Journal Pre-proofs

sRNAminer: a multifunctional toolkit for next-generation sequencing small RNA data mining in plants

Guanliang Li, Chengjie Chen, Peike Chen, Blake C. Meyers, Rui Xia

PII: S2095-9273(23)00935-0
DOI: https://doi.org/10.1016/j.scib.2023.12.049
Reference: SCIB 2535

To appear in: *Science Bulletin*

Received Date: 13 October 2023
Revised Date: 25 November 2023
Accepted Date: 27 December 2023

Please cite this article as: G. Li, C. Chen, P. Chen, B.C. Meyers, R. Xia, sRNAminer: a multifunctional toolkit for next-generation sequencing small RNA data mining in plants, *Science Bulletin* (2023), doi: https://doi.org/10.1016/j.scib.2023.12.049

sRNAminer: a multifunctional toolkit for next-generation sequencing small RNA data mining in plants

Guanliang Li[a,b,c,1], Chengjie Chen[a,b,c,1,*], Peike Chen[a,b,c], Blake C. Meyers[d,e], Rui Xia[a,b,c,*]

[a]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Horticulture, South China Agricultural University, Guangzhou 510640, China

[b]Guangdong Laboratory for Lingnan Modern Agriculture, South China Agricultural University, Guangzhou510640, China

[c]Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in South China, Ministry of Agriculture and Rural Affairs, South China Agricultural University, Guangzhou 510640, China
[d]Donald Danforth Plant Science Center, Saint Louis MO 63132, USA

[e]Division of Plant Science and Technology, University of Missouri–Columbia, Columbia MO 65211, USA

[1] These authors contribute equally to this work.

[*] Correspondence should be addressed to Dr. Chengjie Chen (ccj@scau.edu.cn), Dr. Rui Xia (rxia@scau.edu.cn)

Abstract

Small RNAs (sRNAs), found extensively in plants, play an essential role in plant growth and development. Although various sRNA analysis tools have been developed for plants, the use of most of them depends on programming and command-line environments, which is a challenge for many wet-lab biologists. Furthermore, current sRNA analysis tools mostly focus on the analysis of certain type of sRNAs and are resource-intensive, normally demanding an immense amount of time and effort to learn the use of numerous tools or scripts and assemble them into a workable pipeline to get the final results. Here, we present sRNAminer, a powerful stand-alone toolkit with a user-friendly interface that integrates all common functions for the analysis of three major types of plant sRNAs: microRNAs (miRNAs), phased small interfering RNAs (phasiRNAs), and heterochromatic siRNAs (hc-siRNAs). We constructed a curated or "golden" set of *MIRNA* and *PHAS* loci, which was used to assess the performance of sRNAminer in comparison to other existing tools. The results showed that sRNAminer outperformed these tools in multiple aspects, highlighting its functionality. In addition, to enable an efficient evaluation of sRNA annotation results, we developed IGV-sRNA, a modified genome browser optimized from Integrative Genomics Viewer (IGV) and we incorporated it as a functional module in sRNAminer. IGV-sRNA can display a wealth of sRNA-specific features, enabling a more comprehensive understanding of sRNA data. sRNAminer and IGV-sRNA are both platform-independent software that can be run under all operating systems. They are now freely available at https://github.com/kli28/sRNAminer and https://gitee.com/CJchen/IGV-sRNA.

Keywords

sRNAminer; sRNA analysis; miRNA; phasiRNA; hc-siRNA; IGV-sRNA

Article history:

Received 13 October 2023

Received in revised form 25 November 2023

Accepted 27 December 2023

1. Introduction

Small RNAs play a critical role in plant growth and development. They are typically classified into two classes: microRNAs (miRNAs) and short interfering RNAs (siRNAs) [1,2]. Among siRNAs, phased small interfering RNAs (phasiRNAs) [3,4] and heterochromatic siRNAs (hc-siRNAs) [5] are the main subclasses.

With the rapid development of next-generation sequencing (NGS) technologies and their application in sRNA research, various bioinformatics tools have been developed for the analyses of the NGS sRNA data in plants. MicroRNAs, typically 20 to 22 nt in length, are the most well-studied in plants. Many tools have been constructed for miRNA identification, such as miRDeep-P2 [6], Mirnovo [7], miR-PREFeR [8], Shortstack [9], and miRador [10]. However, when it comes to phasiRNA identification, only a few tools are available, such as PhaseTank [11] and Shortstack [9], and there is still no dedicated tool for hc-siRNA identification. In recent years there have been numerous reports of the involvement of phasiRNAs and hc-siRNAs in plant development pathways. For example, 21-nt phasiRNAs preponderate during early anther development [12,13], and hc-siRNAs are involved in RNA-directed DNA methylation (RdDM), important for transposon silencing and plant development [2,14,15]. Therefore, more sophisticated tools are needed for phasiRNA and hc-siRNA identification, in addition to miRNA characterization. A final justification for new tools is that most of the current and available tools for sRNA data analysis demand computational skills from researchers, not only for running scripts using the command-line but also for assembling multiple scripts into a workable pipeline. This is both time-consuming and challenging for wet-lab biologists who intend to analyze their data on their own but may have weak computational skills.

With the fast and broad application of high-throughput sequencing technologies in sRNA research, diverse sRNAs have been found in plants, posting another challenge for computational analyses of NGS sRNA data. Currently, sRNA annotation tools and pipelines have numerous deficiencies. The outputs of many tools or pipelines used for sRNA annotation yield potential misannotations, especially for miRNA and 24-nt phasiRNAs [16]. This situation is similar to genome-wide annotations of coding genes [17]. Therefore, manual checks of the computationally annotated results from NGS data are a suggested practice and a reliable way to minimize misannotations, often based on visualization of the data [18]. Integrative Genomics Viewer (IGV), an open-source visualization tool, is widely used for deep sequencing data exploration [19]. It is known for its user-friendly interface and superb compatibility with all kinds of genomic data stored in various formats. However, due to the distinct characteristics of the biogenesis and function of plant sRNAs, IGV cannot represent some of the most important features of sRNAs, including, for instance, the length of sRNA reads (a critical feature of sRNA function), the distribution pattern of sRNAs (an indication of potential biogenesis and

function modes), and the secondary structure of sRNA-generating loci (an indispensable feature of *MIRNA* genes) [20].

Here, we developed sRNAminer, an all-in-one toolkit for plant sRNA analyses with a user-friendly interface. It can not only analyze miRNA, phasiRNA, and hc-siRNA with minimal user effort but also provide a variety of common sRNA analysis functions, including sRNA target analysis, degradome analysis, prediction of secondary structure, *PHAS* locus graphing, and so on. In addition, we have developed, on the basis of IGV, a new tool specifically for the visualization of various characteristics of sRNAs and the browsing of NGS sRNA data. We named this functionally improved browser tool IGV-sRNA, which has been incorporated into sRNAminer as a functional module for efficient and effective browsing and visualization of resultant files. This toolkit, sRNAminer, has been widely tested by users and it will be a useful addition to the toolbox of sRNA researchers.

2. Materials and methods

2.1. Dataset

The miRNA annotation results of *Arabidopsis thaliana* and *Oryza sativa* used for constructing the miRNA golden sets were sourced from miRBase [21], PmiREN [22], sRNAanno [23]. For the construction of a golden set of *PHAS* loci, we utilized four datasets of *Oryza sativa* obtained from the National Center for Biotechnology Information (NCBI) (accession#: SRR3955351, SRR3955352, SRR3955353, SRR3955354).

To evaluate the performance of sRNAminer in miRNA identification, for demonstration, we obtained 15 datasets from NCBI, three from *Arabidopsis thaliana* (accession#: SRR3992484, SRR3992485, SRR3992486) and 12 from *Oryza sativa* (accession#: SRR11622382, SRR11622413, SRR11622414, SRR11622415, SRR11622416, SRR12744573, SRR11622383, SRR11622384, SRR11622385, SRR11622386, SRR11622387, SRR11622393). For the performance evaluation in *PHAS* locus annotation, we used the same 12 datasets from *Oryza sativa* used for miRNA annotation and three datasets from *Fragaria vesca* (accession#: SRR1586419, SRR1586420, SRR1586424).

2.2. Pre-processing of sRNA-seq data

Adapters were removed from all the sRNA-Seq data using sRNAminer. Clean reads with

a length of at least 15-nt were kept for further analyses.

2.3. Pre-processing workflow of sRNAminer

For adapter trimming, ten read sequences are randomly selected from the raw sequence file and aligned to search for the longest common substring sequence, which is extracted as a candidate adapter sequence. This process is repeated 1000 times, resulting in 1000 candidate adapter sequences. The frequency of each candidate adapter sequence are then counted, and the candidate adapter sequence with the highest frequency is chosen. After that, the possible extension of the candidate adapter toward the 5' terminal of reads is checked. If an extension is found, the adapter sequence will be elongated; otherwise, the extension process will be halted, and the final adapter sequence will be determined. Using the obtained adapter, adapter trimming is performed on the raw sequencing data with one mismatch allowed.

For data cleaning, sRNA sequencing data often contain reads generated from other sources: ncRNA (rRNA, tRNA, snoRNA, and snRNA), cpDNA/RNA, and mtDNA/RNA. These read data are usually removed before sRNA annotation. Reported ncRNA and organelle sequences are collected from Rfam database and organelle genome from NCBI, respectively. We removed reads that match these sequences with one mismatch allowed.

For sequence collapsing, sRNAminer is designed to support large-scale data analysis on low-memory computing devices. By default, reads are initially divided into 16 files based on the combination of the first two bases of each read sequence. Subsequently, sequence redundancy counting is performed for each sequence file, and redundant sequences are merged with the frequency of each sequence recorded.

For alignment and bam file preparation, bowtie 1 [24] is used to map the reads to the reference genome and bam file is generated by SAMtools [25].

2.4. Parameter calculation of sRNAminer

Identification of miRNA, phasiRNA, and hc-siRNA followed the approaches we used in our previous work [23]. The *P*-value and phasing score were calculated based on the method mentioned in the previous study [26]. The repeat score is calculated for each 15-

mer based on its total hits on genome, with a formula of $\log_2(Hits\ of\ 15\text{-}mer/4)$.

2.5. miRNA and *PHAS* loci identification assessment

To assess the sensitivity and precision of sRNAminer, the F_1 score was utilized as a metric. This metric was mainly calculated based on sensitivity and precision. For a more comprehensive and objective evaluation, we constructed golden sets for miRNA in *Arabidopsis thaliana* and *Oryza sativa*, as well as a golden set for *PHAS* loci in *Oryza sativa*, which were used in the calculation of the F_1 score. The calculation formulas are as follows:

$$\text{Sensitivity} = \frac{\text{Predicted miRNA/}PHAS\text{ loci found in golden miRNA/}PHAS\text{ loci set}}{\text{miRNA/}PHAS\text{ loci in golden miRNA/}PHAS\text{ loci set}}, \quad (1)$$

$$Precision = \frac{\text{Predicted miRNA/}PHAS\text{ loci found in golden miRNA/}PHAS\text{ loci set}}{\text{Predicted miRNA/}PHAS\text{ loci}}, \quad (2)$$

$$F_1 = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity}. \quad (3)$$

2.6. Program implementation

All tools were run on a local server using the default or recommended parameters. The server was equipped with 64 central processing units (CPUs, Intel(R) Xeon(R) Platinum 8370C CPU @ 2.80GHz) and 1007 GB of RAM. The operating system was CentOS7.

3. Results

3.1. Overview of Functions in sRNAminer

Compared with other tools, sRNAminer is a cross-platform software with a Graphical User Interface (GUI) that can be run under Windows, Macintosh, and Linux. Moreover, we also provide a command-line version for users to analyze large data on servers. Notably, sRNAminer covers different types of sRNA analysis with data browsing functions while other tools focus on certain, limited areas (Table 1 and Fig. 1a). The functions of sRNAminer are divided into three main parts: (1) Data pre-processing. This part includes four steps, from Adapter Trimming and Sequence Collapsing to Data Cleaning and Genomic Mapping. (2) sRNA identification and abundance calculation, which can be

performed for three main types of sRNAs, miRNA, phasiRNA, and hc-siRNA. (3) Other common functions, including sRNA target prediction, *PHAS* trigger identification, and degradome analysis, etc. All the sRNA analysis methods were coded from scratch using Java and Python, except for the degradome analysis, which was migrated from a previous open-source tool sPARTA [27].

Table 1. Performance comparison of benchmarked tools

Key Features		miRDeep-P2	Mirnovo	miR-PREFeR	Shortstack	PhaseTank	sRNAminer
Command line		√	√	√	√	√	√
Cross-platform (Linux, Mac OS, and Windows)		×	×	√	×	×	√
GUI (Graphical User Interface)		×	×	×	×	×	√
Adapter Trimming		×	√	×	√	×	√
Noisy Sequence Filtering		√	√	×	×	√	√
miRNA Analysis		√	√	√	√	×	√
phasiRNA Analysis	21-nt	×	×	×	√	√	√
	24-nt	×	×	×	√	√	√
hc-siRNA Analysis		×	×	×	×	×	√
Degradome Analysis		×	×	×	×	√	√
sRNA Data browsing in		×	×	×	×	×	√

real-time						
miRNA secondary structure plot	×	×	×	×	×	√
PHAS locus graph	×	×	×	×	×	√

To enable the easiest and quickest analysis of sRNA data, we offered a“One step analysis mode”(One step sRNAminer) (Fig. 1b). Users can simply input the sRNA sequence data and genome files and click "Start" to identify different types of sRNAs and estimate their abundances.

IGV-sRNA is a powerful sRNA data visualization tool that we developed, on the basis of the original IGV (Integrative Genomics Viewer) software, to provide numerous additional functions specialized for plant sRNA profiling, including secondary structure visualization, calculation of the phasing score on-the-fly, and read length visualization in a color-coded dot plot (Fig. 1a). The same color scheme for sRNA length can also be applied to the visualization of sRNA genomic alignment, allowing users to easily detect the distribution pattern of sRNAs with different lengths. To ensure a more flexible and convenient browsing of sRNA data, the“sRNA viewer”function in sRNAminer is designed to interact with IGV-sRNA.

Fig. 1. Outline of functions in sRNAminer. (a) sRNAminer covers data pre-processing, annotation of different types of sRNAs, and abundance calculation. Several common analysis functions related are also provided. IGV-sRNA, incorporated in sRNAminer, is a genome browsing tool for users to visualize the features of sRNA-generating loci, including phasing score, repeat score, sRNA abundance, etc. (b) sRNAminer provides a “One step sRNAminer” function, which enables users to obtain miRNA, phasiRNAs, and hc-siRNA annotation and abundance calculation results with one click after specifying genome files, raw data, and databases.

3.2. Golden miRNA and *PHAS* loci set construction

Currently, there are several small RNA databases available, such as miRBase [21], PmiREN [22], sRNAanno [23], and plant MPSS databases [28]. However, the lack of manually checked high-quality sRNA loci datasets poses a challenge in evaluating the performance of various sRNA analysis tools. To address this gap, we set out to construct golden sets of miRNAs and *PHAS* loci; that is, sets of manually curated, high-quality, reference data sets.

For miRNAs, we constructed the golden sets using the following process: Firstly, we obtained miRNA sets of *Arabidopsis thaliana* and *Oryza sativa*, which serve as model plants for eudicots and monocots, respectively, from miRBase, sRNAanno, and PmiREN. Subsequently, we filtered for miRNAs that were presented in at least two databases. To ensure the accuracy of these filtered miRNAs, we double-checked their secondary structures manually (Fig. S1 online), resulting in the creation of the golden set of miRNAs (Fig. 2a). As a result, there were 204 miRNAs in the golden set of *Arabidopsis thaliana* (Table S2 online) and 330 in the golden set of *Oryza sativa* (Table S3 online).

For the *PHAS* loci, some monocot species, for instance, rice (*Oryza sativa*), have much larger quantities of *PHAS* loci in the genome, compared to the eudicot model plant, *Arabidopsis* [4]. We analyzed the small RNA data from *Oryza sativa* (Table S1 online) using sRNAminer, PhaseTank [11], and *PHASIS* [29] to obtain a set of 21-*PHAS* loci (generating 21-nt phasiRNAs). Given the well-established role of miR2118 as a trigger for 21-*PHAS* loci [30], we examined whether these loci were targeted by miR2118. If an identified locus contained the target site of miR2118, we checked the number of miR2118 targeting sites and split the tandem 21-*PHAS* loci. Conversely, in cases where miR2118 did not target the loci, we evaluated their phasing patterns manually (Fig. 2b). In total, we identified 2462 21-*PHAS* loci in the golden set of *Oryza sativa* (Table S4 and Fig. S2 online). In addition, using a similar approach, we identified 126 24-*PHAS* loci (generating 24-nt phasiRNA) in the golden set of *Oryza sativa* (Table S5 and Fig. S3a online).

Fig. 2. Workflow of the construction of miRNA/*PHAS* loci golden set. The bottom part of (a) shows the structure of exemplative golden-set miRNAs for *Arabidopsis thaliana (*ath-*)* and *Oryza sativa (*osa-*)*. The bottom part of (b) presents the sRNA distribution and phasing score of a representative golden-set *PHAS* locus for *Oryza sativa.* The dark purple box indicates the miR2118 targeting site on the anti-sense strand, and the dark purple line indicates the miR2118 cleavage site on the anti-sense strand.

3.3. Fast and accurate sRNA locus annotation of sRNAminer

sRNAminer applies well-established criteria to miRNA identification [23]. We benchmarked sRNAminer with two commonly used miRNA annotation pipelines, miRDeep-P2 and ShortStack in *Arabidopsis thaliana* and *Oryza sativa* (Table S1 online). To ensure high accuracy, the software retained only 20- to 22-nt miRNAs for comparison, as miRNAs out of this length range have been rarely reported to be functional. To evaluate the performance of different software with precision and sensitivity, we employed the comprehensive metric known as F_1 score which can provide a balanced measure of a tool's performance on precision and sensitivity. We found that sRNAminer obtained F_1 scores comparable to miRDeep-P2 in both *Arabidopsis thaliana* and *Oryza sativa*, but much higher than ShortStack (Fig. 3). Notably, compared to miRDeep-P2, sRNAminer completed the miRNA identification in only half the time (Fig. 3). Overall, we contend that the miRNA annotation function in sRNAminer can retrieve highly reliable results more efficiently.

Fig. 3. Performance comparison of sRNAminer with miRDeep-P2 and ShortStack on miRNA annotation. F_1 scores and run time of three software packages were compared using NGS sRNA datasets from *Arabidopsis thaliana* (a) and *Oryza sativa* (b).

PhasiRNAs represent another major class of sRNAs. The identification of *PHAS* loci is based on the *P*-value and phasing score [26]. The default cut-off of *P*-value and phasing score in sRNAminer are 10^{-3} and 10, respectively. We used F_1 score as well to compare the performance of sRNAminer and PhaseTank using small RNA data from rice reproductive tissues at four developmental stages [31] (Table S1 online). Despite a slightly longer run time compared to PhaseTank, sRNAminer consistently achieved higher F_1 scores across all stages for 21-*PHAS* loci identification (Fig. 4a). sRNAminer was able to specifically identify 227 21-*PHAS* loci in the golden set among four stages, such as the locus *PHAS21-1616,* while PhaseTank was only able to specifically identify seven loci (Fig. 4a). To demonstrate the generalizability of sRNAminer, we also evaluated the performance of sRNAminer and PhaseTank using sRNA datasets from *Fragaria vesca* (Table S1 online), a eudicot species known for containing a large number of *PHAS* loci [32]. Specially, we define a *PHAS* locus with the *P*-value less than 10^{-5} and the phasing score greater than 15 as a highly confident locus. The result showed that sRNAminer detected much more 21-*PHAS* loci compared to PhaseTank. Among these,

38 loci were manually verified as highly confident ones, such as the locus *PHAS21-56* (Fig. 4b). In contrast, PhaseTank identified 30 extra *PHAS* loci, but none of them were highly confident upon manual evaluation. Moreover, in addition to 21-*PHAS* loci, sRNAminer can also be used for the identification of 24-*PHAS* loci which produce 24-nt phasiRNAs, another main class of phasiRNAs in plants [33]. We also evaluated the performance of sRNAminer and PhaseTank on 24-*PHAS* loci identification with the same dataset used to evaluate the performance of 21-*PHAS* loci identification. Similarly, compared to PhaseTank, sRNAminer consistently achieved higher F_1 scores across all stages and specifically identified 13 24-*PHAS* loci in the golden set among four stages, such as the locus *PHAS24-94* (Fig. S3b online). In conclusion, sRNAminer is a powerful tool for the annotation of *PHAS* locus with high sensitivity and accuracy.

Fig. 4. Performance comparison of sRNAminer with PhaseTank on 21-*PHAS* locus annotation. Results and run time of both tools were compared using NGS sRNA datasets from *Oryza sativa* (a) and *Fragaria vesca* (b). (a) These datasets come from the four stages of rice reproductive tissues which are PMC formation stage (PFS), PMC prophase stage (PPS), PMC meiotic divisions stage (PMS), and early microspore stage (EMS). The bottom panel shows the mapping (dot plot) and phasing score (line plot) of a representative 21-*PHAS* locus (*PHAS21-1616*) in the golden set, which were identified by sRNAminer but not by PhaseTank. The red box indicates the miR2118 targeting site on the sense strand, and the red line indicates the miR2118 cleavage site on the sense strand. (b) The number in red color represents the quantity of highly confident 21-*PHAS* loci (phasing score >= 15 and *P*-value <= 10^{-5}) with manual validation. The bottom panel shows the mapping (dot plot) and phasing score (line plot) of representative high-confident 21-*PHAS* locus (*PHAS21-56*), which were identified by sRNAminer but not by PhaseTank.

Heterochromatic siRNAs (hc-siRNAs) represent the most abundant class of sRNA found in most plant genomes. In sRNAminer, we also developed a method to annotate genomic loci that generate hc-siRNAs. We first identify genomic loci with a predominant representation of 23 and 24 nt small RNAs, which account for 50% or more of the total sRNAs in the loci. As hc-siRNAs are mostly derived from repetitive regions of the genome, mainly transposons, we adapted a metric of the average hits of mapping reads to evaluate the repetitiveness of a sequence; this can distinguish hc-siRNAs from other types of 24-nt siRNAs, for instance, non-repetitive 24-nt phasiRNAs (Fig. S4 online) and 24-nt siRNAs derived from long inverted-repeat regions. sRNA-generating regions with average hits greater than ten were, by default, considered as hc-siRNA loci.

3.4. IGV-sRNA for browsing NGS sRNA data

IGV [19] is a popularly used browser for viewing NGS data generated by all kinds of sequencing techniques. However, current implementations of this browser are not effective for browsing sRNAseq data, because of the unique features of interest to small RNA biologists that are largely irrelevant for RNA-seq data or other datatypes more frequently displayed in IGV. These features relevant to sRNAseq data include high sequence repetitiveness, the critically important variation in length (i.e. differences of a single nucleotide), the importance of one single read (and its exact position and abundance), and other features. Therefore, we developed an enhanced version of IGV, IGV-sRNA, as a function module in sRNAminer; this module is powerful for manually exploring plant sRNA sequencing data. Several sophisticated display features were implemented.

First, IGV-sRNA can automatically resolve the genome-mapped file of collapsed sRNA datasets, which makes it compatible with the high sequence redundancy of raw sRNA data (Fig. 5). Second, sRNA reads can be color-coded according to their length, with the cyan color for 21-nt reads, green for 22-nt, purple for 23-nt, orange for 24-nt, and grey for others (Fig. 5); these color assignments are consistent with other small RNA genome browsers [34]. All the sRNA reads can be shown with these different size-based colors, which is helpful for the quick assessment of the distribution profile of different sRNA. sRNA abundance can also be shown in a color-coded dot plot. Third, a genomic sequence can be folded to show its secondary structure with coverage information indicated simultaneously (Fig. 5a), which is critical for the evaluation of a *MIRNA* locus. Fourth, the phasing score of phasiRNAs is calculated instantly, according to their length, and displayed using a line plot. Whether a sRNA-generating locus is an authentic *PHAS* locus or not can be quickly evaluated by the data track showing the phasing score (Fig. 5b, c). Furthermore, the pre-calculated repeat score from a genome can be shown for hc-siRNA loci. The pre-calculated data of repeat scores can be loaded into IGV-sRNA to quickly assess the repetitiveness of a sequence or region (Fig. 5d).

To help users check their sRNA loci of interest in IGV-sRNA more easily, we developed a function called “sRNA viewer” in sRNAminer. It can automatically import read alignment and genome sequence files into IGV-sRNA and create an interactive table based on the sRNA identification results provided by users. Users could simply click the "GO" button to switch the IGV-sRNA view to the position of the sRNA locus directly (Fig. S5 online). In addition, it is convenient for users to fold a sequence to check the secondary structure and color the miRNA:miRNA* by using the function of “Vienna RNAfold” in sRNAminer (Fig. S6a online). Publication-quality graphs for the *PHAS* locus can be easily prepared in sRNAminer; for this, users just need to input the information of the *PHAS* locus, then the graph will be exported automatically (Fig. S6b online). Overall, sRNAminer in combination with IGV-sRNA provides a great way for users to manually evaluate the

annotation results of sRNA loci.

Fig. 5. Representative sRNA-generating loci viewed with IGV-sRNA. (a) A representative *MIRNA* locus. Users can fold a sequence to view the secondary structure with coverage information indicated synchronously. The abundance of each sRNA is showed in color-coded dot plots (track #2), with the cyan color for 21-nt reads, green for 22-nt, purple for 23-nt, orange for 24-nt, and grey for others. All alignments in IGV-sRNA are displayed with different color codes according to read length (track #3). (b, c) Representative 21-*PHAS* (b) and 24-*PHAS* (c) loci. The phasing score of phasiRNAs can be calculated instantly according to their length and displayed by line plot (track #5). (d) A representative hc-siRNA locus. 23 or 24 nt sRNA are enriched in this region with a high repeat-score (track #6).

3.5. Command-line version of sRNAminer

Local personal computers with limited computational resources are usually not suitable for tasks requiring numerous CPU cores and a large memory footprint, such as performing sRNA analysis for hundreds of datasets or in species with a large genome (>3 Gb). Most resource-intensive tasks rely on high-performance computing servers. Therefore, we also offer a command-line version of sRNAminer which allows users to run sRNAminer in command-line environment on servers. The command-line version of sRNAminer can be easily installed via conda, and its parameters are set in a clear and easy way of “sRNAminer + function + parameter” (Fig. S7 online), allowing users to quickly master the usage of sRNAminer commands. Notably, multithreading is also supported in sRNAminer, which enables the processing of large datasets at high speed.

4. Conclusions

Collectively, sRNAminer can be used for the annotation of all the three major classes of plant sRNAs in a fast and accurate way. sRNAminer coupled with IGV-sRNA provides a convenient and efficient way for the visualization of alignment data, which can help minimize the false-positive rate of annotation results. To help users get started as quickly as possible, we provide instructions for the use of sRNAminer and IGV-sRNA (https://www.yuque.com/u758713/at2327).

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Science Foundation of China (32072547 and 32102320) and the Key-Area Research and Development Program of Guangdong Province (2022B0202070003). This work is also supported by the open competition program of top 10 critical priorities of Agricultural Science and Technology Innovation for the 14th Five-Year Plan of Guangdong Province (2022SDZG05) and the fund from Hainan Yazhou Bay Seed Lab (JBGS-B21HJ0001). Work in the Meyers lab on small RNA classification and miRNA annotation is supported by US National Science Foundation award (2130883).

Author contributions

Chengjie Chen, Rui Xia, Blake C. Meyers, and Guanliang Li initiated the study. Guanliang Li and Chengjie Chen developed the toolkit. Guanliang Li, and Peike Chen set up the golden set of miRNA and *PHAS* loci. Guanliang Li, and Chengjie Chen evaluated the performance of sRNAminer and IGV-sRNA. Guanliang Li, Chengjie Chen, Rui Xia, and Blake C. Meyers wrote the manuscript.

References

[1] D'Ario M, Griffiths-Jones S, Kim M. Small RNAs: Big impact on plant development. Trends Plant Sci 2017;22:1056–68.

[2] Axtell MJ. Classification and comparison of small RNAs from plants. Annu Rev Plant Biol 2013;64:137–59.

[3] Fei Q, Xia R, Meyers BC. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. Plant Cell 2013;25:2400–15.

[4] Liu Y, Teng C, Xia R, et al. PhasiRNAs in Plants: Their biogenesis, genic sources, and roles in stress responses, development, and reproduction. Plant Cell 2020;32:3059–80.

[5] Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 2010;11:204–20.

[6] Kuang Z, Wang Y, Li L, Yang X. miRDeep-P2: Accurate and fast analysis of the microRNA transcriptome in plants. Bioinformatics 2019;35:2521–2.

[7] Vitsios DM, Kentepozidou E, Quintais L, et al. Mirnovo: Genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. Nucleic Acids Res 2017;45:e177.

[8] Lei J, Sun Y. miR-PREFeR: An accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. Bioinformatics 2014;30:2837–9.

[9] Axtell MJ. ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA 2013;19:740–51.

[10] Hammond RK, Gupta P, Patel P, et al. miRador: A fast and precise tool for the prediction of plant miRNAs. Plant Physiol 2023;191:894–903.

[11] Guo Q, Qu X, Jin W. PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades. Bioinformatics 2015;31:284–6.

[12] Zhai J, Zhang H, Arikit S, et al. Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. Proc Natl Acad Sci U S A 2015;112:3146–51.

[13] Liu B, Li G, Chen C, et al. Species-specific regulatory pathways of small RNAs play sophisticated roles in flower development in *Dimocarpus longan* Lour. Horticultural Plant Journal 2023;9:237–49.

[14] Naydenov M, Baev V, Apostolova E, et al. High-temperature effect on genes engaged in DNA methylation and affected by DNA methylation in *Arabidopsis*. Plant

Physiol Biochem 2015;87:102–8.

[15] Popova OV, Dinh HQ, Aufsatz W, et al. The RdDM pathway is required for basal heat tolerance in *Arabidopsis*. Mol Plant 2013;6:396–410.

[16] Polydore S, Lunardon A, Axtell MJ. Several phased siRNA annotation methods can frequently misidentify 24 nucleotide siRNA-dominated *PHAS* loci. Plant Direct 2018;2:1-13.

[17] Dunn NA, Unni DR, Diesh C, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol 2019;15:e1006790.

[18] Xia R, Xu J, Meyers BC. The emergence, evolution, and diversification of the miR390-*TAS3-ARF* pathway in land plants. Plant Cell 2017;29:1232–47.

[19] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics 2013;14:178–92.

[20] Chen C, Zeng Z, Liu Z, et al. Small RNAs, emerging regulators critical for the development of horticultural traits. Hortic Res 2018;5:63.

[21] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res 2019;47:D155–62.

[22] Guo Z, Kuang Z, Zhao Y, et al. PmiREN2.0: from data annotation to functional exploration of plant microRNAs. Nucleic Acids Res 2022;50:D1475–82.

[23] Chen C, Li J, Feng J, et al. sRNAanno—a database repository of uniformly annotated small RNAs in plants. Hortic Res 2021;8:1–8.

[24] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 2009;10:R25.

[25] Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.

[26] Xia R, Meyers BC, Liu Z, et al. MicroRNA superfamilies descended from miR390 and their roles in secondary small interfering RNA biogenesis in eudicots. Plant Cell 2013;25:1555–72.

[27] Kakrana A, Hammond R, Patel P, et al. sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. Nucleic Acids Res 2014;42:e139.

[28] Nakano M, Nobuta K, Vemaraju K, et al. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 2006;34:D731–5.

[29] Kakrana A, Li P, Patel P, et al. *PHASIS*: A computational suite for de novo discovery and characterization of phased, siRNA-generating loci and their miRNA triggers. bioRxiv 2017:158832.

[30] Zhang Y, Waseem M, Zeng Z, et al. MicroRNA482/2118, a miRNA superfamily essential for both disease resistance and plant development. New Phytol 2022;233:2047–57.

[31] Zhang Y, Lei M, Zhou Y, et al. Reproductive phasiRNAs regulate reprogramming of gene expression and meiotic progression in rice. Nat Commun 2020;11:6031.

[32] Xia R, Ye S, Liu Z, et al. Novel and recently evolved MicroRNA clusters regulate expansive *F-BOX* gene networks through phased small interfering RNAs in wild diploid strawberry. Plant Physiol 2015;169:594–610.

[33] Xia R, Chen C, Pokhrel S, et al. 24-nt reproductive phasiRNAs are broadly present in angiosperms. Nat Commun 2019;10:627.

[34] Nakano M, McCormick K, Demirci C, et al. Next-Generation Sequence Databases: RNA and genomic informatics resources for plants. Plant Physiol 2020;182:136–46.

Guanliang Li received his Master's degree from South China Agricultural University in 2023. He mainly focuses on bioinformatics tools development on small RNA and the evolution of small RNA families.

Chengjie Chen, a lecturer at South China Agricultural University, is recognized for his contribution to the development of a plethora of bioinformatics tools and databases, including TBtools, EasycodeML, sRNAanno, and IGV-GSAman.

Rui Xia is a professor at South China Agricultural University. His current research interests focus on flower sex differentiation mechanisms in Sapindaceae plants and flower and fruit development in tropical and subtropical fruits in Southern China, like litchi and longan, using methods such as genomics, bioinformatics, and molecular biology. His group has developed a series of bioinformatics tools and databases, such as TBtools, SapBase, sRNAanno, and IGV-GSAman.