MicroRNA Analysis Package for Solexa

Contents

1 Introduction

MicroRNA Analysis Package for Solexa can process microRNA data produced by Solexa. It is developed by two languages: perl and java.

The micorRNA analysis pipeline contains 5 steps:

1. Cluster & Filter
2. Genome Mapping
3. Non-miRNA Mapping
4. Prediction
5. Target gene preidction

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
- Python: version 2.3.4 or above
- blastall: version 2.2.18 or above
- bwa: version 0.5.8a
- samtools: version 0.1.8
- ViennaRNA: version 1.8.2
- Randfold: version 2.0
- Squid: version 1.9g
- RNAhybrid: version 2.1
- miRanda: version 3.3

2.3 Install

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

```
$ tar xzvf mirna_solexa_pipeline.tar.gz
```

3 Pipeline

The pipeline contains 5 steps:

1. Cluster & Filter
2. Genome Mapping
3. Non-miRNA Mapping
4. Prediction
5. Target gene prediction

3.1 Cluster & Filter

The raw reads are filtered by removing adaptor; then we do cluster for identical reads and only keep one reads and the reads number; subsequently, we only keep the reads among specific lengths which are decided by user at submitting page for analysis.

3.2 Genome Mapping

The selected reads are mapped to genome using bwa.

3.3 Non-miRNA Mapping

The mapped reads in the previous step are mapped to non-miRNA database which including non-miRNA ncRNA and mRNA.

3.4 Prediction

Unmapped reads in the previous step are taken as candidates of miRNA. In this step, we do the following things:

- Predict the potential precusor sequences and structure of the miRNA candidates
- Align the remaining reads to the potential precursors
- Predict the miRNAs
- Extract the conserver miRNA and novel miRNA separately

3.5 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_solexa_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_solexa_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/ s_5_seq.fastq
RESULT_DIRECTORY	The absolute path of directory which the result will be sorted.	/result/
QUEUE_NAME	PBS jobs running queue name.	Ibque

Table 2: Cluster & Filter parameters

Parameter	Description	Example
ADAPTER	Adaptor sequence	TCGTATGCCGTCTTCTGCTTG
MIN_LENGTH	The min valid length of reads.	18
MAX_LENGTH	The max valid length of reads.	35

Table 3: Genome mapping parameters

Parameter	Description	Example
GENOME_REFERENCE_FILE	The multiple *.fa file for genome.	/testdata/ Homo_sapiens.GRCh37.58 .dna.chromosome.22.fa

Table 4: Non-miRNA RNA mapping parameters

Parameter	Description	Example
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_Human_nomiRNA.fa
NON_MIRNA_REFERENCE_LENGTH_FILE	The length file, the length is consistent to the filtered file	/testdata/ filter_ref_Human_nomiRNA.fa.length

Table 5: Prediction parameters

Parameter	Description	Example
MIRNA_REFERENCE_FILE	Mature miRNA file, download from miRBase.	/testdata/mature_miRNA.fa
GENE_LIST_FILE	The gene list file, which contains the following columns: 1. GeneID 2. Chromosome 3. Gene strands 4. Gene start position (1 base) 5. Gene end position (1 base) 6. Exon number 7. Exon start (semicolon delimited) 8. Exon end (semicolon	/testdata/GeneListChrom

	delimited)	
SPECIES_ALIAS_NAME	Species name, abbreviation	Has

Table 6: 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 softwares and configure the environment.

2. Create index for your genome reference sequence using bwa.

For short genome:

bwa index genome.fa

For long genome:

bwa index -a bwtsw genome.fa

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

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

5. Run the command:

java -jar mirna_solexa_analysis.jar config.properties

4.5 Running Time

The running time depends on the environment of the service.

Raw data size	Reference size	Time
613 MB	3.0 GB	47 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/filter_result.html

/html/ncrnamapping_result.html

/html/conserved_mirna.html

/html/novel_mirna.html

4.6.2 Cluster & Filter

/filter/sequence_adaper:

Fastq format file that shows the reads after removing adapter

/filter/sequence_cluster.fa:

Fasta format file that shows the reads after cluster

/filter/sequence_cluster.txt:

File after cluster, which has two fields: reads_sequence and reads_number

/filter/sequence.fa:

Fasta format file that shows reads that are out of the selected length scope

/filter/sequence.stat:

Statistics file that shows the number of reads after each process.

/filter/sequence_clean.stat:

Statistics file that shows the number of reads of different lengths. The file has three fields:

Field	#length	unique	Total
Description	Reads length	The number of reads after cluster	The number of reads before cluster

/filter/filter.png

Picture that shows the reads number after each process.

/filter/total.png

Picture that shows the number of reads of selected lengths before cluster.

/filter/unique.png

Picture that shows the number of reads of selected lengths after cluster.

4.6.3 Genome Mapping

/genomemapping/alignedTogenome.blastparsed

Blastparsed format file that shows the reads mapped to genome.

/genomemapping/alignedTogenome.fa

Fasta format file that shows the reads mapped to genome.

/genomemapping/sequence.bam

/genomemapping/sequence.bam.bai

/genomemapping/sequence.sai

/genomemapping/sequence.sam

/genomemapping/map_to_genome.count

Statistics file that shows the number of reads mapped to genome or not mapped to genome. The file has four fields:

#Class	Total	Mapped	Unmapped
#total	The number of selected reads before cluster	The number of selected reads before cluster, which can map to genome	The number of selected reads after cluster, which cannot map to genome
#unique	The number of selected reads after cluster	The number of selected reads after cluster, which can map to genome	The number of selected reads after cluster, which cannot map to genome

/genomemapping/totalmap.png

Picture that demonstrates the mapping state of total reads.

/genomemapping/uniquemap.png

Picture that demonstrates the mapping state of unique reads.

4.6.4 Non-miRNA Mapping

/ncrnamapping/alignedToncrna.blastparsed:

Blastparsed format file that shows the reads mapped to non-miRNA database.

/ncrnamapping/alignedToncrna.fa:

Fasta format file that shows the reads mapped to non-miRNA database.

/ncrnamapping/alignedToncrna.id:

File that shows reads id and reads family that mapped to non-miRNA database.

/ncrnamapping/ align_to_noMiRNA.count:

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

/ncrnamapping/ mirna_candidate.blastparsed:

Blastparsed format file that shows the miRNA candidates.

/ncrnamapping/ mirna_candidate.fa:

Fasta format file that shows the miRNA candidates.

/ncrnamapping/ ncrna.count:

Statistic file that demonstrates the reads number that mapped to non-miRNA database and not mapped to non-miRNA database.

/ncrnamapping/ no_miRNA.txt:

Statistic file that demonstrates the expressed RNA types and their numbers.

/ncrnamapping/ total.png:

Picture that demonstrates the reads number that mapped to non-miRNA database and not mapped to non-miRNA database before cluster.

/ncrnamapping/ unique.png:

Picture that demonstrates the reads number that mapped to non-miRNA database and not mapped to non-miRNA database after cluster.

/ncrnamapping/ non-coding.png:

Picture that demonstrates the expressed RNA types and their numbers.

/ncrnamapping/ rfam.png:

Picture that demonstrates the non-miRNAs RNA families and their numbers.

4.6.5 Prediction

/prediction/precursors.fa:

File that shows the precursors sequence of predicted miRNA.

/prediction/structures:

File that shows the sequence and structure of predicted miRNA.

/prediction/signatures:

Files that shows the reads that mapped to the precursors file.

/prediction/predictions

File that shows the detail information of predicted miRNAs.

/prediction/miranln.result:

File that shows the predicted miRNAs and reads alignment informations.

/prediction/mirna_prediction.stat:

Statistic file that shows the information of miRNAs, including conserved miRNA and novel miRNAs.

/prediction/mirspe.svg:

Svg file that demonstrates predicted conserved miRNAs and the species information.

/prediction/conserved/:

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

/prediction/novel/:

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

/prediction/loop.fasta:

File that shows loop sequences.

/prediction/remained.fasta:

File that shows remained reads that cannot map.

4.6.6 Target Gene Prediction

/target/target_predict

File that lists miRNAs and their target genes.

5 Test Data

Raw Data	/testdata/s_5_seq.fastq
Cluster & Filter	ADAPTER:TCGTATGCCGTCTTCTGCTTG
Genome mapping	/testdata/Homo_sapiens.GRCh37.58.dna.chromosome.22.fa
Non-coding rna mapping	/testdata/filter_ref_Human_nomiRNA.fa
	/testdata/filter_ref_Human_nomiRNA.fa.length
prediction	/testdata/mature_miRNA.fa
	/testdata/GeneListChrom
Target gene 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.