used tool in plants. ShortStack [14] is a "one-command" script for comprehensive sRNA annotation and quantification, including miRNAs, phased-siRNAs and piRNAs. The University of East Anglia (UEA) sRNA workbench [15] was designed for the comprehensive analysis of NGS sRNA data, such as the identification of miRNAs and their targets as well as expression level comparison for specific sRNA loci. miRDeepFinder [16] provides a comprehensive annotation of plant miRNAs from deep sequencing sRNA datasets. miREvo [17] is an integrated miRNA evolutionary analysis platform for NGS datasets that relies on miRDeep2 as the core algorithm for miRNA prediction. miRanalyzer [18] was developed based on a random forest model and uses support vector machine (SVM) mechanics to annotate and quantify miRNAs. mirTools2 [19] provides detailed annotation for each known miRNA and presents differential miRNA expression in a scatter plot. miRNAkey [20] is a graph-based software for the discovery and quantification of conserved miRNAs. MIReNA [21] was designed for the prediction of miRNAs and pre-miRNAs, and it explores a multidimensional space defined by only five parameters. shortran [22] is a command-line Python-based software package for miRNA annotation and quantification. Semirna [23] uses a putative target sequence as input and allows miRNA searches. microHARVESTER [24] takes a miRNA query and identifies candidate miRNA homologs from a set of sequences. The candidate genes are subsequently subjected to various filters before the final conclusion. MIRcheck [25] uses sequence/structure specification and 20-mer coordinates for miRNA prediction. MaturePred [26] is a machine-learning method based on support vector machines that predicts the positions of plant miRNAs for new plant pre-miRNA candidates. miR-PREFeR [27] uses miRNA expression patterns and follows the criteria for plant miRNA annotation to accurately predict plant miRNAs from one or more sRNA-seq data samples of the same species. However, all these programs above are not fast to identify miRNAs on standard personal computer on most plants, especially on scaffold genome which was not assembled to chromosome level.

Personal computers now commonly contain multiple CPU cores, which allow the computer to process multiple sets of information in parallel. However, the currently available miRNA prediction tools, including miRDeep-P and ShortStack, were implemented for only one CPU core and therefore do not make full use of the resources commonly available on personal computers today. Additionally, among the steps involved in miRNA prediction from deep sequencing data, the prediction of RNA secondary structures is the most time-consuming task. In this study, miR-Island, a “one-command” Perl script, was developed for plant miRNA annotation on standard personal computers that can be equipped in all laboratories at low cost. To achieve fast miRNA annotation, miR-Island incorporates two strategies: (i) a pseudogenomic strategy that greatly accelerates the extraction of miRNA candidate precursors when using a scaffold-level genome assembly and (ii) RNAfold parallel processing, which can increase the speed of secondary structure prediction for miRNA precursors up to 3-fold. The accuracy and speed of miR-Island were compared to that of miRDeep-P and ShortStack on datasets from Arabidopsis and four other plant species. The results showed that miR-Island could annotate plant miRNAs with unprecedented speed and acceptable accuracy on a standard personal computer.

Implementation

Overview

Our software implements the following six steps (Fig. 1).

1) Formatting of the genome. The input sRNA data are transformed into an acceptable format. If the reference genome is a contig- or scaffold-level assembly, then a pseudogenome is generated from the reference genomic sequences (Fig. 2A). Many contigs or scaffolds are tandem linked and spaced with a character string of 20 Ns. The local site of each contig or scaffold in the pseudogenome is recorded, and an index is established for retrieving the real site of an identified miRNA.

2) Mapping sRNA reads to the reference genome. Bowtie (version 0.12.9) [28] is used for this mapping step, and the results/output are in BAM/SAM format.

3) Extraction of potential miRNA precursors. Two different methods are used to identify the potential precursors of known and novel miRNAs. For known miRNAs, two ~ 270 nt fragments of 230 nt or 20 nt genomic sequences flanking the known miRNA at the 5' or 3' end are extracted as the potential precursors. For novel miRNAs, fragments containing two or more islands (Fig. 2B) are excised from the genome as potential precursors, with the precursor length not exceeding the user-set threshold.

4) Secondary structure prediction. miR-Island predicts precursor secondary structures for the precursors with multiple threads through parallel proceeding (Fig. 3).

5) miRNA identification. miR-Island performs miRNA prediction using plant-specific criteria based on the recommendations of Thakur et al. [29], except for the parameter adapted for plant-specific minimum free energy (MFE). The following additional criteria were also adapted: i) loop length should be greater than 6 nt; ii) subroutine for processing input files; iii) miRNA-miRNA* duplex should have fewer than 5 mismatches; iv) mature sequence should not have a continuous string of six or more of the same base; v) miRNA-miRNA* duplex should account for more than 75% of the reads mapping to the precursor locus.

6) Expression analysis for identified miRNA. The reads per million (RPM) method is used to normalize the sRNA expression levels. Moreover, when two sRNA libraries are used for miRNA annotation, miR-Island can compare miRNA expression levels in the two libraries.

Datasets

The raw data for the analyzed sRNA-seq datasets were downloaded in SRA format from the GEO database, and the accession numbers are listed in (Additional file 1). The reference genomes for Arabidopsis, tomato, rice, maize and Salvia miltiorrhiza were downloaded from the ftp sites listed in (Additional file 2). In addition, one S. miltiorrhiza scaffold-level genome assembly [30] was downloaded from the ftp site listed in (Additional file 2).

Performance analysis of miR-Island, miRDeep-P and ShortStack

miRDeep-P and ShortStack were selected to evaluate the performance of miR-Island using sRNA libraries from five plant species, Arabidopsis, tomato, rice, maize and S. miltiorrhiza. To compare the accuracy of the three tools, miRNAs listed in the miRBase database were assumed to be a gold-standard set containing all true miRNAs. The preliminary accuracy metric is a ratio of the number of true miRNAs (i.e., those in miRBase) identified by a given program to the total number of miRNAs predicted by the program. The maximum memory usage of the miRNA prediction tools was determined by checking the memory usage every 1 second during the execution.

Statistical analysis

Statistical analysis was performed using SPSS 22.0 software (Chicago, IL, USA). The results were analyzed using the independent samples t-test or one-way analysis of variance (ANOVA), and $P < 0.05$ was considered statistically significant for all analyses. The strength of agreement between two tools was measured using kappa coefficients defined as $(Po - Pe)/(1 - Pe)$ [31], with the following scale: excellent (1.00–0.81); substantial (0.80–0.61); moderate (0.60–0.41); weak (0.40–0.21); and negligible (0.20–0) [32].

Hardware and software dependencies

In this study, all programs were performed on a standard personal computer with Intel core i3-3240 @ 3.4 GHz, 8 GB RAM and a 500 GB 7200 rpm SATA hard drive; these specifications can be equipped in all laboratories at low cost. miR-Island is a “one-command” Perl script (Perl 5.8 or later versions). To run it, certain software and modules should be installed properly, including Bowtie (version 0.12.9) [28], RNAfold (version 2.0.0 from the Vienna RNA Package) [33], SAMtools (version 0.1.19) [34], BioPerl (version 1.6.923), Perl (version 5.010 or later) and the Perl module "Bio::DB::Sam".

Results

The pseudogenomic strategy accelerates the extraction of miRNA precursors

To reduce the frequency of reading sequences by computer in the precursor extraction step, we first concatenate several scaffold sequences similar to the genome size of plant. Through this transformation, more than 21,000 S. miltiorrhiza genomic scaffold sequences were transformed into 50 pseudogenomic sequences with a maximum length of 60 Mb. Next, the sRNA-seq reads with the adaptor sequence removed were aligned to the pseudogenome for miRNA precursor extraction. It took only 19 seconds to extract 3,957 miRNA candidates from the S. miltiorrhiza pseudogenome. The same extraction step took 85 seconds on the scaffold genome, indicating a > 4-fold processing speed improvement when using the pseudogenome (Fig. 4).

Parallel computing speeds up the secondary structure prediction of miRNA precursors

To reduce the time required for the prediction of precursor secondary structures, we used a parallel computing strategy for RNAfold analysis of miRNA precursor candidates. With parallel processing, the secondary structure predictions for the 3,957 precursors took only 32 seconds. This result corresponded to a 3-fold processing speed improvement over single-thread processing, which required ~ 90 seconds for this step (Fig. 5).

Performance on Arabidopsis data

To evaluate the performance of miR-Island, the accuracy of miRNA annotation was first measured on an Arabidopsis dataset. Of the 128 total miRNAs identified by miR-Island, 68 were known miRNAs registered in miRBase (Fig. 6A), indicating an accuracy of 53.1%. Using ShortStack, 38 of the 55 predicted miRNAs were registered in miRBase (Fig. 6A). The kappa coefficient between the two programs was 0.47 (Table 1), indicating moderate agreement. miRDeep-P predicted 175 miRNAs, but only 57 were registered in miRBase (Fig. 6A), resulting in the lowest prediction accuracy of 32.6% using the independent samples t-test ($P < 0.05$). The kappa coefficients for miR-Island/miRDeep-P and ShortStack/miRDeep-P of 0.54 (Table 2) and 0.43 (Table 3), respectively, indicated moderate agreement. MIREAP identified 214 miRNAs, but none of them were registered in miRBase (Additional file 3). Therefore, all subsequent performance analyses of miR-Island were based on comparisons with ShortStack and miRDeep-P.

Table 1
Agreement of miR-Island and ShortSstack using Cohen's Kappa test.

		miR-Island		
Shortstack		Positive	Negative	total
	Positive	29	9	38
	Negative	39	273	
	total	68		K = 0.47

Table 2
Agreement of miR-Island and miRDeep-P using Cohen's Kappa test.

		miR-Island		
miRDeep-P		Positive	Negative	total
	Positive	39	18	57
	Negative	29	264	
	total	68		K = 0.54

Table 3
Agreement of ShortSstack and miRDeep-P using Cohen's Kappa test.

		miRDeep-p		
		Positive	Negative	total
ShortStack	Positive	24	14	38
	Negative	33	279	
	total	57		K = 0.43

In terms of the efficiency of the three tools on the Arabidopsis dataset, miRDeep-P and ShortStack completed the miRNA annotation in 2,000 and 45 minutes, respectively, whereas miR-Island required only 18 minutes to finish the miRNA annotation and quantification (Fig. 6B). In addition, the memory usage of miR-Island was equivalent to that of ShortStack (2.7 GB) and less than that of miRDeep-P (3.0 GB) (Fig. 6C).

Performance on other plant datasets

For further comparison, the three tools were used for miRNA annotations of tomato, rice and maize datasets. miR-Island was consistently the fastest tool for miRNA annotation (Fig. 7). For the prediction of tomato miRNAs, miR-Island required only about 9% and 2.5% of the time spent by ShortStack and miRDeep-P, respectively (Fig. 7A). For rice miRNA prediction, it required only about 20% and 0.2% of the time spent by ShortStack and miRDeep-P, respectively (Fig. 7B). For maize miRNA prediction, the time spent by miR-Island was only about 50% of the time spent by ShortStack on predicting maize miRNAs (Fig. 7C). miRDeep-P could not finish the maize miRNA prediction because of insufficient memory. Overall, miR-Island was more memory-efficient than the other two programs (Fig. 7D-F). Finally, the analysis indicated that the known miRNAs predited by miR-Island was the most compared to the other two programs (Fig. 7G-I).

Performance on S. miltiorrhiza data

The performance of miR-Island was also evaluated on a scaffold-level S. miltiorrhiza genome assembly, and the results are shown in (Fig. 8). miRDeep-P predicted a total of 118 miRNAs and completed the miRNA analysis in about 11 hours with a maximum memory usage of 2 Gb. Only 8 of the predicted miRNAs were in miRBase, corresponding to an accuracy of 6.8%. ShortStack ran for 40 minutes with a maximum memory usage of > 6 Gb, and a total of 42 miRNAs were predicted, corresponding to an accuracy of 19%. miR-Island completed the miRNA prediction in only 17 minutes with a maximum memory usage of 2 Gb, and 13 of the 69 predicted miRNAs were known miRNAs. In summary, for miRNA prediction using a scaffold-level S. miltiorrhiza reference genome assembly, miR-Island was faster than ShortStack and had similar accuracy, and it was faster and had greater accuracy than miRDeep-P (Fig. 8).

Discussion

At present, almost all of the available tools to analyze sRNA-seq data to identify miRNAs, including miRDeep-P [2], ShortStack [14] and MIREAP, first align the sRNA library reads to the genome then extract the candidate precursors according to certain strategies. For some species, only a scaffold-level genome assembly is available. In these cases, the available reference genome comprises a large number of scaffold sequences, and each individual scaffold is not very long. For example, there are more than 21,000 scaffolds in the S. miltiorrhiza genome [30], and the average scaffold length is only about 5 Kb. In contrast, the 12 rice chromosome sequences are 20–40 Mb in length [35]. A large number of discrete scaffold sequences greatly reduces the efficiency of extracting candidate precursors. To circumvent this inefficiency, miR-Island uses a pseudogenomic method to reduce the number of discrete sequences that must be read by the computer during the precursor extraction step. The pseudogenomic strategy improved the processing speed of miR-Island more than 4-fold compared to the non-pseudogenomic method, in which the scaffold sequences are not transformed into a pseudogenome.

Nowadays, personal computers are generally equipped with dual core processors capable of running two threads concurrently. Because the RNAfold program can only run in a single thread, the available computing power on a typical machine is not fully utilized. In this study, we incorporated a parallel computing method to predict the secondary structures of miRNA precursors. As a result, miR-Island could save 65% of the computing time for the RNAfold step compared to single-thread RNAfold prediction. Along with the pseudogenomic strategy, the parallel computing strategy for the RNAfold step allows for much faster miRNA annotation.

Local scripts have been widely used for plant miRNA annotation from sRNA-seq data (Additional file 4 Ref. 1–62). Plant miRNA identification tools developed in the past decade include miRDeep-P, ShortStack, MIREAP, miRCat, miRDeepFinder, miREvo, miRanalyzer, mirTools2, miRNAkey, MIReNA, Shortran, Semirna, microHARVESTER, MIRcheck, MaturePred and miR-PREFeR. In a search of the PubMed database, there were 62 published reports in which plant miRNAs were analyzed using the tools mentioned above. ShortStack and miRDeep-P were the most widely used for miRNA annotation (19 articles), followed by MIREAP (16 articles). miRCat, miR-PREFeR, miRDeepFinder and miREvo were used in 4, 2, 1 and 1 articles, respectively. The remaining 9 tools were not used for plant miRNA annotation (Additional file 4). In this study, we developed miR-Island, an ultrafast and reliable tool for plant miRNA annotation. Compared to ShortStack, miR-Island had similar accuracy in terms of miRNA prediction, but it was faster than ShortStack and had a more stable prediction result. Compared to miRDeep-P, miR-Island had comprehensive advantages in terms of prediction accuracy, speed and maximum memory usage.

Conclusions

The miR-Island was a single command perl script developed to ultrafast and a memory-efficient tool for plant miRNA annotation and quantification on standard personal computer. The key steps of speedup were as following: 1) excise the precursor candidates with high efficiency, the miR-Island firstly distinguish the provided reference genome. If the provided reference genome has whole chromosome sequences such as Arabidopsis, rice or etc., the genome will be mapped with small RNA reads through call for bowtie program. Contrarily, if the provided reference genome was only scaffold sequences such as S. miltiorrhiza, the several scaffolds of provided genome will be tandem linked and build an index file for reverting the location of finally identified miRNA. Then the tandem genome will be used for mapping with small RNA reads. This strategy could previously speedup the miRNA annotation for miR-Island predict the secondary structures of the excised potential precursors with multiple threads. Compared with the single thread, the miR-Island will increase the RNA fold speed with several times.

Declarations

Availability and requirements

Project name: miR-Island.

Project home page: https://github.com/janeyurigao/miR-Island.

Operating system(s): Unix/Linux.

Programming language: Perl.

Other requirements: 3rd Party Utilities (Install Instructions Provided).

License: GNU GPL 2.

Any restrictions to use by non-academics: None.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Not applicable

Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by Natural Science Foundation of China (grant no.31372075), the Natural Science Foundation of Zhejiang Province (grant no. LY18C Tech University.

Authors' Contributions

WZ and WJ conceived of the study. WZ designed and implemented the software. TG and XM tested the software. TG and XM wrote the manuscript. All authors read, commented on and approved the final manuscript.

Acknowledgments

Not applicable.

References

1.
Meyers BC, Axtell MJ, Bonnie B, Bartel DP, David B, Bowman JL, et al. Criteria for annotation of plant MicroRNAs. Plant Cell. 2008;20(12):3186–90.
2.
Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics. 2011;27(18):2614–5.
3.
Motameny S, Wolters S, Nürnberg P, Schumacher B. Next Generation Sequencing of miRNAs - Strategies, Resources and Methods. Genes (Basel). 2010;1(1):70–84.
4.
Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP. MicroRNAs in plants. Genes Dev. 2002;16(13):1616–26.
5.
Li F, Orban R, Baker B. SoMART: a web server for plant miRNA, tasiRNA and target gene analysis. Plant J. 2012;70(5):891–901.
6.
Xuan P, Guo M, Liu X, Huang Y, Li W, Huang Y. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics. 2011;27(10):1368–76.
7.
Jeongsoo L, Dong-In K, June Hyun P, Ik-Young C, Chanseok S. MiRAuto: an automated user-friendly microRNA prediction tool utilizing plant small RNA sequencing data. Molecules Cells. 2013;35(4):342–7.
8.

Pantano L, Estivill X, Martí E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. Nucleic Acids Res. 2009;38(5):e34.
9.
Huang P, Liu Y, Lee C, Lin W, Gan R, Lyu P, et al. DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res. 2010;38:W385-W91.
10.
Zhang Y, Xu B, Yang Y, Ban R, Zhang H, Jiang X, et al. CPSS: a computational platform for the analysis of small RNA deep sequencing data. Bioinformatics. 2012;28(14):1925–7.
11.
Wu H, Ma Y, Chen T, Wang M, Wang X. PsRobot: a web-based plant small RNA meta-analysis toolbox. Nucleic Acids Res. 2012;40:22–8.
12.
Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic acids research. 2011;39:W112-W7.
13.
An J, Lai J, Sajjanhar A, Lehman ML, Nelson CC. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. Bmc Bioinformatics. 2014;15(1):1–4.
14.
Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. Rna-a Publication of the Rna Society. 2013;19(6):740–51.
15.
Stocks MB, Simon M, Daniel M, Woolfenden HC, Irina M, Leighton F, et al. The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. Bioinformatics. 2012;28(15):2059–61.
16.
Xie F, Xiao P, Chen D, Xu L, Zhang B. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. Plant Mol Biol. 2012;80(1):75–84.
17.
Wen M, Shen Y, Shi S, Tang T. miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. Bmc Bioinformatics. 2012;13(1):140.
18.
Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res. 2011;39:132–8.
19.
Wu J, Liu Q, Wang X, Zheng J, Wang T, You M, et al. mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. Rna Biology. 2013;10(7):1087–92.
20.
Ronen R, Gan I, Modai S, Sukacheov A, Dror G, Halperin E, et al. miRNAkey: a software for microRNA deep sequencing analysis. Bioinformatics. 2010;26(20):2615.
21.
Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics. 2010;26(18):2226–34.
22.
Gupta V, Markmann K, Pedersen CN, Stougaard J, Andersen SU. shortran: a pipeline for small RNA-seq data analysis. Bioinformatics. 2012;28(20):2698–700.
23.
Muñoz-Mérida A, Perkins JR, Viguera E, Thode G, Bejarano ER, Pérez-Pulido AJ. Semirna: searching for plant miRNAs using target sequences. Omics A Journal of Integrative Biology. 2012;16(4):168.
24.
Dezulian T, Remmert M, Palatnik JF, Weigel D, Huson DH. Identification of plant microRNA homologs. Bioinformatics. 2006;22(3):359–60.
25.
Jones-Rhoades MW, Bartel DP. Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA. Mol Cell. 2004;14(6):787–99.
26.
Xuan P, Guo M, Huang Y, Li W, Huang Y. MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs. Plos One. 2011;6(11):e27422.
27.
Lei J, Sun Y. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. Bioinformatics. 2014;30(19):2837–9.
28.
Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
29.
Thakur V, Wanchana S, Xu M, Bruskiewich R, Quick WP, Mosig A, et al. Characterization of statistical features for plant microRNA prediction. Bmc Genomics. 2011;12(1):108.
30.
Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y, et al. Analysis of the Genome Sequence of the Medicinal Plant Salvia miltiorrhiza. Molecular Plant. 2016;9.
31.

Sun K, Xing W, Yu X, Fu W, Xu D. Recombinase polymerase amplification combined with a lateral flow dipstick for rapid and visual detection of Schistosoma japonicum. Parasites Vectors. 2016;9(1):476.
32.
Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
33.
Hofacker IL. Vienna RNA secondary structure server. Nucleic Acids Res. 2003;31(13):3429–31.
34.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map (SAM) Format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
35.
Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (New York). 2013;6(1):p. 4.

Figures

Figure 1

A schematic diagram of miR-Island for plant miRNA annotation and expression analysis.

Figure 2

Transformation of scaffold sequences into a pseudogenome and the extraction of potential miRNA precursors. (A) A pseudogenome was assembled through the tandem linkage of many contigs or scaffold sequences spaced by N20 characterization. Contig 1, 2, …, n indicate the different genome contigs. ‘20Ns’ indicates that two different contigs are separated by 20 consecutive Ns. (B) Fragments containing two or more islands are extracted as potential precursors if their length does not exceed a user-set threshold. ‘min_freq’ is a parameter for island identification, with 15 as the default in miR-Island. An island is a genome region in which the sum of mapped reads is more than min_freq. Island N, N+1, N+2, N+3 indicate the different islands in the genome regions.

Figure 3

Multithreading as a parallel computing strategy for secondary structure prediction. Figure 4 Performance time analysis for the extraction of miRNA precursors from a scaffold-level assembly and a pseudogenome sequence. Results indicate means ± SD of three replicates. Statistical significance was measured by a t-test, and asterisks indicate a significant difference (*P<0.05).

Figure 4

Performance time analysis for the extraction of miRNA precursors from a scaffold-level assembly and a pseudogenome sequence. Results indicate means ± SD of three replicates. Statistical significance was measured by a t-test, and asterisks indicate a significant difference (*P<0.05).

Figure 5

Performance time analysis for miRNA precursor prediction based on multi- and single-thread strategies. Results indicate means ± SD of three replicates. Statistical significance was measured by a t-test, and asterisks indicate a significant difference (P<0.05).

Figure 6

Performance evaluation of miR-Island compared to the miRDeep-P and ShortStack programs on Arabidopsis sRNA data. (A) For each tool, accuracy was calculated as the percentage of known miRNAs (registered in miRBase) among the total number of predicted miRNAs. (B) Performance time. (C) Maximum memory usage. Results indicate means ± SD of three replicates. Different letters indicate significant differences based on post hoc Tukey's HSD test (P < 0.05).

Figure 7

Performance evaluation of miR-Island, miRDeep-P and ShortStack on tomato, rice and maize sRNA data. A-C: Performance time of the three programs on the data from tomato (A), rice (B) and maize (C). D-F: Maximum memory usage of the three programs on the data from tomato (D), rice (E) and maize (F). G-I: Accuracy of the three programs (calculated as the percentage of known miRNAs among the total predicted miRNAs) for the data from tomato (G), rice (H) and maize (I).

Figure 8

Performance evaluation of miR-Island, miRDeep-P and ShortStack on Salvia miltiorrhiza sRNA data. (A) Performance time. (B) Maximum memory usage. (C) For each program, the percentage of known miRNAs (registered in miRBase) among the total predicted miRNAs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2.doc
- TableS3.xlsx
- TableS4.docx
- TableS1.doc