# RNA-seq Analysis Package for Solexa

## Contents

# 1 Introduction

RNA-seq Analysis package for Solexa can process the RNA-seq analysis for Solexa data. It is developed by two languages: perl and java.

The RNA-seq analysis pipeline for Solexa data contains 5 steps:

1. Mapping
2. Annotation
3. DEGseq analysis
4. GO analysis
5. KEGG analysis

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

# 2 Installation

## 2.1 2.1 Hardware Requirement

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

## 2.2 Software Requirement

- Operation system: 64-bit Linux
- JDK: version 1.6
- perl: version 5.8.5 or above
- Python: version 2.3.4 or above
- R: version 2.1.0
- blastall: version 2.2.18
- bwa: version 0.5.8a
- Corona_Lite: Corona_Lite_Plus_4.2.1
- Perl models:
  - SVG model
  - GD model
  - GO model
- DEGseq package
- PBS

## 2.3 Install

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

```
$ tar xzvf rna_solexa_pipeline.tar.gz
```

# 3 RNA-seq Analysis Pipeline for Solexa Data

The pipeline contains 5 steps:

1. Mapping
2. Annotation
3. DEGseq analysis
4. GO analysis
5. KEGG analysis

## 3.1 Mapping

We use bwa-5.08a program to map raw reads to the references selected by users.

## 3.2 Annotation

After mapping, we do annotation for the mapped reads.

## 3.3 DEGseq Analysis

Users could identify DEGs between two samples in their projects, after the corresponding annotation program has finished. To process this analysis, users should provide an annotation file with the same format as the output annotate file in the annotation step. This step is optional, so users could choose whether to do it.

## 3.4 GO Analysis

GO (Gene ontology) provides controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products, which is one of the daily tools for downstream gene functional annotation. This step is optional, so users could choose whether to do it.

## 3.5 KEGG Analysis

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks and also on the structure relationships. This step is optional, so users could choose whether to do it.

# 4 Usage of the Package

The package consists of three parts:

- perl directory
- rna_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 rna_solexa_analysis.jar

We use java programs to do three things:

- Encapsulate 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 6 shows the parameters.

Table 1: General parameters:

| Parameter | Description | Example |
|---|---|---|
| **RAW_DATA_PATH** | The absolute path of the raw data file. | /testdata/sample.fastq |
| **RESULT_DIRECTORY** | The absolute path of directory which the result will be stored. | /result/ |
| **WORK_QUEUE** | PBS jobs running queue name. | ibque |
| **TRACE_QUEUE** | Tracking queue name. | ibque |

Table 2: Mapping parameters

| Parameter | Description | Example |
|---|---|---|
| **SEEDS_LENGTH** | Seed length | 32 |
| **SEEDS_MISMATCH** | Maximum differences in the seed | 2 |
| **TOTAL_MISMATCH** | Maximum differences in total reads | 4 |
| **CHOOSE_GENOME** | Whether or not align with genome | yes/no |
| **CHOOSE_JUNCTION** | Whether or not align with junction | yes/no |
| **GENOME_REFERENCE_FILE** | Multiple *.fa file for genome sequence | /testdata/Homo_sapiens.GRCh37.58.dna.chromosome.22.fa |
| **JUNCTION_REFERENCE_F** | Multiple *.fa file for junction sequence | /testdata/chromos |

| | | |
|---|---|---|
| **ILE** | | ome.22Jun.fa |

Table 3: Annotation parameters

| Parameter | Description | Example |
|---|---|---|
| **BWA_CMAP_FILE** | The cmap file identifies the files comprising the reference genome and the representing junction files. In the cmap: the first column is the chromosome number same as GeneList; the second column is the chromosome number same as sam1<br>the third column is the chromosome number same as sam2.<br>All the three columns stand for the same chromosme and separated by Tab. | /testdata/cmap |
| **GENELIST_CHROM_FILE** | The gene list file, which contains the following columns:<br>1. GeneID<br>2. Chromosome<br>3. Gene strands<br>4. Gene start position (1 base)<br>5. Gene end position (1 base)<br>6. Exon number<br>7. Exon start (semicolon delimited)<br><br>Exon end (semicolon delimited) | /testdata/GeneList Chrom |
| **GENELIST_JUNCTION_FIL E** | The junction list file, which contains the following columns:<br>1. GeneID<br>2. Chromosome<br>3. Gene strands<br>4. Junction sequence start<br>Junction sequence end | /testdata/GeneListJ un |
| **GENE_CLASS_BY_RNA_FI LE** | The gene class file, which contains the following columns:<br>1. GeneID<br>2. Chromosome<br>3. RNA class | /testdata/GeneClas sByRNA |
| **GENE_FUNCTION_FILE** | The gene function file, which contains the following columns:<br>1. GeneID | /testdata/GeneFun ction |

| | | |
|---|---|---|
| | 2. Chromosome<br>3. Gene function | |

Table 4: DEGseq parameters

| Parameter | Description | Example |
|---|---|---|
| **IS_DO_DEGSEQ** | Whether do DEGseq | yes/no |
| **MARK_1** | Sample 1 name (the processed sample ) | sample1 |
| **MARK_2** | Sample 2 name (the sample you want to be compared with the processed sample) | sample2 |
| **COMPARED_ANNOTATE_FILE** | Absolute path of annotate files for both samples, the file should have the same format as the output annotate file in the annotation step. which contains 14 columns:<br>1. ENSEMBL ID, ENSEMBL Gene ID;<br>2. Chromosome, Chromosome number<br>3. Strand, Gene Strand<br>4. GeneLen, Gene Length<br>5. ExonLen, Exon Length<br>6. GReads, Stat. of reads number that mapped to gene sequecnes in chromosome<br>7. JReads, Stat. of reads number that mapped to gene sequecnes in junction sequences<br>8. Exon_Intron, Stat. of reads number that mapped to exon_intron boundary in chromosome<br>9. ExonReads, Stat. of reads number that just mapped to exon sequences which include exon reads in chromosome and junction sequences<br>10. IntronReads, Stat. reads number that just mapped to intron sequences<br>11. RPKMValue, calculated by formula: ((1e+9)*ExonReads)/(ExonLen*Total UniqueMappedReads)<br>12. G_rev, Stat. of reads number that mapped to gene reverse strand in chromosome | |

| | | |
|---|---|---|
| | 13. J_rev, Stat. of reads number that mapped to gene reverse strand in junction sequences<br>14. Function, Functional information | |

Table 5: GO parameters

| Parameter | Description | Example |
|---|---|---|
| **GO_LEVEL** | Go function level | 2 |
| **GO_OPTION** | Which function will be discard. | proc \|func \|comp |
| **GO_NUMBER_FILE** | This file contains go number, one gene one row, the first column is gene ID, the others are Go numbers delimited by tab. | /testdata/Gonumber |
| **CELLULAR_COMPONENT_ONTOLOGY_FILE** | Gene ontology file, http://www.geneontology.org/ontology/ | /testdata/go/component.ontology |
| **MOLECULAR_FUNCTION_ONTOLOGY_FILE** | Gene ontology file, http://www.geneontology.org/ontology/ | /testdata/go/function.ontology |
| **BIOLOGICAL_PROCESS_ONTOLOGY_FILE** | Gene ontology file, http://www.geneontology.org/ontology/ | /testdata/go/process.ontology |

Table 6: KEGG parameters

| Parameter | Description | Example |
|---|---|---|
| GENE_ENSEMBLE_LIST_FILE | Gene relevance file ftp://ftp.genome.jp/pub/kegg/genes/organisms/ | /testdata/kegg/genes/organisms/hsa/hsa_ensembl-hsa.list |
| GENE_KO_LIST_FILE | Gene relevance file ftp://ftp.genome.jp/pub/kegg/genes/organisms/ | /testdata/kegg/genes/organisms/hsa/hsa_ko.list |
| GENE_ENZYME_LIST_FILE | Gene relevance file ftp://ftp.genome.jp/pub/kegg/genes/organisms/ | /testdata/kegg/genes/organisms/hsa/hsa_enzyme.list |
| PATHWAY_FILE_DIRECOR Y | KO map directory ftp://ftp.genome.jp/pub/kegg/pathway/ko/ | /testdata/kegg/pathway/ko/ |

## 4.4 How to run

1. Install the required softwares and configure the environment.

2. Create index for your genome reference sequence and junction 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. Fill the configuration file (config.properties).

5. Run the command:

**java -jar rna_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 |
|---|---|---|
| 233 MB | 3.5 GB | 3.2 hous |

## 4.6 Output Files

### 4.6.1 HTML Files

/html/result_index.html

/html/mapping_result.html

/html/annotation_result.html

/html/degseq_result.html

### 4.6.2 Quality Distribution Picture

/filterpic/_reads1_1.png

/filterpic/_reads1_2.png

/_pic.html

### 4.6.3 Mapping

/mapping/map_to_genome.count: Mapping to genome result statistics

/mapping/map_to_junction.count: Mapping to junction result statistics

/mapping/map.png: PNG file demonstrates the result statistics

/mapping/genome.sai
/mapping/genome.sam
/mapping/jun.sai
/mapping/jun.sam

### 4.6.4 Annotation

/annotation/annotate: The annotation result file

/annotation/statistic: The annotation result statistics file

/annotation/annopie.png: The pie graph of annotation

/annotation/annobar1.png: The bar graph of annotation genes classified by reads number

/annotation/annobar2.png: The bar graph of annotation genes classified by rpkm value

### 4.6.5 DEGseq Analysis

/degseq/output_score.txt

/degseq/output.html

/degseq/output/result.png

/degseq/output/Sample1_hist.png

/degseq/output/Sample1_Sample2_compare.png

/degseq/output/Sample2_hist.png

/degseq/output/SampleS_box.png

### 4.6.6 GO Analysis

/go/go.svg

### 4.6.7 KEGG Analysis

/kegg/pathwayPlot/

# 5 Test Data

| | |
|---|---|
| **Raw Data** | /testdata /solexa_rna.fastq |
| **Mapping** | /testdata/Homo_sapiens.GRCh37.58.dna.chromosome.22.fa |
| | /testdata/chromosome.22Jun.fa |
| **Annotation** | /testdata/bwa_cmap |
| | /testdata/GeneListChrom |
| | /testdata/GeneListJun |
| | /testdata/GeneClassByRna |
| | /testdata/GeneFunction |
| **DEGseq** | /testdata/annotate |
| **GO** | /testdata/GoNumber |
| | /testdata/go/component.ontology |
| | /testdata/go/function.ontology |
| | /testdata/go/process.ontology |
| **KEGG** | /testdata/kegg/genes/organisms/hsa/hsa_ensembl-hsa.list |
| | /testdata/kegg/genes/organisms/hsa/hsa_ko.list |
| | /testdata/kegg/genes/organisms/hsa/hsa_enzyme.list |
| | /testdata/kegg/pathway/ko/ |

# 6 Contact and Support

RNA-seq 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.