User Manual for PhaseTank

Qingli Guo, Northwest A&F University, guoql.karen@gmail.com

Version 1.0

July 31, 2014

1 Introduction

Phased small interfering RNAs (phasiRNAs) are a special subgroup of 21-nt endogenous siRNAs that are involved in regulating genes (e.g. ARF, NB_LRR and MYB) in higher plants (Fei, et al., 2013; Howell, et al., 2007; Zhai, et al., 2011). Even though massive efforts have been made for understanding of their biogenesis pathways, it lacks systematically computational tool for identifying phasiRNAs and their regulatory networks (Eckardt, 2013). Here we first present a standalone software package PhaseTank for characterizing phasiRNAs regulatory cascades from a given organism.

This user manual gives detailed description of the usage of PhaseTank (version1.0) to discover phasiRNAs yielding loci, annotate and quantify these loci, search the miRNAs-directed cleavage, predict targets of phasiRNAs and report phasiRNAs regulatory cascades. Taking data from *Arabidopsis* as example, we show how to use PhaseTank step by step in this manual.

2. License

3. System Requirements and Software Dependencies

PhaseTank is written in Perl (5.8 or the later versions). All the scripts have been tested on two Linux platforms, Ubuntu 12.04 and Fedora 17. Before running it, make sure that the followings are properly installed into your PATH.

1) PhaseTank_v1.0.pl, CleaveLand4_modfied.pl and GSTAr.pl

   Download the tutorial directory named ***PhaseTank_Tutorial.tar.gz*** from https://sourceforge.net/projects/phasetank/files/ and unpack it. Add the three scripts (PhaseTank_v1.0.pl, CleaveLand4_modfied.pl & GSTAr.pl) into your PATH. Note that CleaveLand4_modified.pl is slightly modified from the core program in CleaveLand4 which is a miRNA target prediction software (Addo-Quaye, et al., 2009), and can be downloaded from http://www.bio.psu.edu/people/faculty/Axtell/AxtellLab/Software.html.

2) Bowtie

   Bowtie 0.12.x or 1.x is suitable here (Langmead, et al., 2009), which is available at http://bowtie-bio.sourceforge.net/index.shtml. Do NOT use 'bowtie2'.

(The followings software is needed when predicting targets by CleaveLand4.)

3) Math::CDF. See http://www.cpan.org/modules/INSTALL.html.
4) Samtools (Li, et al., 2009). See http://samtools.sourceforge.net/.index.shtml.
5) RNAplex (Tafer and Hofacker, 2008). See http://www.tbi.univie.ac.at/RNA/index.html.
6) R. See http://www.r-project.org/.

4. Overview of PhaseTank Package

You can find the following files in directory ***PhaseTank_Tutorial.tar.gz***:

1) *ath_genome_TAIR10.21.fa*: the FASTA format genome sequence of *Arabidopsis thaliana* (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/arabidopsis_thaliana/dna/).

2) Reads libraries: six RNA-seq libraries in FASTA format from *Arabidopsis* are

randomly picked on the website, including *GSM1174496.fa, GSM277608.fa, GSM342999.fa* and *GSM709567.fa*, (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1174496), *MTSRNA1.fa* and *RMMT10.fa*(http://mpss.udel.edu/at_sRNA/).

3) *ath_ncRNA.fa*: it is combined of two files. One is the *Arabidopsis* ncRNA sequence (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/arabidopsis_thaliana/ncrna/). The other is tRNA sequence extracted from cDNAs (ftp://ftp.ensemblgenomes.org/pub/release-21/plants/gtf/arabidopsis_thaliana) based on the annotation in GFF file of Arabidopsis (ftp://ftp.ensemblgenomes.org/pub/release-21/plants/gtf/arabidopsis_thaliana).

4) *ath_miRNA.fa*: the present microRNAs (Release 21) in *Arabidopsis thaliana* are downloaded from http://www.mirbase.org/.

5) *de_GSM278335.fa*: the degradome reads in FASTA format from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM278335.

6) ath_cDNA_TAIR10.fa: the FASTA format of cDNA sequences of Arabidopsis (ftp://ftp.ensemblgenomes.org/pub/release-21/plants/fasta/arabidopsis_thaliana/cdna/).

7) The predicted results are put in ***OUTPUT_2014.07.30_07.48*** (note, ‘2014.07.30_07.48’ is the time that run the program), which includes:

a) Run_log_2014.07.30_07.48: the STDERR output of PhaseTank.

b) *Pred_tab_2014.07.30_07.48*: candidate *PHAS* loci predicted by PhaseTank.

c) *Align_2014.07.30_07.48*: the text-based alignments of phasiRNAs to phased RNA clusters and the detail information of each bin for these regions.

d) *PhasiRNAs_2014.07.30_07.48*: the phasiRNAs produced by the predicted PHAS loci with abundance more than 100.

e) *PhasiRNAs_targets_2014.07.30_07.48*: the phasiRNAs targets predicted with the help of CleaveLand4.

f) miRNA_target_2014.07.30_07.48: the miRNAs targets predicting results.

g) *Cascades_2014.07.30_07.48*: the regulatory cascades for each predicted *PHAS* loci.

h) Excised_cluster*_2014.07.30_07.48.fa*: the excised cluster from the genome. Note that if the input reference is cDNA file, there is no such file in OUTPUT directory.

Note that PhaseTank also produces several types of files in the directory in which it is invoked. Some of these files are intermediate and internal files using PhaseTank and will be removed in the end.

At the beginning of running PhaseTank, all the input files will be checked one by one. If the files are not well-formed, PhaseTank would report error messages and abort. In addition, PhaseTank only takes the UNIX format of files as input data. If not, the checking system will report the error message.

5. Detecting phasiRNAs

5.1 Running PhaseTank

Usage: $ perl PhaseTank.pl --genome <*genome_file.fa*> --lib <*read_file_list*> [options]

Or $ perl PhaseTank.pl --cdna <*cdna_file.fa*> --lib <*read_file_list*> [options]

The followings are the detailed descriptions of the arguments and options in the use of PhaseTank:

Arguments:

--genome <string>. Supply PhaseTank with genome sequence in FASTA format as reference sequences. Or --cdna <string>. Also could supply PhaseTank with cdna sequence in FASTA format as reference sequences.

--lib <string>. Supply PhaseTank with a comma-separated list of file(s) containing reads in FASTA format.

Options:

--filter <string>. Supply PhaseTank with FASTA format of other ncRNA sequences. It can help PhaseTank to exclude the reads mapped to other ncRNAs (e.g. tRNA, rRNA, snoRNA).

--miR <string>. Supply PhaseTank with a list of miRNAs in FASTA format for miRNA-directed *PHAS* gene cleavage detection. This option will be ignored without '--trigger_miRNA'.

--degradome <string>. Supply PhaseTank with a set of degradome sequencing reads in FASTA format for phasiRNA targets prediction.

--target <string>. Supply PhaseTank with a FASTA format file containing the interested genes, among which to search the phasiRNA targets.

--trigger_miRNA. Tell PhaseTank to detect miRNA-directed TAS cleavage. It is inactive by default.

--phasiRNA_target. Tell PhaseTank to predict the phasiRNA targeting genes. It is inactive by default.

--ratio <float>. Set phased ratio cutoff value. The default is 0.3.

--number <int>. Set phased number cutoff value. The default is 4.

--abun <int>. The total abundance of phased reads in the phasiRNA cluster. Default is 100. Note, the default normalization level is per twenty millions (20,000,000, can be changed by '--nor <int>'), thus the default abundance value of 100 here is equal to setting 5 of RPM (reads per million).

--READ_abun <int>. The minimum reads abundance to keep for PhaseTank prediction. Default is 1, which means if one read abundance is less than 1, it will be abandoned.

--phasiRNA_abun <int>. Minimum read abundance of phasiRNAs for target prediction. This option will be ignored without '--phasiRNA_target'.

--drift <int>. Maximum phased drift. The default is 2.

--size <int>. Length of phased reads. The default is 21.

--nor <int>. Tell PhaseTank the normalization level for the libraries. Default is 20,000,000.

--island <int>. That is the maximum separation distance of two phasiRNAs in each cluster. The default is 84.

--extendLEN <int>. The length on each side of siRNA cluster (or phasiRNA cluster) will be excised from the reference sequence. The default is 80.

-- max_hits <int>. Tell PhaseTank the '-m' cutoff while using Bowtie ('-m' represent the maximum mapped hits to the reference, if goes out the value, the reads will be filtered out). The default is 5 here. Note that with this parameter changed, the prediction results may fluctuate slightly in big and small dataset due to a few reads may be removed in the big dataset.

--per <float>. Within 0.01-1.00. The top percentage of RSRP value of sequences was put to the later program. The default is 0.05 (5%).

--rsrp <float>. The RSRP value for PhaseTank to filter the sequences. Default is 1.

--CALL_RSRP. Tell PhaseTank to estimate RSRP cutoff from the given reads libraries, which is set from the top 5% (default, can be changed by '--per <float>') of RSRP value of sequences for the later processes. It is inactive by default. You could active this module by '--CALL_RSRP' when you analyze other organisms (should use whole cDNA as input references). Or you also can use the default value instead.

--dir <string>. Set the directory in which PhaseTank will write its output files. The default is 'OUTPUT_run_time/'.

--help. Print the help message and quit.

--version. Print PhaseTank version number and quit.

5.2 Analysis Demonstration for Different Modes

Here we provide four normal analysis modules using PhaseTank. In all of them, if you need to exclude some other ncRNAs you can input the FASTA format of the sequences by '--filter <*ncRNA.fa*>' (for example: '--filter *ath_ncRNA.fa*' in our datasets) to the following commands.

5.2.1 Predict *PHAS* loci from the given organism and read libraries

Irrelevant options: --miR, --degradome, --trigger_miRNA, --phasiRNA_target, --target, --phasiRNA_abun

Example:

```
$ perl PhaseTank_v1.pl --genome ath_genome_TAIR10.fa --lib GSM1174496.fa,GSM277608.fa,GSM342999.fa,GSM709567.fa,MTSRNA1.fa,RMMT10.fa
```

5.2.2 Predict phasiRNAs and search their miRNA-triggered cleavage

Required options: --miR, --degradome, --trigger_miRNA

Irrelevant options: --target, --phasiRNA_target, --phasiRNA_abun

Example:

```
$ perl PhaseTank_v1.pl --genome ath_genome_TAIR10.fa --lib GSM1174496.fa,GSM277608.fa,GSM342999.fa,GSM709567.fa,MTSRNA1.fa,RMMT10.fa --miR ath_miRNA.fa --degradome de_GSM278335.fa –trigger_miRNA
```

5.2.3 Predict phasiRNAs and their targets

Required options: --degradome, --target, --phasiRNA_target

Irrelevant options: --miR, --trigger_miRNA

Example:

```
$ perl PhaseTank_v1.pl --genome ath_genome_TAIR10.fa --lib GSM1174496.fa,GSM277608.fa,GSM342999.fa,GSM709567.fa,MTSRNA1.fa,RMMT10.fa --degradome de_GSM278335.fa --target ath_cDNA_TAIR10.fa --phasiRNA_target
```

5.2.4 Predict phasiRNAs, search the miRNA-triggered cleavage and detect phasiRNAs targets

Required options: --miR, --degradome, --target, --trigger_miRNA, --phasiRNA_target

Example:

```
$ perl PhaseTank_v1.pl --genome ath_genome_TAIR10.fa --lib GSM1174496.fa,GSM277608.fa,GSM342999.fa,GSM709567.fa,MTSRNA1.fa,RMMT10.fa --miR miRNA.fa --degradome de_GSM278335.fa --target ath_cDNA_TAIR10.fa --phasiRNA_target
```

6 Conventions and Recommendations

i. In this manual, all the file names are in italic and the directory names are in bold and italic. Besides, the command lines are listed in the grey backgrounds which start with $. Make sure your files are in **UNIX** format.

ii. In our method, the relative small RNA production (RSRP) for a sequence is calculated as the following steps. First, we calculate the small RNAs production of sequence i (SRP$_i$) using equation (1).

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

where A$_i$ is the abundance of mapped reads onto sequence i, L$_i$ is the length of sequence i;

Second, RSRP$_I$ was calculated as the 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.

Therefore, the default value of RSRP here is set from the given data of *Arabidopsis*, which may be fluctuated in different datasets. However, the default value is recommended to use in your analysis. Because we have analyzed several libraries, the RSRP value will actually fluctuated for different datasets, but it is quite slight. If you still want to estimate RSRP value in your dataset, you need to use whole cDNA sequences as your reference sequences and also should add option '--CALL_RSRP' in your command line.

For example:

```
$ perl PhaseTank_v1.pl --cdna ath_cdna_TAIR10.fa --lib GSM1174496.fa,GSM277608.fa,GSM342999.fa,GSM709567.fa,MTSRNA1.fa,RMMT10.fa –CALL_RSRP
```

iii. If there is a file containing other annotated ncRNAs in your aimed species, you can use this file to filter out the annotated ncRNAs in PhaseTank.

iv. In PhaseTank, the reference could be the genome sequences or any FASTA format of sequences (such as cDNA, EST, or your interested genes). According to the prediction results of our test, the genome sequences contained the richest information for PHAS genes detection. While if there is no complete genome assembly, other sequences data could also be used to predict PHAS loci with good sensitivity and specificity.

v. The running time for PhaseTank mainly depends on the target prediction by CleaveLand4. It takes about 3-4 hours for analysis in 5.2.4 using the listed files and with the default settings. Option like '--phasiRNA_abun' will largely influence the prediction time, because it directly decides the number of phasiRNAs which will be put into targets prediction pipelines.

vi. We used modified CleaveLand4 in our pipeline for searching trigger miRNA and phasiRNA targets. The CleaveLand4 is just modified to remove the screen output and some unimportant files, while the other parts and the core output file remain unchanged. Thus the prediction results will be clear.

REFERENCES

Addo-Quaye, C., Miller, W. and Axtell, M.J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 2009;25(1):130-131.

Eckardt, N.A. The plant cell reviews aspects of microRNA and PhasiRNA regulatory function. *The Plant cell* 2013;25(7):2382.

Fei, Q., Xia, R. and Meyers, B.C. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant cell* 2013;25(7):2400-2415.

Howell, M.D*., et al.* Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *The Plant cell* 2007;19(3):926-942.

Langmead, B*., et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 2009;10(3):R25.

Li, H*., et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

Tafer, H. and Hofacker, I.L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* 2008;24(22):2657-2663.

Zhai, J*., et al.* MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes & development* 2011;25(23):2540-2553.

Addo-Quaye, C., Miller, W. and Axtell, M.J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 2009;25(1):130-131.

Eckardt, N.A. The plant cell reviews aspects of microRNA and PhasiRNA regulatory function. *The Plant cell* 2013;25(7):2382.

Fei, Q., Xia, R. and Meyers, B.C. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant cell* 2013;25(7):2400-2415.

Howell, M.D.*, et al.* Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *The Plant cell* 2007;19(3):926-942.

Langmead, B.*, et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 2009;10(3):R25.

Li, H.*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079.

Zhai, J.*, et al.* MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes & development* 2011;25(23):2540-2553.

Addo-Quaye, C., Miller, W. and Axtell, M.J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 2009;25(1):130-131.

Eckardt, N.A. The plant cell reviews aspects of microRNA and PhasiRNA regulatory function. *The Plant cell* 2013;25(7):2382.

Fei, Q., Xia, R. and Meyers, B.C. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant cell* 2013;25(7):2400-2415.

Howell, M.D.*, et al.* Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *The Plant cell* 2007;19(3):926-942.

Langmead, B.*, et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 2009;10(3):R25.

Zhai, J.*, et al.* MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes & development* 2011;25(23):2540-2553.

Addo-Quaye, C., Miller, W. and Axtell, M.J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* 2009;25(1):130-131.

Eckardt, N.A. The plant cell reviews aspects of microRNA and PhasiRNA regulatory function. *The Plant cell* 2013;25(7):2382.

Fei, Q., Xia, R. and Meyers, B.C. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant cell* 2013;25(7):2400-2415.

Howell, M.D.*, et al.* Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *The Plant cell* 2007;19(3):926-942.

Zhai, J.*, et al.* MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes & development* 2011;25(23):2540-2553.

Eckardt, N.A. (2013) The plant cell reviews aspects of microRNA and PhasiRNA regulatory function, *The Plant cell*, **25**, 2382.

Fei, Q., Xia, R. and Meyers, B.C. (2013) Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks, *The Plant cell*, **25**, 2400-2415.

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, *The Plant cell*, **19**, 926-942.

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 & development*, **25**, 2540-2553.

Eckardt, N.A. (2013) The plant cell reviews aspects of microRNA and PhasiRNA regulatory function, *The Plant cell*, **25**, 2382.