BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 2 2015, pages 284–286
doi:10.1093/bioinformatics/btu628

Data and text mining

Advance Access publication September 21, 2014

PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades

Qingli Guo[1,2], Xiongfei Qu[3] and Weibo Jin[1,2,*]

[1]College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, [2]College of Life Sciences, Northwest A&F University, Yangling, Shaanxi and [3]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: Emerging evidence has revealed phased siRNAs (phasiRNAs) as important endogenous regulators in plants. However, the integrated prediction tools for phasiRNAs are still limited. In this article, we introduce a stand-alone package PhaseTank for systematically characterizing phasiRNAs and their regulatory networks. (i) It can identify phasiRNAs/tasiRNAs functional cascades (miRNA/phasiRNA→*PHAS* loci→phasiRNA→target) with high sensitivity and specificity. (ii) By one command analysis, it generates comprehensive annotation and quantification of the predicted *PHAS* genes from any given sequences. (iii) PhaseTank has no restriction with regards to prior information of sequence homology of unrestricted organism origins.

Availability and implementation: PhaseTank is a free and open-source tool. The package is available at http://phasetank.sourceforge.net/.

Contact: weibojin@gmail.com or guoql.karen@gmail.com.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 8, 2014; revised on September 12, 2014; accepted on September 15, 2014

1 INTRODUCTION

Secondary phased siRNAs (phasiRNAs) play crucial roles in post-transcriptional regulatory networks in plants (Fei *et al.*, 2013). Well-characterized *trans*-acting siRNAs, as a special subgroup of phasiRNAs, are initiated by miRNA-mediated cleavage and converted to dsRNA, yielding siRNAs in a 21-nt phase (Allen and Howell, 2010; Axtell, 2013a; Fei *et al.*, 2013). Currently, eight *TAS* loci have been identified in *Arabidopsis,* and *TAS*3 is experimentally validated to suppress the juvenile-to-adult transition (Hunter *et al.*, 2006). Moreover, phasiRNA-mediated regulation of *NB-LRR* encoding disease-resistance proteins appears to be widespread in dicots (Zhai *et al.*, 2011). However, the significance and broad roles of phasiRNAs remain unclear in plants (Fei *et al.*, 2013).

Next-generation sequencing technology provides a powerful tool for genome-wide screening of phasiRNAs (Wang *et al.*, 2009), and several methods were proposed correspondingly (Axtell, 2010, 2013b; Chen *et al.*, 2007; Gupta *et al.*, 2012; Li *et al.*, 2012). However, no methods could systematically identify phasiRNAs and their regulatory cascades, and none reports the detailed annotation and quantification of candidate *PHAS* loci (Fei *et al.*, 2013).

In this study, we present a novel tool PhaseTank to identify *PHAS* loci, which contains our new sights and also incorporates the advantages of the previous methods. We propose that the relative small RNAs production (RSRP) of *PHAS* loci is higher than that of the transcripts processed by random cleavage, inasmuch as the *PHAS* loci should generate considerable phasiRNAs to maintain their biological functions (Allen *et al.*, 2005). In our method, the phased ratio, abundance and number of a phasiRNA cluster are considered as three main contributors to score a candidate *PHAS* loci. Additionally, PhaseTank could detect both triggered miRNAs and phasiRNA targets for predicted *PHAS* loci. Therefore, PhaseTank employs a new filter and scoring system to perform *de novo* prediction of *PHAS* loci and their regulatory cascades on a genome-wide scale.

2 APPROACH

2.1 Computing RSRP

The RSRP of a sequence is calculated as the following steps. First, we calculate the small RNAs production of sequence *i* (SRP_i) using Equation (1).

$$SRP_i = A_i / L_i \tag{1}$$

where A_i is the abundance of mapped reads onto sequence *i*, and L_i is the length of the sequence *i*. Second, $RSRP_i$ is computed as Equation (2):

$$RSRP_i = \ln\left(\frac{SRP_i}{\frac{1}{N}\sum_{i=1}^{N} SRP_i}\right) \tag{2}$$

where *N* is the total number of the sequences.

2.2 Definition of siRNA and phasiRNA clusters

The siRNA cluster is referred to the genomic region that contains at least four sRNA hits with a maximum separation distance of 100-nt (Supplementary Fig. S1) (Johnson *et al.*, 2009). Similarly, we define phasiRNA cluster as a specific region that contains at least four phased reads with a maximum separation distance of 84-nt. To set the phasiRNA cluster, every position in a given sequence is successively assigned as a '21-bin cycle' (Axtell, 2010) (Supplementary Fig. S2). We then search for the most

[*]To whom correspondence should be addressed.

abundant bin and the 21-nt reads mapped to this bin. These reads are used to search phasiRNA clusters.

2.3 Computing phased score

Phased score is the core indicator in PhaseTank, and it consists of three main factors, namely, phased ratio, number and abundance. Larger phased score of a sequence indicates higher possibility that it will be a true *PHAS* gene. Specifically, in a phasiRNA cluster, phased ratio is the abundance of the highest abundant bin_x divided by the total abundance (Axtell, 2010). Phase drift for 1 ~ 2 position(s) is considered if the second abundant bin is bin_y ($y = x \pm 1$ or 2). Phased number is the distinct number of 21-nt phased reads, and the abundance of these reads represents phased abundance. The phased score for each phasiRNA cluster is calculated as Equation (3):

$$P\_Score = P\_Ratio * P\_Number * \ln\,(P\_Abundance) \qquad (3)$$

where P is the abbreviation of 'phased'.

2.4 Searching triggered miRNAs

MiRNA-directed *TAS* loci cleavage often occurs at 9–11nt positions from the 5′ terminal (Allen and Howell, 2010). According to the prediction results of CleaveLand4 (Addo-Quaye *et al.*, 2009), one miRNA is defined as a phase-trigger, if its cleavage site occurs almost at the phased positions (one position shift).

3 WORKFLOW AND IMPLEMENTATION

The workflow of PhaseTank is illustrated in Supplementary Figure S3. First, the reads are mapped to the references using bowtie (Langmead *et al.*, 2009). Then PhaseTank excises siRNA clusters based on the mapping information and keeps the clusters with top 5% RSRP value. For these clusters, PhaseTank searches the most abundant bin as phased sites. The phasiRNA clusters are excised accordingly from these sites and then processed into the filtering system. The phased score of the filtered clusters are then computed as Equation (3). By setting advanced options, PhaseTank can predict miRNA-mediated cleavage and phasiRNA targets using CleaveLand4 (Addo-Quaye *et al.*, 2009). Finally, the phasiRNA cascades of each predicted *PHAS* gene are reported in a text file.

PhaseTank is written in Perl (5.8 or later versions), and has been tested on Ubuntu 12.04 and Fedora 17. To run it, users should properly install the software, such as bowtie (Langmead *et al.*, 2009), samtools (Li *et al.*, 2009), RNAplex (Tafer and Hofacker, 2008), Math::CDF and R (see http://phasetank.sourceforge.net/).

4 RESULTS

The RSRP values of the *Arabidopsis* transcripts are shown in Supplementary Figure S4. Consequently, the RSRP values of the known 22 *PHAS* loci ranked top 1539 of the 41 391 transcripts. Herein PhaseTank processes clusters with the top 5% RSRP value for further prediction.

We found that 22 *PHAS* genes reported by previous studies (Chen *et al.*, 2007; Howell *et al.*, 2007) were supported by the reads libraries (Supplementary Table S5). PhaseTank was then performed in *Arabidopsis*, with 21 and 18 *PHAS* genes being detected using genome and cDNA data, respectively, including all eight known *TAS* loci (Supplementary Tables S6 and S7). To estimate the sensitivity, PhaseTank was compared with the recent published software ShortStack (Axtell, 2013b) in different tissues using cDNA sequences of *Arabidopsis*. Evidently, the average sensitivity of PhaseTank (77.90%) was significantly higher than that of ShortStack (26.92%) (Fig. 1A and Supplementary Table S8). Meanwhile, the sensitivity of PhaseTank reached 95.45% with the use of genome sequence, whereas that of ShortStack was only 54.54% (Fig. 1B, Table 1, Supplementary Tables S6 and S9). Moreover, we selected 210 other annotated ncRNAs as true negatives to assess the specificity of the two programs (Supplementary Table S5). Among the ncRNAs, 2 and 17 were predicted as *PHAS* genes by PhaseTank and ShortStack, respectively, with corresponding 99.04% and 91.04% specificity (Table 1 and Supplementary Table S5). The comparisons of other characteristics of both tools are listed in Table 1. Moreover, PhaseTank also identified a rarely reported miR393b triggered cascades (Fig. 1C) (Si-Ammour *et al.*, 2011).

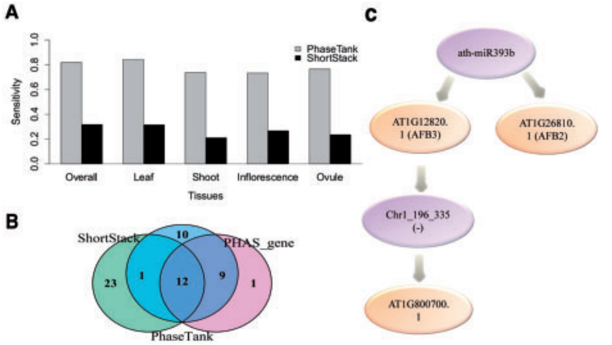

Fig. 1. Comparison of PhaseTank versus ShortStack. (**A**) PhaseTank shows higher sensitivities compared with ShortStack in different *Arabidopsis* tissues using cDNA data; (**B**) PhaseTank detects a greater number of true *PHAS* genes than that predicted by ShortStack using *Arabidopsis* genome data; (**C**) A rarely reported regulatory cascade is discovered by PhaseTank

Table 1. Comparison of phasiRNA prediction programs

Program	Detailed annotation	Phased score	PhasiRNA alignment	Triggered miRNA	Sensitivity	Specificity	Regulatory cascades	Reference
PhaseTank	Yes	Yes	Yes	Yes	95.45%	99.04%	Yes	This paper
ShortStack	No	No	No	No	54.54%	91.90%	No	Axtell (2013b)

We further tested PhaseTank in tomato (*Solanum lycopersicum*) using the genome data as well. Among the 19 reported tomato *PHAS* genes (Shivaprasad *et al.*, 2012), PhaseTank was able to detect 13 of them (Supplementary Table S10). Furthermore, PhaseTank also discovered 10 miRNA-initiated cascades in tomato (Supplementary Fig. S11), and two of which have been reported by Shivaprasad *et al*. (2012).

5 CONCLUSION

In conclusion, our results demonstrate that PhaseTank is an effective and highly applicable tool for comprehensive annotation and quantification of regulatory cascades involved in phasiRNA pathways. Therefore, PhaseTank would help to elucidate the function and evolution of this special class of small RNAs in land plants.

Funding: This work was supported by the grants from Natural Science Foundation of China [No. 31372075 and No. 31000913]; and the Zhejiang Provincial Top Key Discipline of Biology.

Conflict of interest: none declared.

REFERENCES

Addo-Quaye,C. *et al*. (2009) CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics*, **25**, 130–131.

Allen,E. and Howell,M.D. (2010) miRNAs in the biogenesis of trans-acting siRNAs in higher plants. *Semin. Cell Dev. Biol*., **21**, 798–804.

Allen,E. *et al*. (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, **121**, 207–221.

Axtell,M.J. (2010) A method to discover phased siRNA loci. *Methods Mol. Biol*., **592**, 59–70.

Axtell,M.J. (2013a) Classification and comparison of small RNAs from plants. *Ann. Rev. Plant Biol*., **64**, 137–159.

Axtell,M.J. (2013b) ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, **19**, 740–751.

Chen,H.M. *et al*. (2007) Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in *Arabidopsis*. *Proc. Natl Acad. Sci. USA*, **104**, 3318–3323.

Fei,Q. *et al*. (2013) Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell*, **25**, 2400–2415.

Gupta,V. *et al*. (2012) Shortran: a pipeline for small RNA-seq data analysis. *Bioinformatics*, **28**, 2698–2700.

Howell,M.D. *et al*. (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell*, **19**, 926–942.

Hunter,C. *et al*. (2006) Trans-acting siRNA-mediated repression of ETTIN and ARF4 regulates heteroblasty in *Arabidopsis*. *Development*, **133**, 2973–2981.

Johnson,C. *et al*. (2009) Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome Res*., **19**, 1429–1440.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*., **10**, 25.

Li,F. *et al*. (2012) SoMART: a web server for plant miRNA, tasiRNA and target gene analysis. *Plant J*., **70**, 891–901.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Shivaprasad,P.V. *et al*. (2012) A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. *Plant Cell*, **24**, 859–874.

Si-Ammour,A. *et al*. (2011) miR393 and secondary siRNAs regulate expression of the TIR1/AFB2 auxin receptor clade and auxin-related development of Arabidopsis leaves. *Plant Physiol*., **157**, 683–691.

Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev*., **10**, 57–63.

Zhai,J. *et al*. (2011) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev*., **25**, 2540–2553.