2017

Sagnik Banerjee

Iowa State University

1/8/2017

Small RNA



Contents

[Chapter 1 Introduction 4](#_Toc482160519)

[Chapter 2 Documentation 5](#_Toc482160520)

[2.1 Introduction to <> 5](#_Toc482160521)

[2.2 Data 5](#_Toc482160522)

[2.3 Deliverables 5](#_Toc482160523)

[2.4 Methods implemented 5](#_Toc482160524)

[2.5 Flowcharts 5](#_Toc482160525)

[2.6 Results 5](#_Toc482160526)

[Chapter 3 Timeline 6](#_Toc482160527)

[3.1 6](#_Toc482160528)

[3.1.1 Text 6](#_Toc482160529)

[3.2 Text 6](#_Toc482160530)

[Chapter 4 Binarization 7](#_Toc482160531)

[4.1 7](#_Toc482160532)

[4.1.1 Text 7](#_Toc482160533)

[4.2 Text 7](#_Toc482160534)

[Chapter 5 Recognition 8](#_Toc482160535)

[5.1 8](#_Toc482160536)

[5.1.1 Text 8](#_Toc482160537)

[5.2 Aaa 8](#_Toc482160538)

[5.3 aaa 9](#_Toc482160539)

[Chapter 6 Indexing and Retrieval 10](#_Toc482160540)

[6.1 10](#_Toc482160541)

[6.1.1 Text 10](#_Toc482160542)

[6.2 Text 10](#_Toc482160543)

[Chapter 7 A. Test 10](#_Toc482160544)

List of Figures

[Figure 5.1 Sample image 9](#_Toc356647999)

List of Tables

[Table 1.1 Sample table 4](#_Toc482158032)

[Table 2.1 abcd 5](#_Toc482158033)

Chapter 1

# Introduction

This document serves to provide required information about the progress of the project. This document is organized into chapters. The first chapter basically outlines the structure of this document. The second chapter will contain the entire documentation of the project. I have decided to restrict documentation to a single chapter at least for the time being. This may change in the future or according to the needs of the project. Sections and subsections in the second chapter will cater to most of the required information of the project. A basic outline of what is considered important has been provided in this template. One can add more to the list and alter it as it seems fit. The third chapter is a timeline chapter which will store daily activities related to the project. This will be organised in the form of a diary which will outline the work done each day. This chapter will also contain the Minutes of meeting. The subsequent chapters will contain information related to the project topic. This could be some biological information or related paper review. The main objective of this document is to keep everything written. Later this document can be used to prepare manuscripts for submission to journal. This will serve as a very nice document to prepare the supplementary document. In the second chapter all information related to methods need to be provided. Please note that this section is designed to contain the most updated method which is implement. So it has to be altered as and when required.

Chapter 2

# Documentation

## Introduction to phasiRNAs

Plant genomes encode abundant but diverse populations of small non-coding RNAs, which can be broadly divided into microRNAs (miRNAs) and endogenous small interfering RNAs (siRNAs). Endogenous siRNAs can be further grouped into several sub-classes such as heterochromatic small interfering RNAs (siRNAs), natural antisense siRNAs (nat-siRNAs) and trans-acting siRNAs (ta-siRNAs). The role of miRNAs as post-transcriptional regulators is well known. Among siRNAs, tasiRNAs and natsiRNAs are known to act as guide molecules for post-transcriptional gene regulation, and heterochromatic siRNAs in transcriptional gene silencing, but the role of phasiRNAs in gene regulation is still unclear.

The biogenesis of miRNAs consists of several steps. In plants, the primary miRNA transcripts, normally forming hairpin structures, are transcribed by RNA polymerase II (Pol II). The hairpin-like structures are then processed by DICER-LIKE 1 (DCL1) to produce miRNA:miRNA\* duplex. Mature miRNAs will be loaded into an RNA-induced silencing complex (RISC), normally with an Argonaute (AGO) protein. Plant miRNAs guide the RISC to their targets on the basis of perfect or nearly perfect complementarities, which normally induce cleavages of their targets at the centers of the complementarities.

miRNAs, which are generated by cleavage of single-stranded RNA stem–loop structures by DICER-Like 1, play pivotal roles in development and stress responses by targeting the transcripts of transcription factors that mediate the transition between various developmental stages or regulate responses to biotic or abiotic stimulation. Heterochromatic siRNAs trigger DNA methylation and histone modification and are involved in heterochromatin formation of repeated genomic regions and the silencing of other genomic regions, including the control of transposon movement. nat-siRNAs consist of two types of siRNAs, i.e. 24-nt nat-siRNAs and 21-nt nat-siRNAs. They are processed from pairs of natural cis-antisense transcripts by cleavage with DICER-Like 2 and DICER-Like 1, respectively. The 24-nt nat-siRNAs play a key role in environmental stress response and in reproduction by targeting unlinked transcripts and guiding cleavage to produce 21-nt nat-siRNAs whose function is still not clear. [2]

Similar to miRNAs, trans-acting small interfering RNAs (tasiRNAs) are a class of siRNAs that represses their target transcripts at post-transcriptional level. The primary transcripts of tasiRNAs are used to generate double strand RNAs (dsRNAs) by RDR6 (RNA-dependent RNA polymerase 6). The dsRNAs are then cleaved by DCL4 to form phased 21 nt segments or by DCL5 to form 24 nt phased segments. The precise phasing of tasiRNAs is guided by miRNAs through either two or one miRNA binding site. Four families of tasiRNA generating loci, named TAS1 to TAS4, have been identified in Arabidopsis thaliana. Among these 4 TAS genes, TAS3 is a well conserved gene. In addition to these four typical non-coding genes, accumulating evidences suggest that coding genes, especially PPR NB-LRR disease resistance proteins, MYB transcription factors, also generate phased siRNAs. These phased siRNAs do not necessarily function in trans, thus are named as phasiRNAs, and their corresponding generating loci are called as PHAS genes by Zhai et al. Similar to TAS, PHAS from coding genes are also targeted by miRNAs, such as miR161 targeting PPR; miR428 and miR2118 targeting NB-LRR and miR828 targeting MYB transcripts. [3]

In plants, microRNAs (miRNAs) comprise a class of predominantly 20–22 nt small RNAs (sRNAs), which play critical roles in gene regulation in plant development and stress responses. miRNA biogenesis relies on the formation of a stem-loop structure of a precursor mRNA that is subsequently cleaved by DICER-LIKE 1 (DCL1). This results in the release of a miRNA/miRNA\* duplex with a 3′ 2-nt overhang. The functional strand of the duplex (the miRNA) is loaded into an Argonaute (AGO) protein, the key component of the RNA-induced silencing complex, directing interactions with target genes and resulting in either mRNA cleavage or translation inhibition. The cleavage of target genes by some miRNAs can trigger the biogenesis of another class of sRNAs; an RNA-dependent RNA polymerase is recruited to convert an upstream or downstream cleaved fragment into double-strand RNA, subsequently chopped by DICER-LIKE 4 into 21-nt sRNAs. These sRNAs are “in phase” with the miRNA cleavage site, hence named phased, secondary small interfering RNAs (phasiRNAs), some of which function in *trans* (*trans*-acting siRNAs, tasiRNA) or in *cis*. The tasiRNAs from the *(TRANS-ACTING SIRNA GENE 3)* transcripts, targeting AUXIN RESPONSE FACTOR (ARF)-encoding mRNAs and triggered by miR390, are among the well-characterized tasiRNAs. *TAS3* is also perhaps the earliest-evolved tasiRNA known in plants, a pathway that expanded in flowering plants to include many phasiRNA-generating loci.

PhasiRNAs are produced from both protein-coding and noncoding genes. In many eudicots, three large gene families generate the majority of phasiRNAs, including those encoding nucleotide binding leucine-rich repeat proteins (*NB-LRR* genes), pentatricopeptide repeat proteins (*PPR* genes), and *MYB* transcription factors (*MYB* genes). Monocot genomes have thus far yielded a contrasting picture in which phasiRNAs are produced predominantly from noncoding transcripts. In rice, >1,000 noncoding loci distributed in all the chromosomes give rise to the production of appreciable 21-nt phasiRNAs. In addition, monocots produce a class of 24-nt phasiRNAs, apparently absent in eudicots.

## Formula used

Here we discuss about the two formulas used.

### Algorithm for computation of p-value:

**Step 1:** Map sRNAs to reference genome using Bowtie1, since it is faster than Bowtie2 or HiSAT for small sequences like sRNAs. Perform exact match and remember to get rid of all information about those reads which do not map to any coordinate. This needs to be done to keep the alignment file size in check

**Step 2:** Add a two-nucleotide positive offset to all the sRNAs mapped on to the anti-sense strand to account for the 3’ overhang of tasiRNAs. Combine two reads which appear on both strands in the same location.

**Step 3:** Now we define some terms:

*L*: length of the sRNA you are interested in. This value is typically between 20 and 22.

*Phase Register*: A region of the genome where an sRNA could cleave. In some papers, this same thing is called a cycle, a phase cycle or a register. Typically, a phase register is as long as ‘L’.

*Window*: A sequence of the genome which will be probed for presence of sRNAs. This has a typical length of 9L, 10L or 11L. Length of a window (mL) is often represented as a multiple of the ‘m’ – the number of phase registers it contains.

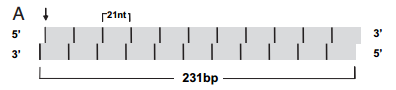


Figure 2.1 Diagram to explain sRNA p-value computation

The above figure depicts a window which in this case is 231bp long. Addition of 2 nt to coordinates of the antisense strand will align the lower strand end-to-end with the strand above it. Here L=21 and m=11.

*Positive Windows*: Windows that abide by the 3 following rules are called positive windows:

* Contains at least 10 unique reads
* More than half of the reads should be ‘L’ nt long
* At least three unique reads falling into the phase registers (Not sure how important this point is – may choose to ignore it)

*Phased and non-phased locations*: The vertical arrow indicates the start site for the small RNA used to determine the phased and non-phased positions. 21 phased sites relative to the start site are indicated as black vertical bars. Four hundred forty non-phased sites relative to the start site are indicated as grey. In this paper [4], they have considered the two strands separately which is why there are more phased sites. In our case there will be ‘m’ phased sites in a window.

‘*n’:* Total number of possible locations where phasing can occur. Hence, in our case, a window can have maximum ‘m’ such positions.

*‘k’*: Number of phased locations in the window which is covered by at least one sRNA.

*phased sRNA:* reads which have mapped onto one of the phased locations.

**Step 4**: Calculate the Phasing score

Phasing score will be computed for each location (loc) of the genome using the following formula.

*‘k’*: Number of phased locations in the window which is covered by at least one sRNA. Or this calculation we will consider k>=3.

*Pi*: Number of phased reads at the ith phase from the position *loc*

*Ui*: total number of reads for all small RNAs with start coordinates out of the *ith* phase

## Point of Contact(s) for this project

Dr. Matthew Hunt

Dr. Roger Wise

## Data

90-sample data from Big Seq experiment

## Workflow

### Agenda 1: Generate a distribution for read length for each of the 90 samples.

### Agenda 2: Create bowtie indices for mapping of small RNA reads.

The genome used for this project was downloaded from [EnsemblPlants](http://plants.ensembl.org/Hordeum_vulgare/Info/Index). To find more information about this version please visit the website. I decided to use this genome instead of the one we have on LSS since this has better annotation than in version 1.

nohup bowtie-build Hordeum\_vulgare.Hv\_IBSC\_PGSB\_v2.dna.toplevel.fa bowtie1\_index &

### Agenda 3: Perform mapping of the reads.

Map the reads back to the genome index.

nohup bowtie -v 0 -S -a -p 15 /home/bigdata/sagnik/data/barley/genome/bowtie1\_index /home/bigdata/sagnik/small\_rna/data/quality\_trimmed\_reads/11526\_rar3/T1\_0\_R1.fastq > /home/bigdata/sagnik/small\_rna/data/quality\_trimmed\_reads/11526\_rar3/T1\_0\_R1\_bowtie1\_barley.sam 2> /home/bigdata/sagnik/small\_rna/data/quality\_trimmed\_reads/11526\_rar3/T1\_0\_R1\_bowtie1\_barley.error &

## Deliverables

## Methods implemented

## List of softwares used

## Flowcharts

## Results

## Discussion

Chapter 3

# Timeline

Matt wants me to compute p-values for the positions.

## May 1st Week, 8th – 14th

### Matt and I discussed the project at length. Here are the attached MOMs.

Table 3.1 MOM, Matt, 8th May 2017

|  |  |  |
| --- | --- | --- |
|  | Agenda | Response |
| 1 | What are the samples used? | The 90-sample data is used here |
| 2 | Where are all the required files on isilon? | Genome file  rpwise-lab/general\_lab\_use/blast\_databases/fasta\_files\_only/ 2-13-2017\_20031\_CI-16151-WT\_ref\_v3.fasta  Trimmed fastq files  rpwise-lab/  Mla6\_Bln1\_Rar3\_panel\_smRNA\_seq\_exp\_20239\_July2014/  backup\_2-1-2017/  Mla6\_Bln1\_Rar3\_panel\_smRNA\_seq\_exp\_20239\_July2014/  data/  fastq\_files/  trimmed\_fastq |
| 3 | Which software to use for aligning? | Bowtie1 and not Bowtie2 since in Matt’s experience Bowtie2 works better with slightly longer reads. I found some pipelines which have used STAR aligner as well. I will start with Bowtie1 but will also try out HiSAT2 and STAR aligner. |
| 4 | How about peak calling? | Peak calling will probably not be able to detect such narrow peaks but this will be worth a try with MACS2. |
| 5 | Do I have to design a software for an end user? | No, we are the sole user. |
| 6 | When would you expect the results? | By one month |

Table 3.2 MOM, Matt, 11th May 2017

|  |  |  |
| --- | --- | --- |
|  | Agenda | Response |
| 1 | I have generated the alignments with Bowtie1 without any mismatches. Roughly ~21% reads map to Barley and 1% to Blumeria. These are not unique mappings. Is this what you expect? | This is quite natural. |
| 2 | Are secondary alignments reported by bowtie those alignments which exceed the set cut-off score but are worse than the best alignment? If yes, then we should not have any secondary alignment and I checked it for a file and could not find any. | Yes, we should not have those |
| 3 | Where are the transcriptome gtf file for the SNP genome on isilon? | rpwise-lab/  general\_lab\_use/  Barley\_transcriptome\_files/transcriptome\_gtfs/9-8-2016\_barley\_aligned\_BigSeq\_WT\_transcripts.gtf |
| 4 | How about relaxing the mapping conditions a bit, like one mismatch per read? | We can try that but downstream programs expect exact matches. Moreover, we will be doing stats with this data so we need to be careful about it. |
| 5 | What is the degradome data? | This data is required for validation of our predicted phases. |
| 6 | What activity does PARE do? | Does something with the monophosphate that is left behind after microRNA cuts the transcript |
| 7 | Do you want me to correlate biological replicates in the 90-sample data? | No, not at least for the time being. We will combine data from all the 90 samples and check for phaseRNA |

### Text

## Text

Chapter 4

# Binarization

Aaaaaaaaaaaaaaaaaaaaaaa

## 

### Text

## Text

Chapter 5

# Recognition

Aaaaaaaaaaaaaaaaaa

## 

### Text

## Aaa

Aaaaaa

Is illustrated in Figure 5.1 Test par

|  |  |
| --- | --- |
| Blue hills.jpg | Blue hills.jpg |
|  |  |
| Blue hills.jpg | Blue hills.jpg |
|  |  |

Figure 5.1 Sample image

## aaa

Chapter 6

# Indexing and Retrieval

## 

### Text

[5] [6]

## Text

# A. Test

**References**

[1] Y. Zheng, S. Wang, and R. Sunkar, “Genome-wide discovery and analysis of phased small interfering RNAs in Chinese sacred lotus,” *PLoS One*, vol. 9, no. 12, p. e113790, 2014.

[2] C. Zhang, G. Li, J. Wang, and J. Fang, “Identification of trans-acting siRNAs and their regulatory cascades in grapevine,” *Bioinformatics*, vol. 28, no. 20, pp. 2561–2568, 2012.

[3] R. Xia, J. Xu, S. Arikit, and B. C. Meyers, “Extensive families of miRNAs and PHAS loci in Norway spruce demonstrate the origins of complex phasiRNA networks in seed plants,” *Mol. Biol. Evol.*, vol. 32, no. 11, pp. 2905–2918, 2015.

[4] H.-M. Chen, Y.-H. Li, and S.-H. Wu, “Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 9, pp. 3318–3323, 2007.

[5] J. Sauvola and M. Pietikäinen, “Adaptive document image binarization,” *Pattern Recognit.*, 2000.

[6] A. Mollah, S. Basu, M. Nasipuri, and D. Basu, “Text/graphics separation for business card images for mobile devices,” *arXiv Prepr. arXiv1004.0766*, 2010.