

smallRNA-Seq for Not-Annotated Species Analysis Report

TBD180101

December 11, 2018

Contents

| | |
|----------------------------------|----------|
| General Information | 2 |
| Sample Information | 2 |
| Reference Information | 2 |
| Analysis Method | 3 |
| Pipeline Descriptions | 3 |
| Analysis Result | 4 |
| Preprocessing | 4 |
| Reference mapping | 4 |
| Extract consensus | 4 |
| Candidate of smallRNAs | 5 |

General Information

In traditional small RNA analysis methods using small RNA-Seq data, a small RNA database such as miRBase is essential. Therefore, there was a difficulty in analyzing the small RNA of a species that does not have a database. The species that do not have such a small RNA database will be called non-annotated species. So, we had developed small RNA analysis methods using small RNA-Seq data in these species. This method uses the genome sequence and gene coordination information to identify the loci of small RNAs and calculate their expression levels. It also provides information on which genes can be biologically affected using genomic location information of identified small RNAs. This data is expected to provide insight into newly studied species.

Sample Information

Table 1: Sample Informations

| SampleName | No.Seqs | Residues | GC |
|--------------|------------|-------------|------|
| OE348GFP-2_1 | 18,071,398 | 903,569,900 | 54.7 |
| R7491_3_1 | 18,302,875 | 915,143,750 | 53.6 |

- SampleName : Sample ID provided by customer
- No.Seqs : Number of sequences (reads or contigs)
- Residues : Total number of base-pairs
- GC : Ratio of GC

Reference Information

- Species name : *Fusarium graminearum*
- Reference source : BioMax

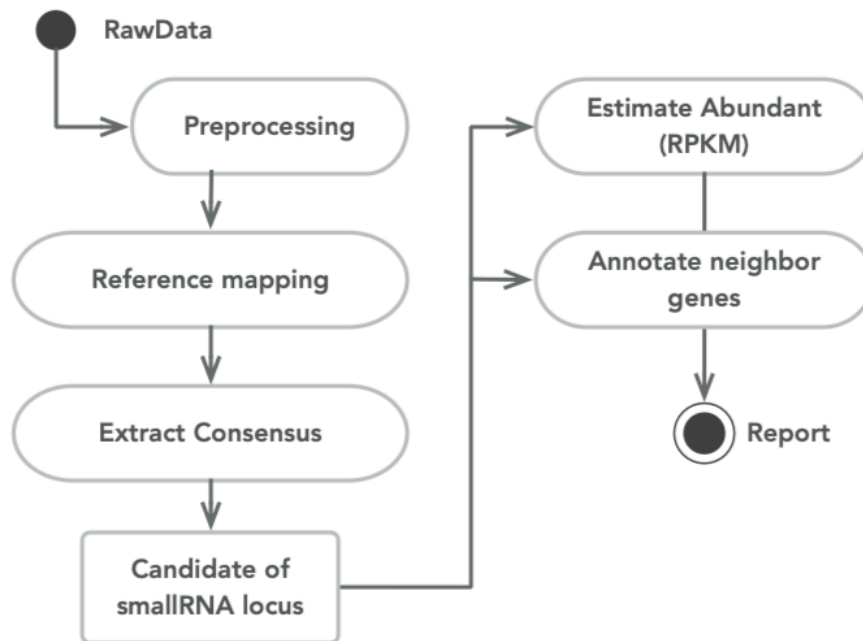


Figure 1: analysis pipeline.

Analysis Method

Pipeline Descriptions

1. Preprocessing : Low quality reads were filtered according to the following criteria
 - Sequencing quality $< Q20$
 - minimum read length $< 17\text{bp}$
 - trim 4bp from 5' and 3'
2. Reference mapping : Map on to the genome
3. Extract consensus : Extract consensus sequence based on alignment information of all samples.
4. Estimate Abundant : Align the extracted consensus sequence once again for each sample. Thereafter, RPKM is calculated based on the alignment information.
5. Annotate neighbor genes : Based on the alignment information, annotation is made of which genes are present around the confirmed small rna locus.

Analysis Result

Preprocessing

Table 2: Preprocessing Results

| SampleName | No.Seqs | Residues | GC |
|------------|------------|-------------|------|
| OE348GFP-2 | 17,677,404 | 363,544,054 | 53.2 |
| R7491_3 | 18,123,364 | 414,788,514 | 51.6 |

- No. Seqs : Number of reads remaining after preprocessing
- Residues : The remaining base pairs after preprocessing
- GC : GC ratio after preprocessing

Reference mapping

Table 3: Mapping Statistics

| Sample | TotalReads | MappedReads | MappedRate |
|------------|------------|-------------|------------|
| OE348GFP-2 | 17677404 | 14507939 | 82.07 |
| R7491_3 | 18123364 | 16347680 | 90.20 |

- TotalReads : The total number of reads used in the reference mapping
- MappedReads : Number of reads aligned to reference genome
- MappedRate : $(\text{MappedReads}/\text{TotalReads}) \times 100$

Extract consensus

Table 4: Extract consensus sequence result

| SampleName | No.Seqs | Residues | Average | Minimum | Maximum | N50 | Npct | GC |
|------------|---------|------------|---------|---------|---------|-----|------|------|
| merged | 314,346 | 14,012,677 | 44.58 | 17 | 300 | 52 | 0.5 | 51.6 |

- No.Seqs : The number of consensus extracted from the alignment information in all samples. This means the entire candidate smRNA locus.
- Residues : The number of consensus base-pairs
- Average/Minimum/Maximum/N50 : The average/min/max/N50 length of the consensus sequence (candidate smRNA locus).
- Npct : ratio of N base in consensus sequence

- GC : ratio of GC base in consensus sequence

Candidate of smallRNAs

| smRNA_candidate | contig | start | end | length | ... |
|--------------------------------|------------------|--------|--------|--------|-----|
| supercontig_3.10_100074-100096 | supercontig_3.10 | 100074 | 100096 | 23 | ... |
| supercontig_3.10_10019-10035 | supercontig_3.10 | 10019 | 10035 | 17 | ... |
| supercontig_3.10_100305-100328 | supercontig_3.10 | 100305 | 100328 | 24 | ... |
| supercontig_3.10_100331-100359 | supercontig_3.10 | 100331 | 100359 | 29 | ... |
| supercontig_3.10_100362-100383 | supercontig_3.10 | 100362 | 100383 | 22 | ... |

- smRNA_candidate : The ID of the candidate small RNA locus, consisting of [contig] _ [start] - [end].
- contig : The contig ID where the candidate small RNA locus is located.
- start/end : The genomic location where the candidate small RNA locus is located.
- length : Length(bp) of candidate small RNA locus
- ReadCnt:[SampleName] : The mapped read count on the candidate small RNA locus
- type : The relationship between the locus of the candidate small RNA and the location of the gene.
- gap : the distance between genes involved in the candidate small RNA
- gene : genes involved in the candidate small RNA
- conseq : consensus sequence of smRNA locus