

---

**Prokaryotes RNA-seq Analysis Report**

**Demo Report**

**May 1, 2016**

---

## Contents

1. Library Preparation and Sequencing.....	1
1.1 Total RNA Sample QC.....	1
1.2 Library Construction and Quality Assesment.....	2
1.3 Library QC.....	2
1.4 Sequencing.....	3
2. Analysis Workflow.....	3
3. Project Results.....	4
3.1 Raw Data.....	4
3.2 Data Quality Control.....	5
3.2.1 Error Rate.....	5
3.2.2 GC Content Distribution.....	5
3.2.3 Data Filtering.....	6
3.2.4 Data Quality Control Summary.....	7
3.3 Mapping to Reference Genome.....	8
3.3.1 Overview of Mapping Status.....	8
3.3.2 Mapped Regions in Reference Genome.....	8
3.3.3 Distribution of Mapped Reads in Chromosomes.....	9
3.3.4 Visualization of Mapping Status of Reads.....	10
3.4 Expression Quantification.....	10
3.4.1 Expression Quantification.....	10
3.4.2 Distributions of gene expression levels.....	11
3.5 RNA-seq Advanced QC.....	12
3.5.1 Correlation of RNA-seq samples.....	12
3.6 Differential Expression Analysis.....	13
3.6.1 List of differentially expressed genes.....	13
3.6.2 Screening of differentially expressed genes.....	14
3.6.3 Cluster Analysis of Gene Expression Differences.....	14
3.6.4 Venn diagram of differentially expressed genes.....	15
3.7 GO Enrichment Analysis.....	16
3.7.1 GO Enrichment Result List of DEGs.....	16
3.7.2 Gene Ontology functional classification.....	16
3.7.3 GO Enrichment DAG Figure.....	17
3.8 KEGG Enrichment Analysis.....	18
3.8.1 KEGG Enrichment List.....	18
3.8.2 KEGG Enrichment Scattered Plot.....	19
3.8.3 KEGG Enrichment Pathway.....	19
3.9 SNP and InDel.....	20
3.10 Novel Gene Prediction.....	21
3.11 Gene Structure Analysis.....	21
3.11.1 Operon prediction.....	22

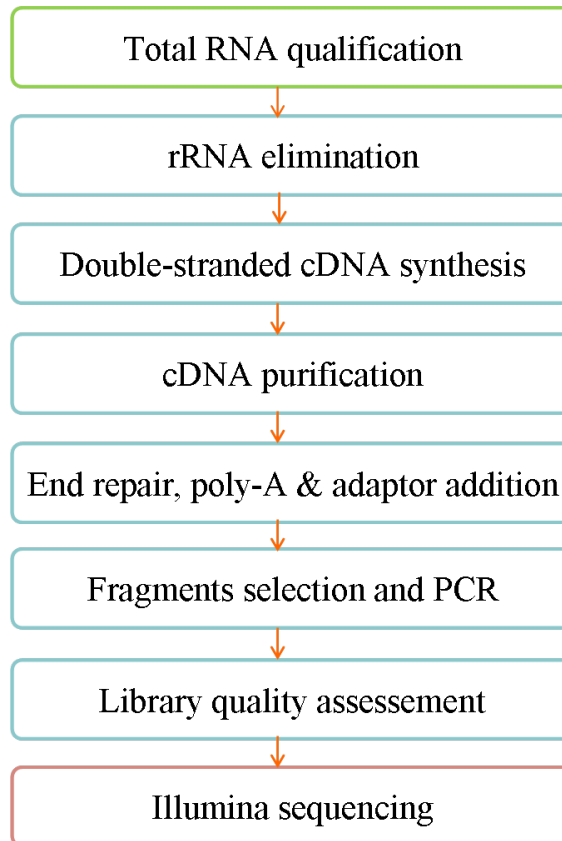
---

3.11.2 TSS and TTS prediction.....	22
3.11.3 Promoter prediction.....	23
3.12 UTR Analysis.....	23
3.12.1 UTR prediction and UTR length distribution.....	23
3.12.2 5' UTR SD sequece prediction.....	24
3.12.3 3' UTR $\rho$ -independent terminator prediction.....	24
3.13 Antisense Transcripts Prediction.....	24
3.14 sRNA Analysis.....	25
3.14.1 sRNA prediction and length distribution.....	25
3.14.2 sRNA secondary structure prediction.....	26
3.14.3 sRNA targeted gene prediction.....	26
3.14.4 sRNA Expression Quantification.....	26
4 Appendix.....	27
4.1 Result Directory Lists.....	27
4.2 Software List.....	28
5 References.....	30

---

## 1. Library Preparation and Sequencing

From the RNA sample to the final data, each step, including sample test, library preparation, and sequencing, influences the quality of the data, and data quality directly impacts the analysis results. To guarantee the reliability of the data, quality control (QC) is performed at each step of the procedure. The workflow is as follows:



### 1.1 Total RNA Sample QC

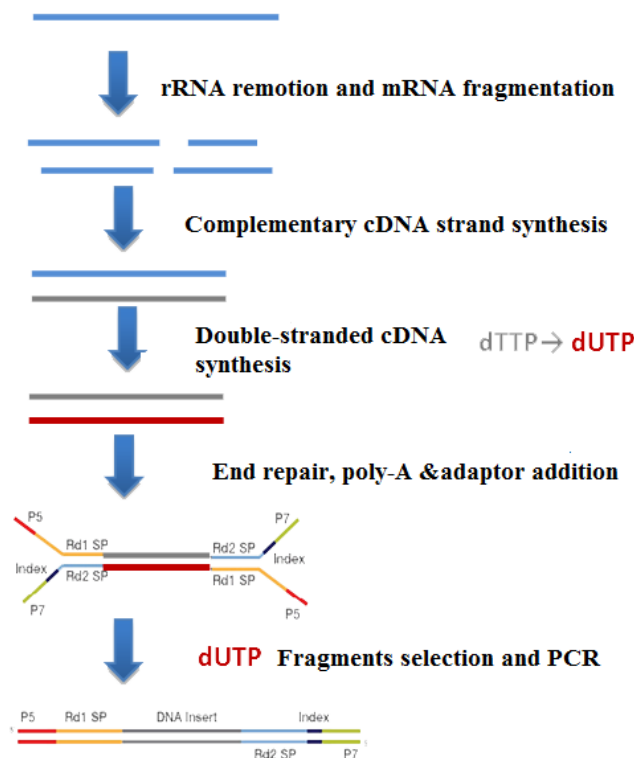
All samples need to pass through the following four steps before library construction:

- (1) Agarose Gel Electrophoresis: tests RNA degradation and potential contamination
- (2) Nanodrop: tests RNA purity (OD260/OD280)
- (3) Qubit: quantifies the RNA (determines concentration)
- (4) Agilent 2100: checks RNA integrity

---

## 1.2 Library Construction and Quality Assesment

After total RNA sample QC, Ribo-zero kits were used to remove rRNAs, then fragmentation buffer were added to fragment the mRNAs. Using random hexamers as primers, the mRNA fragments were reverse-transcribe to single stranded cDNAs. Atfer the systhesis of single stranded cDNAs, buffer, dNTPs (in which dTTP were replaced by dUTP), DNA polymerase I and RNase H were applied to systhesis the complementary cDNA strands. The double stranded cDNAs were purified using AMPure XP beads. The double stranded cDNAs were end-repaired, polyadenylated, ligated with adaptor sequences and size-selected using AMPure XP beads. Then the uracil containing strands were degraded by USER Enzyme, and the remained strands were amplified using PCR and purified using AMPure XP beads. The workflow chart is as follows:



## 1.3 Library QC

Library concentration was first quantified using a Qubit 2.0 fluorometer (Life Technologies), and then diluted to 1 ng/μl before checking insert size on an Agilent 2100 and quantifying to greater accuracy by quantitative PCR (Q-PCR) (library activity >2 nM).

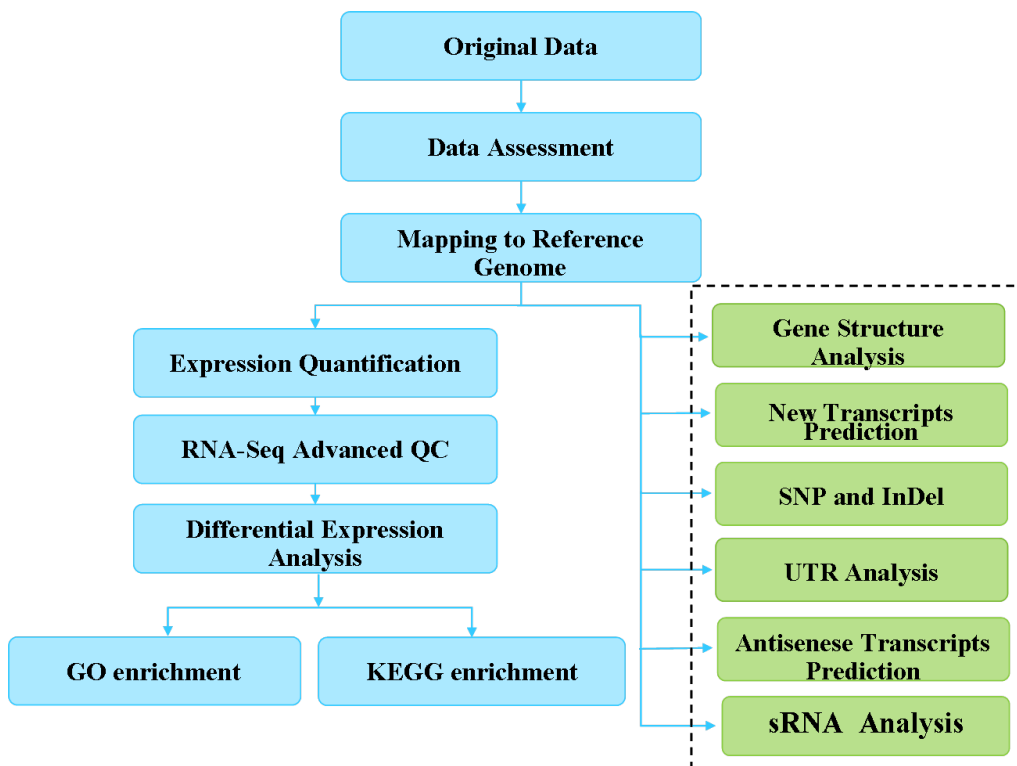
---

## 1.4 Sequencing

Libraries are fed into HiSeq machines according to activity and expected data volume.

## 2. Analysis Workflow

The analysis workflow for data with a reference genome is as follows:



Note: PPI (Protein-protein interaction) network analysis is only available for species with known PPI network data.

---

## 3. Project Results

### 3.1 Raw Data

The original raw data from Illumina HiSeq™ are transformed to Sequenced Reads by base calling. Raw data are recorded in a FASTQ file, which contains sequence information (reads) and corresponding sequencing quality information.

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT
NAAGAACACGTTCCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHFAE0ECFFD/AEHH
```

Line 1 begins with a '@' character and is followed by the Illumina Sequence Identifiers and an optional description.

Line 2 is the raw sequence read.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier and description.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of characters as there are bases in the sequence (Cock et al.).

Illumina Sequence Identifier details:

Identifier	Meaning
HWI-ST1276	Instrument – unique identifier of the sequencer
71	run number – Run number on instrument
C1162ACXX	FlowCell ID – ID of flowcell
1	LaneNumber – positive integer
1101	TileNumber – positive integer
1208	X – x coordinate of the spot. Integer which can be negative
2458	Y – y coordinate of the spot. Integer which can be negative
1	ReadNumber - 1 for single reads; 1 or 2 for paired ends
N	whether it is filtered - NB: Y if the read is filtered out, not in the delivered fastq file, N otherwise
0	control number - 0 when none of the control bits are on, otherwise it is an even number
CGATGT	Illumina index sequences

---

## 3.2 Data Quality Control

### 3.2.1 Error Rate

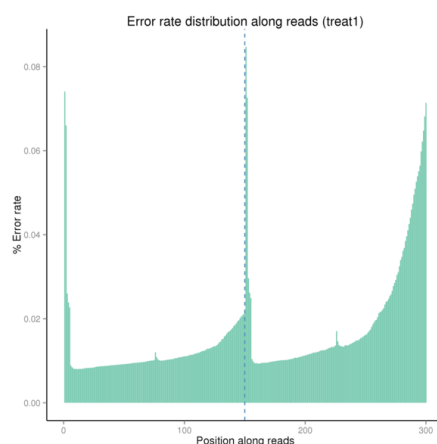
The error rate for each base were transformed by the Phred score as in equation 1 (equation 1:  $Q_{\text{phred}} = -10\log_{10}(e)$ ). The relationship between Phred quality scores  $Q$  and base-calling error “ $e$ ” is given below:

Relationship between sequencing quality for a single base and Phred score given by the Illumina CASAVA v1.8 software:

Phred score	Base Calling error rate	Base Calling correct rate	Q-score
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

Sequencing error rate and base quality are affected by the sequencing machine, reagent and the samples. The error rate distribution has two features :

- (1) Error rate increases as the sequencing reads are extended and sequencing reagents are consumed.
- (2) The first six bases have relatively high error rates due to the random hexamers used in priming cDNA synthesis (Jiang et al.).



**Figure 3.2.1 Error Rate Distribution**

The x-axis shows the base position along each sequencing read and the y-axis shows the base error rate.

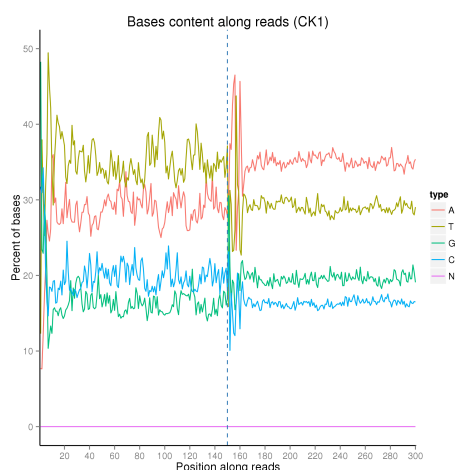
### 3.2.2 GC Content Distribution

AT and GC content distributions were evaluated to detect potential AT or GC separation, which may affect subsequent gene expression quantification.



---

Theoretically, G should equal to C, and A should equal to T throughout the whole sequencing process for non-stranded libraries, whereas AT/GC separation is normally observed in stranded libraries. For DGE (Digital Gene Expression) libraries, a large variation of sequencing error in the first 6-7 bases is allowed due to the usage of random primers in library construction.



**Figure 3.2.2 GC content distribution**

The x-axis shows each base position within a read, and the y-axis shows the percentage of each base, with each base represented by a different color.

### 3.2.3 Data Filtering

Raw reads are filtered to remove reads containing adapters or reads of low quality, so that downstream analyses are based on clean reads. The filtering process is as follows:

- (1) Discard reads with adaptor contamination.
- (2) Discard reads when uncertain nucleotides constitute more than 10 percent of either read ( $N > 10\%$ ).
- (3) Discard reads when low quality nucleotides (base quality less than 20) constitute more than 50 percent of the read.

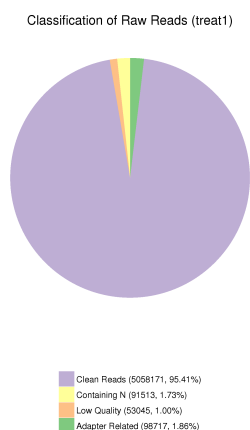
RNA-seq Adapter sequences (Oligonucleotide sequences of adapters from TruSeq™ RNA and DNA Sample Prep Kits):

NEBNext® Ultra™ RNA Library Prep Kit

RNA 5' Adapter (RA5):

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA  
TCT-3'

RNA 3' Adapter (RA3):  
 5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC ( 6-nucleotide index ) ATCTCGTATGCCGTCTTCTGCTTG-3'



**Figure 3.2.3 Raw Data**

Results are shown as percentage of total raw reads.

- (1) Adapter related, reads that had adapter contamination.
- (2) Containing N, reads in which uncertain nucleotides constituted more than 10 percent of the read.
- (3) Low quality, reads in which low quality nucleotides constituted more than 50 percent of the read.
- (4) Clean reads, reads that passed quality control

### 3.2.4 Data Quality Control Summary

**Table 3.2.4 Data Production**

Sample name	Raw reads	Clean reads	Clean bases	Error rate(%)	Q20(%)	Q30(%)	GC content(%)
CK1	12408246	12124130	1.52G	0.03	96.88	93.45	53.56
CK2	11792200	11447320	1.44G	0.03	96.72	93.1	55.88
CK3	9249114	9023924	1.12G	0.03	96.64	93.03	53.37
treat1	10337270	10116342	1.26G	0.03	96.53	92.82	54.69
treat2	16764978	16430586	2.05G	0.01	98.75	96.94	45.59
treat3	14732668	14126118	1.77G	0.01	97.11	92.81	46.5

- (1) Sample name: the names of samples
- (2) Raw Reads: the original sequencing reads counts
- (3) Clean Reads: number of reads after filtering
- (4) Clean Bases: clean reads number multiply read length, saved in G unit
- (5) Error Rate: average sequencing error rate, which is calculated by  $Q_{phred} = -10 \log_{10}(e)$
- (6) Q20: percentages of bases whose correct base recognition rates are greater than 99% in total bases
- (7) Q30: percentages of bases whose correct base recognition rates are greater than 99.9% in total bases
- (8) GC content: percentages of G and C in total bases

### 3.3 Mapping to Reference Genome

Algorithm for mapping sequences: appropriate software is chosen according to the characteristics of the reference genome. In general, Bowtie2 is chosen for the genomes of bacteria and other species with a high gene density. The mismatch parameter is set to two, and other parameters are set to default. In general, the total mapped rate should be more than 70%, and the percentage of reads that can be mapped to multiple sites in the reference genome should be less than 10%, if there is no contamination and a correct reference genome is chosen.

#### 3.3.1 Overview of Mapping Status

Table 3.3.1 Overview of Mapping Status

Sample_name	CK1	CK2	CK3	treat1	treat2	treat3
Total reads	12124130	11447320	9023924	10116342	16430586	14126118
Total mapped	6185606 (51.02%)	10450242 (91.29%)	4805319 (53.25%)	6813186 (67.35%)	15320669 (93.24%)	13251878 (93.81%)
Multiple mapped	644876 (5.32%)	1075470 (9.39%)	447253 (4.96%)	712078 (7.04%)	794433 (4.84%)	904200 (6.4%)
Uniquely mapped	5540730 (45.7%)	9374772 (81.89%)	4358066 (48.29%)	6101108 (60.31%)	14526236 (88.41%)	12347678 (87.41%)
Read-1	2770450 (22.85%)	4687301 (40.95%)	2179090 (24.15%)	3050440 (30.15%)	7254266 (44.15%)	6166795 (43.66%)
Read-2	2770280 (22.85%)	4687471 (40.95%)	2178976 (24.15%)	3050668 (30.16%)	7271970 (44.26%)	6180883 (43.76%)
Reads map to '+'	2770280 (22.85%)	4687285 (40.95%)	2179063 (24.15%)	3050664 (30.16%)	7264591 (44.21%)	6175339 (43.72%)
Reads map to '-'	2770450 (22.85%)	4687487 (40.95%)	2179003 (24.15%)	3050444 (30.15%)	7261645 (44.2%)	6172339 (43.69%)

Details of Data Quality Control

(1) Sample name: name of sample. '\_1' and '\_2' stands for reads starting from different ends, respectively. The total clean reads for each sample are the sum of reads starting from both ends. (2) Raw reads: counts of raw reads.

(3) Clean reads: counts of clean reads.

(4) Clean bases: sum of nucleotides of all reads in clean data, in denomination of gigabyte.

(5) Error rate: calculated from Equation 1.

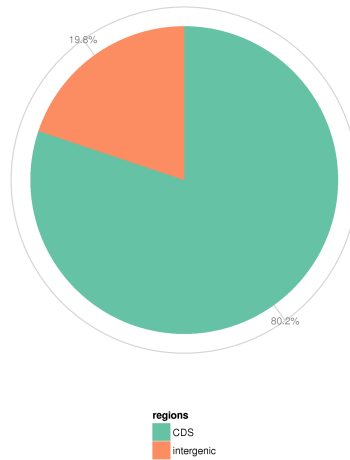
(6) Q20 and Q30: percentages of nucleotides with Phred value larger than 20 and 30 in total nucleotides, respectively.

(7) GC content: percentages of G and C in total nucleotides.

#### 3.3.2 Mapped Regions in Reference Genome

Mapped regions are classified base on the structural annotation of the reference genome. Exon-mapped reads should be abundant if the reference genome is well-annotated. Reads map to intergenic regions may due to weak annotation of the reference genome, noncoding RNA or background noise.

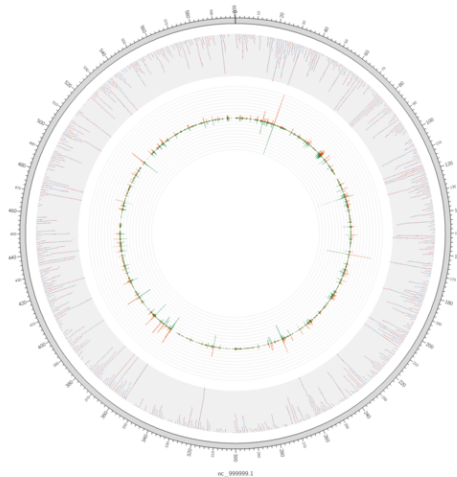
Percent of reads mapped to genome regions (treat1)



**Figure 3.3.2 Classification of Reads According to Mapped Region.**

### 3.3.3 Distribution of Mapped Reads in Chromosomes

The distribution of mapped reads in chromosomes was estimated by statistics on the genomic locations of total mapped reads. A random selected sample of total mapped reads and their genomic location distribution is displayed in below figure:



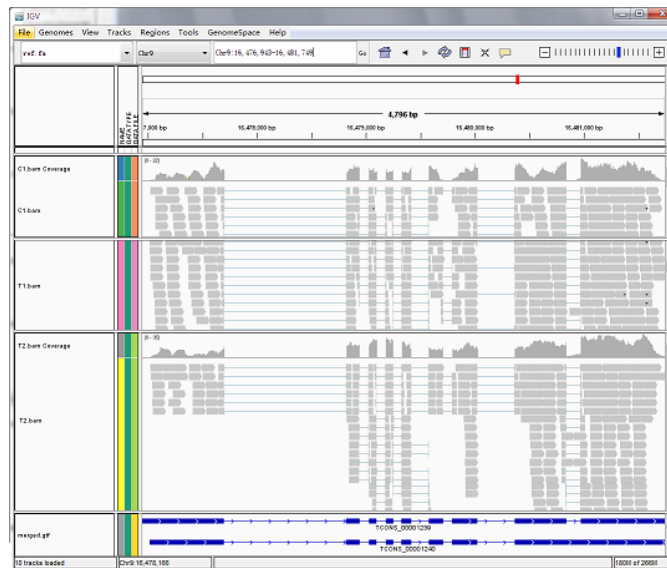
**Figure 3.3.3 Distribution Plot of Mapped Reads in Chromosomes.**

The outermost circle represents the reference genome. The region with grey background shows the distribution of the random selected sample of reads, in which the reads mapped to positive and negative chains are in red and blue, respectively. The innermost circle shows the positive chain coverage distribution in orange and the negative chain coverage distribution in green, respectively. Strange dots were discarded if from the mean of coverages more than 3 times the standard deviation.

---

### 3.3.4 Visualization of Mapping Status of Reads

Files are provided in BAM format, a standard file format that contains mapping results, and the corresponding reference genome and gene annotation file for some species. The Integrative Genomics Viewer (IGV) is recommended for visualizing data from BAM files. The IGV has several features: (1) it displays the positions of single or multiple reads in the reference genome, as well as read distribution between annotated exons, introns or intergenic regions, both in adjustable scale; (2) displays the read abundance of different regions to demonstrate their expression levels, in adjustable scale; (3) provides annotation information for both genes and splicing isoforms; (4) provides other related annotation information; (5) displays annotations downloaded from remote servers and/or imported from local machines.



## 3.4 Expression Quantification

### 3.4.1 Expression Quantification

Gene expression level is measured by transcript abundance. The greater the abundance, the higher is the gene expression level. In our RNA-seq analysis, the gene expression level is estimated by counting the reads that map to genes or exons. Read count is not only proportional to the actual gene expression level, but is also proportional to the gene length and the sequencing depth. In order for the gene expression levels estimated from different genes and experiments to be comparable, the FPKM is used. In RNA-seq, FPKM, short for the expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced, is the most commonest method of estimating gene expression levels, which takes into account the effects of both

sequencing depth and gene length oncounting of fragments(Trapnell, Cole, et al., 2010).

HTSeq software was used to analyze the gene expression levels in this experiment, using the union mode. The result files present the number of genes with different expression levels and the expression level of single genes. In general, an FPKM value of 0.1 or 1 is set as the threshold for determining whether the gene is expressed or not.

**Table 3.4.1 The number of genes with different expression levels**

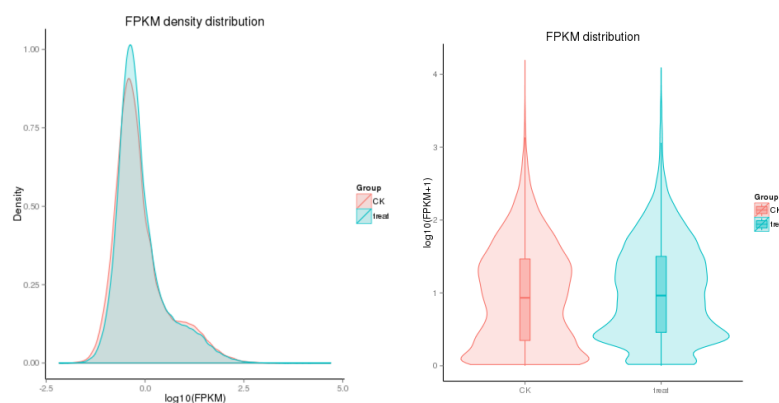
FPKM Interval	CK1	CK2	CK3	treat1	treat2	treat3
0~1	854(15.83%)	829(15.36%)	923(17.11%)	905(16.77%)	616(16.63%)	633(17.09%)
1~3	248(4.60%)	212(3.93%)	245(4.54%)	274(5.08%)	36(0.97%)	40(1.08%)
3~15	832(15.42%)	767(14.21%)	1029(19.07%)	1032(19.13%)	217(5.86%)	224(6.05%)
15~60	1428(26.46%)	1413(26.19%)	1423(26.37%)	1365(25.30%)	659(17.79%)	662(17.87%)
>60	2034(37.69%)	2175(40.31%)	1776(32.91%)	1820(33.73%)	2177(58.76%)	2146(57.92%)

**Table 3.4.2 Gene expression levels**

Gene_id	CK1	CK2	CK3	treat1	treat2	treat3
PSYRH_RS00175	4398.278343	2136.658857	2251.261063	4110.437105	462.3299003	342.9179606
PSYRH_RS15710	35.35451423	31.50135285	27.96744984	14.38204746	27.20701432	30.11574595
PSYRH_RS17820	25.98862455	39.31446147	26.88670611	21.79325744	400.5975882	161.7983214
PSYRH_RS20900	145.4266875	87.0431443	132.1648198	82.07482199	58.56553557	41.31620085

### 3.4.2 Distributions of gene expression levels

An FPKM distribution diagram and violin plot were applied to compare gene expression levels of different group pf samples. For biological replicates, the mean FPKM values were used.



**Figure 3.4.2 Distributions of gene expression levels**

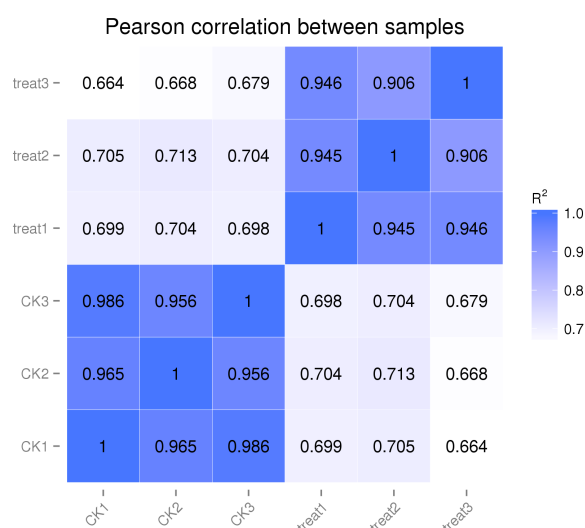
Upper panel: FPKM distribution, the x-axis shows the  $\log_{10}(\text{FPKM}+1)$  and the y-axis shows gene density. Lower panel: FPKM violin Plot, the x-axis shows the sample names and the y-axis shows the  $\log_{10}(\text{FPKM}+1)$ . Each violin has five statistical magnitudes (max value, upper quartile, median, lower quartile and min value). Lower right panel : FPKM distribution, the x-axis shows the  $\log_{10}(\text{FPKM}+1)$  and the y-axis shows gene density.

---

## 3.5 RNA-seq Advanced QC

### 3.5.1 Correlation of RNA-seq samples

Biological replicates are necessary for any biological experiment, including those involving RNA-seq technology (Hansen et al.). In RNA-seq, replicates have a two-fold purpose. First, they demonstrate whether the experiment is repeatable, and secondly, they can reveal differences in gene expression between samples. The correlation between samples is an important indicator for testing the reliability of the experiment. The closer the correlation coefficient is to 1, the greater the similarity of the samples. ENCODE suggests that the square of the Pearson correlation coefficient should be larger than 0.92, under ideal experimental conditions. In this project, the  $R^2$  should be larger than 0.8.



**Figure 3.5.1 Correlation of RNA-seq samples**

Heat maps of the correlation coefficient between samples are shown. (If the samples are more than 4 groups, then only present the scatter diagrams between biological replicates. The scatter diagrams demonstrate the correlation coefficient between samples;  $R^2$ , the square of the Pearson coefficient.

## 3.6 Differential Expression Analysis

### 3.6.1 List of differentially expressed genes

The input data for differential gene expression analysis are readcounts from the gene expression level analysis. The differential gene expression analysis contains three steps:

- 1) Readcounts Normalization;
- 2) Model dependent p-value estimation;
- 3) FDR value estimation based on multiple hypothesis testing.

Different softwares and parameter sets are applied in different situations. The analysis methods are listed below:

Type	Software	Normalization method	p-value estimation model	FDR estimation method	Differentially expressed gene screening standard
With biological duplicates	DESeq(Anders et al, 2010)	DESeq	Negative binomial distribution	BH	padj < 0.05
Without biological duplicate	DEGseq(Wang et al, 2010)	TMM	Poisson distribution	BH	$ \log_2(\text{FoldChange})  > 1$ & $q\text{value} < 0.005$

The readcount value of the  $i$ th gene in the  $j$ th sample is  $K_{ij}$ , then

Negative binomial distribution:  $K_{ij} \sim \text{NB}(\mu_{ij}, \sigma_{ij}^2)$

Poisson distribution:  $K_{ij} \sim P(\mu_{ij})$

Table 3.6.1 List of differentially expressed genes

Gene Id	CK	treat	log2FoldChange	pval	p-adjusted
Novel00034	710.053852	1484.110192	-1.0636	2.47E-49	9.42E-48
Novel00050	8.885473304	30.72281534	-1.7898	0.00090134	0.0035153
Novel00051	25.87240756	60.95010139	-1.2362	0.00059057	0.0024209
Novel00055	36.58724302	115.9538514	-1.6641	1.1012E-09	9.5009E-09

Each column stands for:

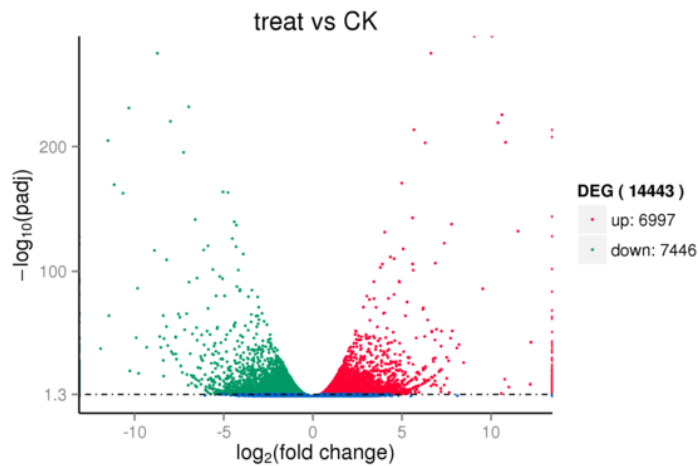
- (1) Gene id: id of differentially expressed gene
- (2) readcount\_Sample1: normalized readcount of Sample1
- (3) readcount\_Sample2: normalized readcount of Sample2
- (4) log2FoldChange:  $\log_2(\text{Sample1}/\text{Sample2})$
- (5) pvalue(pval): p-value of statistical hypothesis testing
- (6) qvalue(padj): adjusted p-value. The lower the qvalue is, the more significant the gene differentially expressed.



---

### 3.6.2 Screening of differentially expressed genes

Volcano plots are used to illustrate the overall distribution of differentially expressed genes.



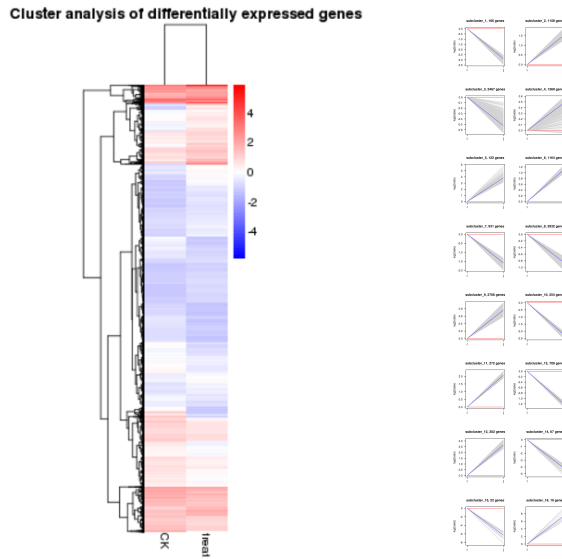
**Figure 3.6.2 Volcano plot for differentially expressed genes**

The x-axis shows the fold change in gene expression between different samples, and the y-axis shows the statistical significance of the differences. Significantly up and down regulated genes are highlighted in red and green, respectively. Genes do not show significant differential expression are in blue.

### 3.6.3 Cluster Analysis of Gene Expression Differences

Cluster analysis is used to find genes with similar expression patterns under various experimental conditions. By clustering genes with similar expression patterns, it may be possible to discern unknown functions of previously characterized genes or the function of unknown genes. In hierarchical clustering, areas of different colors denote different groups (clusters) of genes, and genes within each cluster may have similar functions or take part in the same biological process.

In addition to the FPKM cluster, the H-cluster, K-means and SOM are also used to cluster the log<sub>2</sub>(ratios). Genes within the same cluster exhibit the same trends in expression levels under different conditions.

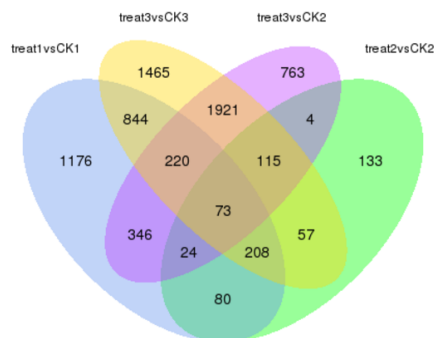


**Figure 3.6.3 Cluster analysis.**

Upper panel: the overall results of FPKM cluster analysis, clustered using the  $\log_{10}(\text{FPKM}+1)$  value. Red denotes genes with high expression levels, and blue denotes genes with low expression levels. The color range from red to blue represents the  $\log_{10}(\text{FPKM}+1)$  value from large to small. Lower panel:  $\log_2(\text{ratios})$  line chart. Each grey line in a subline chart represents the relative expression value of a gene cluster under different experimental conditions, and the blue line represents the mean value. The x-axis shows the experimental condition and the y-axis shows the relative expression level.

### 3.6.4 Venn diagram of differentially expressed genes

The Venn diagram presents the counts of differential expressed genes of all compares and the counts of genes in the intersections of different compares.



**Figure 3.6.4 Venn diagram of differentially expressed genes**

The sum of numbers in each circle is the counts of differentially expressed genes of the certain compare, and the numbers in the intersets are the counts of genes differentially expressed in all overlapped compares.

---

### 3.7 GO Enrichment Analysis

Gene Ontology (GO, <http://www.geneontology.org/>) is a major bioinformatics initiative to unify the presentation of gene and gene product attributes across all species. DEGs refer to differentially expressed genes.

GO enrichment analysis is used by Goseq (Young et al, 2010), which is based on Wallenius non-central hyper-geometric distribution. Its characteristics are: the probability of drawing an individual from a certain category is different from that of drawing it from outside of the category, and this difference is obtained from estimating of the preference of gene length.

#### 3.7.1 GO Enrichment Result List of DEGs

**Table 3.7.1 Significantly Enriched GO Terms in DEGs**

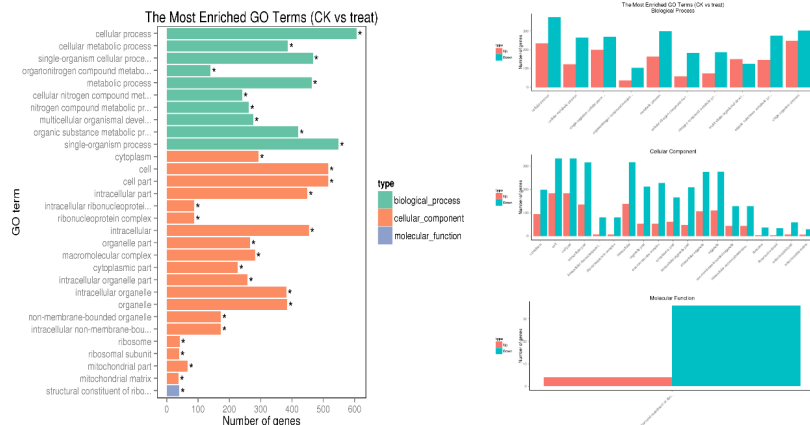
GO accession	Description	Term type	Over represented p-Value	Corrected p-Value	DEG item	DEG list
GO:0043900	regulation of multi-organism process	biological_process	0.00031841	0.46106	3	221
GO:0046999	regulation of conjugation	biological_process	0.00031841	0.46106	3	221
GO:0006544	glycine metabolic process	biological_process	0.0007502	0.52094	6	221
GO:0003735	structural constituent of ribosome	molecular_function	0.00089942	0.52094	13	221

Each column stands for:

- (1) GO accession: Gene Ontology entry
- (2) Description: Detailed description of Gene Ontology.
- (3) Term\_type: GO types, including cellular component, biological process, and molecular function.
- (4) Over\_represented\_pValue: p-value in hypergeometric test.
- (5) Corrected\_pValue: Corrected P-value; GO with corrected p-values < 0.05 are significantly enriched in DEGs.
- (6) DEG\_item: Number of DEGs with GO annotation.
- (7) DEG\_list: Number of all reference genes with GO annotation.

#### 3.7.2 Gene Ontology functional classification

There are two graphs in each group. Fig 1: The x-axis is GO terms enriched and the y-axis is the number of differential expression genes. Different colors are used to distinct biological process, cellular component and molecular function, in which the enriched GO terms are marked by “\*”. Fig 2: The GO terms in the Figure 1, which are drawn in subsets of graph based on biological process, cellular component, molecular function and differential expression genes.

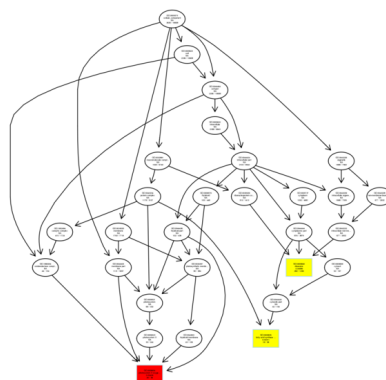


**Figure 3.7.2 Gene Ontology functional classification**

There are two graphs in each group. Fig 1: The x-axis is GO terms enriched and the y-axis is the number of differential expression genes. Different colors are used to distinct biological process, cellular component and molecular function, in which the enriched GO terms are marked by "\*". Fig 2: The GO terms in the figure 1, which are drawn in subsets of graph based on biological process, cellular component, molecular function and differential expression genes.

### 3.7.3 GO Enrichment DAG Figure

Directed Acyclic Graph (DAG) is a way to show the results of GO enrichment of DEGs. The branches represent the containment relationships, and the range of functions gets smaller and smaller from top to bottom. Generally, the top ten of GO enrichment results are selected as the master nodes in directed acyclic graph, showing the associated GO terms together via the containment relationship, and the degree of colours represent the extent of enrichment. In the project, DAG figures of biological process, molecular function and cellular component are drawn, respectively.



**Figure 3.7.3 Illustration of topGO DAG.**

Each node represents a GO term, and TOP10 GO terms are boxed. The darker the color is, the higher is the enrichment level of the term. The name and p-value of each term are present on the node.

### 3.8 KEGG Enrichment Analysis

The interactions of multiple genes may be involved in certain biological functions. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of manually curated databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. KEGG is utilized for bioinformatics research and education, including data analysis in genomics, metagenomics, metabolomics and other omics studies. Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways associated with differentially expressed genes compared with the whole genome background. The formula is:

$$p = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

Here, N is the number of all genes with a KEGG annotation, n is the number of DEGs in N, M is the number of all genes annotated to specific pathways, and m is number of DEGs in M.

#### 3.8.1 KEGG Enrichment List

Table 3.8.1 KEGG Enrichment List

#Term	Database	ID	Sample number	Background number	P-Value	Corrected P-Value
Ribosome	KEGG PATHWAY	psb03010	15	70	1.51316397372e-05	0.000741450347125
Starch and sucrose metabolism	KEGG PATHWAY	psb00500	7	25	0.000827054978654	0.020262846977
Glycine, serine and threonine metabolism	KEGG PATHWAY	psb00260	9	48	0.00184904201267	0.0302010195403
Protein export	KEGG PATHWAY	psb03060	4	17	0.0187419418178	0.186638292562

Each column stands for:

- (1) #Term: description of KEGG pathways.
- (2) ID: KEGG ID.
- (3) Sample number: number of DEGs with pathway annotation.
- (4) Background number: number of all reference genes with pathway annotation.
- (5) P-value: P-value in hypergenometric test.
- (6) Corrected P-value: Pathways with corrected p-values < 0.05 are significantly enriched in DEGs.

### 3.8.2 KEGG Enrichment Scattered Plot

Scatter diagram is a graphical display way of KEGG enrichment analysis results. In this plot, enrichment degree of KEGG can be measured through Rich factor, Qvalue and genes counts enriched to this pathway. Rich factor is the ratio of DEGs counts to this pathway in the annotated genes counts. The more the Rich factor is, the higher is the degree of enrichment. Qvalue is the adjusted p-value after multiple hypothesis testing, and its range is [0,1]. The more the qvalue is close to zero, the more significant is the enrichment. Top 20 most significant enriched pathways are chosen in KEGG scatter plot, and if the enriched pathways counts is less than 20, then put all of them into the plot. KEGG enrichment scatter diagram is as follows.

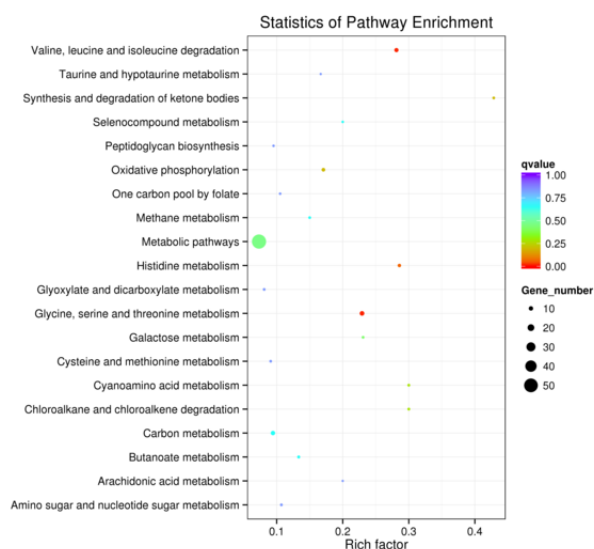


Figure 3.8.2 KEGG enrichment scatter plot of DEGs.

The y-axis shows the name of the pathway and the x-axis shows the Rich factor. Dot size represents the number of different genes and the color indicates the q-value.

### 3.8.3 KEGG Enrichment Pathway

KEGG enrichment pathway shows the DEGs significantly enriched pathways. In the diagram, nodes containing only up-regulated genes are labeled in red; nodes containing only down-regulated genes are labeled in green; nodes containing both up and down-regulated genes are in yellow.

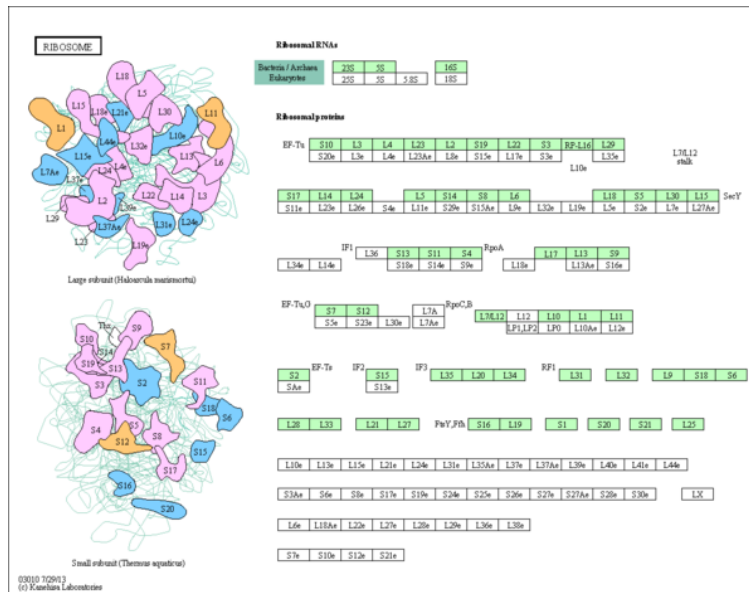


Figure 3.8.3 KEGG Enrichment Pathway

### 3.9 SNP and InDel

A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a single nucleotide in the genome, or other shared sequence, differs between members of a biological species or paired chromosomes. Two types of variation occur with SNPs, namely transitions and transversions, with a probability ratio of 1:2. SNPs occur most often in CG sequences, resulting in C to T transitions, which are associated with the tendency of C to be methylated in CG sequences. In general, a canonical SNP should be present in more than 1% of the whole population. In contrast to SNPs, INDEL refers to insertions or deletions of small fragments (one or more nucleotides) comparing to the reference genome.

Analysis tools, such as Samtools and Picard, are used to sort the reads according to the genome coordinates, followed by screening out repeated reads. Finally, GATK2 is used to carry out SNP calling and INDEL calling. After filtration, results such as those shown in the following table are obtained, in which INDEL and SNPs share the same columns. In addition, we used SnpSift for functional annotations and enrichment analysis.

**Table 3.9 SNP analysis results**

#CHROM	POS	REF	ALT	CK1	CK2	CK3	treat1	treat2	treat3
NC_999999.1	129	G	C	0,250	0,238	1,235	3,216	1,231	2,254
NC_999999.1	180	T	C	0,243	0,173	0,166	0,141	0,134	0,143
NC_999999.1	264	C	G	1,109	0,48	0,41	0,33	0,44	0,45
NC_999999.1	315	C	T	0,111	0,50	0,42	0,37	0,32	0,35

#CHROM: Chromosome/Scaffold ID.

POS: Position of SNP in corresponding chromosome/scaffold.

REF: Reference genotype.

ALT: SNP genotype (Alternative genotype).

other coloums: sample names, show the counts of reads with reference genotype and SNP genotype in each sample, respectively.

### 3.10 Novel Gene Prediction

The RNA-seq reads were assembled according to the reference genomes using Rockhopper(R. McClure, et al, 2013), and then compared to known gene structures, so that novel gene transcripts were predicted. The novel transcripts were aligned to sequences in NCBI NR database using Blastx (cutoff: evalue < 1e-5). Novel transcripts with NR annotations were considered as novel potential protein coding transcripts.

**Table 3.10 Novel Gene Prediction**

Gene_ID	Start	Stop	Strand	NR_GI	NR_ID
Novel00001	5982	6225	-	330898604	EGH30023.1
Novel00002	63915	64059	+	515849134	WP_017279887.1
Novel00003	161099	161289	-	520937825	WP_020348691.1
Novel00004	219113	219299	+	520937825	WP_020348691.1

(1) Gene\_ID: ID of novel transcript

(2) Start: transcription start position

(3) Stop: transcription end position

(4) Strand: strand specification

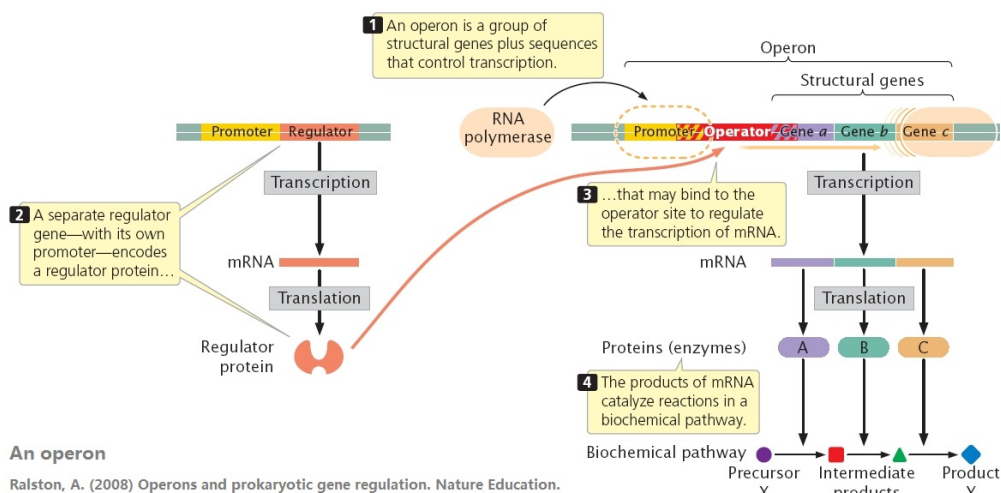
(5) NR\_GI: Genbank ID of the gene in NR database

(6) NR\_ID: serial number of the gene in NR database

### 3.11 Gene Structure Analysis

In prokaryotic genomes, functionally related genes are usually clustered and regulated by a single upstream promoter and a single downstream terminator. Such genetic structure is called operon. The genes in the same operon are transcribed together into an mRNA strand and then translated to different proteins. A typical prokaryotic operon and its regulation mechanism is shown below:





According to the locations of reads in the reference genome, transcription start sites (TSS) and transcription termination sites (TTS) of operons are predicted using Rockhopper. Then promoters prediction were applied using 700-bp sequences in the upstream of TSS, by time-delay neural network (TDNN) method.

### 3.11.1 Operon prediction

**Table 3.11.1 Results of operon prediction**

Start	Stop	Strand	Number of Genes	Genes
1	6226	+	4	PSYRH_RS00005, PSYRH_RS00010, PSYRH_RS00015, PSYRH_RS00020
7160	11549	-	4	PSYRH_RS00030, PSYRH_RS00035, PSYRH_RS00040, PSYRH_RS00045
11630	13107	+	2	PSYRH_RS00050, PSYRH_RS00055
15694	19338	-	3	PSYRH_RS00065, PSYRH_RS00070, PSYRH_RS00075

- (1) Start: the start position of the first gene
- (2) Stop: the end position of the last gene
- (3) Strand: strand specificity
- (4) Number of Genes: number of genes in the operon
- (5) Genes: list of gene names

### 3.11.2 TSS and TTS prediction

**Table 3.11.2 Results of TSS and TTS prediction**

TSS	TTS	Strand	Gene
1575	2678	+	PSYRH_RS00010
2692	3804	+	PSYRH_RS00015
3805	6245	+	PSYRH_RS00020
6225	5982	-	rna00001

- (1) TSS: position of TSS
- (2) TTS: position of TTS
- (3) Strand: strand specificity
- (4) Gene: gene name

### 3.11.3 Promoter prediction

**Table 3.11.3 Results of promoter prediction**

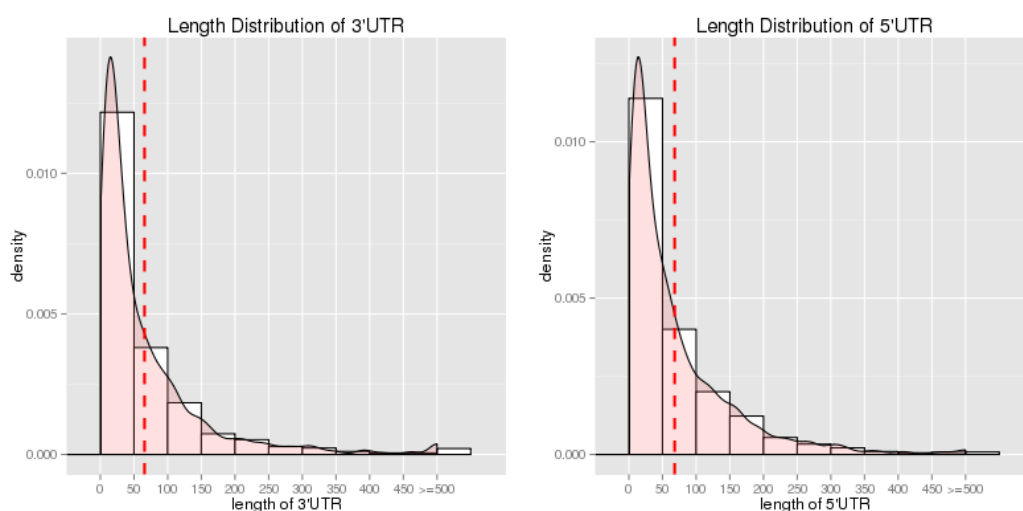
Sequence ID	strand	Start Position	End Position	Score	Sequence
PSYRH_RS21755.head	-	582	627	0.85	AGGTTGCCACCTTTTGCAGAGTCAAAGCCATGATTGTTTATCTCCTTT
sRNA00116.head	-	556	601	0.85	AGGTTGCCACCTTTTGCAGAGTCAAAGCCATGATTGTTTATCTCCTTT
PSYRH_RS21765.head	-	186	231	0.99	GGTTTCAAGGCCTCGGCACGACCAATTTCAAATCGAACTCGAACTTTG
PSYRH_RS21770.head	-	612	657	0.85	ACTGATGGAAGCACGGGCTTGTGCGAGCGTAATCTTGCTCAGACTGACGG

- (1) Sequence ID: sequence ID from input
- (2) strand: strand specificity
- (3) Start Position: start position of predicted promoter
- (4) End Position: end position of predicted promoter
- (5) Score: score of accuracy
- (6) Sequence: promoter sequence

### 3.12 UTR Analysis

#### 3.12.1 UTR prediction and UTR length distribution

The 5' and 3' UTR sequences were extracted based on the start and end positions of transcription and translation. The length distributions of both 5' and 3' UTR sequences were plotted, respectively. For 5' UTR sequences, SD sequences were predicted using RBSfinder (rbs region length = 50). For 3' UTR sequences,  $\rho$ -independent terminators were predicted using TransTermHP.



**Figure 3.12.1 UTR length distributions.**

The x axis shows length intervals of UTRs, and y axis shows densities of UTRs in different length intervals. The red dashed line shows the averaged length.

### 3.12.2 5' UTR SD sequence prediction

Table 3.12.2 Predicted SD sequences in 5' UTR region

Gene_ID	Start	Stop	Strand	Pattern	Position
PSYRH_RS00035	8538	7996	-	AGGAG	8548
PSYRH_RS00050	11630	12181	+	GGGAG	11621
PSYRH_RS00075	19338	18394	-	ATGAG	19349
PSYRH_RS00080	19898	19392	-	TGCAG	19914

- (1) gene\_id: gene id
- (2) Start: start position of gene
- (3) Stop: end position of gene
- (4) Strand: strand specificity
- (5) Pattern: SD sequence pattern
- (6) Position: start position of SD sequence

### 3.12.3 3' UTR $\rho$ -independent terminator prediction

Table 3.12.3 Predicted  $\rho$ -independent terminators in 3' UTR region

Gene_ID	Term_start	Term_end	Strand	5'_tail	5'_stem	Loop	3'_stem	3'_tail
PSYRH_RS00230	46382	46355	-	GCTGCGGGGGAAGCA	GGGACAAGCGGT	GAGC	ACCACTTGTC	TGGCAGGCAAACGTC
PSYRH_RS00435	96421	96408	-	CTAAAGCAAAAAAAAA	CCCGT	CGAA	ACGGG	CTTTTTTGTCTGTT
PSYRH_RS00440	97913	97881	-	TGCGGTACAACAAAA	GCCGATGACACGTCG	CTCT	CG-CGCGCCATCGGC	TGCAATACCCGAGTG
PSYRH_RS00760	163070	163041	-	CAGTATTCCAGACAA	GCAAACGCCCGA	CAAG	TCGAGGCITTTGC	TTGAGCATTTCGAG

- (1) gene\_id: gene id
- (2) Term\_start: start position of terminator
- (3) Term\_end: end position of terminator
- (4) strand: strand specificity
- (5) 5'\_tail: 5' tail sequence
- (6) 5'\_stem: hairpin 5' stem sequence
- (7) loop: hairpin loop sequence
- (8) 3'\_stem: hairpin 3' stem sequence
- (9) 3'\_tail: 3' tail sequence

## 3.13 Antisense Transcripts Prediction

Natural antisense transcripts (NATs) are a group of endogenous RNAs with sequences complementary to other RNA transcripts. Cis-NATs are transcribed from the opposite DNA strand but the same genomic loci of their sense transcripts. Cis-NATs widely exist in both prokaryotes and eukaryotes and play important roles in gene regulation. According to orientations in the genome, Cis-NATs are classified to three types: enclosed (full overlapped by the sense transcript), convergent (3'-3' overlap) and divergent (5'-5' overlap). The genomic loci, type and numbers of cis-NATs were identified using the strand-specific RNA-seq data.

**Table 3.13 Results of Antisense Transcripts Prediction**

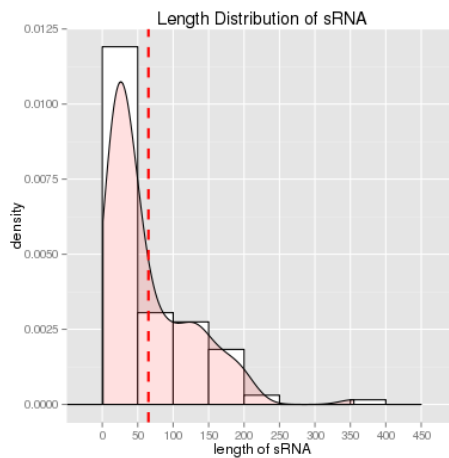
Plus_gene_id	Plus_start	Plus_end	Plus_length	Plus_name(description)	Minus_gene_id	Minus_start	Minus_end	Minus_length
Novel00005	317022	317206	185	(novelGene)	PSYRH_RS01520	317018	317821	804
Novel00008	418580	418673	94	(novelGene)	PSYRH_RS01975	416522	419037	2516
Novel00009	506418	506596	179	(novelGene)	PSYRH_RS02370	506218	506934	717
Novel00010	553781	554131	351	(novelGene)	PSYRH_RS02520	553414	554166	753

- (1) plus\_transcript\_id: id of sense transcript
- (2) plus\_start: start position of sense transcript
- (3) plus\_end: end position of sense transcript
- (4) plus\_length: length of sense transcript
- (5) plus\_name(description): name (description) of sense transcript
- (6)~(10): The same information as (1)~(5) but for anti-sense transcript
- (11)types: type of anti-sense transcript (including enclosed, convergent and divergent)
- (12)overlap\_start: start position of overlapped region
- (13)overlap\_end: end position of overlapped region
- (14)overlap\_length: length of overlapped region

### 3.14 sRNA Analysis

In prokaryotes, non-coding RNAs with length between 50 and 500 nt are defined as small RNA (sRNA). Novel intergenic transcripts were discovered by Rockhopper, and aligned with sequences in NCBI NR database using Blastx. Novel transcripts without NR annotation were sRNA candidates. RNAfold and IntaRNA were applied to predict the secondary structures and targeted genes of sRNAs, respectively.

#### 3.14.1 sRNA prediction and length distribution



**Figure 3.14.1 sRNA length distribution**

The x axis shows length intervals of sRNAs, and y axis shows densities of sRNAs in different length intervals. The red dashed line shows the averaged length.

### 3.14.2 sRNA secondary structure prediction

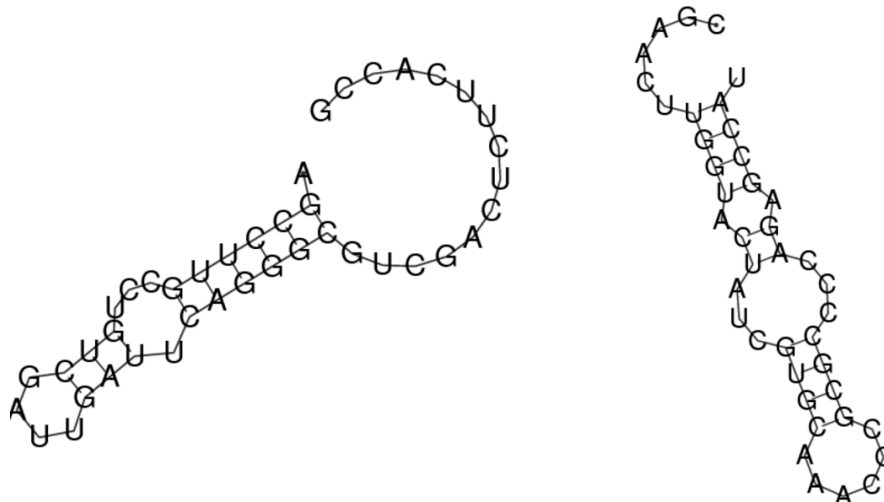


Figure 3.14.2 sRNA secondary structures

### 3.14.3 sRNA targeted gene prediction

Table 3.14.3 Predicted sRNA targeted genes

sRNA_id	mRNA_id	energy(kcal/mol)	sRNA_position	mRNA_position
sRNA00001	PSYRH_RS06180	-11.8829	28 -- 42	457 -- 473
sRNA00002	PSYRH_RS06180	-13.885	2 -- 18	716 -- 729
sRNA00003	PSYRH_RS06180	-9.29651	30 -- 47	161 -- 179
sRNA00004	PSYRH_RS06180	-10.8061	1 -- 16	942 -- 957

- (1) sRNA\_id: sRNA id
- (2) mRNA\_id: target gene id
- (3) energy(kcal/mol): Free energy
- (4) sRNA\_position: sRNA binding position
- (5) mRNA\_position: target gene binding position

### 3.14.4 sRNA Expression Quantification

Table 3.14.4 Results of sRNA expression quantification

sRNA_id	CK1	CK2	CK3	treat1	treat2	treat3
sRNA00110	997.9491536	1350.219897	1015.237987	1101.540375	1015.237987	1101.540375
sRNA00105	969.4659137	690.8136415	915.0549281	831.8828702	915.0549281	831.8828702
sRNA00103	933.9579215	1478.629578	462.4471142	389.2721515	462.4471142	389.2721515
sRNA00007	92.61158118	139.7485272	233.1034234	459.15125	233.1034234	459.15125

sRNA\_id: sRNA id

other columns: fpkm of sRNA in each sample

---

## 4 Appendix

### 4.1 Result Directory Lists

Click to open the result directory.(Note: Please make sure the report directory and the result directory is under the same directory). Result Directory Lists: [html](#)

```
../NHHWXXXXXX_Prokaryotes_results
├── 1. OriginalData: Raw Data (fastq format)
├── 2. QC: Data Quality Control
│   ├── 2.1. ErrorRate: RNA-seq Error Rate
│   ├── 2.2. GC:GC Content Distributions
│   ├── 2.3. ReadsClassification: Classification of reads
│   └── 2.4. DataTable: Data Quality Control Summary
├── 3. Mapping: Results of Mapping to a Reference Genome
│   ├── 3.1. MapStat: Overview of Mapping Results
│   ├── 3.2. MapReg: Mapped Regions in Reference Genome (exons,
introns, or intergenic regions)
│   ├── 3.3. ChrDen: Distribution of Mapped Reads in Chromosomes
│   └── 3.4. IGV: Visualization of Mapping Status of Reads using IGV
├── 4. GeneExprQuatification: Expression Quantification
│   ├── 4.1. GeneExprQuatification: Gene Expression Quantification
│   └── 4.2. GeneExpContrast: Contrast of Gene Expression Levels
├── 5. AdvancedQC: RNA-seq Advanced QC
│   └── 5.1. Correlation: RNA-Seq Correlations
├── 6. DiffExprAnalysis: Gene Differential Expression Analysis
│   ├── 6.1. DEGsList: List of Differentially Expressed Genes (all,
up-regulated, down-regulated)
│   ├── 6.2. DEGsFilter: Volcano plot
│   ├── 6.3. DEGcluster: Cluster Analysis of Gene Expression Differences
│   │   └── Subcluster
│   └── 6.4. VennDiagram: The Venn Diagrams
├── 7. DEG_GOEnrichment: GO Enrichment Analysis of DEGs
│   ├── 7.1. DEG_GOList: GO Enrichment Result List of DEGs
│   ├── 7.2. DAG: GO Enrichment DAG Figure
│   └── 7.3. BAR: GO Enrichment Bar Chart of DEGs
├── 8. DEG_KEGGenrichment: KEGG Pathway Enrichment Analysis of
DEGs
│   ├── 8.1. DEG_KEGGList: KEGG Enrichment List
│   ├── 8.2. DEG_KEGGScat: KEGG Enrichment Scattered Plot
│   └── 8.3. DEG_KEGGPath: KEGG Enrichment Pathways
```

- |           |—— ALL
- |           |—— DOWN
- |           |—— UP
- |—— 9. SNP: SNP and Indel Results
- |—— 10. NovelGene: Novel Gene Prediction
- |—— 11. GeneStruct
- |—— 12. UTR
- |—— 13. AntiTrans: Antisense Transcript Analysis
- |—— 14. sRNA
  - |—— 14.1 sRNA\_Length
  - |—— 14.2 Secondary\_Structure
  - |—— 14.3 sRNA\_Inta
  - |—— 14.4 sRNA\_ExprQuantification

## 4.2 Software List

### Software and Parameter

Analysis	Software	Version	Remarks
Mapping	Bowtie2	2.2.3	Mapping to a reference genome.
Expression Quantification	HTSeq	v0.6.1	
Novel Gene Prediction	Rockhopper	1.2.1	
SNP detection	GATK2	v3.2	
Differential Expression Analysis	DEGSeq	1.12.0	For sample with bio-replicate using DESeq; samples without bio-replicate using DESeq, EdgeR for specific conditions.
	DESeq	1.10.1	
	edgeR	3.0.8	
GO Enrichment	GOSeq topGO,hmmscan	Release2.12	Hmmscan is used to get GO annotation for novel genes.
KEGG Enrichment	KOBAS	v2.0	
Operon and TSS/TTS Prediction	Rockhopper	1.2.1	
SD Sequence prediction	RBSfinder	v1.0	
Rho-independent Terminator Sequence Prediction	TranstermHP	v2.0.9	
sRNA Secondary Structure Prediction	RNAfold	2.0	
sRNA Targeted Gene Prediction	IntaRNA	v1.2.5	





---

## 5 References

Marioni, J. C., C. E. Mason, et al. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*.

Mortazavi, A., B. A. Williams, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*.

Wang, Z., M. Gerstein, et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*.

Busch, A., A. S. Richter, et al. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*. (IntaRNA)

Hofacker, I. L. and P. F. Stadler (2006). Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*. (RNAfold)

Kingsford, C. L., K. Ayanbule, et al. (2007). Rapid, accurate, computational discovery of Rho-independent transcription terminators Illuminates their relationship to DNA uptake. *Genome biology*. (TransTermHP)

McClure, R., D. Balasubramanian, et al. (2013). Computational analysis of bacterial RNA-Seq data. *Nucleic acids research*. (Rockhopper)

Suzek, B. and M. Ermolaeva S. Schreiber and SL Salzberg. (2001). A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*. (RBSfinder)

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. (Bowtie)

Langmead, B. and S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*. (Bowtie 2)

Anders, S. (2010). HTSeq: Analysing high-throughput sequencing data with Python. (HTSeq)

McKenna, A, Hanna, M, Banks, E, Sivachenko, A, Cibulskis, K, Kernytzky, A, Garimella, K, Altshuler, D, Gabriel, S, Daly, M, DePristo, MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. (GATK)

---

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* (DESeq)

Anders, S. and Huber, W. (2012). Differential expression of RNA-Seq data at the gene level-the DESeq package. (DESeq)

Wang, L.Feng, Z.Wang, X.Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics.* (DEGseq)

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (edgeR)

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology.* (GOseq)

Kanehisa, M., M. Araki, et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic acids research.* (KEGG)

Mao, X., Cai, T., Olyarchuk, J.G., Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics.* (KOBAS)

Waibel, A. H., Hanazawa, T., Hinton, G. E., Shikano, K., Lang, K. J.. (1989). Phoneme Recognition Using Time-Delay Neural Networks *IEEE Transactions on Acoustic, Speech, and Signal Processing* Vol. 37 no.3, 328-339. (TDNN)