RNA-seq Analysis with Reference Genome

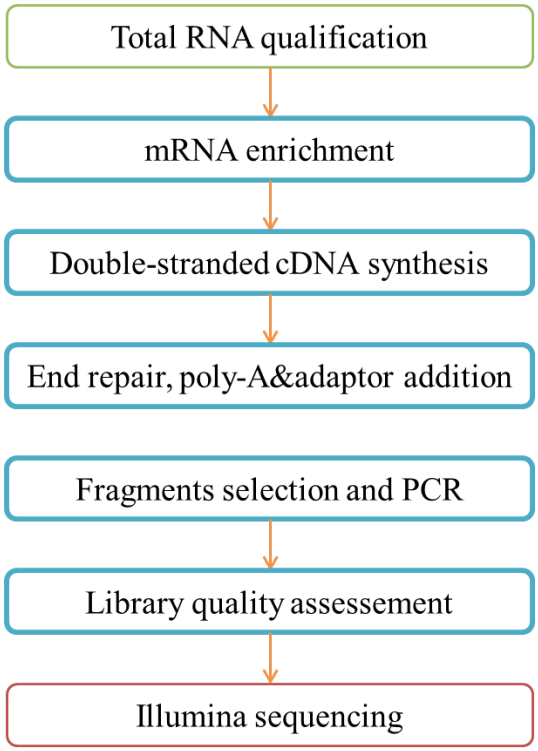
Demo Report

May 1, 2016

Contents

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

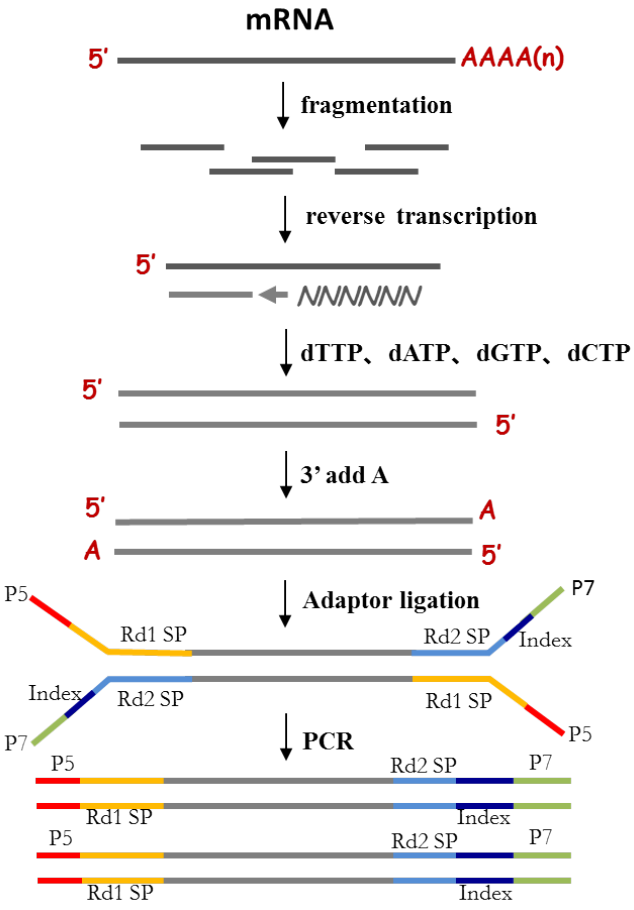
(2) Nanodrop: tests RNA purity (OD260/OD280)

(3) Qubit: quantifies the RNA (determines concentration)

(4) Agilent 2100: checks RNA integrity

1.2 Library Construction

After the QC procedures, mRNA from eukaryotic organisms is enriched using oligo(dT) beads. For prokaryotic samples, rRNA is removed using a specialized kit that leaves the mRNA. The mRNA from either eukaryotic or prokaryotic sources is then fragmented randomly in fragmentation buffer, followed by cDNA synthesis using random hexamers and reverse transcriptase. After first-strand synthesis, a custom second-strand synthesis buffer (Illumina) is added with dNTPs, RNase H and Escherichia coli polymerase I to generate the second strand by nick-translation. The final cDNA library is ready after a round of purification, terminal repair, A-tailing, ligation of sequencing adapters, size selection and PCR enrichment. The workflow chart is as follows:



1.3 Library QC

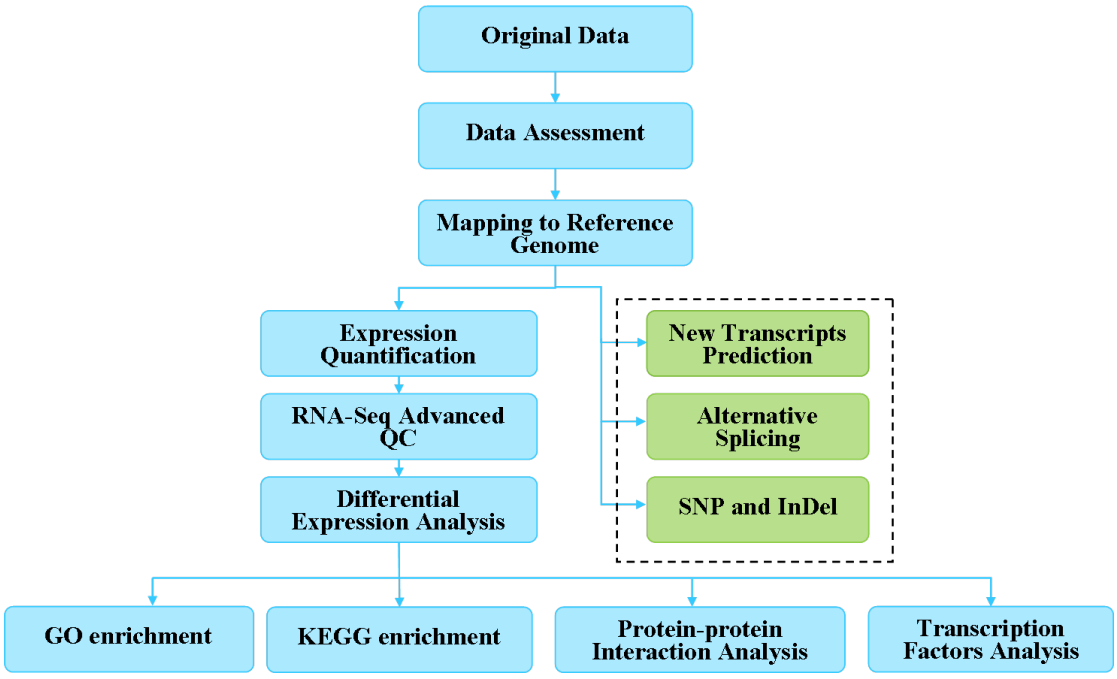
Library concentration was first quantified using a Qubit 2.0 fluorometer (Life Technologies), and then diluted to 1 ng/μl before checking insert size on an Agilent 2100 and quantifying to greater accuracy by quantitative PCR (Q-PCR) (library activity >2 nM).

1.4 Sequencing

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

2 Analysis Workflow

The analysis workflow for data without a reference genome is as follows:



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.

Every read in FASTQ format is stored in four lines as follows:

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHHFAE0ECFFD/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 can be transformed by the Phred score as in equation 1(equation 1: Qphred = -10log10(e)). The relationship between Phred quality scores Q and base-calling error "e" is given below:

Base Quality and Phred score relationship with the Illumina CASAVA v1.8 software:

Phred score	Base Calling error rate	Base Calling correct rate	Q-sorce
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 depend on the sequencing machine, reagent availability, and the samples.

(1) Error rate increases as the sequencing reads are extended and sequencing reagents become more and more scarce.

(2) The first six bases have a relatively high error rate due to the random hexamers used in priming cDNA synthesis (Jiang et al.).

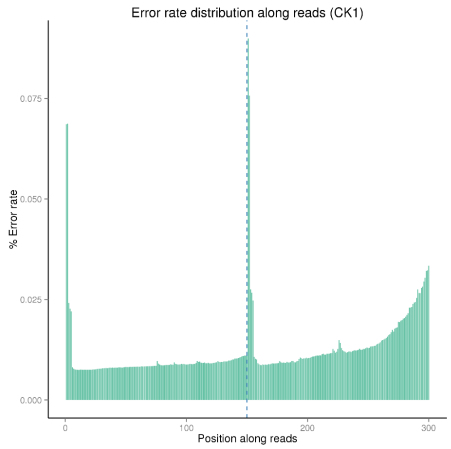


Figure 3.2.1 Error Rate Distribution

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

3.2.2 A/T/G/C Content Distribution

GC content distribution is evaluated to detect potential AT/GC separation, which affects subsequent gene expression quantification. Theoretically, G should equal C, and A should equal 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 use of random primers in library construction.

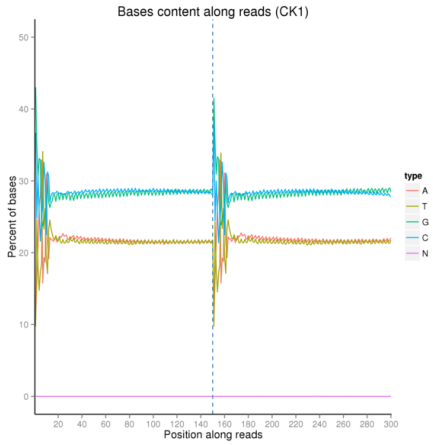


Figure 3.2.2 GC content distribution

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^{TM} RNA and DNA Sample Prep Kits):

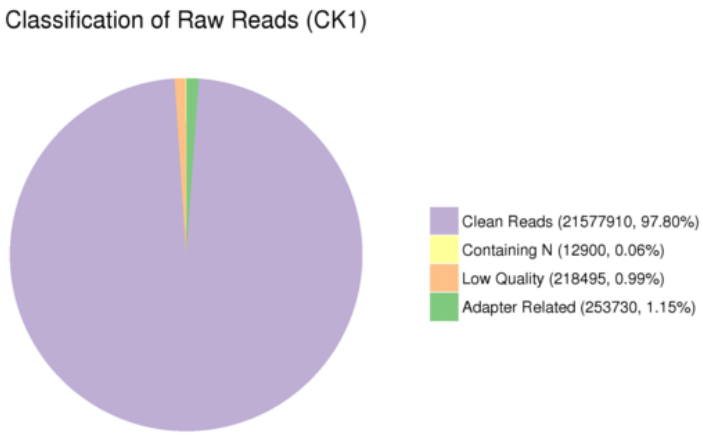


Figure 3.2.3 Raw Reads Components

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 2.1 Data Production

Sample name	Raw reads	Clean reads	Clean bases	Error rate(%)	Q20(%)	Q30(%)	GC content(%)
CK1	64702586	61323654	9.2G	0.01	97.48	93.96	52.14
CK2	59175436	56052998	8.41G	0.01	97.58	94.16	52.11
CK3	57777524	54885692	8.23G	0.01	97.58	94.18	52.14
treat1	46226950	43922506	6.59G	0.01	97.53	94.06	51.73
treat2	51332556	48744902	7.31G	0.01	97.56	94.14	51.94
treat3	46276058	43921706	6.59G	0.01	97.44	93.86	51.79

Detail statistics of sequencing data:

(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 $Qphred=-10\log_{10}(e)$

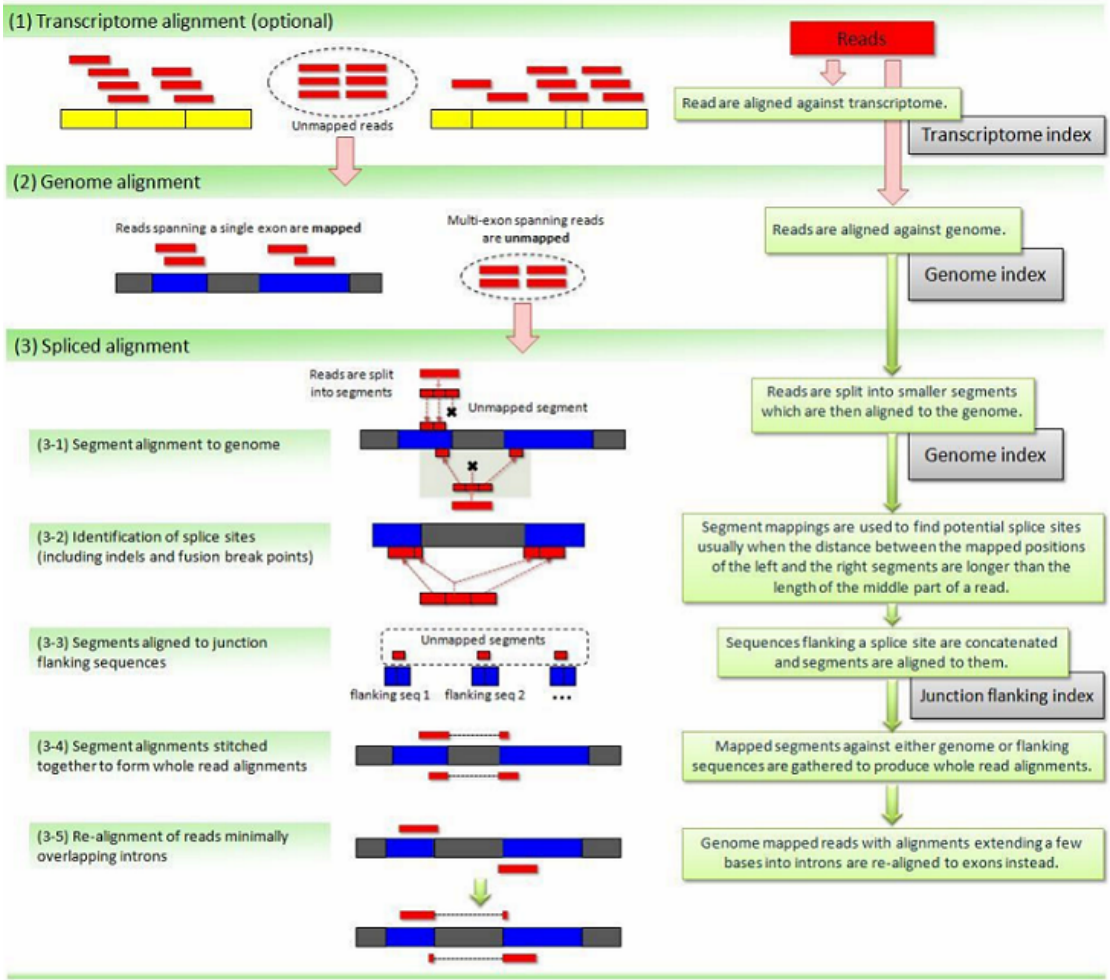
(6) Q20: percentages of bases whose correct base recognition rates are greater than 99% in total bases

(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 a Reference Genome

Algorithm for mapping sequences: appropriate software is chosen according to the characteristics of the reference genome. In general, TopHat2 is chosen for animal and plant genomes, and 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. Appropriate parameters are also set, such as the longest intron length. Only filtered reads are used to analyze the mapping status of RNA-seq data to the reference genome. This process is shown by the following figure:



The TopHat2 algorithm can be divided into three parts:

(1) Align reads to a reference transcriptome (optional).

(2) Map reads to the exons.

(3) Reads are segmented and then mapped to the adjacent exons.

When the reference genome is appropriate and the experiment is contamination-free, the TMR (Total Mapped Reads or Fragments) should be larger than 70% and MMR (Multiple Mapped Reads or Fragments) should be no more than 10%.

3.3.1 Overview of Mapping Status

Table 3.1 Overview of Mapping Status

Sample_name	CK1	CK2	CK3	treat1	treat2	treat3
Total reads	61323654	56052998	54885692	43922506	48744902	43921706
Total mapped	52075296 (84.92%)	47491790 (84.73%)	46740224 (85.16%)	37360618 (85.06%)	41258260 (84.64%)	37088412 (84.44%)
Multiple mapped	755991 (1.23%)	753703 (1.34%)	637373 (1.16%)	576324 (1.31%)	535660 (1.1%)	642777 (1.46%)
Uniquely mapped	51319305 (83.69%)	46738087 (83.38%)	46102851 (84%)	36784294 (83.75%)	40722600 (83.54%)	36445635 (82.98%)
Reads map to '+'	25633110 (41.8%)	23350492 (41.66%)	23033558 (41.97%)	18375086 (41.84%)	20343808 (41.74%)	18211638 (41.46%)
Reads map to '-'	25686195 (41.89%)	23387595 (41.72%)	23069293 (42.03%)	18409208 (41.91%)	20378792 (41.81%)	18233997 (41.51%)
Non-splice reads	37477100 (61.11%)	34145053 (60.92%)	33975152 (61.9%)	26548912 (60.44%)	29472150 (60.46%)	26693312 (60.77%)
Splice reads	13842205 (22.57%)	12593034 (22.47%)	12127699 (22.1%)	10235382 (23.3%)	11250450 (23.08%)	9752323 (22.2%)

Mapping Results Details:

(1) Total number of filtered reads (Clean data).

(2) Total number of reads that can be mapped to the reference genome. In general, this number should be larger than 70% when there is no contamination and the correct reference genome is chosen.

(3) Number of reads that can be mapped to multiple sites in the reference genome. This number is usually less than 10% of the total.

(4) Number of reads that can be uniquely mapped to the reference genome.

(5) Number of reads that map to the positive strand (+) or the minus strand (-).

(6) Splice reads can be segmented and mapped to two exons (also named junction reads), whereas non-splice reads can be mapped entirely to a single exon. The ratio of splice reads depends on the insert size used in the RNA-seq experiments.

3.3.2 Mapped Regions in Reference Genome

Mapped regions can be classified as exons, introns, or intergenic regions. Exon-mapped reads should be the most abundant type of read when the reference genome is well-annotated. Intron-reads may be derived from pre-mRNA contamination or intron-retention events from alternative splicing. Reads mapped to intergenic regions are mainly because of weak annotation of the reference genome.

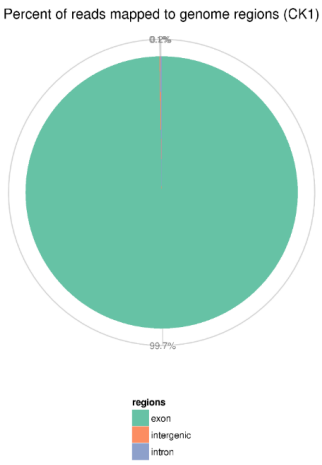


Figure 3.1 Classification of Reads According to Mapped Region.

3.3.3 Distribution of Mapped Reads in Chromosomes

To obtain an overview of the distribution of mapped reads in chromosomes, the "window size" is set to 1K, the median number of reads mapped to the genome inside the window is calculated, and transformed to the log2value. In general, the longer the whole chromosome, the more total number of mapped reads within it would be(Marquez et al.).

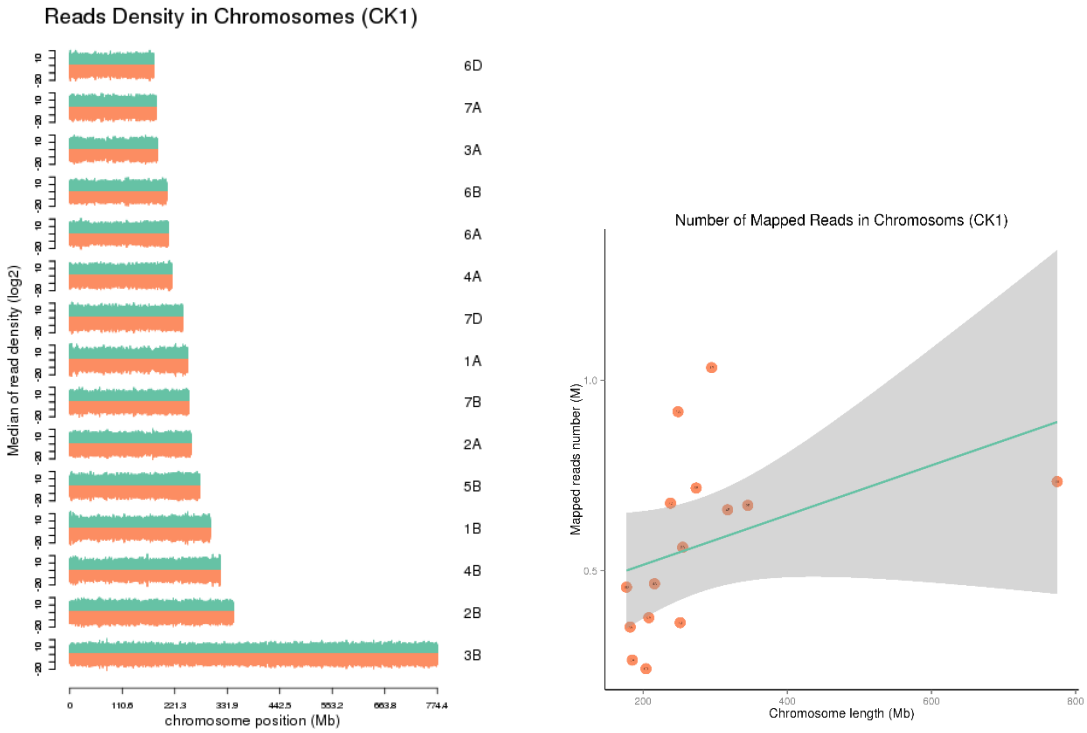


Figure 3.2 Distribution Plot of Mapped Reads in Chromosomes.

Two panels are shown for each sample. In the left panel, the X-axis shows the length of the chromosomes (in Mb), and the Y-axis indicates the $\log_2$ of the median of read density. Green and red indicate, respectively, the positive and negative strands. In the right panel, the X-axis shows the length of the chromosomes, and the Y-axis indicates the number of mapped reads in each chromosome. The grey region indicates the 95% confidence interval.

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.

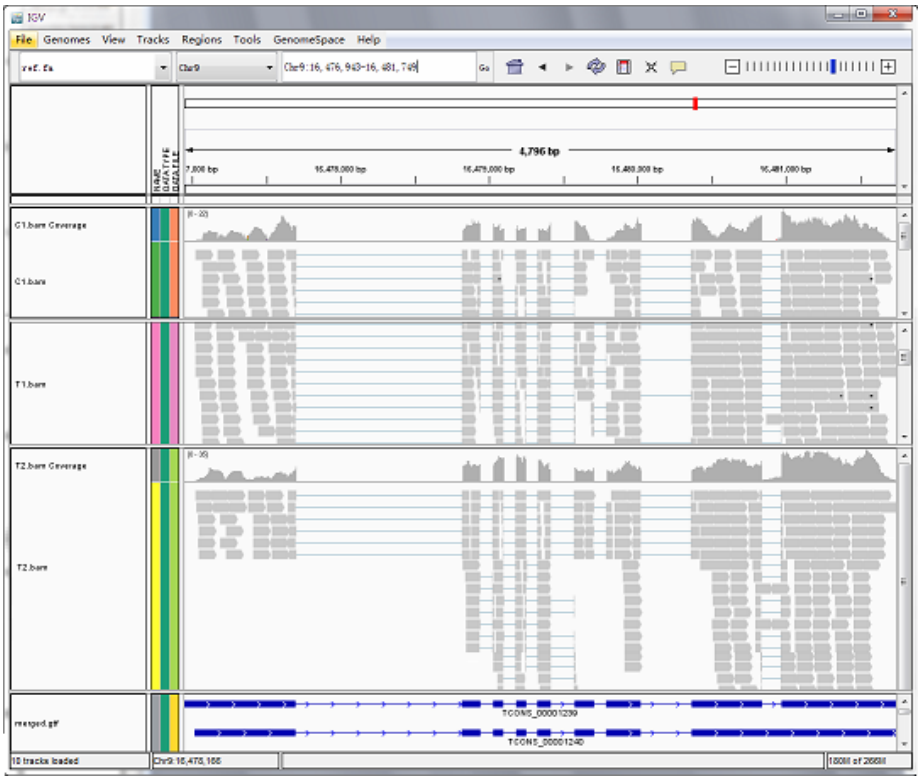
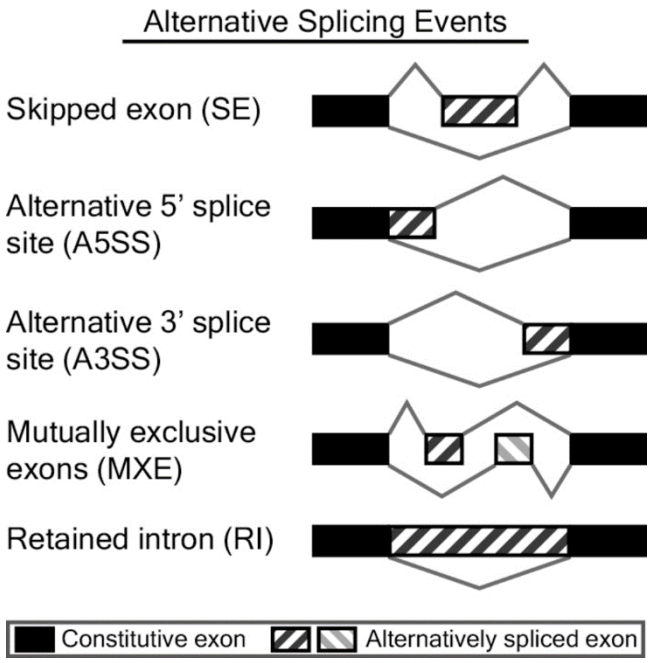


Figure 3.3 IGV interface

3.4 Alternative Splicing Analysis

Alternative splicing (AS) is a universal gene regulation mechanism in most eukaryotes. Eukaryotic gene sequences consist of intronic and exonic regions. During RNA processing, the exons are retained in mature mRNA while introns are excluded by spliceosome. Some pre-mRNAs may have different splicing patterns in different condition and yeild different protein isoforms, which increases the biologocial complexity and adaptability of eukaryotic species.

rMATS (replicate multivariate analysis of transcript splicing) is designed for detection of differential alternative splicing from RNA-seq data. rMATS uses a hierarchical model to simultaneously account for sampling uncertainty in individual replicates and variability among replicates. The classification of alternative splicing events by rMATS are defined below:



The 5 AS types can be defined as:

(1) SE: Skipped exon

(2) MXE: Mutually exclusive exon

(3) A5SS: Alternative 5' splice site

(4) A3SS: Alternative 3' splice site

(5) RI: Retained intron

3.4.1 Classification and statistics of AS Events

Classification and statistics of AS events were applied to each group of RNA-seq data with biological replicates. Then the quantitative level of each class of alternative

splicing events was estimated, and differential AS analysis between treatment and control groups are applied. rMATS adopts two quantification methods parallelly, namely evaluating splicing with reads span splicing junctions only, and with both reads on target and reads span splicing junctions. The difference between two methods is that reads targeting alternatively spliced exons (striped regions in the above figure) are eliminated by the method of evaluating splicing with only reads span splicing junctions. Customers may choose one of the evaluating methods for further studies accordingly.

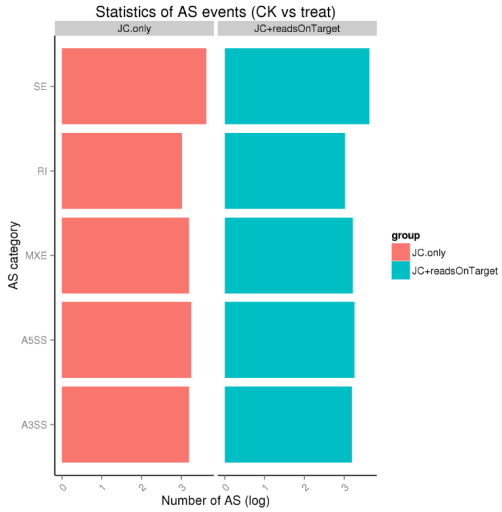


Figure 4.1 Classification of AS Events.

The Y-axis illustrates the 5 types of AS events, and the X-axis illustrates the counts for each type of AS events, respectively. JC. only: only the reads span splicing junctions are taken into account; JC + reads On Target: both the reads span splicing junctions and the reads on target are taken into account.

Table 4.1 Statstistics of AS Events

EventType	NumEvents.JC.only	SigEvents.JC.only	NumEvents.JC+readsOnTarget	SigEvents.JC+readsOnTarget
SE	2717	4 (2:2)	2765	7 (4:3)
MXE	690	2 (1:1)	727	2 (1:1)
A5SS	1536	4 (2:2)	1585	6 (3:3)
A3SS	1361	0 (0:0)	1391	0 (0:0)
RI	914	2 (1:1)	922	2 (1:1)

(1) event_type: AS event types(SE,MXE,A5SS,A3SS,RI).

(2) NumEvents.JC.only: the total number of AS events, with only reads span splicing junctions taken into account.

(3) SigEvents.JC.only: the total number of differential AS events, with only reads span splicing junctions taken into account(up:down).

(4) NumEvents.JC+readsOnTarget: the total number of AS events, with both reads span splicing junctions and reads on target exons taken into account.

(5) SigEvents.JC+readsOnTarget: the total number of differential AS events, with both reads span splicing junctions and reads on target exons taken into account.

3.4.2 Statistics on expression level of different AS types for individual genes

We analyze the expression level of different AS types for individual genes in both treatment and control groups of samples. The threshold of differential AS analysis is set as FDR < 0.05. The AS events in the below table were evaluated with only reads span splicing junctions.

Table 4.2 Overview of expression level of different AS types for individual genes

ID	GeneID	geneSymbol	chr	strand	IJC_SAMPLE_1	SJC_SAMPLE_1	IJC_SAMPLE_2	SJC_SAMPLE_2	PValue	FDR
1409	"FBgn0027950"	"MBD-like"	3R	+	430	127	191	134	1.15E-05	0.009071344
2682	"FBgn0027580"	"CG1516"	2R	-	7	195	19	52	1.34E-05	0.009071344
3029	"FBgn0034504"	"CG8929"	2R	+	308	226	138	241	4.31E-06	0.009071344
3088	"FBgn0034051"	"Mlf"	2R	-	72	213	137	139	1.25E-05	0.009071344

(1) ID: Unique AS event ID given by rMATS.

(2) GeneID: gene ID for genes with AS events.

(3) gene symbol: gene symbol, 'NA' for none.

(4) chr: Chromosome ID.

(5) strand: Strand specificity.

(6) IJC_SAMPLE_1: inclusion junction counts for SAMPLE_1, replicates are separated by comma.

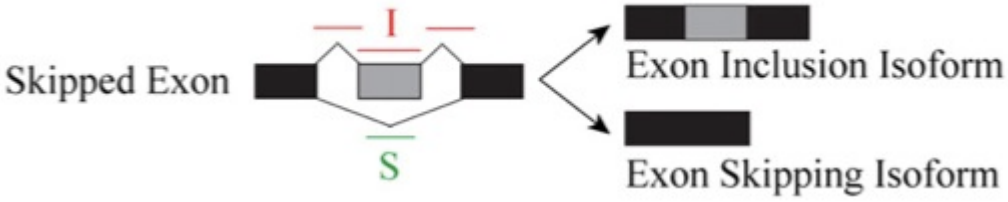
(7) SJC_SAMPLE_1: exclusion junction counts for SAMPLE_1, replicates are separated by comma.

(8) IJC_SAMPLE_2: inclusion junction counts for SAMPLE_2, replicates are separated by comma.

(9) SJC_SAMPLE_2: exclusion junction counts for SAMPLE_2, replicates are separated by comma.

(10) P value: p-value.

(11)FDR: adjusted p-value.



3.5 Novel Gene Prediction

3.5.1 Novel Gene Prediction

Mapping information from all samples is combined and placed as input into the regular Cufflinks assembler. The assembled transcriptomes are then compared to the reference transcripts to determine if they are sufficiently different to be considered novel. In brief, in this process we can (1) identify novel genes, (2) identify novel exons of novel genes, and (3) optimize the start and end information of known transcripts. The outputs are provided as GTF files; more information about GTF format is available at (http://mblab.wustl.edu/GTF22.html).

Table 5.1 Annotation for novel transcripts

seqname	source	feature	start	end	score	strand	frame	attributes
211000022278312	novelGene	exon	931	1010	.	-	.	gene_id "Novel00001"; transcript_id "Novel00001.1"; exon_number "2";
211000022278312	novelGene	exon	1120	1258	.	-	.	gene_id "Novel00001"; transcript_id "Novel00001.1"; exon_number "3";
211000022279056	novelGene	exon	465	472	.	+	.	gene_id "Novel00002"; transcript_id "Novel00002.1"; exon_number "1";
211000022279056	novelGene	exon	540	560	.	+	.	gene_id "Novel00002"; transcript_id "Novel00002.1"; exon_number "2";

1) seqname: Chromosome ID.

(2) source: Source ID.

(3) feature: Structure type.

(4) start: Start coordinate

(5) end: End coordinate

(6) score: Not related

(7) strand: Strand specificity

(8) frame: Not related

(9) attributes: Includes gene ID, transcript ID etc.

3.5.2 Optimization of known gene attributes

Table 5.2 Optimization of known gene attributes

ID	GeneID	geneSymbol	chr	strand	IJC_SAMPLE_1	SJC_SAMPLE_1	IJC_SAMPLE_2	SJC_SAMPLE_2	PValue	FDR
1409	"FBgn0027950"	"MBD-like"	3A	+	430	127	191	134	1.15E-05	0.009071344
2682	"FBgn0027580"	"CG1516"	4A	-	7	195	19	52	1.34E-05	0.009071344
3029	"FBgn0034504"	"CG8929"	7D	+	308	226	138	241	4.31E-06	0.009071344
3088	"FBgn0034051"	"Mlf"	6B	-	72	213	137	139	1.25E-05	0.009071344

(1) Gene_id: Unique gene ID from the reference GTF file.

(2) Chromosome: Chromosome/scaffold ID

(3) Strand: Strand specificity

(4) Original_span: Gene start and end positions in reference.

(5) Assembled_span: Gene start and end positions in assembled transcriptome.

3.6 SNP & InDel

3.6.1 SNP & 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) when 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, GATK3 is used to carry out SNP calling and INDEL calling. After filtering, results such as those shown in the following table are obtained, in which INDEL and SNPs share the same columns.

Table 6.1 SNP results

#CHROM	POS	REF	ALT	CK1	CK2	CK3	treat1	treat2	treat3	Gene_id
3A	10610	G	T	0,131	0,137	0,94	303,0	359,0	275,0	FBgn0085664
4A	1122087	T	A	NA	2,0	3,0	1,0	3,2	1,2	FBgn0050428
7D	66513	G	A	152,309	492,585	598,916	322,493	158,320	208,742	FBgn0039959
6B	3353682	G	A	7,6	NA	NA	1,32	NA	NA	FBgn0002645

#CHROM: Chromosome/Scaffold ID of SNPs.

POS: Position of SNPs on corresponding chromosome/scaffold.

REF: Reference genotype.

ALT: SNP genotype (Alternative genotype).

Gene_id: Gene ID from reference GTF file.

other coloums: Lettered columns show the number of reads supporting either the reference genotype or SNP genotype in each sample.

3.7 Expression Quantification

3.7.1 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 on counting 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 3.7.1 The number of genes with different expression levels

FPKM Interval	CK1	CK2	CK3	treat1	treat2	treat3
0~1	10133(57.46%)	10089(57.21%)	10025(56.85%)	10134(57.47%)	10104(57.30%)	9971(56.54%)
1~3	789(4.47%)	795(4.51%)	848(4.81%)	852(4.83%)	832(4.72%)	951(5.39%)
3~15	2224(12.61%)	2233(12.66%)	2222(12.60%)	2296(13.02%)	2284(12.95%)	2244(12.72%)
15~60	3114(17.66%)	3139(17.80%)	3139(17.80%)	2954(16.75%)	2978(16.89%)	3041(17.24%)
>60	1375(7.80%)	1379(7.82%)	1401(7.94%)	1399(7.93%)	1437(8.15%)	1428(8.10%)

Table 3.7.2 Gene expression levels

Gene_id	CK1	CK2	CK3	treat1	treat2	treat3
FBgn0085309	34.6327893	99.32356815	0	59.26433452	69.23846564	0
FBgn0267012	23.32232532	0	0	0	78.32945691	0
FBgn0061492	45.55843284	48.95929976	46.77138648	74.0914727	71.87907232	63.49794895
FBgn0053795	0.068977629	0.303770898	0.153895206	0.288996229	0.347646041	0.194675997

3.7.1 Comparison between Gene Expression Levels

To compare gene expression levels under different conditions, an FPKM distribution diagram and violin Plot are used. For biological replicates, the final FPKM would be the mean value.

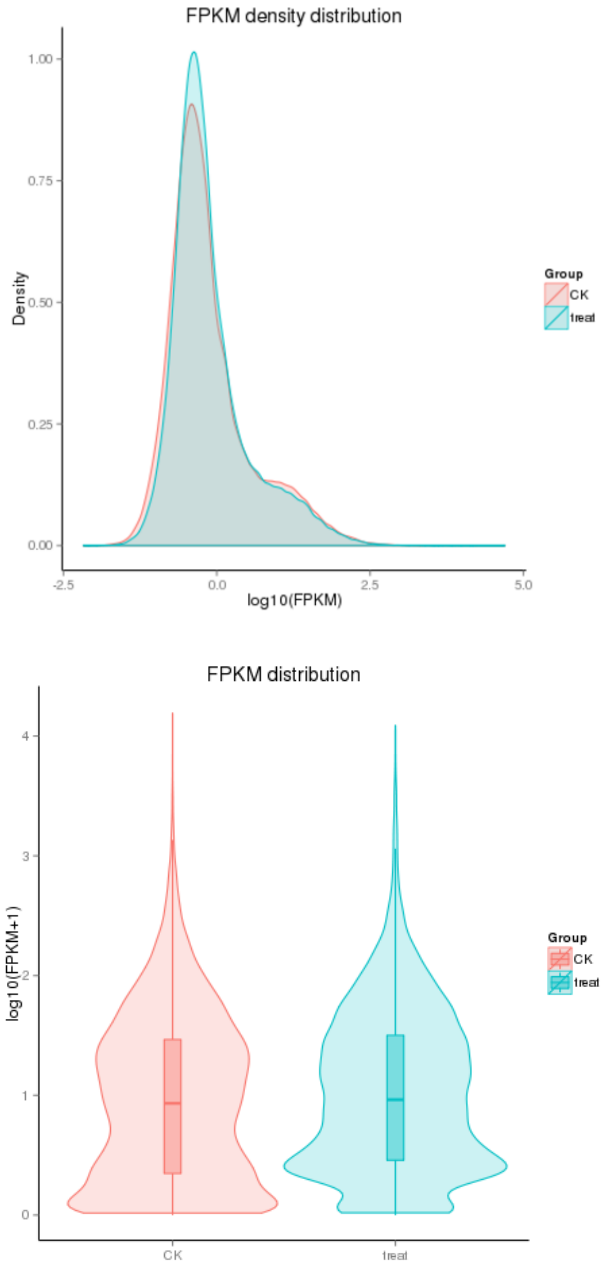


Figure 3.7.1 Different gene expression levels under different experiment conditions.

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

3.8 RNA-seq Advanced QC

3.8.1 RNA-Seq Correlation

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 R2 should be larger than 0.8.

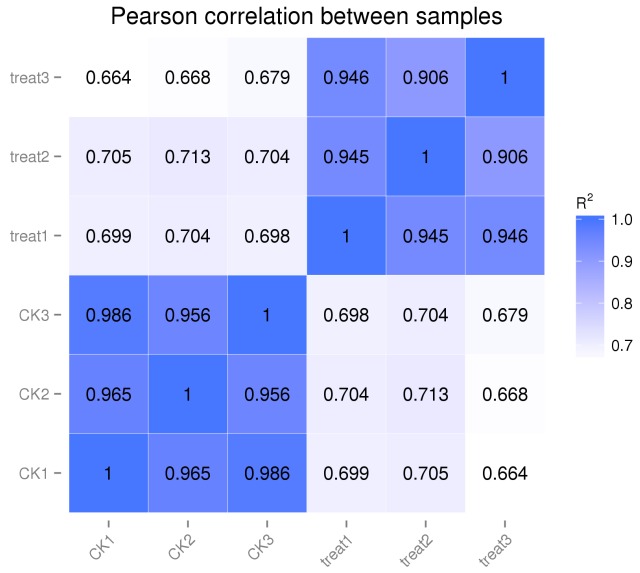


Figure 8.1 RNA-Seq correlation.

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; R2, the square of the Pearson coefficient.

3.9 Differential Gene Expression Analysis

3.9.1 List of Differentially Expressed Genes

The input data for differential gene expression analysis are readcounts from 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 software and parameter sets are applied in different situations. The analysis methods are listed below:

Type	Software	Normalzation method	p-value estimation model	FDR estimation method	Differentially expressed gene screening stardard
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	$\lvert\log_2(\text{FoldChange})\rvert > 1$&qvalue < 0.005

The readcount value of the ith gene in the jth sample is Kij, then

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

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

Table 3.9.1 Differentially Expressed Genes

Gene Id	CK	treat	log2FoldChange	pval	p-adjusted
FBgn0000639	36.17650266	252.4500919	-2.8029	2.83E-40	4.93E-37
FBgn0027611	14.36797269	54.31676236	-1.9185	1.00E-06	9.69E-05
FBgn0028400	39.36531292	6.325628725	2.6376	2.05E-07	2.23E-05
FBgn0028533	28.91921033	4.600457255	2.6522	7.99E-06	0.00064263

Differentially Expressed Genes List includes:

(1) Gene id

(2) Sample1: The readcount values of sample1 after normalization

(3) Sample2: The readcount values of sample2 after normalization

(4) log2FoldChange: $\log_2$(Sample1/Sample2)

(5) pvalue (pval): The p-value.

(6) qvalue (p-adjusted): the p-value after normalization. The smaller the q-value is, the more significant is the difference

3.9.2 Screening of differentially expressed genes

Volcano plots are used to infer the overall distribution of differentially expressed genes. For experiments without biological replicates, the threshold is normally set as: |log2(Fold Change) | > 1 and q-value < 0.005. For experiments with biological replicates, as the DESeq already eliminates the biological variation, our threshold is normally set as: padj < 0.05.

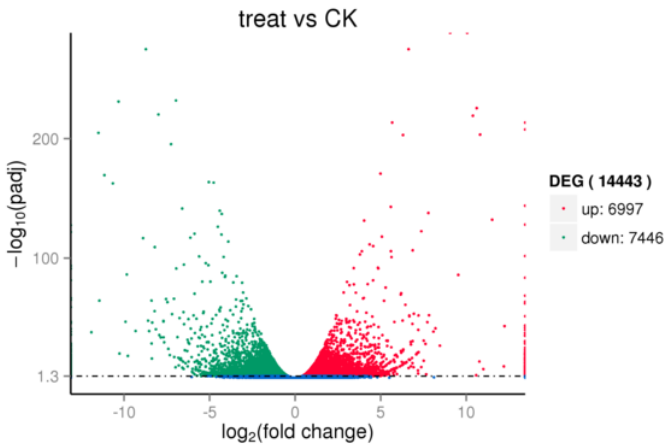


Figure 9.1 Volcano plot for differentially expressed genes

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 did not express differently between treatment group and control group are in blue.

3.9.3 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 is possible to discern unknown functions of previously characterized genes or functions 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 log2(ratios). Genes within the same cluster exhibit the same trends in expression levels under different conditions.

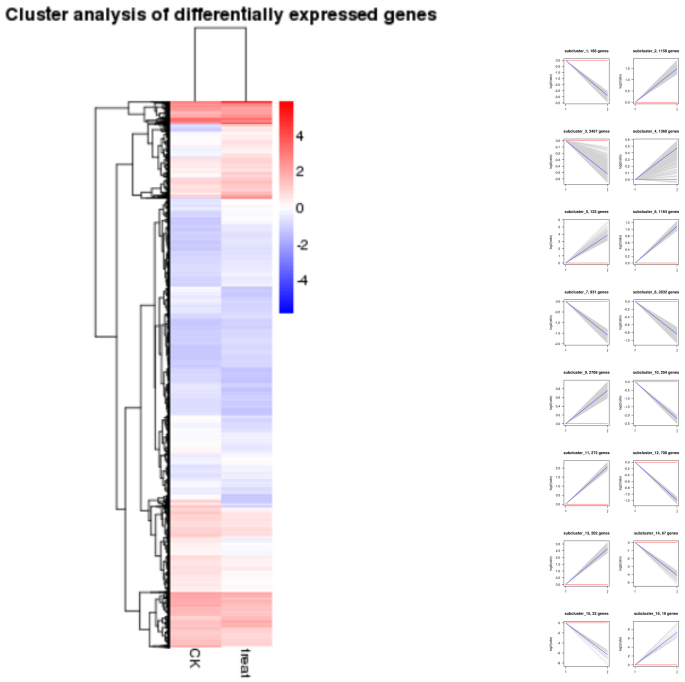


Figure 9.2 Cluster analysis.

Figure I: the overall results of FPKM cluster analysis, clustered using the log10(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 log10(FPKM+1) value from large to small. Figure II: log2(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 value.

3.9.4 The Venn Diagram of Gene Expression Differences

The Venn diagram presents the number of genes that are uniquely expressed within each group, with the overlapping regions showing the number of genes that are expressed in two or more groups. (The diagram depicts only the results for groups 2, 3 , 4 and 5).

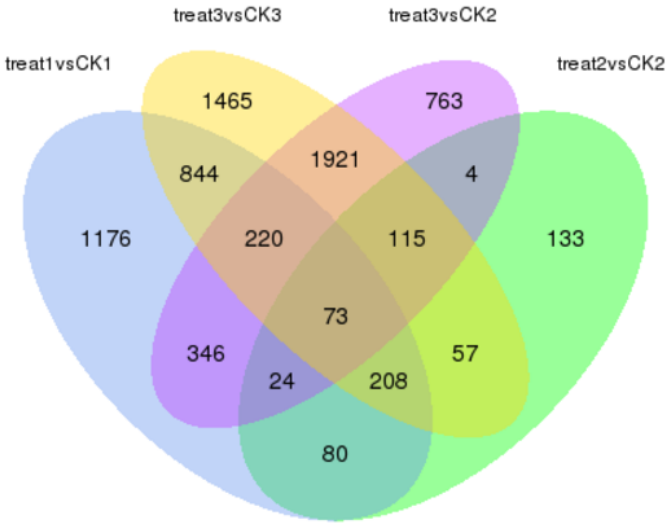


Figure 9.3 Venn diagram of differentially expressed genes

The sum of the numbers in each circle is the total number of genes expressed within a group, and the overlap represents the genes expressed in common between groups.

3.10 GO Enrichment Analysis of DEGs

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 outsides of the category, and this difference is obtained from estimating of the preference of gene length.

3.10.1 GO Enrichment Result List of DEGs

Table 10.1 Significantly Enriched GO Terms in DEGs

GO accession	Description	Term type	Over represented p-Value	Corrected p-Value	DEG item	DEG list
GO:0004097	catechol oxidase activity	molecular_function	4.97E-06	0.016248	2	18
GO:0036263	L-DOPA monooxygenase activity	molecular_function	4.97E-06	0.016248	2	18
GO:0036264	dopamine monooxygenase activity	molecular_function	4.97E-06	0.016248	2	18
GO:0016682	oxidoreductase activity, acting on diphenols and related substances as donors, oxygen as acceptor	molecular_function	2.71E-05	0.066506	2	18

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 p-Value: p-value in hypergenometric test.

(5) Corrected p-Value: 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.

3.10.2 GO Enrichment Bar Chart of DEGs

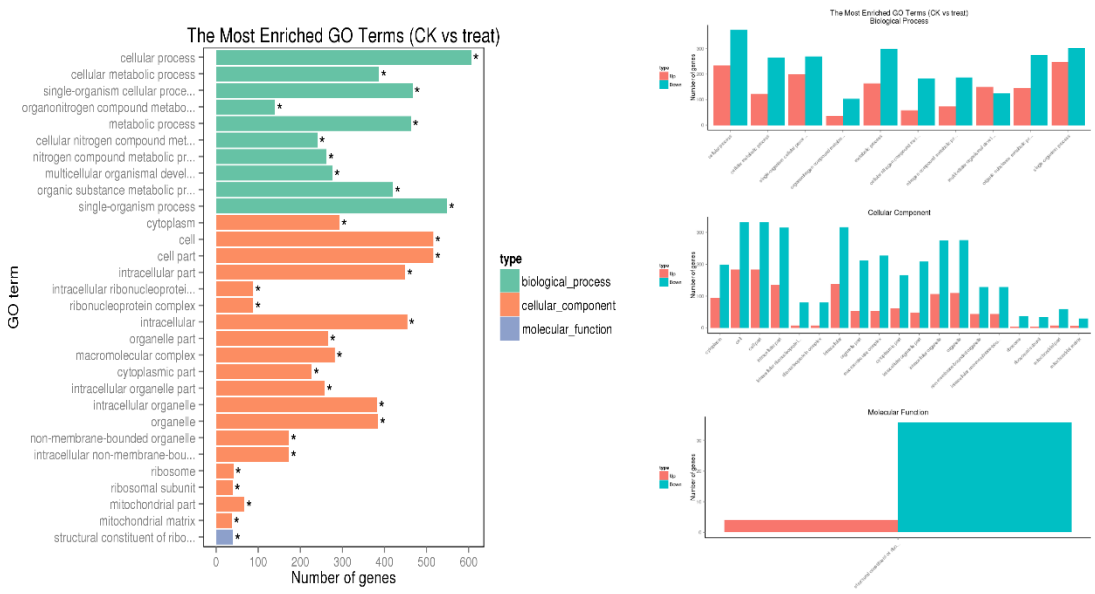


Figure 10.1 Gene Ontology functional classification

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.

3.10.3 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.

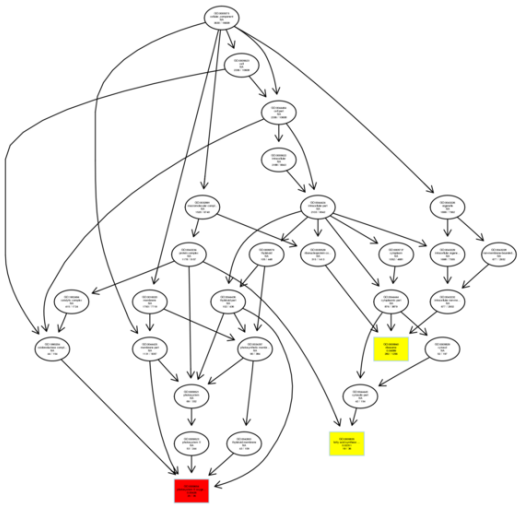


Figure 10.2 Illustration of topGO DAG.

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.

3.11 KEGG Pathway Enrichment Analysis of DEGs

3.11.1 KEGG Enrichment List

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 genomics 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.

Table 11.1 KEGG Enrichment List

#Term	Database	ID	Sample number	Background number	P-Value	Corrected P-Value
ECM-receptor interaction	KEGG PATHWAY	dme04512	1	11	0.021906934	0.093617632
Riboflavin metabolism	KEGG PATHWAY	dme00740	1	13	0.025515797	0.093617632
Tyrosine metabolism	KEGG PATHWAY	dme00350	1	17	0.032697594	0.093617632
Other glycan degradation	KEGG PATHWAY	dme00511	1	22	0.041607836	0.093617632

(1) #Term: Description of KEGG pathways.

(2) Database:

(3) ID: KEGG ID.

(4) Sample number: Number of DEGs with pathway annotation.

(5) Background number: Number of all reference genes with pathway annotation.

(6) P-value: P-value in hypergenometric test.

(7) Corrected P-value: Pathways with corrected p-values 0.05 are significantly enriched in DEGs.

3.11.2 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.

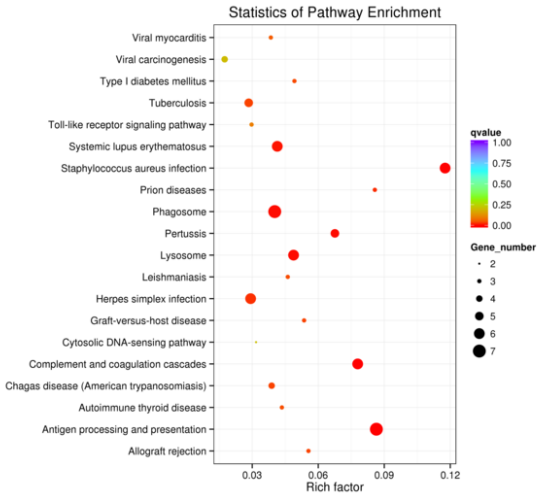


Figure 11.1 KEGG enrichment scatter plot of DEGs

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.

3.11.3 KEGG Enrichment Pathway

KEGG enrichment pathway shows the DEGs significantly enriched pathways. In the diagram, if this node contains up-regulated genes, the KO node is labeled in red. If the node contains up-regulated genes, the KO node is labeled in green. If the node contains both up and down-regulated genes, the labeled color is yellow.

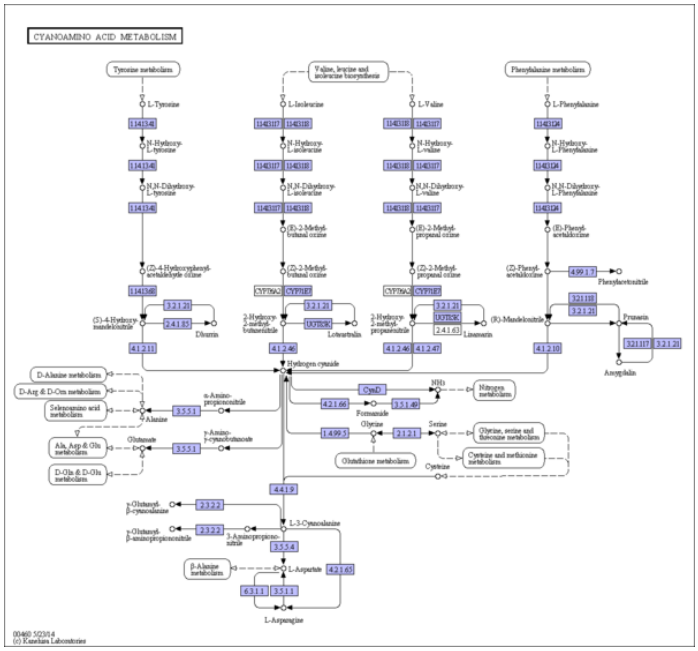


Figure 3.11.2 Diagram showing significantly enriched KEGG pathways

3.12 Protein-Protein Interaction Network Analysis

The STRING database (http://string-db.org/) is used for the analysis of PPI (predicted protein-protein interactions). The results document can be imported into Cytoscape software, and then visualized and edited. The Cytoscape manual can be found at: http://wiki.cytoscape.org/Cytoscape_3/UserManual.

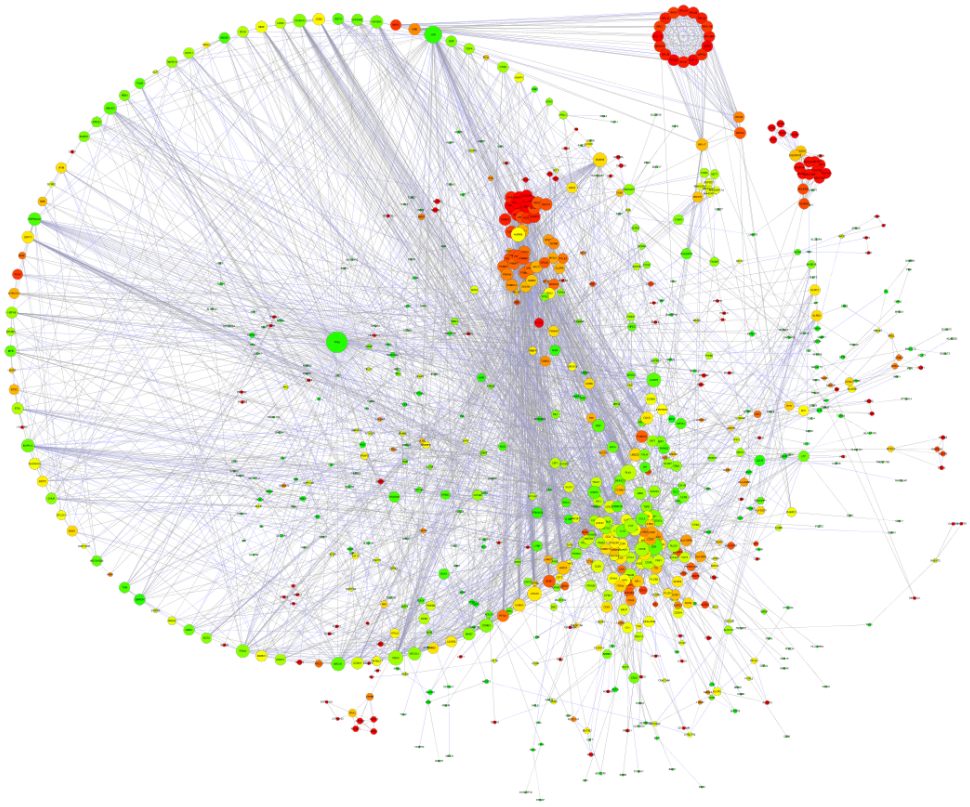


Figure 3.12.1 Cytoscape UI

3.13 The Transcription Factor Analysis Results

iTAK is used to perform the transcription factor analysis of plants and Animal TFDB database is used to perform the transcription factor analysis of animals.

Table 3.13.1 The Transcription Factor Analysis Results

Ensembl ID	Gene ID	Symbol	Family
FBgn0041111	33496	lilli	AF-4
FBgn0039411	43174	dys	bHLH
FBgn0003270	35110	amos	bHLH
FBgn0085432	43769	pan	HMG

For plant:

First Column: the Gene ID

Second Column: the family name of the transcription factors

For Animal:

Ensembl ID: The Ensembl Gene ID

Gene ID: The NCBI Gene ID

Symbol: The Gene Name

Family: The Family Name of Transcription Factor

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_species_results

├── 1. OriginalData: Raw Data（ fastq format ）

├── 2. QC: Data Quality Control
│ ├── 2.1. ErrorRate: Error Rate
│ ├── 2.2. GC: GC Content Distribution
│ ├── 2.3. ReadsClassification: Data Filtering
│ └── 2.4. DataTable: Data Quality Control Summary
├── 3. Mapping: Mapping to a Reference Genome
│ ├── 3.1. MapStat: Overview of Mapping Status
│ ├── 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. AS: Alternative Splicing Analysis
├── 5. NovelGene: Novel Gene Prediction
├── 6. SNP: SNP & InDel
├── 7. GeneExprQuatification: Expression Quantification
│ ├── 7.1. GeneExprQuatification: Expression Quantification
│ └── 7.2. GeneExpContrast: Comparison between Gene Expression Levels
├── 8. AdvancedQC: RNA-seq Advanced QC
│ └── 8.1. Correlation: RNA-Seq Correlation
├── 9. DiffExprAnalysis: Gene Expression Difference Analysis
│ ├──9.1. DEGsList: List of Differentially Expressed Genes (all,up-regulated,down-regulated)
│ ├── 9.2. DEGsFilter: Volcano plot
│ ├── 9.3. DEGcluster: Cluster Analysis of Gene Expression Differences
│ │ └── Subcluster
│ └── 9.4. VennDiagram: The Venn Diagram
├── 10. DEG_GOEnrichment: GO Enrichment Analysis of DEGs
│ ├── 10.1. DEG_GOList: GO Enrichment Result List of DEGs
│ ├── 10.2. DAG: GO Enrichment DAG Figure
│ └── 10.3. BAR: GO Enrichment Bar Chart of DEGs
├── 11. DEG_KEGGEnrichment: KEGG Pathway Enrichment Analysis of DEGs

```
│   ├── 11.1. DEG_KEGGList: KEGG Enrichment List
│   ├── 11.2. DEG_KEGGScat: KEGG Enrichment Scattered Plot
│   └── 11.3. DEG_KEGGPath: KEGG Enrichment Pathway
│       ├── ALL
│       ├── DOWN
│       └── UP
├── 12. DEG_PPI: Protein-Protein Interaction Network Analysis
└── 13. DEG_Trans_Factor: The Transcription Factor Analysis Results
```

4.2 Software List

Software and Parameter

Analysis	Software	Version	Parameters	Remarks
Mapping	Tophat	v2.0.12	mismatch = 2	mapping to a reference
Quantification	HTSeq	v0.6.1	-m union	
Alternative Splicing	rMATS	3.0.8	default parameter	
Novel Gene Prediction	cufflinks	2.1.1	default parameter	
SNP detection	GATK3	v3.4	MQ < 40.0 and QD < 2.0	
Differential Expression Analysis	DEGSeq	1.12.0	|log2foldchang|>1 && qvalue<0.005	For sample with bio-replicate using DESeq, samples without bio-replicate using DEGSeq. EdgeR for specific conditions.
	DESeq	1.10.1	padj<0.05	
	edgeR	3.0.8	padj<0.05	
GO Enrichment	GOSeq, topGO,hmmscan	Release2.12	Corrected P-Value<0.05	hmmscan
KEGG Enrichment	KOBAS	v2.0	Corrected P-Value<0.05	
Protein-Protein Interaction Analysis	BLAST	v2.2.28	e-value = 1e-10 && string score >700	Using blast, String database
The Transcription Factor Analysis	Plants: iTAK;Animals: AnimalTFDB	1.2	default parameter	

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.

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)

Trapnell C., Pachter L., and Salzberg S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. (TopHat)

Kim D., G. Pertea, et al. (2012). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. (TopHat2)

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis.

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

Shen S., Park JW., Lu ZX., Lin L., Henry MD., Wu YN., Zhou Q., Xing Y. rMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data. (rMATS)

Trapnell C., A. Roberts, et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. nature protocols. (Tophat & Cufflinks)

Trapnell C. et al. (2010).Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. (Cufflinks)

Van der Auwera et al., (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics. (GATK3)

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)

Perez-Rodriguez et al (2010).PlnTFDB: updated content and new features of the plant transcription factor database. (TF)

Hong-Mei Zhang, Hu Chen, et al. (2012).AnimalTFDB: a comprehensive animal transcription factor database. (TF)