PhasiRNA Prediction Pipeline (PPP) User Manual

Pingchuan Li

li@dbi.udel.edu

Blake Meyers Lab, University of Delaware

Newark, DE, 19716

**Legal**: This program is free software; you can redistribute it and/or modify it without the permission of authors. This program can be distributed in the hope that it will be useful for the sRNA research community.

To acknowledge PPP in your work, please cite:

Pingchuan Li et al., (2013) *in prep.*

**Alignment Subjects Support**:

Genome data

EST/cDNA/transcriptome (developing)

**Overview**:

Xxx

Xxx

**Versions:**

# 2013-05-30 normalized the phased siRNA abundance

# 2013-05-31 adding the 231bp sliding window to calculate the *p*-value, which give all the phased siRNA located in the same cluster pretty similar score, which benefit the boundary finding of cluster

# 2013-05-31 adding the hits number for each of reads in the output of cluster

# 2013-06-01 v.22 adding the sliding windows calculate support for scoring a certain sRNA

# 2013-06-01 v.23 adding the filter for the interested phased siRNA abundance -n (noise)

# 2013-06-02 v.24 adding more levels for the cutoff

# 2013-06-05 v.25 adding the filter for the average of hits for phased siRNA

# 2013-06-10 v.26 adding 2 filter in the command line, t and ht

# 2013-06-14 v.27 enforce the k value to 21 when it's bigger than 21 otherwise the script will get error info.

# 2013-06-24 v.28 add a qualification filter for the cluster in terms of the sRNA amount, default >=8

# 2013-06-28 v.29 finished the whole thing

# 2013-06-29 v.291 bug fixes; increase the minimal requirement of small RNA amount in a certain cluster to 8.

**Motivation**: a) In order to maximally identify the phasiRNA cluster (PHAS) within a species, b) In order to more accurately identify the boundaries of the phasiRNA clusters c) to identify more phasiRNA cluster guided by other microRNA besides miR2118/miR2275 etc. d) to improve the efficiency of phasiRNA prediction, including the sensitivity by employing the PARE/degradome data. e) for those species without the whole genome sequences but tons of EST or cDNA sequence.

**Dependencies**:

1. Perl 5.12, one of the built-in functions ‘each’ will be available for the Array data extraction overn the version 5.12. In addition, the Perl should be compiled to support the multiple threads. Query this availability by Perl –v on the command line of Linux.

|  |
| --- |
| [pingchuan@Taiji]$ perl -v  This is perl 5, version 18, subversion 0 (**v5.18.0**) built for x86\_64-linux-**thread-multi** |

1. With the following CPAN modules in its search environment.
   1. Scalar::Util
   2. Data::Dumper
   3. Parallel::ForkManager
   4. Getopt::Long
   5. BerkleyDB, optional for machine with low memory, this will drop the efficiency of script but get it through the RAM torture. (developing)

Mac OS user please refers to the following page on how to fix the error of CPAN url by install the Xcode command line tools. (http://www.virtuin.com/2012/10/fix-perl-cpan-on-os-x-mountain-lion.html)

1. Bowtie package, a short sequence alignment program developed by Ben Langmead and Cole Trapnell (<http://bowtie-bio.sourceforge.net>).
   1. Version 0.12.8 will be okay, we have not tested the higher version of Bowtie, but if the output format by default is same with the version 0.12.8, we don’t think it would be any problem.
   2. Current the script only supports the Bowtie given that this software is very popular and stable for its output
   3. Bowtie-build, used to make the indexing of the subject sequences.

**Workflow in details**:

Data prepares

1. sRNA preparation: All the sRNA data should be trimmed the adapters in both the 5’ and 3’ ends. The file format supports the tagcount and fasta with only distinct reads.
   * The fasta file should be looks like below:

>seq\_id|amount

xxxxxxxxxx

for example:

>seq\_1666|123

atcgatcgatcgatcg

Here the 1666 is the 1666the distinct reads in the fasta file and 123 is its total sequenced times by HiSeq, 454 or similar NGS sequencer. Though the hypergeometric distribution won’t depend on the small RNA abundance, but we will use it to generate the summary table and filter out the trimmed reads with low abundance.

* + The tagcount file is a file with minimal meta information like below:

atcgatcgatcgatcg 123

ctcgatcgatcgatc 125

ttcgatcgatcgatcgg 124

The sequence and abundance should be split by the tab.

1. (**optional**) PARE/degradome preparation: All the PARE/degradome will only support the tagcount format, with adapters removed.
2. Genome preparation:

We attempt to develop two versions of PPP, one is specific developed for the genome scan, another one is for the EST, cDNA or multiple scaffolds (>=10,000).

* + Genome preparation: “***Life was like a box of chocolates. You never know what you are gonna get***”. Thus please follow our rules to simplify the analysis and enjoy the results. The recommended Fasta header will be appreciated with short length, no whitespace and unusual characters, in additional, the **most important thing** is to make it ended with unique digital number like : >chr1 , >chr\_1 or >chr01. Please DON’T add other additional information following the header. We never tested the other format, so please keep your chromosome header consistent with our suggestions.
  + EST/cDNA/transcriptome: the rule sharing for header with the Genome Preparation on common: shorter is better. In additional, uniqueness. For example >AT1G44020, >LOC\_OS12G16350 etc. are the typical header format. No whitespace locates in the headers.

1. Subjects indexing:
   * The Genome and transcriptome can be easily created the indexing file by following the Bowtie’s user manual by default: here is what we performated for the genome:

|  |
| --- |
| [pingchuan@Taiji]$ bowtie-build -f RICE\_MSU7\_genome.fasta RICE\_MSU7\_genome |

Your formatted genome will be looks like below:

|  |
| --- |
| -rw-r--r-- 1 pingchuan pingchuan 107M Feb 17 2012 RICE\_MSU7\_genome.1.ebwt  -rw-r--r-- 1 pingchuan pingchuan 45M Feb 17 2012 RICE\_MSU7\_genome.2.ebwt  -rw-r--r-- 1 pingchuan pingchuan 8.7K Feb 17 2012 RICE\_MSU7\_genome.3.ebwt  -rw-r--r-- 1 pingchuan pingchuan 90M Feb 17 2012 RICE\_MSU7\_genome.4.ebwt  -rw-r--r-- 1 pingchuan pingchuan 107M Feb 17 2012 RICE\_MSU7\_genome.rev.1.ebwt  -rw-r--r-- 1 pingchuan pingchuan 45M Feb 17 2012 RICE\_MSU7\_genome.rev.2.ebwt |

**Please note**: we have no condition to test the SOLiD data, which is a typical basecalling on the color space. Thus the above indexing way might not be able to adapt to the SOLiD sRNA alignment.

**Usage**:

|  |
| --- |
| Scripts: PhasiRNA Prediction Pipeline (PPP). The sRNA readswere mapped  by bowtie by default.  Version: 0.292, Programmed by Pingchuan Li @ Blake Meyers Lab 06/2013  Arguments:  Mandatory:  -i sRNA input file name  -f two files formats are supported,[t] for tagcount, [f] for fasta.  -d indexed genome by bowtie, indicating the prefix of the genome  -rl register len(rl), such as 21 or 24, separated by comma like: 21, 24 or  21,22,23,24 etc  -px prefix of all the output file, in order to distinguish for each  other  optional:  -q degradome or PARE data in ONLY tagcount format  -m mismatches for both the sRNA and PARE/degradome alignment by bowtie,  default = 0  -p p-value in decimal number, defult = 0.005  -cpu cpu numbers for the bowtie alignment  -n noise, default = 1, for those which have abundance less equal than 1,  properly increase noise value for union dataset  -g gap between two separate cluster, 300bp by default  -t minimal proportation of the interested register small RNA abundance,  default = 50  -ht the maximal average of hits for the small RNA of a certain cluster,  defaut = 10 |

The current prediction pipeline supports two options as the inputs

1. sRNA (tagcount)

|  |
| --- |
| **phasing\_cluster\_analysis\_PNAS\_version.v2.92.pl -i 9311f\_tag\_count.txt -f t -d ~/database/rice/9311/9311 -rl 21 -p 5e-3 -cpu 12 -px demo** |

1. sRNA (tagcount )+ PARE/degradome (tagcount)

|  |
| --- |
| **phasing\_cluster\_analysis\_PNAS\_version.v2.92.pl -i 9311f\_tag\_count.txt -f t -d ~/database/rice/9311/9311 -q 9311f.degradome\_chopped.txt -rl 21 -p 5e-3 -cpu 12 -px demo** |

To make sure that the script located in the path where the Linux can find it, and permission of the script has been changed to **755** by **chmod**

**Output description**s:

Xxx

Xxx