MicroRNA Analysis Package for SOLiD

contents

1 Introduction

Micro-RNA Analysis Package for SOLiD can process microRNA data produced by SOLiD. It is developed by two languages: perl and java.

The micro-RNA analysis pipeline contains 4 steps:

1. Cut reads
2. Mapping
3. miRNA prediction
4. Target gene prediction

The tool uses PBS to schedule the job queues, so users should install PBS on your server first.

2 Installation

2.1 Hardware Requirement

- CPU: 64 bit Intel or AMD CPU
- RAM: 2 GB to 4 GB per CPU

2.2 Software Requirement before installation

- Operation system: 64-bit Linux
- JDK: version 1.6
- perl: version 5.8.5 or above
- MaToGff: version 0.2.06
- Corona_Lite_Plus_4.2.1
- RNA_pipeline_0.5.0
- ViennaRNA: version 1.82
- RNAhybrid: version 2.1
- miRanda: version 3.3

2.3 Install

Unzip the mirna_solid_pipeline.tar.gz package into the desired location:

```
$ tar xzvf mirna_solid_pipeline.tar.gz
```

3 Pipeline

The pipeline contains 4 steps:

1. Cut reads
2. Mapping
3. miRNA prediction
4. Target gene prediction

3.1 Cut Reads

The pipeline expects a rawdata of length 35, so if the raw data has reads of length higher than 35, we will first trim reads to 35 length.

3.2 Mapping

In this step, we do three things:

First, filter non-miRNA reads.

Second, align reads to miRBase.

Third, align reads to genome.

All the above jobs are done by RNA-MAP software.

3.3 miRNA Prediction

The candidate novel miRNAs predicated using miRDeep was classified according to the miRBase annotation. That is, if the seed sequence (2-8 base pair) of a novel miRNA is same to miRBase mature sequences, this novel miRNA was denoted to be conserved.

3.4 Target Gene Prediction

We integrated two popular methods: miRanda and RNAHybrid for miRNA target predication.

4 Usage

The package consists of three parts:

- perl directory
- mirna_solid_analysis.jar
- config.properties

4.1 Perl Directory

We use perl scripts to process the analysis. The scripts are located in the perl directory.

4.2 mirna_solid_analysis.jar

We use java programs to do three things:

- Encapsulate the invoking of the perl scripts so users can easily perform the process by running the java program.
- Do some counting and draw pictures after each step is finished.
- Generate html files which demonstrate the results after the pipeline is finished.

All the java programs are packaged in this executable jar file.

4.3 config.properties

This is the configuration file which has listed all the parameters needed by the program, so users should fulfill this file before running the program.

Each of the steps has a number of parameters. Table 1 to Table 5 shows the parameters.

Table 1: General parameters

Parameter	Description	Example
RAW_DATA_PATH	The absolute path of the raw data file.	/testdata/test.csfasta
RESULT_DIRECTORY	The absolute path of directory which the result will be sorted.	/result/
CORONA_PATH	The absolute path of directory of corona software	/software/corona
QUEUE_NAME	PBS jobs running queue name.	Ibque

Table 2: Cut Reads parameters

Parameter	Description	Example
READS_LENGTH	The reads length of raw data	50

Table 3: Mapping parameters

Parameter	Description	Example
ADAPTER	The adapter sequence	CGCCTTGGCCGTACAGCAG
GENOME_REFERENCE_FILE	The multiple *.fa file for genome.	/testdata/ Mus_musculus.NCBIM37.62.dna.chromosome.18_19.fa
NON_MIRNA_REFERENCE_FILE	The multiple fasta format file containing (base) sequences that need to be removed from the reads file. These sequences can be ribosomal RNA, tRNA, mRNA.	/testdata/ filter_ref_Mouse_nomiRNA.fa
PRECURSOR_GFF_REFERENCE_FILE	The microRNA precursor .gff formate file for this species. Download from miRBase	/testdata/mmu_pre_pos.gff
PRECURSOR_FASTA_REFERENCE_FILE	The microRNA precursor .fasta formate file for this species. Download from miRBase	/testdata/mmu.fa

Table 4: miRNA Prediction parameters

Parameter	Description	Example
MATURE_GFF_REFERENCE_FILE	The microRNA mature .gff file for this species. Just convert precursor positions to mature positions in the precursor .gff file.	/testdata/mmu_mature_pos.gff
SPECIES_ALIAS_NAME	Species name, abbreviation	mmu

Table 5: Target Gene Prediction parameters

Parameter	Description	Example
UTR3_SEQUENCE_FILE	The sequence of 3'UTR file	/testdata/GeneUTR3
RNAHYBRID_3UTR	data set name used for RNAhybrid –s option	3utr_human

4.4 How to run

1. Install the required software and configure the environment.

2. Switch to the directory where the package is installed.

3. Fulfill the configuration file (config.properties).

4. Run the command:

java -jar mirna_solid_analysis.jar config.properties

4.5 Running Time

The running time depends on the environment of the service.

Raw data size	Reference size	Time
60 KB	339 MB	5.8 hours

4.6 Output Files

All the output files are located in the result directory users set.

4.6.1 HTML Files

/html/result_index.html

/html/conserved_mirna.html

/html/novel_mirna.html

4.6.2 Cut Reads

/cutreads/rawdata_trimed.csfasta

Csfasta format file that shows raw data after cutting off sequence length to 35.

4.6.3 Mapping

/mapping/filter/*.csfasta.ma.35.2

Ma file that shows reads mapped to non-miRNA reference.

/mapping/miRBase/*.csfasta_extend.ma.35.6

Ma file that lists reads mapped to miRBase.

/mapping/genome/*.csfasta_extend.ma.35.2.all

Ma file that lists reads mapped to genome.

/mapping/genome/*.csfasta_extend.ma.35.2.unmatched

Ma file that lists reads unmapped.

/mapping/miRNA_matching_summary

Statistics file that show the number of reads mapped to each reference. The file has five fields:

#Step	Total	Multiple	Percent	Unique	Percent
filter	Total reads number	The number of reads that multiply mapped to non-miRNA	The percent of reads that multiply mapped to non-miRNA	The number of reads that uniquely mapped to non-miRNA	The percent of reads that uniquely mapped to non-miRNA
miRBase	Total reads	The number of reads that	The percent of reads that	The number of reads that	The percent of reads that

	number	multiply mapped to miRBase	multiply mapped to miRBase	uniquely mapped to miRBase	uniquely mapped to miRBase
genome	Total reads number	The number of reads that multiply mapped to genome	The percent of reads that multiply mapped to genome	The number of reads that uniquely mapped to genome	The percent of reads that uniquely mapped to genome
total_map	Total reads number	The number of reads that multiply mapped to above references	The percent of reads that multiply mapped to above references	The number of reads that uniquely mapped to above references	The percent of reads that uniquely mapped to above references

/mapping/filter_stat

Statistics file that demonstrates the non-miRNA RNAs and their number.

/mapping/filter.png

Picture that demonstrates the non-miRNA RNAs and their number.

/mapping/genome.png

Picture that demonstrates the reads number that multiply mapped to each references.

4.6.4 Conserve miRNA Prediction

/conserve/ known_microRNA.fa

Fa format file that lists the predicted conserved miRNAs.

/conserve/known_microRNA.list

File that lists the features of conserved miRNAs, including:

miRNA ID, chromosome ID that located, strand, precursor start, precursor end, mature start, mature end, star start, star end, mature copy, star copy, mature sequence, star sequence and precursor sequence.

/conserve/ reads_length_distribution

File that shows the reads length distribution of conserved miRNA.

/conserve/ conserve_output.Mapped

File that shows alignment information of reads mapped to conserved miRNA.

4.6.5 Novel miRNA Prediction

/novel/novel_microRNA.fa:

Fa format file that lists the predicted novel miRNAs.

/novel/novel_microRNA.list:

File that lists the features of novel miRNAs, including:

miRNA ID, chromosome ID that located, strand, precursor start, precursor end, mature start, mature length, star start, mature copy, star copy and mature sequence.

/novel/novel_microRNA.txt:

File that shows structure of novel miRNAs and alignment information of reads mapped to novel miRNA.

/novel/readslen.png

Picture that demonstrates reads length distribution of conserve and novel miRNAs.

/novel/structure/

A directory that locates all the novel miRNA structure SVG files.

4.6.6 Target Gene Prediction

/target/target_predict

File that lists miRNAs and their target genes.

5 Test Data

Raw Data	/testdata/test.csfasta
Mapping	/testdata/ Mus_musculus.NCBIM37.62.dna.chromosome.18_19.fa /testdata/ filter_ref_Mouse_nomiRNA.fa /testdata/ mmu_pre_pos.gff /testdata/ mmu.fa
Conserve miRNA Prediction	/testdata/ mmu_mature_pos.gff
Target Prediction	/testdata/ GeneUTR3

6 Contact and Support

MicroRNA Analysis Package for Solexa is developed and maintained by Beijing Institute of Genomics(BIG), Chinese Academy of Sciences. If you have feedback or questions, please feel free to contact us at rnatap@big.ac.cn.