Transcriptome Sequencing and Analysis Report (with reference)

GENEWIZ biotechnology co. LTD

NGS.Service@azenta.com

Contract no.: ProjectID

Species: Homo sapiens(hg38)

Customer name: Customer

Institute: Institute

BI: BI

Date: 2021-10-27

1 Experimental workflow

2 Bioinformatics analysis workflow

3 Bioinformatics Analysis Result

3.1 Raw data

3.2 Sequencing data quality assessment

3.2.1 Sequencing data quality analysis

3.2.2 Data Filtering

3.3 Alignment to a reference genome

3.3.1 Aligning the clean data to the reference genome

3.3.2 Distribution of reads in the reference genome

3.3.3 Read density distribution on the chromosomes

3.3.4 Visualization of the alignment results

3.4 Alternative splicing analysis

3.4.1 Alternative splicing classification and quantification

3.4.2 Alternative splicing annotation

3.5 Novel transcript prediction

3.6 SNV and InDel analysis

3.6.1 SNV and InDel analysis

3.6.2 Genomic distribution of SNV / InDel

3.7 Gene expression analysis

3.8 RNA-seq overall quality assessment

3.8.1 Quantitative saturation curve

3.8.2 RNA-Seq correlation examination

3.8.3 Sequencing homogeneity examination

3.9 PCA analysis

3.10 Gene differential expression analysis

3.10.1 Gene expression comparison

3.10.2 List of differentially expressed genes

3.10.3 Determination of differentially expressed genes

3.10.4 Cluster analysis of differentially expressed genes

3.10.5 Venn diagrams of differentially expressed genes

3.11 Differential gene GO enrichment analysis

3.11.1 List of Differences Gene GO Enrichment

3.11.2 DAG of differential gene GO enrichment
3.11.3 Histogram of differential gene GO enrichment
3.12 DEU analysis
3.13 PPI analysis
3.14 Gene fusion analysis
3.15 RNA editing analysis
3.16 LncRNA prediction analysis
3.16.1 Removal of other types of RNAs
3.16.2 Removal of based on the characteristics of lncRNA
3.16.3 Removal of transcripts containing protein domains
3.16.4 Removal of transcripts with protein-coding potential
3.16.5 lncRNAs Statistics
3.16.6 lncRNA description and statistical information
3.16.6.1 Classification of known and unknown lncRNAs
3.16.6.2 Distribution of lncRNA according to length, exon count and classification
3.17 Co-expression network analysis
3.17.1 Construction of co-expression network
3.17.2 Cluster Analysis for gene expression module identification
3.17.3 Core module selection
3.17.3.1 Module feature gene selection
3.17.3.2 Association analysis between gene modules and known biological features
3.17.3.3 Function analysis of gene expression modules
3.18 Differential alternative splicing analysis
3.18.1 Differential alternative splicing filtering
3.18.2 Differential alternative splicing results
3.19 Short time series gene expression analysis
3.20 GSEA analysis
3.21 Transcription factor analysis

4 Appendix

5 Reference

AZENTA
LIFE SCIENCES

Experimental Workflow

Experimental Workflow

Transcriptome sequencing experiments include RNA extraction and QC, library construction, purification, library QC and quantitation, as well as sequencing cluster generation and high through-put sequencing. Each step is important for data quality and quantity, which in turn affect the data analysis. To ensure the accuracy and reliability of the analysis results, every step is under strict monitoring and quality control. After mixing libraries based on their effective concentration and the required sequencing data volume, Illumina platform is used for high through-put sequencing.

AZENTA
LIFE SCIENCES

Bioinformatics Workflow

Bioinformatics Workflow

The transcriptome includes all RNAs a cell transcribes in a certain functional state, including the protein-encoding mRNA and non-coding RNA. Transcriptome sequencing uses high throughput sequencing platforms to capture and sequence the entire mRNA pool of specific tissues or cells at a given time, and therefore to obtain almost all the transcript information of a specific species or organ in a certain state. Transcriptome studies, which has been greatly facilitated by next generation sequencing, has transformed gene functional and structural research and has been used widely in basic research, clinical diagnosis and drug development. Text_2 = Bioinformatics analysis is performed after obtaining the original sequence data (Pass Filter Data). The workflow of the analysis is summarized in the Figure 2.1.

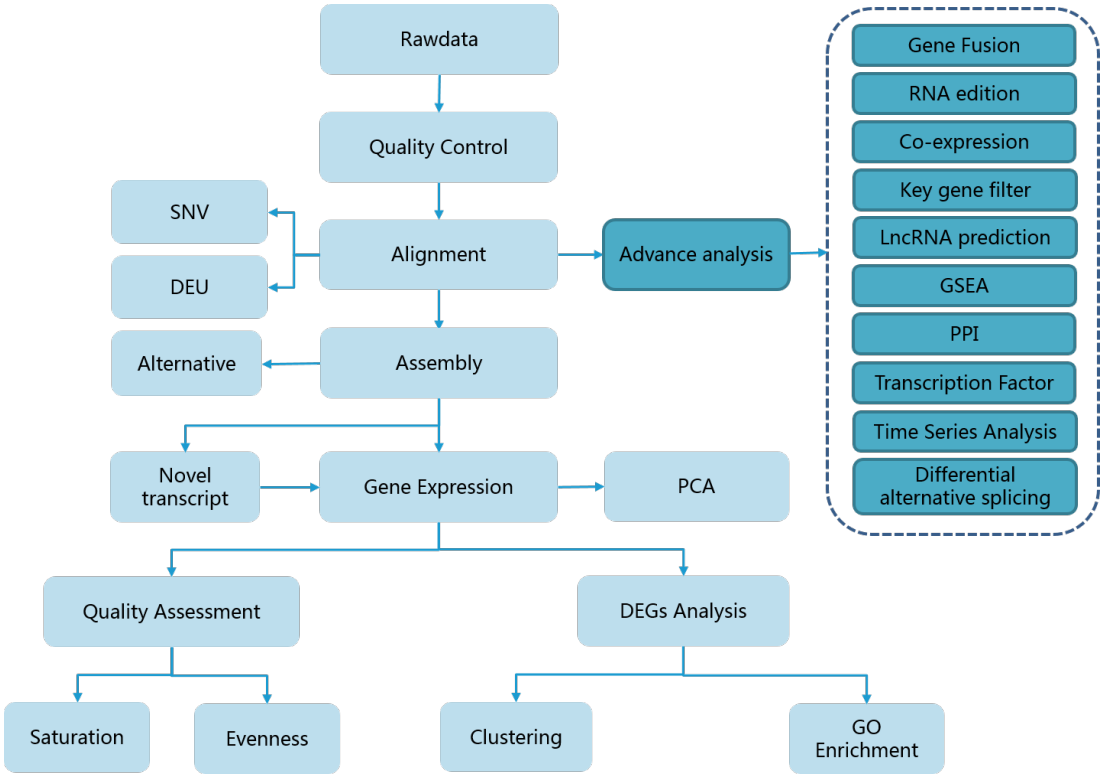


Figure 2.1 Transcriptome data analysis workflow

AZENTA
LIFE SCIENCES

Analysis Result

Analysis Result

3.1 Raw data

Sample information and group information.

Table 3.1.1 Sample list.

SampleName	SampleName	SampleName	SampleName
Sample1-1	Sample2-1	Sample1-2	Sample2-2
Sample1-3	Sample2-3		

Table 3.1.2 Group information.

Control Group	Sample List	Experimental Group	Sample List
Sample1	Sample1-1,Sample1-2,Sample1-3	Sample2	Sample2-1,Sample2-2,Sample2-3

Column explain:

(1) Control Group: control group

(2) Sample List: control sample list

(3) Experimental Group: experimental group

(4) sample List: experimental sample list

Bcl2fastq (v2.17.1.14) was used to processed the original image data for base calling and preliminary quality analysis. The Illumina built-in software determines whether to keep or discard each of the sequencing fragments (namely reads) based on the quality of the first 25 bases. The raw data (Pass Filter Data) obtained from this step is stored in FASTQ format, which contains the base sequence (the second line of a FASTQ record) and the corresponding sequencing quality information (the fourth line of a FASTQ record).

In FASTQ format, each sequence contains four lines of information as shown below:

@GWZHISEQ01:289:C3Y96ACXX:6:1101:1704:2425 1:N:0:GGCTAC

GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT

+

@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF

The first and third lines contain sequence identifier information produced by the sequencer (some fastq files omit name information and leaves it empty after the “+”sign on the third line to save space). The second line contains the sequence information. The fourth line depicts the quality information of each corresponding base on the second line. The fourth line contains sequence quality information, and the quality score is the ASCII value of the corresponding character minus 33. For example, the ASCII value of '@' is 64, and therefore the corresponding base quality score is 31 (64-33). Starting with Illumina GA Pipeline v1.8 (currently v1.9), base quality scores range from 0 to 41.

Table 3.1.3 Explanation of the components in the sequence identifiers (including in the first line of fastq format)

Type	Description
GWZHISEQ01	Unique instrument name
289	Run ID
C3Y96ACXX	Flowcell ID
6	Flowcell lane
1101	Tile number within the flowcell lane
1704	'x'-coordinate of the cluster within the tile
2424	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
GGCTAC	Index sequence

3.2 Sequencing data quality assessment

3.2.1 Sequencing data quality analysis

Sequencing base quality is affected by sequencer, reagents, samples and other factors. The first few bases from the 5'-end are usually of higher error rate and the error rate drops afterward. With long read sequencing platforms (e.g. 150+bp), sequencing error rate might rise again close to the 3' end. Since the first six bases usually have a higher than average error rate, and this is also the length of the random primer, it is suggested that the high error rate is due to the annealing between imperfectly matched primers and template (Jiang et al.). Statistics of sequencing error rate across all base positions can be used to identify abnormally high error rates. For example, it would raise concern if the base error rate in the middle of the sequence is significantly higher than that of the positions towards the end. In general, the sequencing error rate for each base position is less than 0.5%. An error in the sequence is indicated by letter ‘e’. The base quality scores of Illumina sequencing platforms are expressed in QPhred. The formula to calculate QPhred based on error rate is:

Formula: $Q_{phred} = -10\log_{10}(e)$

Table 3.2.1.1 The correlation between Illumina Bcl2fastq base call error rate and Qphred scores is as follows:

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	100%

Quality assessment of the sequencing data was evaluated using FastQC (v0.10.1).

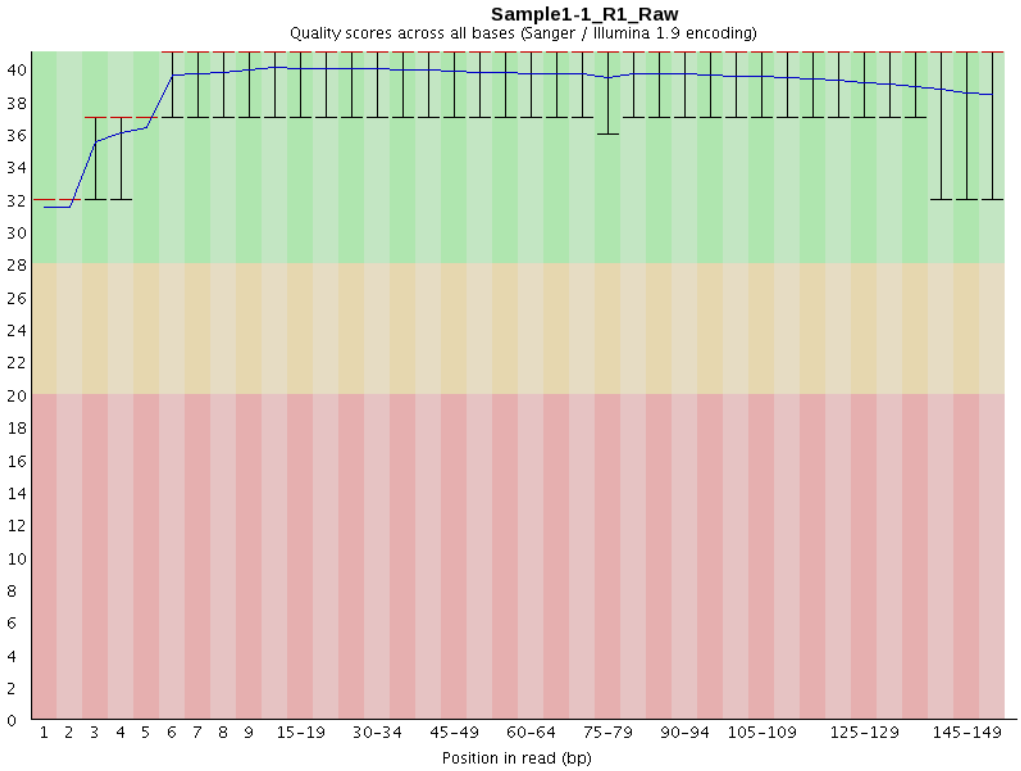


Figure 3.2.1.1 Quality scores across all bases X axis: base position in reads. Y axis: quality score. The higher the score the more reliable the base calling is. In general, if the base quality value is 13, the error rate is 5%, if the quality value is 20, the error rate is 1%, if the quality value is 30, the error rate is 0.1%.

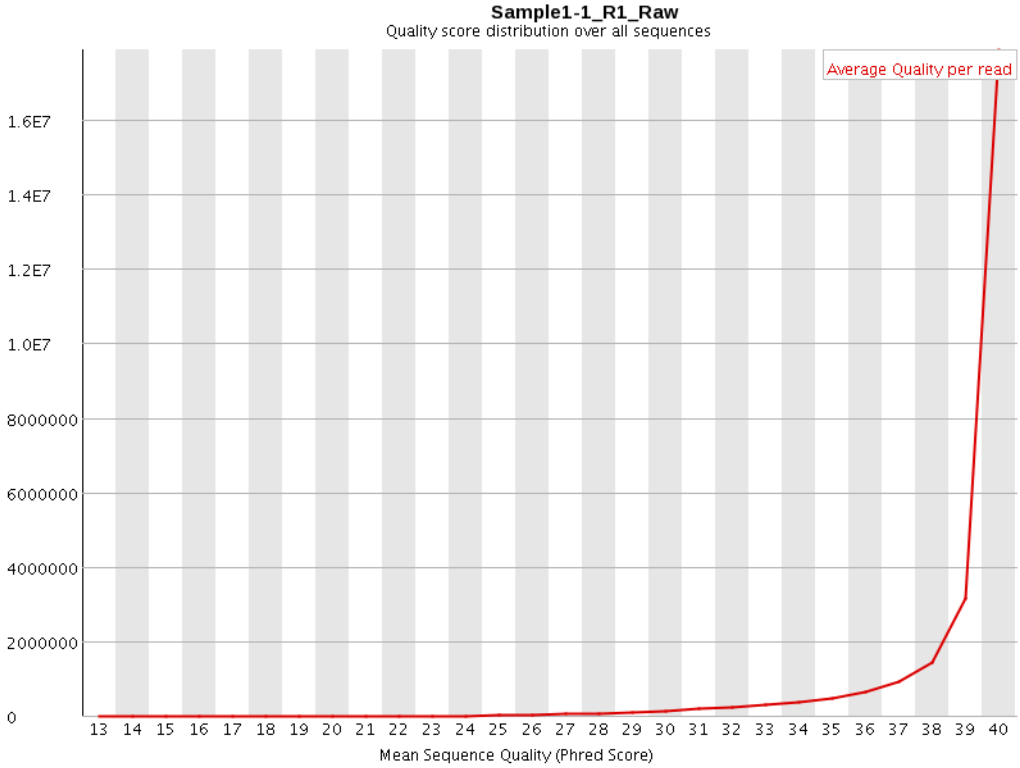


Figure 3.2.1.2 Quality score distribution overall sequences. X axis: the average value of the quality score of the corresponding base. Y axis: the number of sequences. In general, if the peak of the average value is greater than 30 is an indication of high quality.

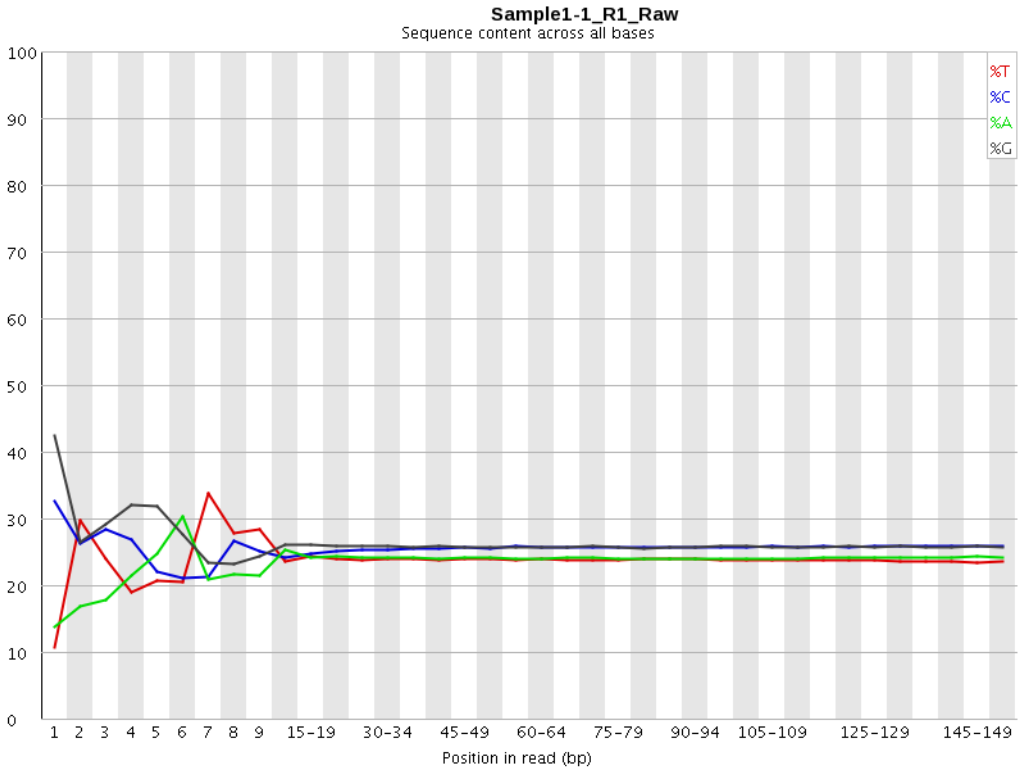


Figure 3.2.1.3 Sequence ATCG content distribution across all bases. This is used to determine whether there is difference in the percentage of A/T and G/C. X axis: base position in reads. Y axis: the percentage of each single base (ATCG) at the corresponding position.

3.2.2 Data Filtering

During sequencing, quality concerns may arise. A small number of target sequences might be reads into adapter sequences, and bases toward the 3'-end might have low quality due to the lengthy sequencing cycles. To eliminate the negative effect of these technical issues, low-quality reads and contaminations were filtered out before data analysis. In addition, adapter sequences were removed at this step.

Software: Cutadapt (version 1.9.1)

Method description:

(1) remove the adapter sequences

(2) remove the 5' or 3' end bases that contains N's or of quality values below 20

(3) remove reads that are less than 75 bp long after trimming

The statistics of raw data are in Table 3.2.2.1.

Table 3.2.2.1 Raw data statistics

Sample	length	Reads	Bases	Q20 (%)	Q30 (%)	GC (%)	N (ppm)
Sample1-1	150.00	52792600	7918890000	97.46	94.01	51.93	43.73
Sample2-1	150.00	53795532	8069329800	97.57	94.32	51.67	46.35
Sample1-2	150.00	49022754	7353413100	97.52	94.22	51.92	45.84
Sample2-2	150.00	55493200	8323980000	97.54	94.26	51.72	46.25
Sample1-3	150.00	55743832	8361574800	97.59	94.33	51.75	45.23
Sample2-3	150.00	61756336	9263450400	97.82	94.84	51.68	43.20

The statistics of processed data are in Table 3.2.2.2.

Table 3.2.2.2 Filtered data statistics

Sample	length	Reads	Bases	Q20 (%)	Q30 (%)	GC (%)	N (ppm)
Sample1-1	148.17	52379256	7760931714	97.80	94.47	52.00	7.33
Sample2-1	148.23	53366960	7910802955	97.92	94.78	51.74	7.53
Sample1-2	148.12	48623220	7201898080	97.88	94.69	52.00	7.57
Sample2-2	148.10	55054404	8153437796	97.89	94.72	51.80	7.51
Sample1-3	148.29	55309086	8201996566	97.91	94.78	51.83	7.44
Sample2-3	148.18	61301206	9083835489	98.16	95.30	51.77	2.55

Column explain:

(1) Sample: Sample name

(2) length: Average read length

(3) Reads: Number of reads

(4) Bases: Number of bases

(5) Q20, Q30: The percentage of bases with quality scores (Qphred) higher than 20 or 30

(6) GC%: The percentage of G+C in the read

(7) N(ppm): The number of base 'N' per million bases

3.3 Alignment to a reference genome

3.3.1 Aligning the clean data to the reference genome

Filtered data were subsequently aligned to the reference genome. It is important to select the appropriate reference genome for optimal analysis. The alignment rate reflects the compatibility between the selected genome and data to be analyzed.

Short-read alignment was performed using Hisat2 (v2.0.1) (Nat Methods. 2015 Apr; Kim D, et al.) with default parameters.

Reference genome and annotation files

Table 3.3.1.1 Data alignment statistics

Samples	Total reads	Total mapped	Multiple mapped	Uniquely mapped	Read1	Read2	Reads map to '+'	Reads map to '-'	Non_splice reads	Splice reads	Reads mapped in proper pairs
Sample1-1	52379256	47468697 (90.625%)	4431804 (8.46099%)	43036893 (82.164%)	21888612	21148281	21542789	21494104	25030486	18006407	41061578
Sample1-2	48623220	44312083 (91.1336%)	4148482 (8.53189%)	40163601 (82.6017%)	20415925	19747676	20107932	20055669	23635293	16528308	38410348
Sample1-3	55309086	50531781 (91.3625%)	4643956 (8.39637%)	45887825 (82.9662%)	23323125	22564700	22967657	22920168	26052405	19835420	43823418
Sample2-1	53366960	48434183 (90.7569%)	4562851 (8.54995%)	43871332 (82.2069%)	22292573	21578759	21954375	21916957	25517804	18353528	42003456
Sample2-2	55054404	50139522 (91.0727%)	4761139 (8.64806%)	45378383 (82.4246%)	23061817	22316566	22710287	22668096	26824423	18553960	43387224
Sample2-3	61301206	56222670 (91.7154%)	5472138 (8.92664%)	50750532 (82.7888%)	25659057	25091475	25397953	25352579	31225227	19525305	48706922

Column explain:

(1) Total reads: number of the total reads that passed the filtering step

(2) Total mapped: number of reads that were successfully aligned to the reference. In general, this number should be greater than 70% given that there is no contamination and an appropriate reference was selected

(3) Multiple mapped: the number of sequences with multiple alignment positions on the reference. This value is generally less than 10%

(4) Uniquely mapped: number of sequences with only one alignment positions on the reference

(5) Reads map to '+' , Reads map to '-': number of sequences mapped to the plus / minus strand of the reference

(6) Splice reads: number of reads mapped to more than one exon (also known as junction reads). Non-splice reads: number of reads mapped to only one exon. The percentage of splice reads is determinedly partly by the length of sequencing reads

(7) Reads mapped in proper pairs: In paired-end sequencing, the number of sequence pairs with proper alignment to the reference evaluated by the alignment direction and location

3.3.2 Distribution of reads in the reference genome

Mapped reads were assigned to genomic features – exons, introns and intergenic regions. Sequences located in the intron region may be due to immature mRNA contamination or incomplete genomic annotation, whereas sequences mapped to intergenic region may be due to incomplete genomic annotations and background noise.

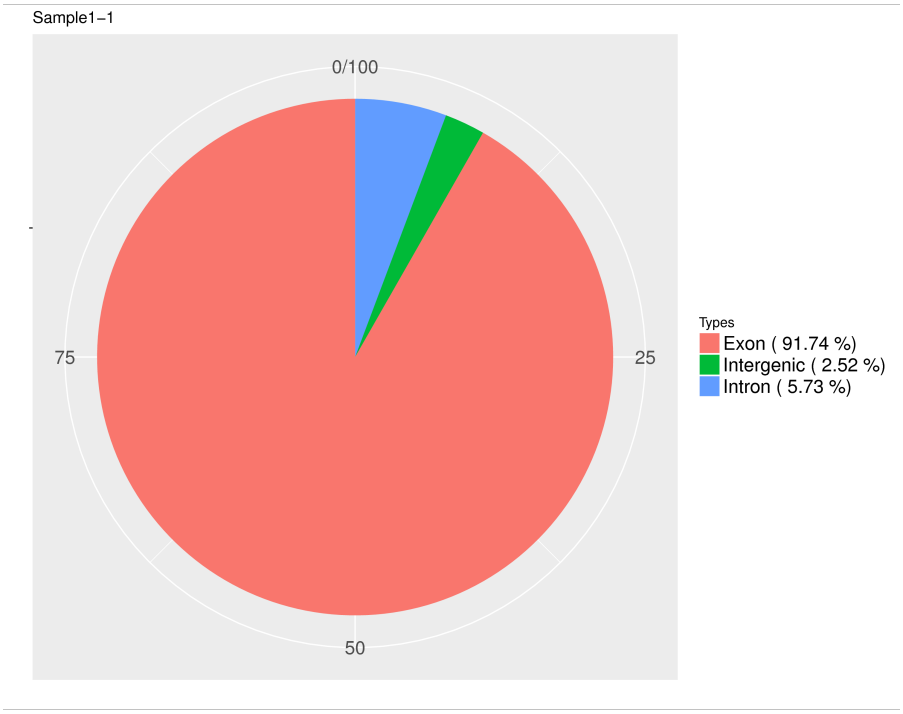


Figure 3.3.2.1 The distribution of reads in different genomic regions

3.3.3 Read density distribution on the chromosomes

Read density on each chromosome is calculated as the $\log_2$ values of read counts in 1kb windows and is illustrated in Figure 3.3.3.1. Normally, the longer the chromosome is, the more the total read count mapping to the chromosome (Marquez et al.). The homogeneity of the sequencing can be assessed from the relationship between read count and chromosome length.

Figure 3.3.3.1 Read depth (log2) across chromosome, the abscissa is the length of the chromosome.

3.3.4 Visualization of the alignment results

For data visualization, we provide the bam files of RNA-seq alignment, and recommend to use IGV (Integrative Genomics Viewer) browser to visualize the bam files. IGV browsers can display the abundance of reads to reflect the level of transcriptional activity.

IGV download

IGV manual

IGV application instruction:

(1) To upload reference genome: Select Genomes -> Load Genome From File. If you are to use a widely-studied genome such as human genome, you could directly select the reference genome from the drop-down list. For references not listed, the customer can download the references from the link we provided and go to File -> Load to upload the genome.

(2) Upload alignment files in bam format: Select File -> Load From File, and then upload bam files in desired order.

(3) To view the results, select the reference genome of interest from the second menu on the toolbar of the IGV interface. The result in the genome browser is as follows:

Figure 3.3.4.1 Snapshot of IGV browser interface

3.4 Alternative splicing analysis

Alternative splicing allows a single gene to produce multiple mRNA transcripts, and different mRNA transcripts are translated into different forms of proteins to increase diversity (Black, 2003; Stamm, 2005; Lareau, 2004). Although alternative splicing is known to be prevalent in eukaryotes, its scope and significance may still be underestimated. Recently, alternative splicing studies based on high-throughput sequencing have been published from human (Pan, 2008；Wang, 2008；Sultan, 2008), mouse (Tang, 2009; Mortazavi, 2008), and Arabidopsis (Filichkin) studies, resulting in the discovery of novel alternative splicing events.

We use StringTie (v1.3.3b) (Nature Biotechnology 2015; Pertea M, et al.) to do Assembly and predict alternative splicing and ASprofile (V1.0.4) for classification and quantification. The classification categorization of alternative by ASprofile is shown below:

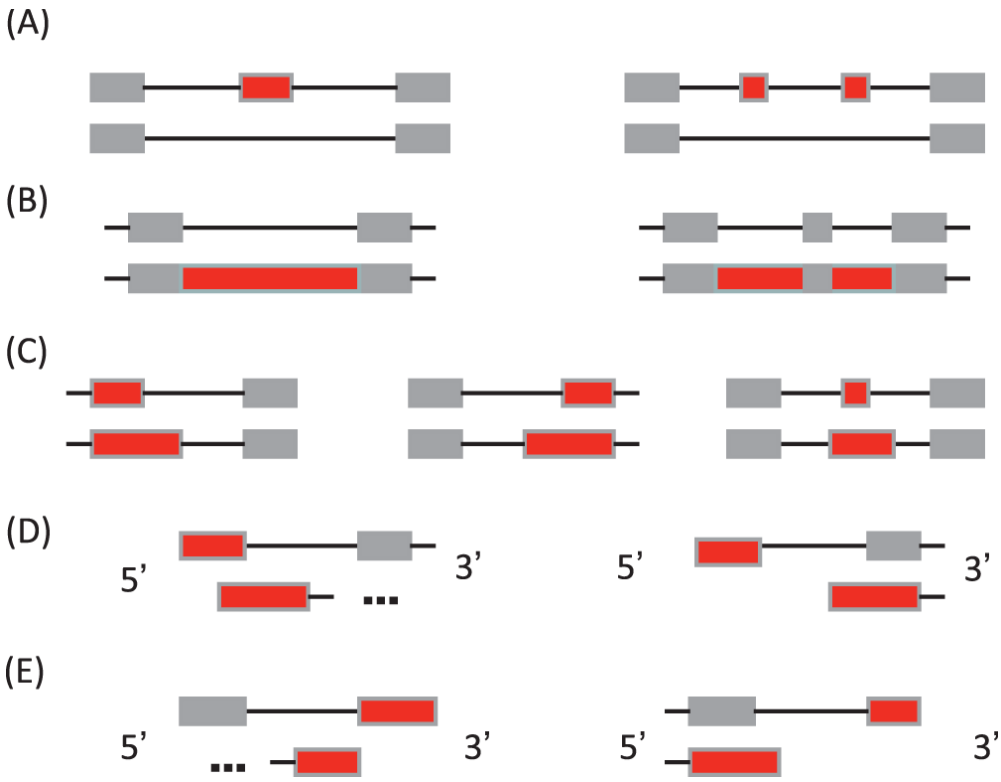


Figure 3.4.1 Basic alternative splicing categories: (A) SKIP, MSKIP (B) IR, MIR (C) AE (D) TSS (E) TTS. The site of alternative splicing is in red.

(A) SKIP: Skipped exon (SKIP_ON,SKIP_OFF pair)

XSKIP: Approximate SKIP (XSKIP_ON,XSKIP_OFF pair)

MSKIP: Multi-exon SKIP (MSKIP_ON,MSKIP_OFF pair)

XMSKIP: Approximate MSKIP (XMSKIP_ON,XMSKIP_OFF pair)

(B) IR: intron retention (IR_ON, IR_OFF pair)

XIR: Approximate IR (XIR_ON, XIR_OFF pair)

MIR: Multi-IR (MIR_ON, MIR_OFF pair)

XMIR: Approximate MIR (XMIR_ON, XMIR_OFF pair)

(C) AE: Alternative exon ends (5', 3', or both)

XAE: Approximate AE

(D) TSS: Alternative 5' first exon (transcription start site)

(E) TTS: Alternative 3' last exon (transcription terminal site)

3.4.1 Alternative splicing classification and quantification

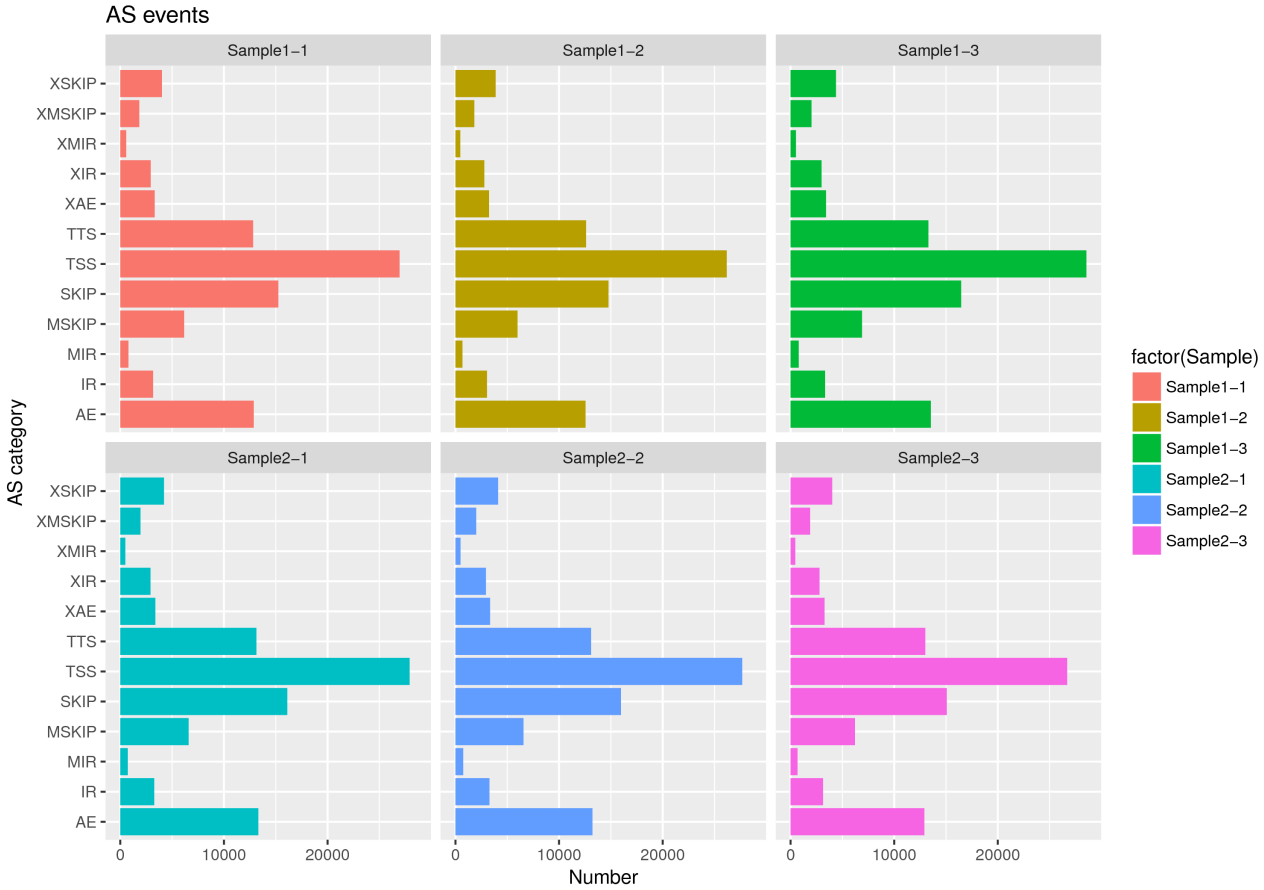


Figure 3.4.1.1 Statistic summary of alternative splicing events of each sample. X axis: number of splicing events. Y axis: types of splicing.

3.4.2 Alternative splicing annotation

Table 3.4.2.1 AS annotation and quantification (Partial results are shown. For complete results please see: *.anno.fpkm.xls)

event_id	event_type	gene_id	chrom	event_start	event_end	event_pattern	strand	fpkm	ref_id
1000065	TSS	XLOC_000022	1	817371	818202	818202	+	0.3576110000	FAM87B
1000066	TTS	XLOC_000022	1	818723	819837	818723	+	0.3576110000	FAM87B
1000068	TSS	XLOC_000023	1	827608	827775	827775	+	1.1195810000	LINC01128
1000071	TSS	XLOC_000023	1	851348	852110	852110	+	0.0015300000	LINC01128
1000075	SKIP_OFF	XLOC_000023	1	847654	847806	829104,851927	+	0.8705390000	LINC01128

Column explain:

(1) event_id: AS event ID

(2) event_type: AS event type (TSS, TTS, SKIP_ {ON , OFF}, XSKIP_ {ON, OFF}, MSKIP_ {ON, OFF}, XMSKIP_ {ON, OFF}, IR_ the ON {, OFF}, {XIR_ the ON, OFF}, the AE, XAE)

(3) gene_id: gene ID from cuffmerge assembly

(4) chrom: chromosome ID

(5) event_start: AS event starting position

(6) event_end: AS event end position

(7) event_pattern: AS event characteristics (for TSS, TTS - inside boundary of alternative marginal exon; for *SKIP_ON,the coordinates of the skipped exon(s), for *SKIP_OFF, the coordinates of the enclosing introns, for *IR_ON, the end coordinates of the long, intron-containing exon, for *IR_OFF, the listing of coordinates of all the exons along the path containing the retained intron, for *AE, the coordinates of the exon variant)

(8) strand: reference strand information of the AS event

(9) fpkm: expression of the gene with the corresponding AS type in FPKM

(10) ref_id: corresponding gene ID in the reference (gene annotation)

3.5 Novel transcript prediction

The existing transcript annotation databases may not cover all the genes in the transcriptome. New genes or transcripts can be discovered by leveraging high-throughput sequencing (Mortazavi, 2008). For that purpose, StringTie (v1.3.3b) (Nature Biotechnology 2015; Pertea M, et al.) was employed for de novo transcript assembly using the alignment bam files, the result of which was then subject to comparison with existing annotation reference (gtf file) using Cuffcompare (V2.2.1). This purpose of this workflow is to:

(1) discover novel genes (by comparing to the given annotation reference)

(2) discovered novel exons of existing genes

(3) optimize the boundary of the existing genes

The result of novel gene or exon prediction is in GTF format. For details of GTF format, please see: GTF format.

Table 3.5.1 annotation of novel transcript structure results (Partial results are shown. For complete results please see: *_comp.combined.gtf)

seqname	source	feature	start	end	score	strand	frame	attributes
1	StringTie	exon	131025	134836	.	+	.	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; gene_name "CICP27"; old "STRG.11.1"; nearest_ref "ENST00000442987"; class_code "="; tss_id "TSS1";
1	StringTie	exon	629062	629433	.	+	.	gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "1"; gene_name "MTND1P23"; old "STRG.15.1"; nearest_ref "ENST00000416931"; class_code "="; tss_id "TSS2";
1	StringTie	exon	629640	630683	.	+	.	gene_id "XLOC_000003"; transcript_id "TCONS_00000003"; exon_number "1"; gene_name "MTND2P28"; old "STRG.16.1"; nearest_ref "ENST00000457540"; class_code "="; tss_id "TSS3";
1	StringTie	exon	631074	632616	.	+	.	gene_id "XLOC_000004"; transcript_id "TCONS_00000004"; exon_number "1"; gene_name "MTCO1P12"; old "STRG.17.1"; nearest_ref "ENST00000414273"; class_code "="; tss_id "TSS4";
1	StringTie	exon	632757	633438	.	+	.	gene_id "XLOC_000005"; transcript_id "TCONS_00000005"; exon_number "1"; gene_name "MTCO2P12"; old "STRG.18.1"; nearest_ref "ENST00000427426"; class_code "="; tss_id "TSS5";

Column explain:

(1) seqname: name of the chromosome or scaffold, chromosome names can be given with or without the 'chr' prefix

(2) source: name of the program that generated this feature, or the data source (database or project name)

(3) feature: feature type name, e.g. Gene, Variation, Similarity

(4) start: Start position of the feature, with sequence numbering starting at 1

(5) end: End position of the feature, with sequence numbering starting at 1

(6) score: A floating point value

(7) strand: defined as + (forward) or - (reverse)

(8) frame: One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on

(9) attributes: A semicolon-separated list of tag-value pairs, providing additional information about each feature

Table 3.5.2 Structure optimization of existing genes (Partial results are shown. For complete results please see: *_novel.xls)

Gene_id	Chromosome	Strand	Original_span	Assembled_span
XLOC_000007	1	+	MTATP6P1: 633696-634376	633535-634922
XLOC_000007	1	+	MTATP8P1: 633535-633741	633535-634922
XLOC_000007	1	+	MTCO3P12: 634376-634922	633535-634922
XLOC_000008	1	+	AL669831.7: 781937-782050	778770-810060
XLOC_000009	1	+	FAM87B: 817371-819837	817371-820116

Column explain:

(1) Gene_id: gene ID

(2) Chromosome: chromosome ID

(3) Strand: reference strand information

(4) Original_span: gene and its start position - end position according to the original reference

(5) Assembled_span: gene start position - end position according to the novel assembly

3.6 SNV and InDel analysis

3.6.1 SNV and InDel analysis

Samtools (v0.1.19) was used for mpileup to compare each sample with the reference genome for SNV detection. Annovar (v2016.05.11) was then used for annotation. Correlation between mutation information and gene information can be derived based on the annotated gene information in the database, enabling annotation of the mutation site. Information including amino acid level mutation and mutation frequencies is shown in Table 3.6.1.1.

The dbSNP database (version 147), one of the NCBI databases, includes all the SNV and InDel information that have been reported (note: the annotated results should be viewed in the same version of the database).

The 1000 genome database (version 1000g2015aug) records information about the frequency of mutations at the relevant mutation sites.

Table 3.6.1.1 SNV analysis (Partial results are shown. For complete results please see: All.xls)

Type	Chr	Start	END	Ref	Obs	Func	Gene	ExonicFunc	AAChange	1000genome	dbsnp	Sample1-1/(hom/het)	Qual	Depth	F
SNV	CHR_HG30_PATCH	179617537	179617537	A	G	intergenic	.	.	.	.	.	-	-	-	-
SNV	CHR_HG30_PATCH	179617607	179617607	C	G	intergenic	.	.	.	.	.	-	-	-	-
SNV	CHR_HG30_PATCH	179621710	179621710	G	A	intergenic	.	.	.	.	.	hom	.	19	1
SNV	CHR_HG30_PATCH	179623896	179623896	G	A	intergenic	.	.	.	.	.	-	-	-	-
SNV	CHR_HSCHR1_2_CTG31	155248423	155248423	T	C	intergenic	.	.	.	.	.	-	-	-	-

Column explain:

(1) Type: point mutation classification (SNV / InDel)

(2) Chr: chromosome ID

(3) Start: starting position

(4) End: end position

(5) Ref: reference base

(6) Obs: mutant base

(7) Func: functional classification

(8) Gene: gene name corresponding to functional classification

(9) ExonicFunc: functional classification of exon mutation

(10) AAChange: nucleotide and amino acid mutation information (NCBI SEQ ID NO: mutation : amino acid mutation)

(11) 1000genome: mutation frequency according to 1000genome database

(12) dbsnp: SNP annotation database ID

(13) Sample*: sample information, four elements:

---hom/het mutation type (homozygous or heterozygous)

---Qual ---quality

---Depth base depth

---Freq mutation frequency

AAChange Example Description:

Table 3.6.1.2

SNV example	NM_177987:c.G729A:p.P243P
NM_177987	Gene marking
c.	chromosome ID
G	reference base
729	position in gene
A	mutant base
p.	peptide
P	reference amino acid
243	amino acid position in peptide
P	mutant amino acid

Table 3.6.1.3

InDel example	NM_014696:c.720_721insATGAGGGAG:p.E240delinsEMRE
NM_014696	gene marking
c.	chromosome ID
720_721	bases insert between 720 and 721
ins	insert
p.	peptide
E240	amino acid MRE insert after amino acid E at position 240 of peptide
delins	insert
EMRE	amino acid changes situation after insertion

3.6.2 Genomic distribution of SNV / InDel

Based on the annotation information in the reference genome, the distribution of SNV / InDel events in genomic compartments is summarized in Table below.

Table 3.6.2.1 SNV / InDel genomic distribution

Sample	Sample1-1	Sample2-1	Sample1-2	Sample2-2	Sample1-3	Sample2-3
exonic	6602	7244	6325	7022	7352	6439
intergenic	2050	2198	1884	2196	2030	2484
intronic	3588	4379	3085	4131	4090	4171
splicing	102	106	81	99	98	88
exonic;splicing	8	9	8	10	8	7
ncRNA_exonic	937	897	871	898	971	994
ncRNA_intronic	842	911	754	870	893	910
ncRNA_splicing	3	3	2	2	2	3
ncRNA_UTR3	0	0	0	0	0	0
ncRNA_UTR5	0	0	0	0	0	0
upstream	127	192	114	157	172	168
downstream	450	500	418	505	527	573
upstream;downstream	13	19	13	15	16	20
UTR3	10866	11353	10444	11290	11506	11342
UTR5	1022	1193	975	1137	1221	1013
UTR5;UTR3	4	5	4	6	4	6
Total	26614	29009	24978	28338	28890	28218

Column explain:

(1) Sample: sample name

(2) exonic: exon region

(3) intergenic: intergenic region

(4) intronic: intron

(5) splicing: splice site

(6) exonic;splicing: exon; splice site

(7) ncRNA_exonic: ncRNA exon region

(8) ncRNA_intronic: ncRNA intron

(9) ncRNA_splicing: ncRNA splicing site

(10) ncRNA_UTR3: 3' UTR of ncRNA

(11) ncRNA_UTR5: 5' UTR of ncRNA

(12) upstream: gene upstream region

(13) downstream: gene downstream region

(14) upstream;downstream: gene upstream region; other gene downstream region

(15) UTR3: gene 3' UTR region

(16) UTR5: gene 5' UTR region

(17) UTR5;UTR3: gene 5' UTR region; 3' UTR of other genes

(18) Total: Total number of mutations

SNV / InDel distribution across all functional regions is illustrated in the pie chart as below:

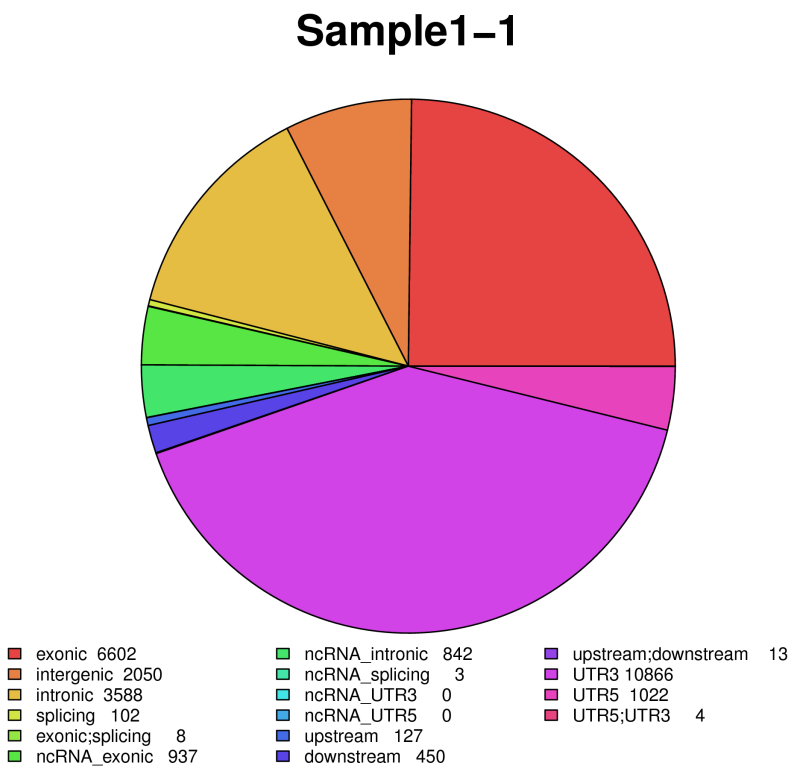


Figure 3.6.2.1 SNV / InDel genomic distribution

3.7 Gene expression analysis

The level of gene expression is measured by read density, the higher the read density, the higher the level of gene expression. Gene expression calculation was performed with the formula below, which calculates FPKM (Fragments per kilo bases per million reads) based on read counts from HT-seq (V 0.6.1) (Mortazavi, 2008).

The formula is:

$$FPKM = \frac{\text{total exon Fragments}}{\text{mapped reads (Millions)} \times \text{exon length (KB)}}$$

Figure 3.7.1

The ratio of (total exon fragments / mapped reads [millions]) is the read count mapped to the gene normalized to total read counts. The value is then normalized to gene length (exon length [KB]), so that the expression of genes with different sequencing depths and length are comparable.

The numbers of genes with different expression levels are summarized in Table 3.7.1. In general, FPKM threshold for gene expression is set between 0.1-1, although there is no absolute standard and various thresholds have been used in the literature.

Table 3.7.1 Distribution of genes expression levels

Sample	0-0.1	0.1-1	1-3	3-15	15-60	>60
Sample1-1	4511(19.92%)	7341(32.42%)	3691(16.30%)	4916(21.71%)	1637(7.23%)	545(2.41%)
Sample2-1	4721(20.58%)	7234(31.53%)	3710(16.17%)	5000(21.79%)	1723(7.51%)	556(2.42%)
Sample1-2	4346(19.38%)	7291(32.51%)	3742(16.69%)	4844(21.60%)	1638(7.30%)	565(2.52%)
Sample2-2	4703(20.61%)	7230(31.69%)	3648(15.99%)	4916(21.55%)	1739(7.62%)	578(2.53%)
Sample1-3	4835(20.79%)	7485(32.18%)	3736(16.06%)	5052(21.72%)	1617(6.95%)	534(2.30%)
Sample2-3	5076(21.92%)	7314(31.59%)	3645(15.74%)	4754(20.53%)	1750(7.56%)	616(2.66%)

Table 3.7.2 Gene expression results across all samples (Partial results are shown. For complete results please see: all.fpkm_anno.xls)

gene_id	Exonic.gene.sizes	Sample1-1	Sample1-1_FPKM	Sample2-1	Sample2-1_FPKM	Sample1-2	Sample1-2_FPKM	Sample2-2	Sample2-2_FPKM	Sample1-3	Sample1-3_FPKM	Sample2-3	Samp 3_FP
ENSG00000000003	4535	450	5.17	435	4.91	427	5.24	417	4.55	550	5.93	484	4.72
ENSG00000000005	1610	4	0.13	7	0.22	3	0.10	1	0.03	3	0.09	6	0.16
ENSG00000000419	1207	801	34.57	855	36.26	754	34.80	818	33.57	854	34.58	1018	37.31
ENSG00000000457	6883	252	1.91	404	3.00	218	1.76	371	2.67	289	2.05	431	2.77
ENSG00000000460	5967	33	0.29	41	0.35	28	0.26	47	0.39	50	0.41	53	0.39

Column explain:

(1) gene_id: gene ID

(2) Exonic.gene.sizes:exon length

(3) Sample:count count of each gene

(4) Sample_FPKM: FPKM of each gene

(5-8) Chr，Start，End，Strand: gene location, including chromosome, start position, end position and strand

(9) GeneSymbol: gene symbol

(10) Description: gene description

(11-13) GO_BP,GO_CC,GO_MF: gene ontology, including biological process, cellular component and molecular function

3.8 RNA-seq overall quality assessment

3.8.1 Quantitative saturation curve

The quality saturation curve gives a good indication of the amount of data required to quantify gene expression. The higher the expression of the gene, the fewer reads it needs for quantification. On the other hand, the lower the expression, the more reads it requires for accurate quantification. RSeQC (V2.6.3) was used to generate saturation curves. RPKM at each sequencing depths were calculated and compared to the real RPKM. The accuracy of the gene expression was evaluated by the percent relative error using the following formula:

$$\text{Percent Relative Error} = \frac{|RPKM_{obs} - RPKM_{real}|}{RPKM_{real}} \times 100$$

Figure 3.8.1.1

$RPKM_{obs}$ is the RPKM calculated using the read count at the corresponding sampling percentage. $RPKM_{Real}$ is the real RPKM from gene expression analysis.

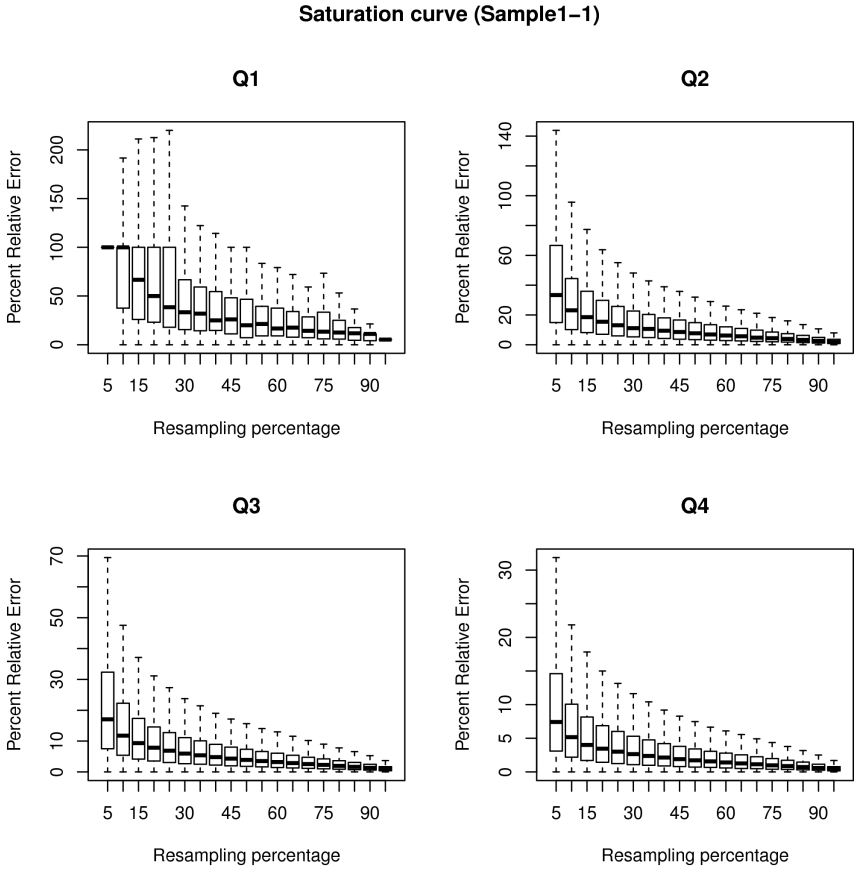


Figure 3.8.1.2 Percent error rate saturation curve. X axis: percentage of sampling reads. Y axis: Percent relative error. Q1 is a saturation box plot with transcript expression levels below 25%, Q2 is a saturation box plot with transcript expression levels between 25% and 50%, Q3 is a saturation box plot with transcript expression levels between 50%, and 75% Q4 is the saturation box plot of transcript expression levels above 75%.

3.8.2 RNA-Seq correlation examination

Biological duplication has two main purposes – one is to prove that the biological experiments are repeatable with reasonable variation, and the other is for subsequent differential gene expression analysis. The correlation of gene expression between samples is an important index

of the reliability of experiment. Pearson correlation is a correlation coefficient that indicates the degree of linear relationship between two variables. The greater the absolute value (ranging between 0 and 1), the stronger the linear relationship. Normally R^2 greater than 0.8 is seen as reasonable. Otherwise, further explanation is needed or the experiment needs to be repeated. In this section, we also calculated spearman rank correlation coefficient and kendall-tau rank correlation coefficient for your reference.

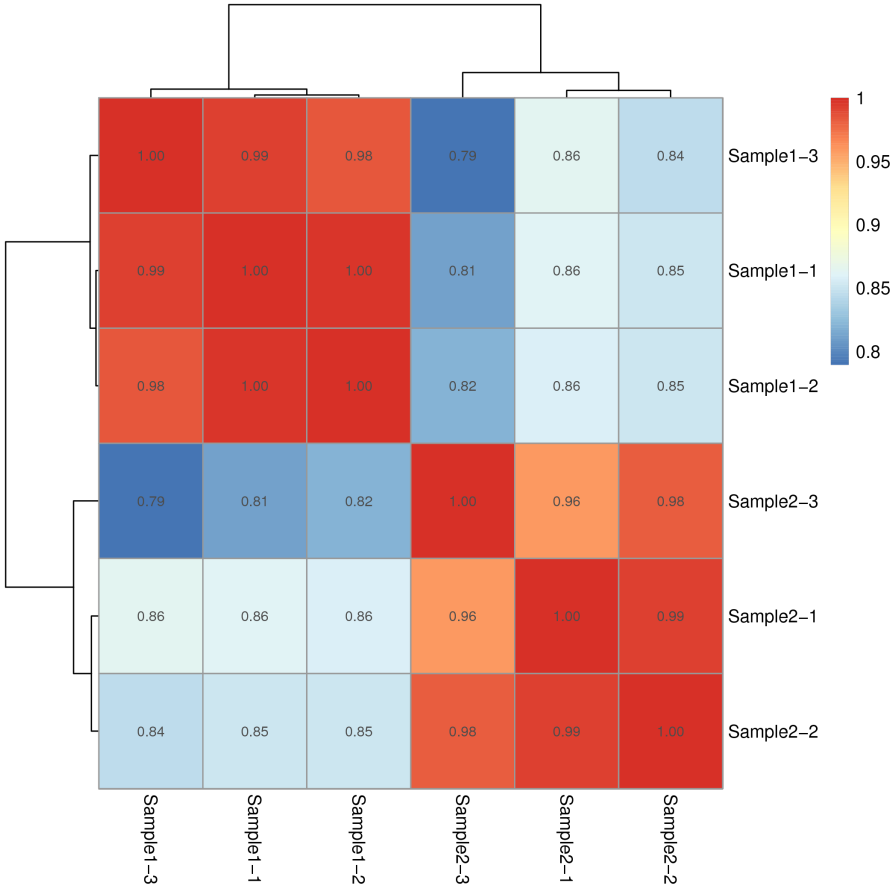


Figure 3.8.2.1 RNA-Seq correlation examination

R^2:pearson correlation; rho:spearman rank correlation; tau:kendall-tau rank correlation.

3.8.3 Sequencing homogeneity examination

Ideally, RNAseq reads are independently sampled and uniformly distributed across the transcriptome. However, many studies have identified factors that may affect read distribution (Dohm et al., 2008). For example, fragmentation and RNA reverse transcription during library construction may result in severe 3 'bias in RNA-seq results. Other factors including differences in GC content, random primer, RNA degradation also result in uneven coverage. Python script geneBody_coverage.py from RSeQC (V2.6.3) is used to assess sequencing homogeneity.

The algorithm for homogeneity calculation:

(1) Divide each transcript into 100 bins, from 5 'to 3'

(2) Calculate the average sequencing depth of each bin and normalize to maximum value.

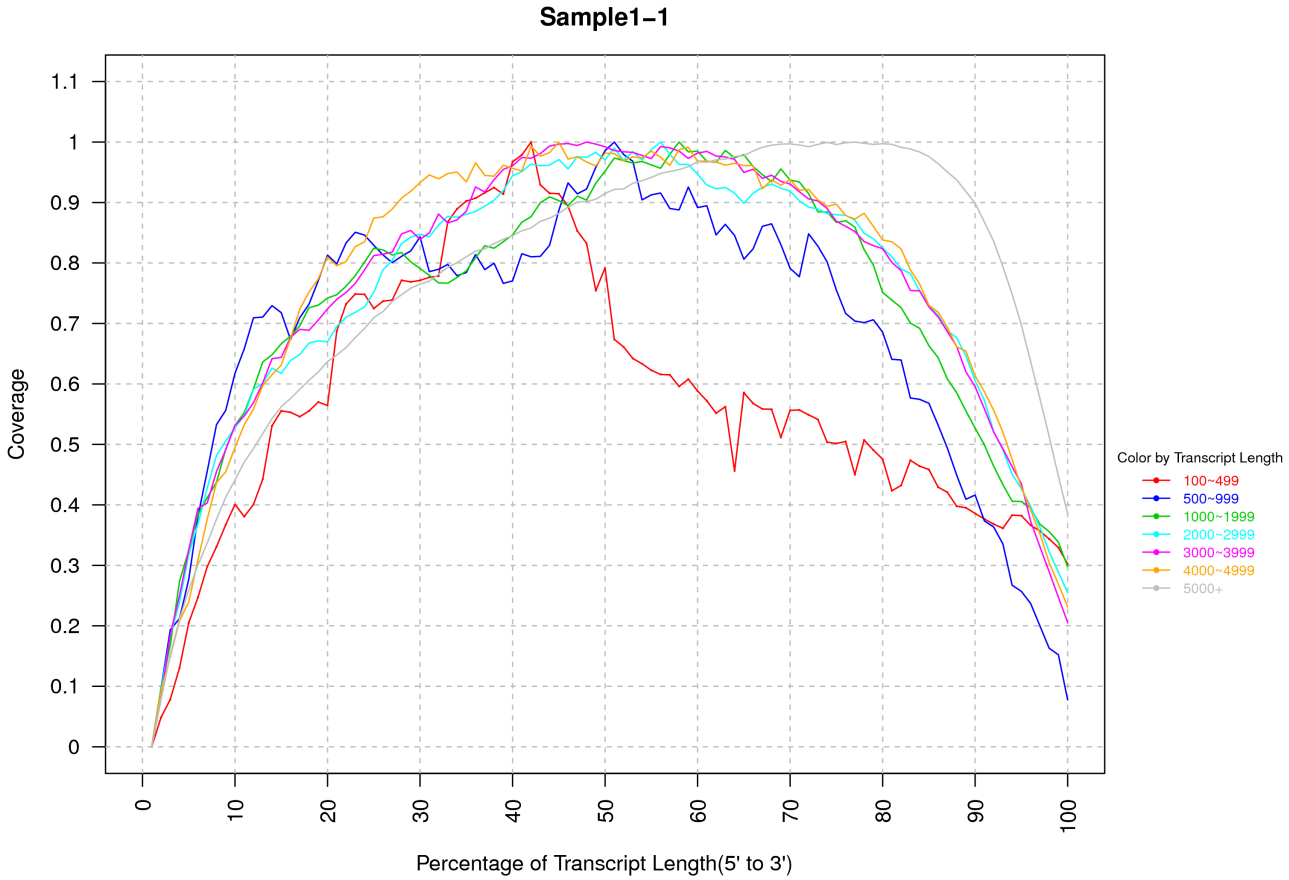


Figure 3.8.3.1 Read density distribution of transcripts of different lengths. X axis: the percentage of the length of the transcript. Y axis: the average sequencing depth. Color code is to specify transcripts of different lengths.

3.9 PCA analysis

PCA (Principal Component Analysis) reduces data complexity and is helpful to analyze sample relationship and the scales of the difference. The basic principle of PCA is to convert the original variables into a new set of independent variables (i.e. the principal components). All factors are ranked based on significance; minor factors and noise are eliminated, and thereby simplifies the data. Usually diagrams are made using two or three principal components as axes and conclusions on sample relationships can be drawn based on the distance between the various samples. Samples of close relationship tend to cluster together. The following figure shows the clustering relationships between samples:

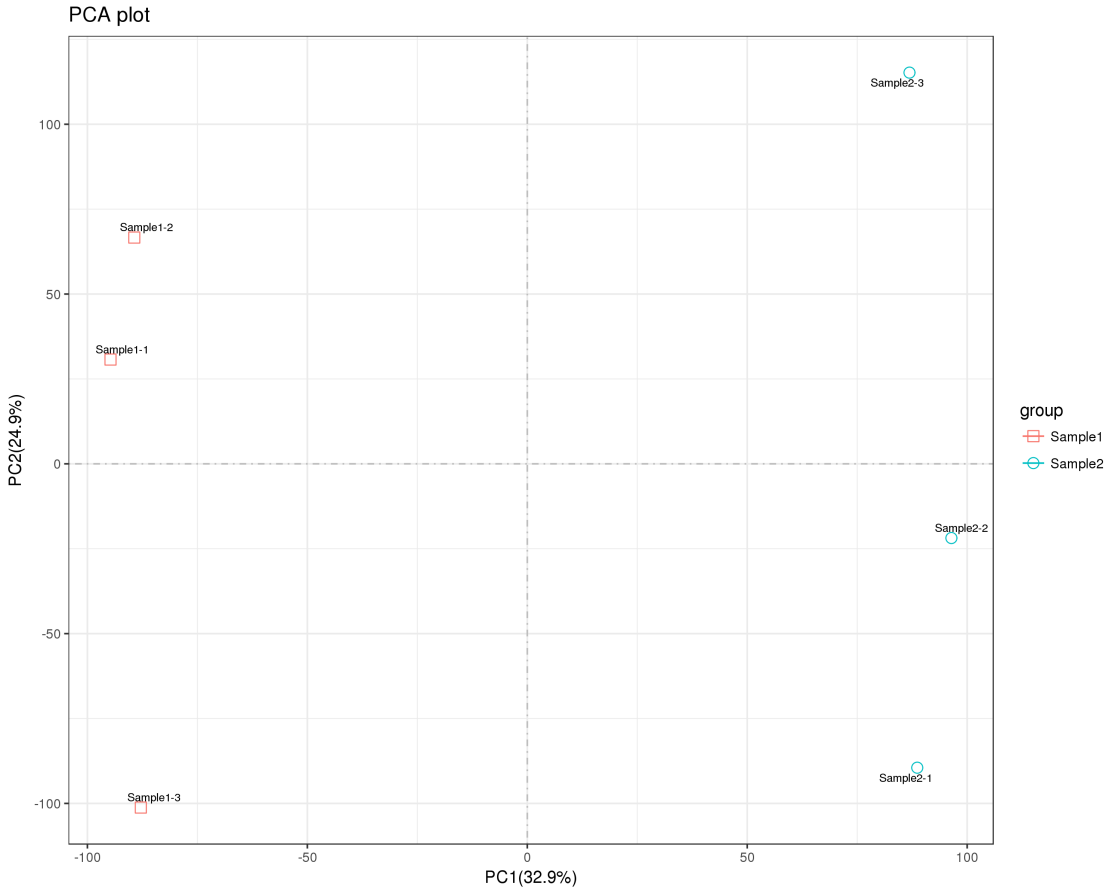


Figure 3.9.1 Principal component analysis chart, position the sample representative of the value of each dot on each of the main components

3.10 Gene differential expression analysis

3.10.1 Gene expression comparison

Expression levels of all genes under different experimental conditions were compared by FPKM profiles.

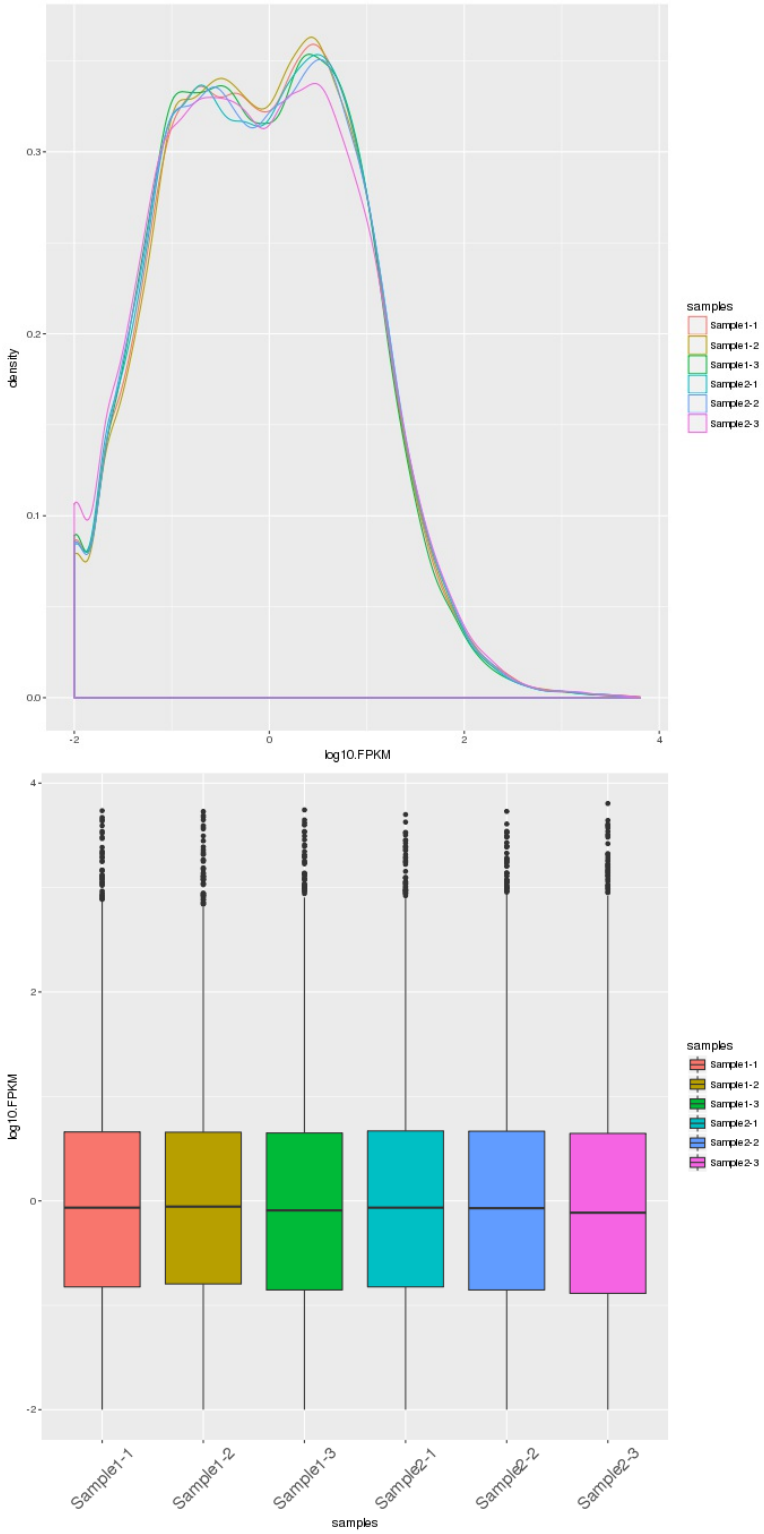


Figure 3.10.1.1 Comparison of gene expression levels under different experimental conditions Figure 1: FPKM distribution. X axis: log10FPKM. Y axis: number of genes in density of each FPKM value. Figure 2: FPKM box plot, X axis: sample names, Y axis: Log10 values of FPKM. Each of the five elements in each box plot, from top to bottom, specifies the maximum, upper quartile, median, lower quartile and the minimum value, respectively.

3.10.2 List of differentially expressed genes

The input data for gene differential expression is the read count data obtained from gene expression analysis.

For samples with biological replicates, gene differential analysis was performed using the Bioconductor package DESeq2 (V1.6.3), which was based on a model with a negative binomial distribution. If the read count of the i^{th} gene in the j^{th} sample is K_{ij}, then:

$$K_{ij} \sim NB(\mu_{ij},\alpha_{ij})$$

$$\mu_{ij}=s_j q_{ij}$$

$$\log_2 = x_j \beta_i$$

In special cases, gene differential analysis is performed using the bioconductor package edgeR (V3.4.6).

DESeq2 is used for this analysis.

Table 3.10.2.1 List of differentially expressed genes (Partial results are shown. For complete results please see: *_DE.xls)

gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Chr	Start	End	Strand	GeneSymbo
ENSG00000085662	7499.22131724389	2.87519660890771	0.0762793904888239	37.6929677922501	0	0	7	134442350	134459284	-	AKR1B1
ENSG00000090339	4336.81431701823	4.33511972742747	0.113754123487362	38.1095611704053	0	0	19	10270835	10286615	+	ICAM1

gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Chr	Start	End	Strand	GeneSymbo
ENSG00000105825	14227.7322883093	3.77388125235295	0.0748764387498663	50.401452250688	0	0	7	93885397	93890991	-	TFPI2

gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Chr	Start	End	Strand	GeneSymbo
ENSG00000108691	17364.4634728589	4.3723697719581	0.09288599819695	47.0724313333769	0	0	17	34255218	34257203	+	CCL2

gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Chr	Start	End	Strand	GeneSymbo
ENSG00000111331	3557.6827636357	4.49448039093373	0.114589304546654	39.2225121595344	0	0	12	112938352	112973249	+	OAS3

Column explain:

(1) gene_id: gene ID

(2) baseMean: the average of the normalized count values, dividing by size factors

(3) log2FoldChange: the effect size estimate

(4) lfcSE: the standard error estimate for the log2 fold change estimate

(5) stat: Wald test

(6) pval: calculated probability

(7) padj: p-value adjusted for multiple testing using Benjamini-Hochberg to estimate the false discovery rate

(8-11) Chr, Start, End, Strand: gene location, including chromosome, start position, end position and strand

(12) GeneSymbol: gene symbol

(13) Description: gene description

(14-16) GO_BP,GO_CC,GO_MF: gene ontology, including biological process,cellular component and molecular function

3.10.3 Determination of differentially expressed genes

The results from DESeq2 analysis was further analyzed to determine genes with significant differential expression according to the criteria of fold change greater than 2 and qvalue(fdr, padj) less than 0.05. The number of up- and down-regulated genes are summarized in Table below.

Table 3.10.3.1 Summary of gene numbers that are significantly up- or down-regulated between groups

Sample-VS-Sample	UPs	Down
Sample1-VS-Sample2	875	401

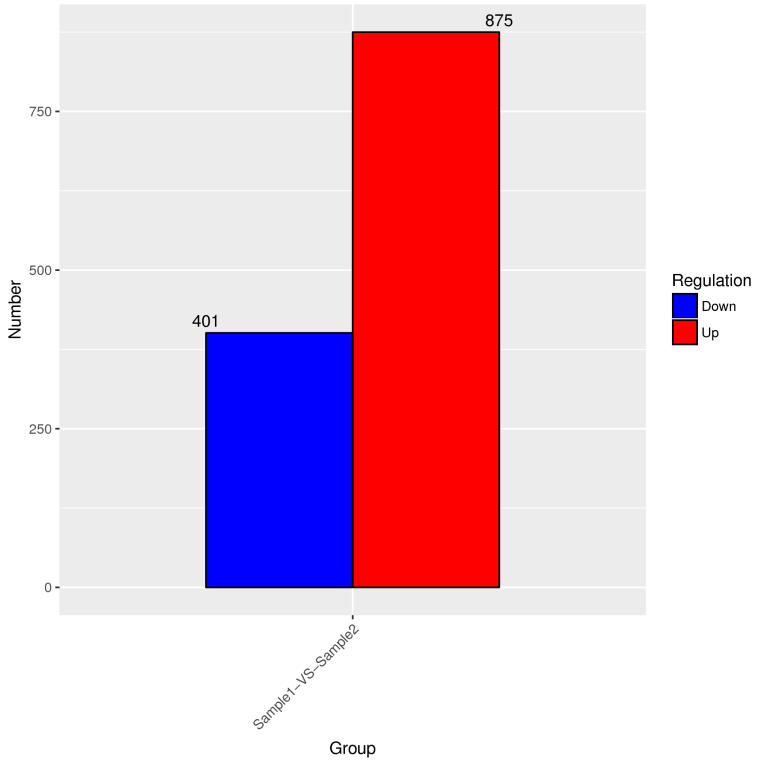


Figure 3.10.3.1 Bar graph of genes significantly up- or down-regulation between groups

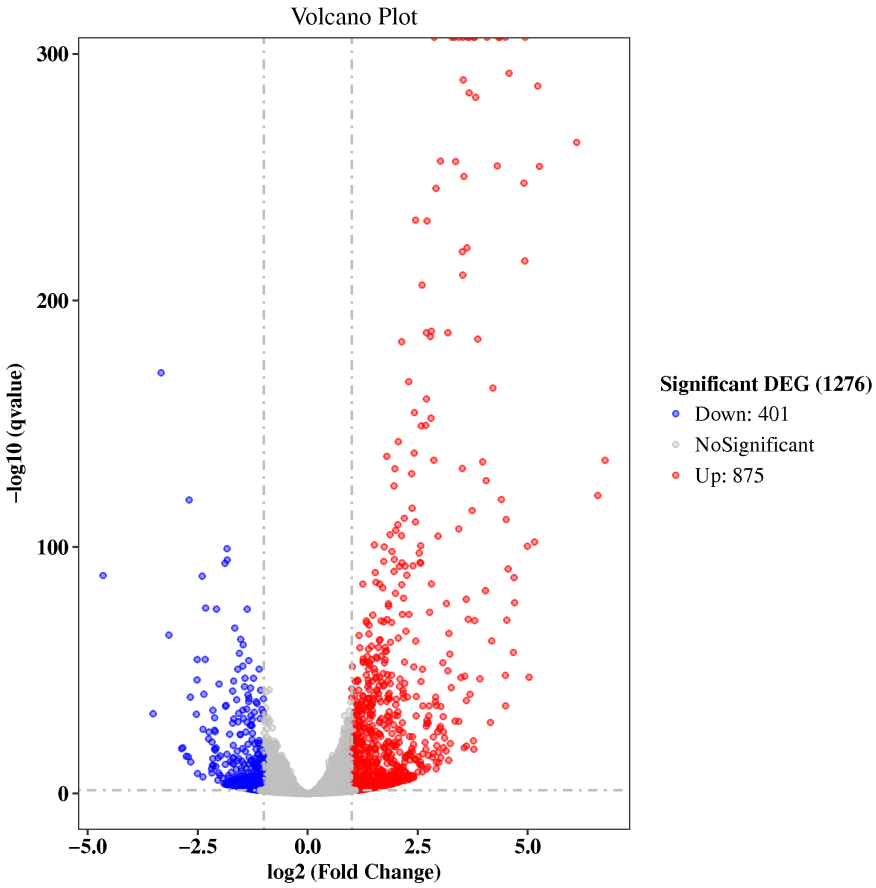


Figure 3.10.3.2 Differential expression volcano plot, red dots represent genes that are significantly up-regulated and blue dots represent those that are significantly down-regulated. X axis: log2 fold change of gene expression. Y axis: statistical significance of the differential expression in $\log_{10}$(qvalue(fdr, padj)).

3.10.4 Cluster analysis of differentially expressed genes

Clustering analysis is to calculate and classify data according to similarity, so that samples or genes with similar expression patterns can be grouped together. This can assist to predict the function of unknown genes, and to predict whether they participate in the same metabolic process or cellular pathway. The FPKM value of different genes under different experimental conditions was taken as the expression level and used for hierarchical clustering. The most obvious feature of this method is the generation of dendrogram. The regions of different colors represent different clusters. Genes with similar expression patterns are within the same cluster and close to each other, and they may have similar functions or participate in the same biological processes.

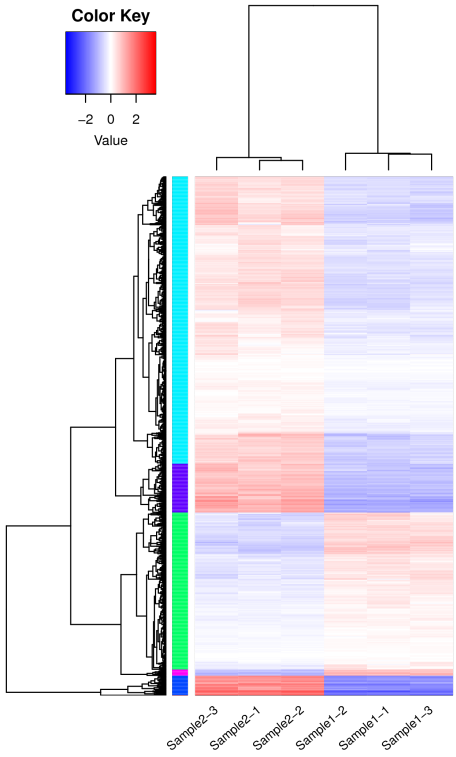


Figure 3.10.4.1 Cluster analysis of differentially expressed genes

Log_{10}(FPKM + 1) values are used for clustering. Genes of high expressed are in red, and low expression in blue.

3.10.5 Venn diagrams of differentially expressed genes

The Venn diagram shows the number of genes differentially uniquely expressed in each group or differentially expressed in multiple groups. Venn diagram are generated only when the number of groups is between 2 and 5.

3.11 Differential gene GO enrichment analysis

Gene Ontology (GO, Gene Ontology database) is an international standardized gene classification system, which provides a set of dynamically updated standard vocabulary to describe the properties of genes and gene products in the organism. GO contains three ontologies that describe the molecular function, cellular component, and biological process of the gene.

The GO functional enrichment analysis returns the GO terms that are enriched among differentially expressed genes against the genomic background, and thus provides information on how the differentially expressed genes are related to certain biological functions. The software we used here is GOSeq(Young et al, 2010), which based on an extensiton of the hypergeometric distribution known as the Wallenius non-central hyper-geometric distribution. This method is able to account for gene length bias and read counts bias when performing GO analysis. Threshold for filtering here is: over_represented_pvalue <= 0.05.

3.11.1 List of Differences Gene GO Enrichment

Table 3.11.1.1 GO enrichment of differentially expressed genes (Partial results are shown. For complete results please see: Sample1-VS-Sample2.xlsx)

category	term	ontology	numDEInCat	numInCat	over_represented_pvalue	over_represented_FDR	GeneNumber(Up)	GeneNumber(Down)
GO:0005615	extracellular space	CC	124	1056	3.15475155077932E-40	4.2800514289423E-36	100	24

Column explain:

(1) category: GO term ID

(2) term: description of gene ontology term

(3) ontology: gene ontology(CC: cellular_component; BP: biological_process; MF: molecular_function)

(4) numDEInCat: number of significant differential expression genes in the category

(5) numInCat: number of genes in the category

(6) over_represented_pvalue: pvalue, the smaller, the more significant

(7) over_represented_FDR: pvalue adjust

(8-9) GeneNumber(Up/Down): numbers of up or down reguation genes. These two columns are hyperlinks , which can link to their own significant differential expression genes

3.11.2 DAG of differential gene GO enrichment

Directed Acyclic Graph (DAG) is a graphical representation of the results of enrichment analysis of the differentially expressed genes. The branch represents the inclusion relation, and the functional range defined from the top to the bottom is in decreasing order. TopGO is applied to do the analysis, and generally, the top 5 enriched GO terms were selected as primary nodes of the directed acyclic graphs, and the associated GO terms are displayed by the inclusion relation. The color scale represents the degree of enrichment. DAG shows the enrichment of biological processes, molecular functions as well as cellular components.

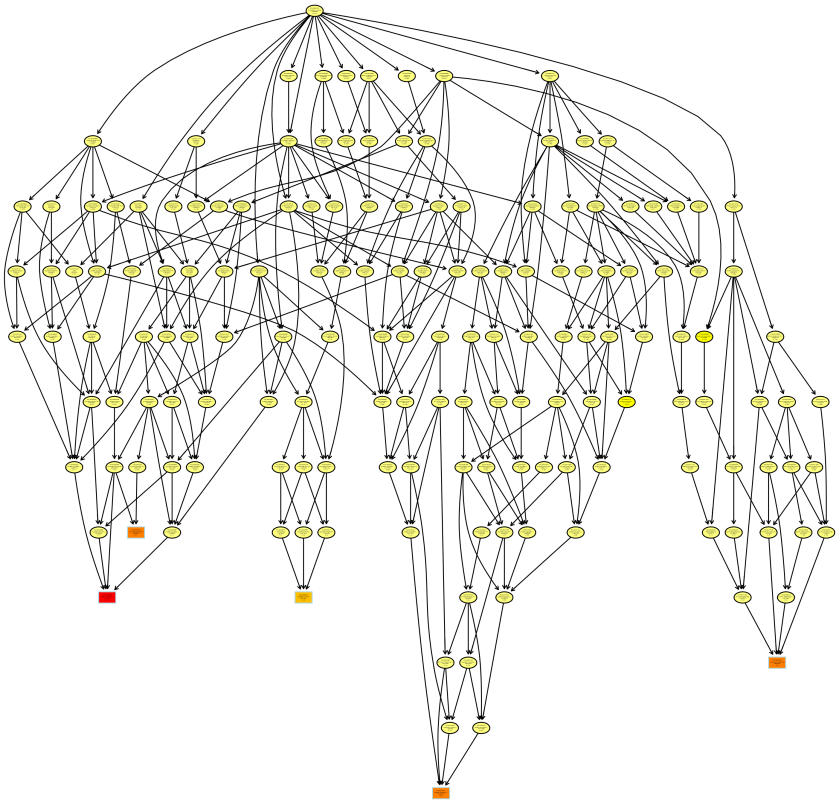


Figure 3.11.2.1 Directed Acyclic Graph

The subgraph induced by the top 5 GO terms identified by the classic algorithm for scoring GO terms for enrichment. Boxes indicate the 5 most significant terms. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). Black arrows indicate is-a relationships and red arrows part-of relationships.

3.11.3 Histogram of differential gene GO enrichment

The number differentially expressed genes in each GO term is shown in the histogram with the specification of the relevant biological process, cellular component and molecular function. Shown is the top 30 most prominent GO categories. All the output is shown if the analysis returns less than 30 outputs.

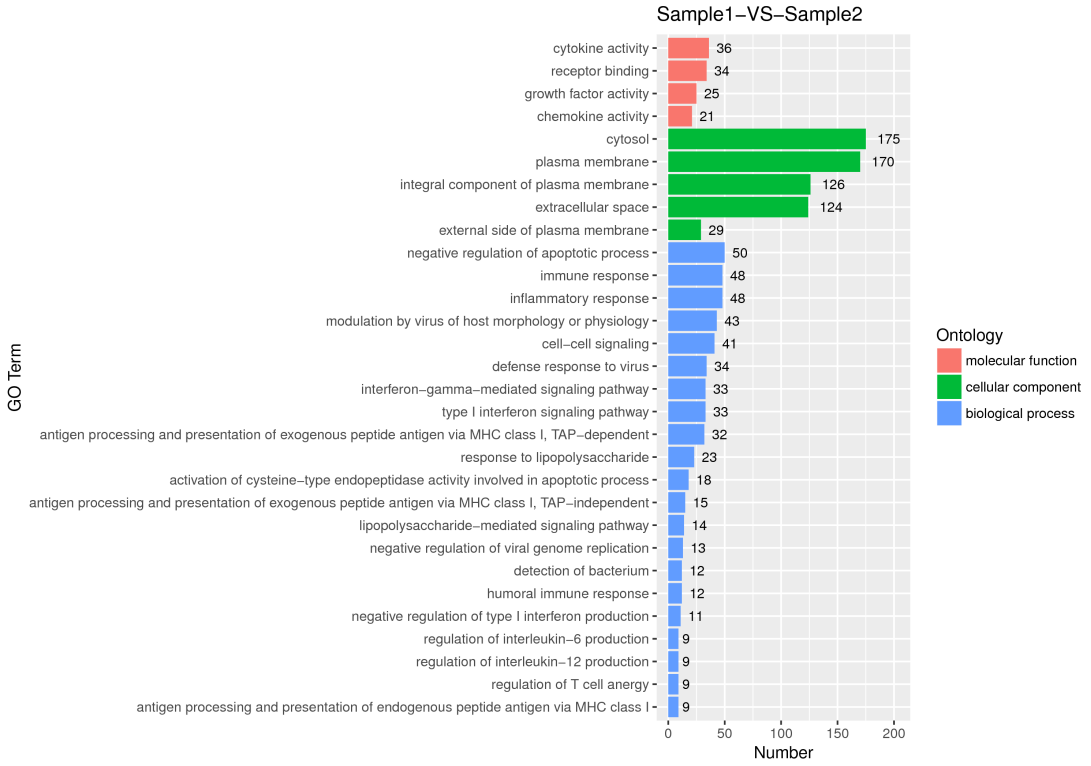


Figure 3.11.3.1 Figure 1: GO enrichment histogram. X axis: number of differentially expressed gene in this GO category. Color code is to distinguish the categories - biological processes, cellular components and molecular functions. Figure 2: GO enrichment pvalue histogram. X axis: -log10(p-value) of each term. Y axis: siginificant enriched go term.

3.12 DEU analysis

In addition to differential expression analysis of gene levels, differential exon usage (DEU) analysis was also performed to assess differential expression of exons. DEU analysis is currently the best method for studying alternative exon usage in alternative splicing. DEXSeq was used for DEU analysis (V1.18.4). DEXSeq uses a generalized linear model to detect the differential expression of genes on exon level. Padjusted < 0.05 was considered significant during DEU analysis. The exon differential expression result from DEU is shown in Table 3.12.1 and Figure 3.12.1. In Figure 3.12.1, exons differentially expressed are highlighted in light purple.

This analysis is only for samples with biological duplication. If there is no biological repeat, this analysis is not performed.

Table 3.12.1 DEU gene lists

groupID	GeneID	featureID	exonBaseMean	dispersion	stat	pvalue	padj	Sample1	Sample2	log2fold_M2.PA_M
ENSG00000075391:E032	ENSG00000075391	E032	13.3485109	0.007891411	29.23547058	6.41E-08	0.021770267	8.47910923	17.64304901	-1.057115288

Column explain:

(1) groupID: gene ID and exon ID

(2) GeneID: Gene ID

(3) featureID: exon ID

(4) exonbasemean: average expression after correction

(5) dispersion: statistic deviation

(6) stat: LRT statistics

(7) pvalue: statistical significance level

(8) padj: adjusted p value by BH

(9) ctrl: expression values of control group

(10) expr: expression values experimental group

(11) log2fold_ctrl_expr: log2 value of fold difference between control and experimental group

(12) genomicData.seqnames: chromosome ID

(13) genomicData.start: starting site of the gene

(14) genomicData.end: termination site gene

(15) genomicData.width: gene length

(16) genomicData.strand: the direction of strand

(17) countData.Sample: read count of each sample

(18) transcripts: gene transcript ID

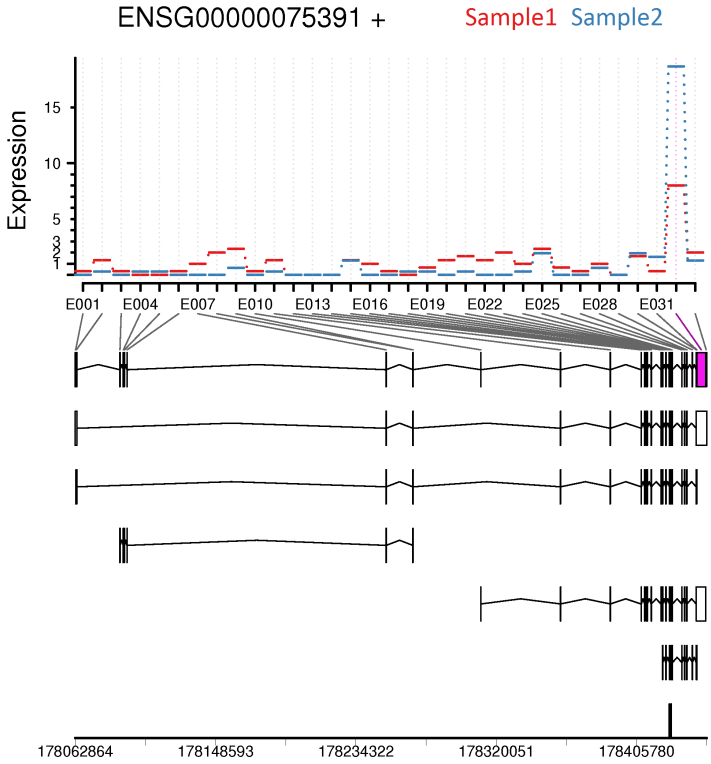


Figure 3.12.1 DEU analysis result. Y axis: expression level of each exon of the two sample groups. X axis: corresponding exons. All transcripts are of different splicing of the same gene. Exons that are differentially expressed are in bright purple.

3.13 PPI analysis

We performed analysis of the differential gene protein interaction network using the STRING protein interaction database (http://string-db.org/). The STRING database is a system to look for known protein interactions and predicted protein interactions. This interaction includes both the direct physical interaction and the indirect functional interaction. For species in the database, we extract the target gene set (such as the differential gene list) from the database to construct the network. For species not included in the database, we first used blastx to align enriched sequences in target genes to protein sequences of reference species in STRING database, and then construct the interaction network based on the aligned sequences from the reference species.

We provide differential gene network data files, which can be directly imported into Cytoscape (http://cytoscape.org/) for visualization and exploration. Users can graph and label the topological properties of the network. For example, the size of a node in an interaction network graph is proportional to the degree of the node - the more the edges connected to the node, the greater of its degree, and the bigger the node, and these nodes may be at key positons in the network. The color of the node is related to the clustering coefficient, and the color gradients from green to red correspond to the values of the aggregation coefficients from low to high. The aggregation coefficient indicates the connectivity between adjacent nodes of this node - the higher the aggregation coefficient, the better the connectivity between the adjacent nodes. According to different research purposes and needs, users can also adjust the network map node location and color, label

expression and perform other analysis. It should be noted that the results obtained by blast cannot guarantee accuracy, and the analysis in this section is only for preliminary scientific exploration to help users discover candidate genes. After importing the file into the Cytoscape, the result is as in Figure below.

Table 3.13.1 Differentially expressed gene protein interaction network data (Partial results are shown. For complete results please see: Sample1-VS-Sample2.interaction.xls)

protein1	protein2	combined_score	mode	action
ENSP00000000233	ENSP00000211287	190	activation	activation
ENSP00000000233	ENSP00000211287	190	binding	
ENSP00000000233	ENSP00000211287	190	catalysis	
ENSP00000000233	ENSP00000211287	190	inhibition	inhibition
ENSP00000000233	ENSP00000211287	190	ptmod	

Column explain:

(1) protein1: target gene protein ID

(2) protein2: interacting protein ID

(3) combined_score: combined score

(4) mode: relationship ("reaction", "expression", "activation", "ptmod"(post-translational modifications), "binding", "catalysis")

(5) action: type of effect ("inhibition", "activation")

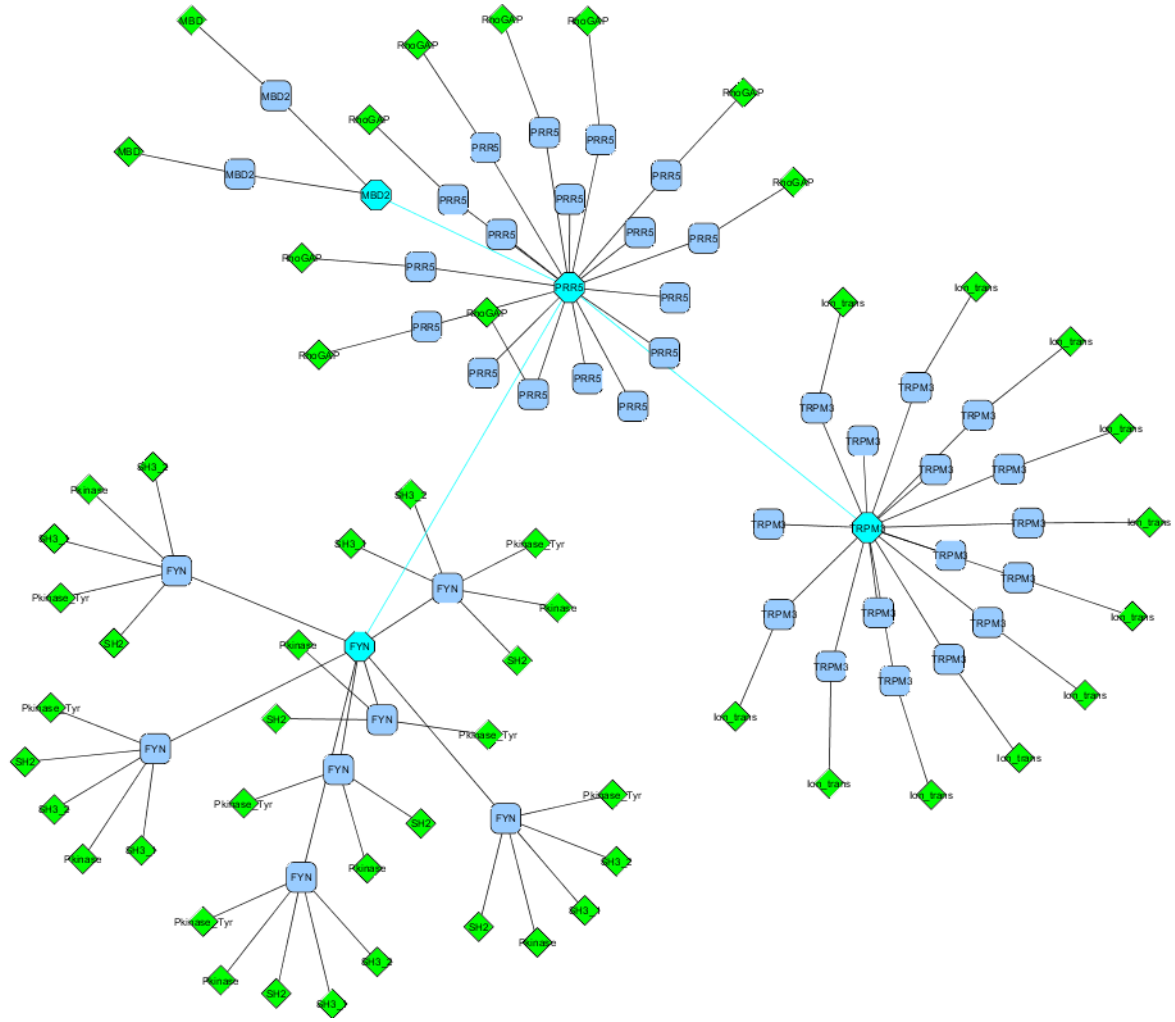


Figure 3.13.1 PPI example plot

3.14 Gene fusion analysis

Gene fusion gene refers to the chimeric combination of two or more genes under the control of the same transcriptional regulatory elements (including promoter, enhancer, ribosome binding sequence, terminator, etc.). The product of a fusion gene is a fusion protein. We used STAR-Fusion (v1.4.0) to study gene fusion events in the transcriptome. STAR-Fusion identifies fusion gene candidates by the genomic location of the aligned paired reads. Table 3.14.1 is an illustration of the fusion gene results in a tabular fashion. Figure 3.14.1 shows the fusion gene of selected by the customer.

Table 3.14.1 Gene fusion analysis result (Partial results are shown. For complete results please see: Sample.star-fusion.fusion_predictions.abridged)

#FusionName	JunctionReadCount	SpanningFragCount	SpliceType	LeftGene	LeftBreakpoint	RightGene	RightB
MIPEPP3--ABHD12	129	0	INCL_NON_REF_SPLICE	MIPEPP3^ENSG00000233325.3	chr13:21872464:+	ABHD12^ENSG00000100997.14	chr20:
HNRNPUL2--C11orf49	81	20	ONLY_REF_SPLICE	HNRNPUL2^ENSG00000214753.2	chr11:62494091:-	C11orf49^ENSG00000149179.9	chr11:
HNRNPUL2-BSCL2--C11orf49	81	20	ONLY_REF_SPLICE	HNRNPUL2-BSCL2^ENSG00000234857.2	chr11:62494091:-	C11orf49^ENSG00000149179.9	chr11:
GNB1--NADK	64	31	ONLY_REF_SPLICE	GNB1^ENSG00000078369.13	chr1:1822259:-	NADK^ENSG00000008130.11	chr1:1
FSD1L--SLC44A1	30	3	ONLY_REF_SPLICE	FSD1L^ENSG00000106701.7	chr9:108210516:+	SLC44A1^ENSG00000070214.11	chr9:1

Column explain:

(1) fusionName: name of the fusion as geneA--geneB

(2) JunctionReadCount: number of split RNA-Seq reads that map and define the fusion breakpoin

(3) SpanningFragCount: number of paired-end reads that span the fusion breakpoint but the reads do not directly overlap the breakpoint

(4) SpliceType: category of support at the fusion breakpoint: { ONLY_REF_SPLICE: fusion breakpoint occurs at reference (known) splice junctions. INCL_NON_REF_SPLICE: fusion breakpoint occurs at a breakpoint that does not involve all reference (known) exon junctions. NO_JUNCTION_READS_IDENTIFIED: only spanning fragments support the fusion. (Only happen if --min_junction_reads is set to zero) }

(5) LeftGene: identifier of the gene represented by the left section of the fusion transcript

(6) LeftBreakpoint: position of the left fusion breakpoint in the context of the genome

(7) RightGene: identifier of the gene represented by the right section of the fusion transcript

(8) RightBreakpoint: position of the right fusion breakpoint in the context of the genome

(9) LargeAnchorSupport: YES|NO, indicates whether there are at least 25 aligned bases on each side of the fusion breakpoint

(10) FFPM: normalized measure of the quantity of RNA-Seq fragments supporting the fusion event as fusion fragments per total million RNA-Seq fragments

(11) LeftBreakDinuc: the genomic dinucleotides found at the left breakpoint (putative splice site if splicing is involved)

(12) LeftBreakEntropy: entropy calculation for the 15 bases immediately upstream from the fusion junction

(13) RightBreakDinuc: the genomic dinucleotides found at the right breakpoint (putative splice site if splicing is involved)

(14) RightBreakEntropy: entropy calculation for the 15 bases immediately downstream from the fusion junction

(15) Annots: a simplified annotation for fusion transcript. For human source, the fusion annotation info based on CTAT_HumanFusionLib

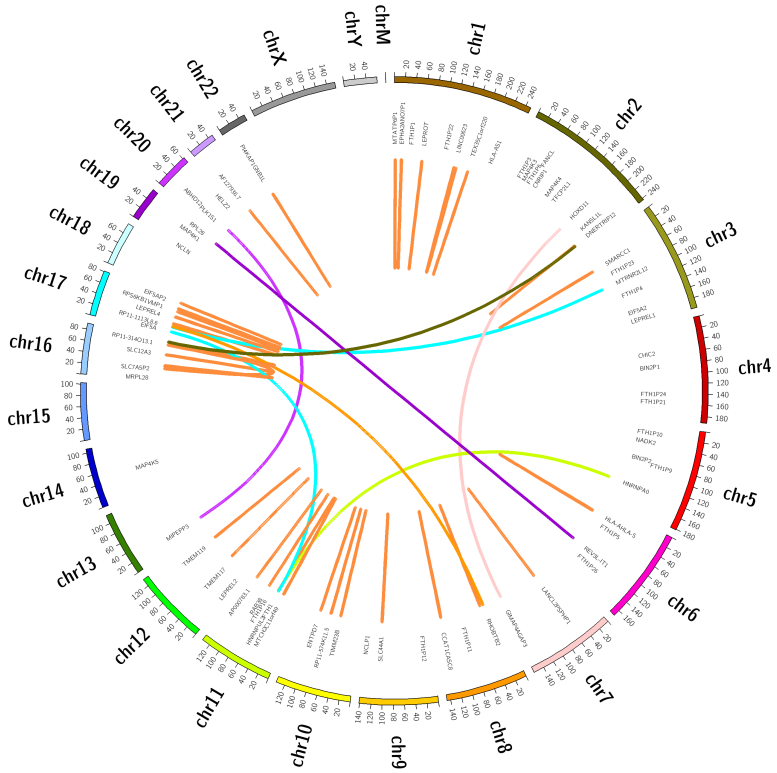


Figure 3.14.1 Gene fusion network diagram. Each line indicates a gene fusion event between the two genes connected by the line.

3.15 RNA editing analysis

RNA editing refers to the process of altering the genetic information on the mRNA level. Specifically, it refers to the deletion, insertion, or chemical modification of a nucleotide in the mRNA molecule. This type of modification affects the expression of genes, the production of different amino acids and the formation of open reading frames. In mammalian cells, one mechanism is the hydrolytic deamination at the C6 position of adenosine, which replaces the adenyl-amino group with an oxygen group, converting adenosine to inosine. Since both I and G are complementary to C in the same manner, this editing changes the coding information of the codon. This type of RNA editing is mediated by RNA-dependent adenine deaminase. RNA editing analysis was performed using GIREMI(0.3.1).

Table 3.15.1 RNA editing results (Partial results are shown. For complete results please see: Sample.RNAEdit.xls)

chr	coordinate	gene	reference_base	upstream_1base	downstream_1base	major_base	major_count	tot_count	major_ratio	MI	pvalue_MI	estimated
10	104500103	Inte	A	C	A	A	7	12	0.583333	0.02965	8.952776e-07	0.5
10	76777359	KAT6B	A	C	A	A	9	15	0.6	0.335419	0.00541161	0.623932
10	127599482	FANK1	A	T	G	G	5	6	0.833333	-1	-1	0.61991
10	24644288	KIAA1217	A	T	G	G	8	12	0.666667	0.309097	0.003071077	0.601836
10	76932138	SAMD8	A	T	G	A	6	10	0.6	0.693285	0.5233355	0.5

Column explain:

(1) chr: Name of the chromosome

(2) coordinate: Position of the SNVs in the chromosome (1-based)

(3) gene: Name of the gene harboring this SNV

(4) reference_base: The nucleotide of this SNV in the reference chromosome

(5) upstream_1base: The upstream neighboring nucleotide of this SNV in the reference chromosome

(6) downstream_1base: The downstream neighboring nucleotide of this SNV in the reference chromosome

(7) major_base: The major nucleotide of the SNV in the RNA-seq data

(8) major_count: Number of reads with the major nucleotide

(9) tot_count: Total number of reads covering this SNV in the RNA-Seq data

(10) major_ratio: The ratio of major nucleotide (major_count/tot_count)

(11) MI: The mutual information of this SNV if a value exists

(12) pvalue_MI: P-value from the MI test if applicable

(13) estimated_allelic_ratio: Estimated allelic ratio of the gene harboring this SNV

(14) RNAE_t: Type of RNA editing or RNA-DNA mismatches (A-to-G, etc)

(15-18) [A, C, G, T]: Numbers of reads with specific nucleotides at this site

(19) ifRNAE: 1: the SNV is predicted as an RNA editing site based on MI analysis. 2: the SNV is predicted as an RNA editing site based on GLM. 0: the SNV is not predicted as an RNA editing site

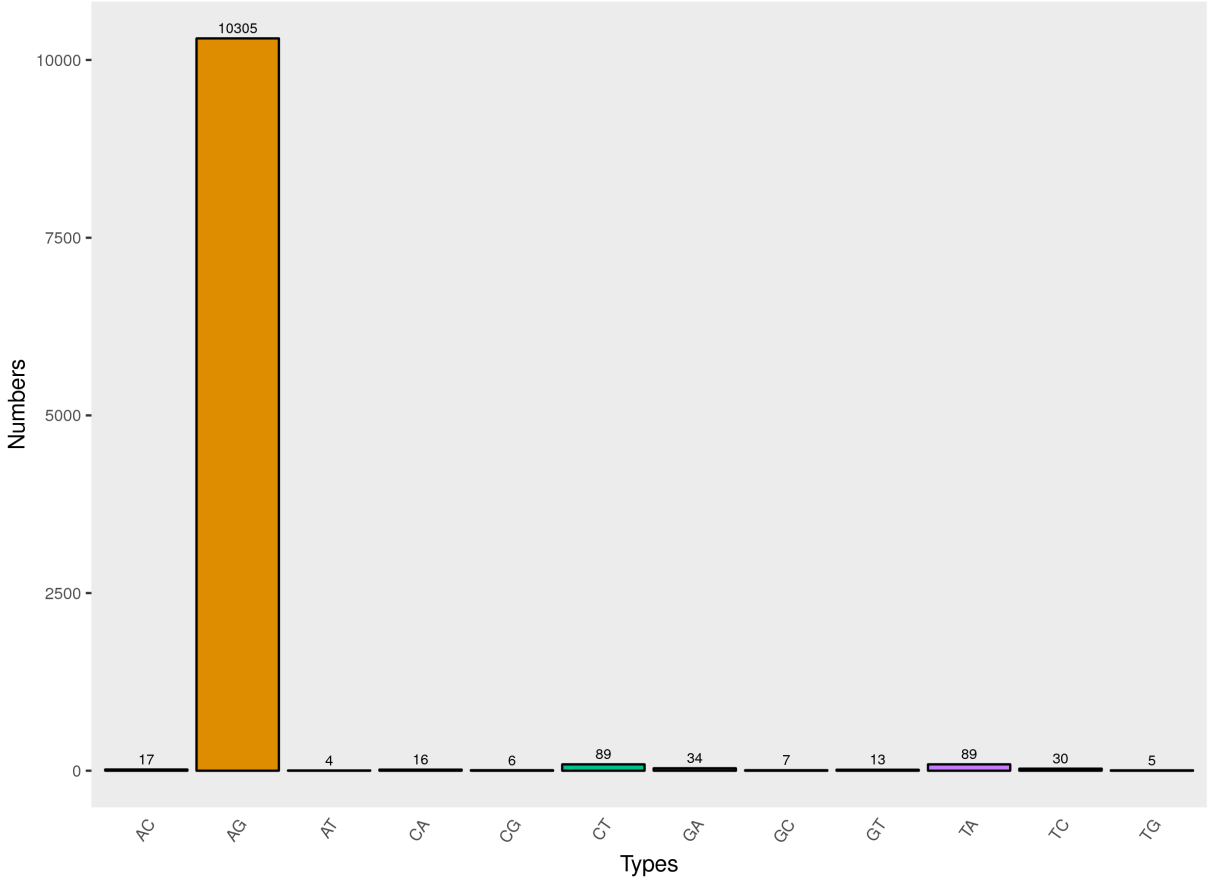


Figure 3.15.1 Statistic summary of RNA editing types

3.16 LncRNA prediction analysis

LncRNA is a type of non-coding RNA with a length greater than 200 nt. It does not encode proteins, but has a wide range of regulatory functions of organisms. Novel transcripts can be discovered from RNA-seq results and potentially novel lncRNA can be identified. Identification and prediction of lncRNA involves the following steps:

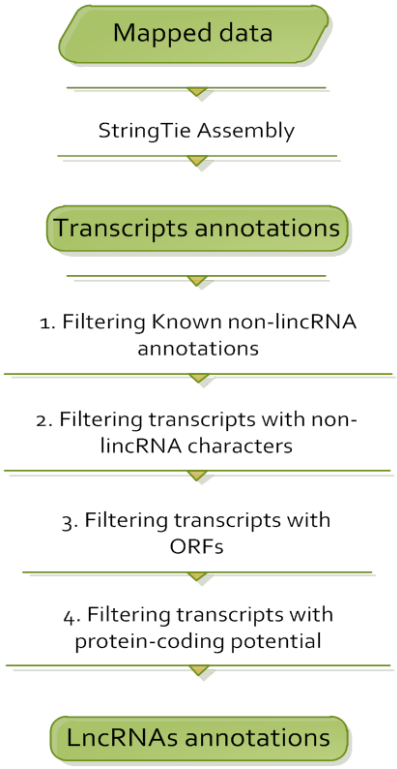


Figure 3.16.1 lncRNA identification and prediction workflow

3.16.1 Removal of other types of RNAs

The first step of filtering was to remove mRNA, miRNA, tRNA, snoRNA, rRNA and pseudogenes based on annotation provided by UCSC, Ensembl and GENCODE.

3.16.2 Removal of based on the characteristics of lncRNA

According to the characterization of lncRNA by GENECODE v7, lncRNAs tend to contain two exons and are no less than 200 bp. In addition, low sequencing coverage is usually associated with higher error rate, and therefore we also removed sequences with low sequencing coverage. The workflow is as follows:

(1) Filter out transcripts containing only one exon

(2) Filter put transcripts less than 200 bp

(3) Filter out transcripts with read count less than 3

3.16.3 Removal of transcripts containing protein domains

A protein domain is a conserved part of a protein with specific structure and independent function. Different domains of a protein are encoded by different exons of the gene. Therefore, ORF prediction of all the possible coding regions of transcripts was further used as a filtering criteria to reduce false-positive lncRNA sequences.

HMMER-3 was used to evaluate all possible open reading frames of transcripts. HMMER-3 aligns all transcripts with possible amino acid sequences to all known protein family members in the Pfam database to identify protein domains that the transcript may contain. Evalue is set to the default value (1e-5) and the result is in tab format. The result is shown in Figure 3.16.3.1.

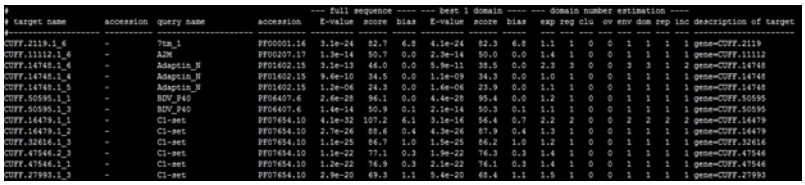


Figure 3.16.3.1 ORF prediction by hmmsearch

3.16.4 Removal of transcripts with protein-coding potential

CPC (Coding Potential Calculator) was used to predict protein-coding sequences. CPC utilizes the support vector machine algorithm to establish protein coding potential classification models based on features including the length of peptide chain, amino acid composition, protein homology, secondary structure and expression. The model predicts all possible translations of the input transcript sequences. Based on the predicted results, transcripts containing the protein coding potential are filtered out.

3.16.5 lncRNAs Statistics

Subsequently, the structure and sequence information of all lncRNA were summarized in following table:

Table 3.16.5.1

Samples	Sequences	Bases	Min	Max	Average	N50	(A+T)%	(C+G)%
Transcript	125775	81178494	201	22166	645.43	847	58.28	41.72

Column explain:

(1) Sequence: number of transcripts

(2) Bases: number of bases of all transcripts

(3) Min/Max/Average: transcript minimum, maximum and average length

(4) N50: transcript N50 value

(5) (A+T)%: percentage of A and T bases

(6) (C+G)%: percentage of C and G bases

3.16.6 lncRNA description and statistical information

3.16.6.1 Classification of known and unknown lncRNAs

We integrated Ensembl, Gencode UCSC databases for the annotation of known lncRNA and employed Cuffcompare for annotation. The results are as follows:

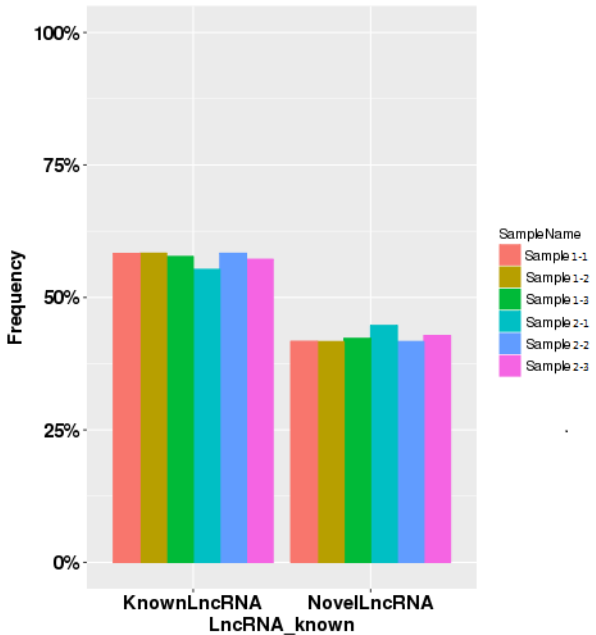


Figure 3.16.6.1.1 Information on the known and unknown lncRNAs

3.16.6.2 Distribution of lncRNA according to length, exon count and classification

The length of lncRNA and the number of exons contained in the lncRNA were analyzed. The lncRNA was further divided into three groups according to their genomic location: intergenic lncRNA, intronic lncRNA, and antisense lncRNA. The number of lncRNAs in each category was also summarized.

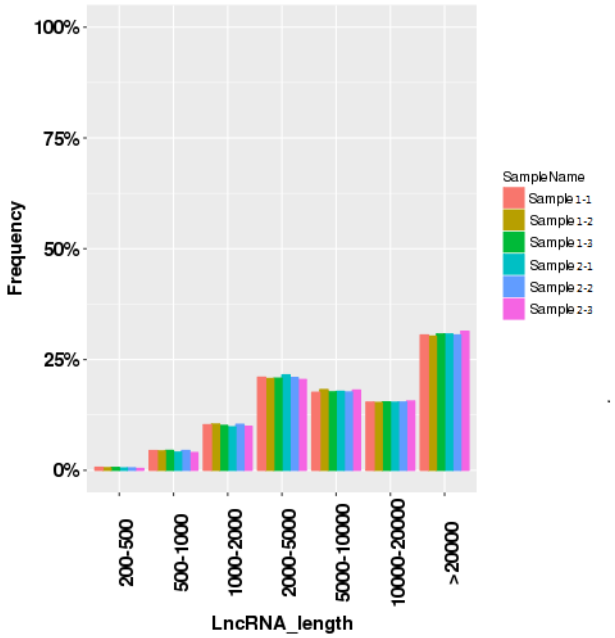


Figure 3.16.6.2.1 LncRNAs length distribution

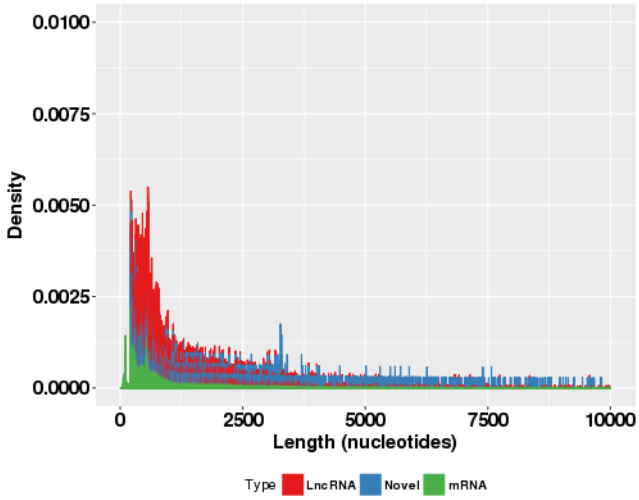


Figure 3.16.6.2.2 Length distribution of different types of RNA transcripts

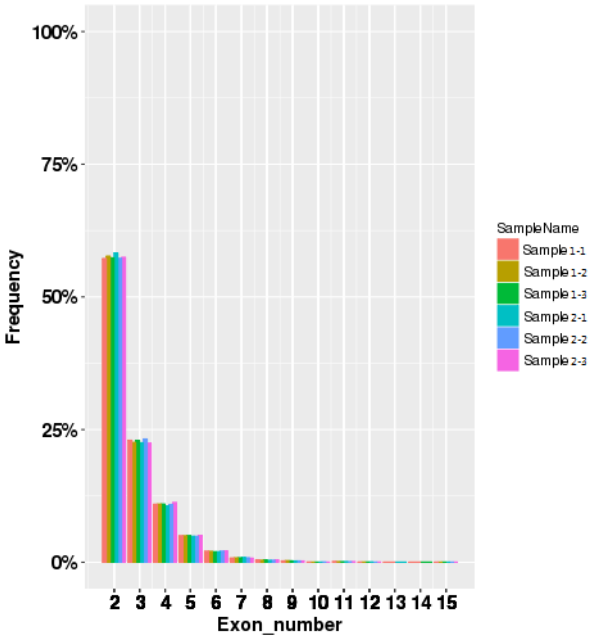


Figure 3.16.6.2.3 Distribution of exon numbers in lncRNAs

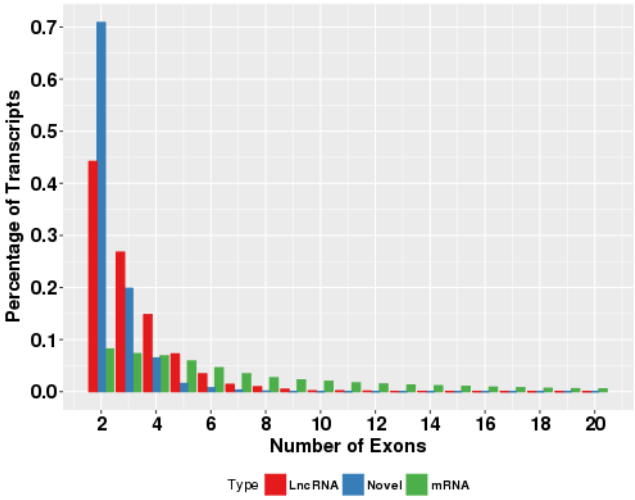


Figure 3.16.6.2.4 Distribution of exon numbers in different type of RNA transcripts

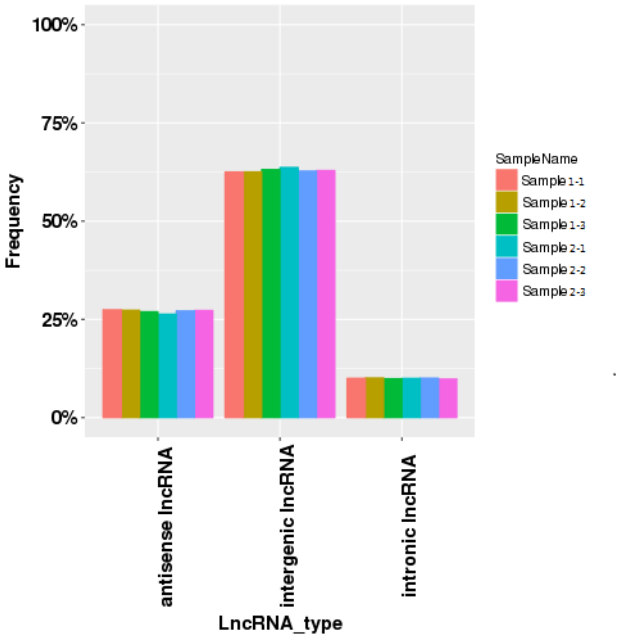


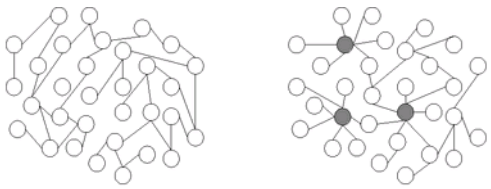
Figure 3.16.6.2.5 Distribution of exon numbers in different types of lncRNAs

3.17 Co-expression network analysis

Gene co-expression refers to the phenomenon that some genes have similar expression profiles. This similarity suggests that they are regulated by similar factors and mechanisms. Gene co-expression network is a scale-free network in which the nodes represent genes, and the edges between genes are determined by the expression levels of two related genes. Co-expressed genes are in the same gene co-expression network. Using the co-expression network, researchers can analyze the gene regulatory activity and identify key regulators of gene expression. Construction of gene co-expression network is based on gene expression data, which is obtained from gene expression analysis of the RNA-Seq data. R package WGCNA was used to construct gene co-expression network and the result display was through Cytoscape. For reliable analysis, we recommend to include at least 15 DGE samples in the analysis.

3.17.1 Construction of co-expression network

Appropriate soft-thresholding was first determined for the construction of scale-free network. Next TOMSimilarity was employed to calculate the co-expression similarity coefficient between genes to realize the functional connection based on the soft-thresholding and gene expression information. The schematic diagram is as follows:



(a) Random network (b) Scale-free network

Figure 3.17.1.1 Co-expression scale-free network diagram, from the web.

3.17.2 Cluster Analysis for gene expression module identification

Sometimes co-expression networks can be complicated for application due to the large number of genes. Cluster analysis is very helpful to identify the main effect genes. WGCNA uses unsupervised clustering for gene clustering and classification according to functional similarity. TOM diagram was generated using TOMplot () as shown below:

In the figure below, different modules are represented by different color codes. A module is defined as a set of genes with similar expression patterns. If certain genes always have a similar trend of expressional regulation in a physiological process or in different tissues, it is

reasonable to assume that these genes are functionally related and they are classified into one module. The biological meaning of each gene expression model is further explored using GO and biological pathway analysis.

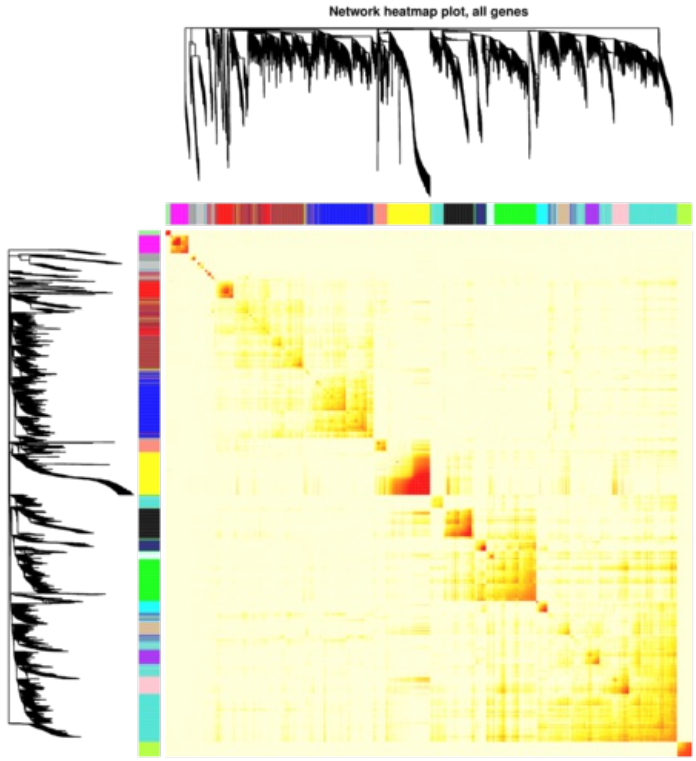


Figure 3.17.2.1 Cluster analysis heatmap the map above and to the left are the phylogenetic trees of the transcriptome, different modules are represented by different color code. Each cross point indicates the relationship of a particular gene with others. The brighter the point, the stronger the relationship.

3.17.3 Core module selection

3.17.3.1 Module feature gene selection

To facilitate the correlation analysis of the module with other data sets such as phenotype information, it is necessary to define a feature gene in each module. This feature gene can be representative of the module's feature with acceptable degree of information loss. A big advantage of doing so is to simplify the calculation and obtain results in a timely manner even if dealing with very large amount of data. A co-expression network is constructed for each module, and the gene at the node is the feature gene of the module.

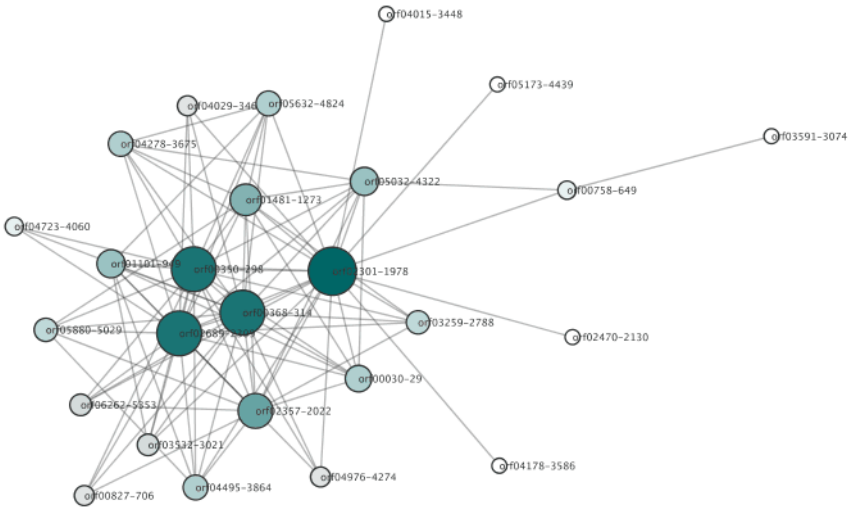


Figure 3.17.3.1.1 Gene expression module network diagram

3.17.3.2 Association analysis between gene modules and known biological features

There are different ways to determine the association between gene module and known biological features (1) calculated the eigenvalues of the gene network / module and the correlation coefficient between the eigenvectors of the module and the phenotype of interest. (2) for the grouped phenotype data such as disease status, p values of the differential expression of each individual gene in various groups (e.g. disease group vs control group) is calculated and the gene significance (GS) is defined as $\log_{10}$-transformed p value. Module significance (MS) is defined as the average GS value of genes in the module. Then MS values are compared. In general, a higher than the average MS value is an indication of an association between the module and the disease. (3) predict the gene network/module formation based on the key genes in the networks that have high degrees of connections with other genes.
Based on the biological characteristics of the samples and the gene expression profiles of the modules, correlation coefficient and p value were calculated and modules with biometric correlations were selected.

3.17.3.4 Function analysis of gene expression modules

Function analysis of gene expression modules was performed using WGCNA and the associated databases. GO and KEGG enrichment was analyzed to predict the function of each module. Modules enriched in relevant biological functions are selected as the core modules.

3.18 Differential alternative splicing analysis

Alternative splicing allows a single gene to produce multiple mRNA transcripts, and different mRNA transcripts are translated into different forms of proteins to increase diversity (Black, 2003; Stamm, 2005; Lareau, 2004).
We use rMATS (v3.2.5) to detect alternative splicing event. The classification categorization of alternative by rMATS is shown below:

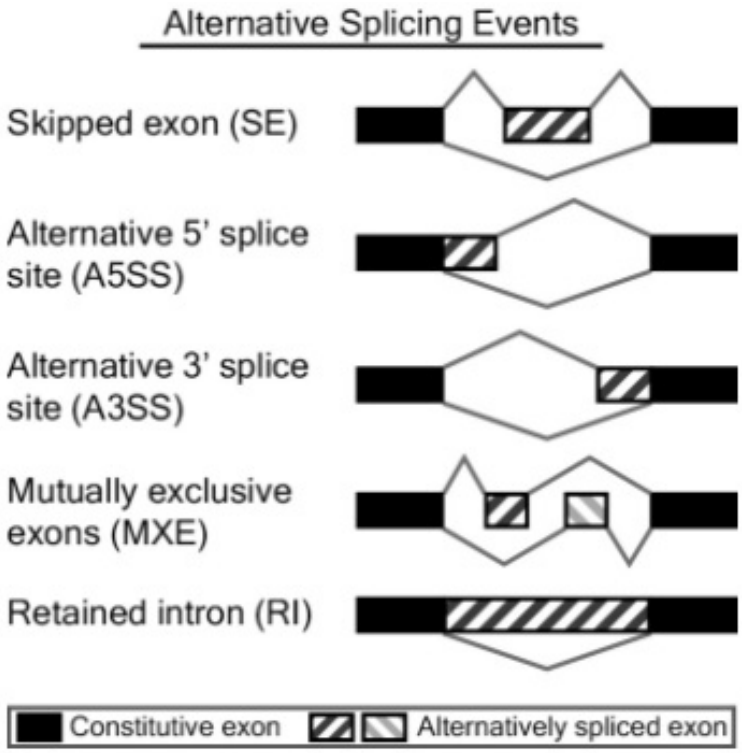


Figure 3.18.1 Basic alternative splicing categories

3.18.1 Differential alternative splicing filtering

The results was further analyzed to determine alternative splicing with significant differential expression according to the criteria of IncLevelDifferenc greater than 0.02 and FDR less than 0.05. The number of alternative splicing categories are summarized in Table below.

Table 3.18.1.1 Differentially alternative splicing statistic

Group	A3SS	A5SS	MXE	RI	SE
Sample1-VS-Sample2	141	29	46	35	23

3.18.2 Differential alternative splicing results

Differential alternative splicing results are shown in table below.

Table 3.18.2.1 Differential alternative splicing results (Partial results are shown. For complete results please see: *.AS_anno.xls)

GeneID	geneSymbol	chr	strand	longExonStart_0base	longExonEnd	shortES	shortEE	flankingES	flankingEE	IC_SAMPLE_1	SC_SAMPLE_1
ENSG00000181163	NPM1	5	+	170819916	170819982	170819917	170819982	170819713	170819820	0	5277
ENSG00000243678	NME1-NME2	17	+	49248847	49248969	49248865	49248969	49246743	49247410	5282	0
ENSG00000179218	CALR	19	+	13054533	13054704	13054647	13054704	13054026	13054443	6465	127
ENSG00000179218	CALR	19	+	13054530	13054704	13054647	13054704	13054026	13054443	6653	127
ENSG00000179218	CALR	19	+	13054613	13054704	13054647	13054704	13054026	13054443	2386	127

Column explain:

(1) GeneID: Gene ID

(2) geneSymbol: Gene symbol

(3) chr: Chromosome ID

(4) strand: reference strand information of the AS event

(5) longExonStart_0base: start of the long exon (0-base)

(6) longExonEnd: end of the long exon(1-base)

(7) shortES: start of the short exon (0-base)

(8) shortEE: end of the short exon (1-base)

(9) flankingES: start of the flanking exon (0-base)

(10) flankingEE: end of the flanking exon (1-base)

(11) IC_SAMPLE_1: Inclusion junction count for first sample

(12) SC_SAMPLE_1: Skipping junction count for first sample

(13) IC_SAMPLE_2: Inclusion junction count for second sample

(14) SC_SAMPLE_2: Skipping junction count for second sample

(15) IncFormLen: length of inclusion form, used for normalization

(16) SkipFormLen: length of skipping form, used for normalization

(17) PValue: Significance of splicing difference between two sample groups

(18) FDR: False Discovery Rate calculated from p-value

(19) IncLevel1: inclusion level for SAMPLE_1 replicates (semicolon separated) calculated from normalized counts

(20) IncLevel2: inclusion level for SAMPLE_2 replicates (semicolon separated) calculated from normalized counts

(21) IncLevelDifferenc: average(IncLevel1)- average(IncLevel2)

(22) significant: Up/Down regulation，'no' indicates the gene is not significant differential.

rMATS outputs can be converted into sashimiplots shown below by rmats2sashimiplot.

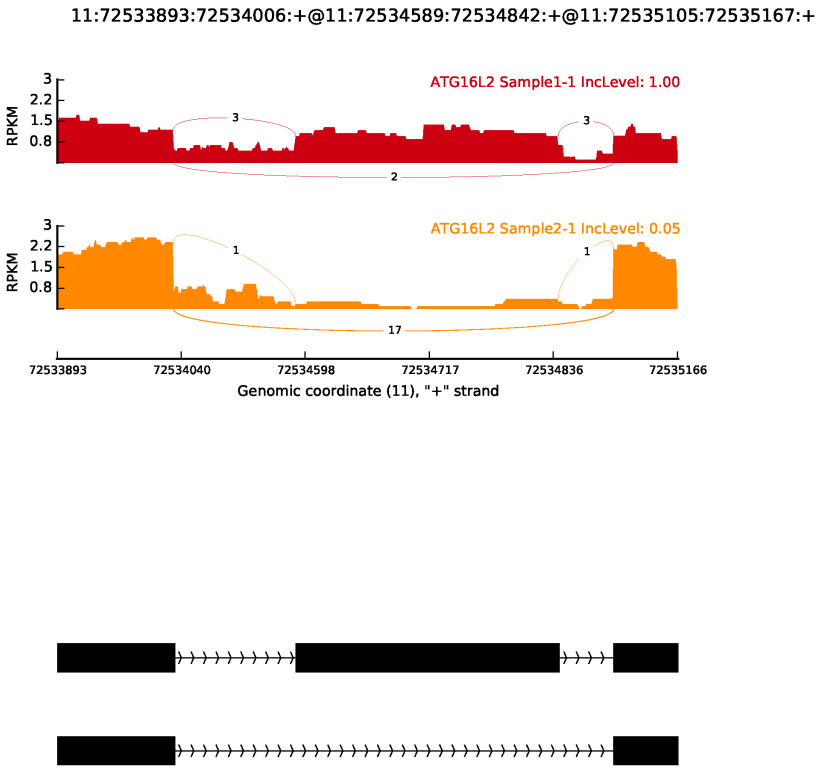


Figure 3.18.2.1 Differentail alternative splicing visualization. Sashimi plot (stand-alone) for alternatively spliced exon and flanking exons in group samples (colored by experimental condition). Per-base expression is plotted on y-axis of Sashimi plot, genomic coordinates on x-axis, and mRNA isoforms quantified are shown on bottom (exons in black, introns as lines with arrow heads).

3.19 Short time series gene expression analysis

The analysis of time series gene expression has enabled insights into development, response to environmental stress, cell cycle progression, pathogenic infection, cancer, circadian rhythm, and other biomedically important processes. Gene expression is a tightly regulated spatiotemporal process. Genes with similar expression dynamics have been shown to share biological functions. Clustering reduces the complexity of a transcriptional response by grouping genes into a small number of response types. Given a set of clusters, genes are often functionally annotated by assuming guilt by association, sharing sparse functional annotations among genes in the same cluster. Furthermore, regulatory mechanisms characterizing shared response types can be explored using these clusters by, for example, comparing sequence motifs or other features within and across clusters.

Short Time-series Expression Miner (STEM) is applied in the analysis, which is for clustering, comparing, and visualizing short time series gene expression data from experiments (3–8 time points).

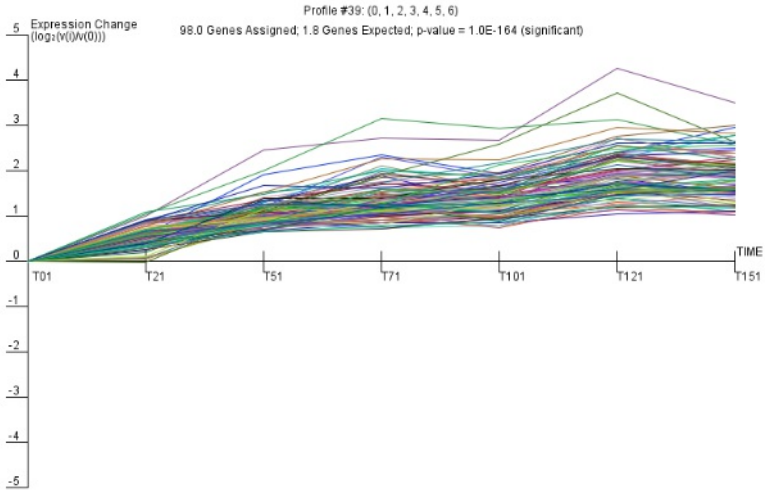


Figure 3.19.1 Example of detailed model profile information windows. The window plots a graph of all genes assigned to the profile, the x-axis scaled to be based on real time and the y-axis to be uniform. The text at top gives information about the profile including the number of genes assigned, the number of genes expected, and the p-value significance.

Table 3.19.1 STEM result table

gene_id	T01	T21	T51	T71	T101	T121	T151
ENSMUSG00000000031	0	0.29	0.65	0.96	0.82	1.54	1.51
ENSMUSG00000000093	0	0.09	0.08	0.66	0.64	1.13	0.97
ENSMUSG00000000094	0	1.08	0	1.08	1.16	2.56	2.23
ENSMUSG00000000120	0	0.62	0.29	0.55	0.53	0.82	1.24
ENSMUSG00000000125	0	1	1	1.09	1.42	1.05	2.14

Column explain:

(1) gene_id: gene id

(2) Sample: expression value after transform of each sample

3.20 GSEA analysis

Gene set enrichment analysis (GESA) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes. Microarray and proteomics results often identify thousands of genes which are used for the analysis. GSEA analysis steps:

(1) Sorting all genes according to certain indicators. We can pre-sort them manually, for example, according to the P value of the difference analysis results. The software itself also provides several sorting methods, such as correlation expression with phenotype.

(2) Marking a particular type of gene in a sort, the target gene can be a pathway or a GO term, etc.

(3) Using the weighting method to calculate the change in ES (Enrichment Score) value. If you encounter a identified gene, increase ES, and

vice versa.

(4) After running the statistics, the ES curve maximum seat can be enrichment score.

(5) Making permutation test， calculating P value and FDR according to enrichment score.

GSEA enrichment result showed below:

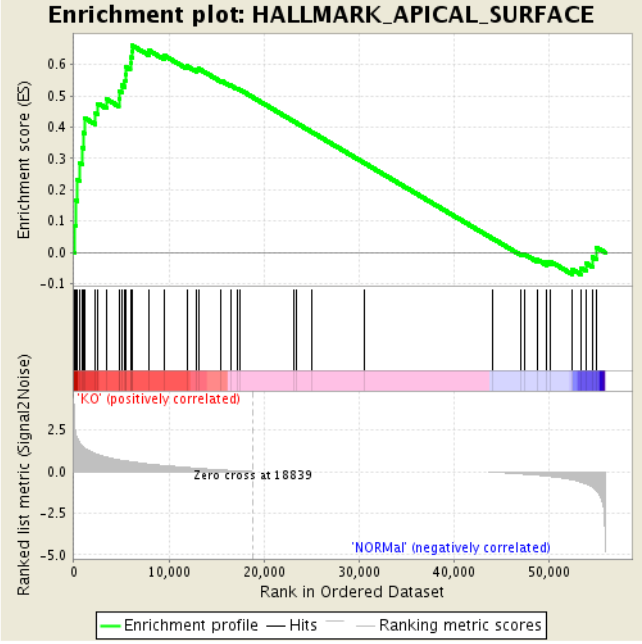


Figure 3.20.1 GSEA enrichment map. The top figure shows the increase and decrease curve of the ES value accumulation process, the middle figure shows the position of the target gene set members in all ordering gene (black vertical line), and the bottom figure shows the sorted genes from high to low, the true value of the index used for ranking (there is corrected by standard deviation, considering the relative difference value of genes, the effect is similar to P value).

3.21 Transcription factor analysis

Transcription regulation is an important part of the regulation of gene expression, and transcription factor (TF) regulated gene by combining gene upstream specific nucleotide sequence.

Plant transcription factor identification was conducted using the plant transcription factor database PlantTFDB4.0 and hmmsearch according to the pfam file of the transcription factor family.

Animal transcription factor identification was conducted using the animal transcription factor database AnimalTFDB2.0.

TF annotation results are shown below:

Table 3.21.1 TF annotation example

gene_id	TF_Family	TF_ID
Glyma.10G071700	bZIP	Glyma.10G071700.3.p
Glyma.12G116900	C3H	Glyma.12G116900.1.p
Glyma.07G038200	AP2	Glyma.07G038200.1.p
Glyma.07G132400	MYB	Glyma.07G132400.2.p
Glyma.12G236800	NF-YA	Glyma.12G236800.5.p

Column explain:

(1) gene_id: gene id

(2) TF_Family: transcription factor family

(3) TF_ID: transcription factor id

Appendix

Appendix

1 Document description

readme.pdf -- Analysis results directory description

method.pdf -- Experiment and analysis method description

software.pdf -- Analyze the software list

FAQ.pdf -- After sale FAQ document

2 Notes

We suggest the result files be opened with a professional text editor such as Excel or EditPlus.
When opening the report with Internet Explorer, if it returns "for security reasons, Internet Explorer has restricted this page from running scripts or ActiveX Controls that can access your computer. Click here for options ..." Please select ‘Allow’ to view the report.

AZENTA
LIFE SCIENCES

Reference

Reference

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

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

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

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

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

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

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

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

[9] Love, M. I.,Anders, S., and Huber, W.(2015). Differential analysis of count data-the DESeq2 package. (DESeq2)

[10] Kim, D., Langmead, B., and Salzberg, S. L.,(2015). HISAT: a fast spliced aligner with low memory requirements. (HISAT)

[11] Pertea, M., Pertea, Geo. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L.,(2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. (StringTie)

[12] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), 10-12. (Cutadapt)

[13] Andrews, S. (2016). FastQC: a quality control tool for high throughput sequence data. 2010. (FastQC)

[14] Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nature biotechnology, 29(1), 24. (IGV)

[15] Florea, L., Song, L., & Salzberg, S. L. (2013). Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. F1000Research, 2. (ASprofile)

[16] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079. (Samtools)

[17] Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research, 38(16), e164-e164. (Annovar)

[18] Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics, 28(16), 2184-2185. (RSeQC)

[19] Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl_1), D258-D261. (GO)

[20] Li, Y., Rao, X., Mattox, W. W., Amos, C. I., & Liu, B. (2015). RNA-seq analysis of differential splice junction usage and intron retentions by DEXSeq. PLoS One, 10(9), e0136653. (DEXSeq)

[21] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., ... & Jensen, L. J. (2010). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research, 39(suppl_1), D561-D568. (STRING)

[22] Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., ... & Jensen, L. J. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, gkw937. (STRING)

[23] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504. (Cytoscape)

[24] Haas, B., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., ... & Sun, J. (2017). STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq. BioRxiv, 120295. (STAR-Fusion)

[25] Zhang, Q. (2018). Analysis of RNA Editing Sites from RNA-Seq Data Using GIREMI. In Transcriptome Data Analysis (pp. 101-108). Humana Press, New York, NY. (GIREMI)

[26] Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research,

39(suppl_2), W29-W37. (HMMER)
[27] Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic acids research, 35(suppl_2), W345-W349. (CPC)
[28] Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9(1), 559. (WGCNA)
[29] Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., ... & Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences, 111(51), E5593-E5601. (rMATS)
[30] Ernst, J., & Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. BMC bioinformatics, 7(1), 191. (STEM)
[31] Subramanian, A., Tamayo, P., Mooth a, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550. (GSEA)
[32] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., ... & Houstis, N. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics, 34(3), 267. (GSEA)
[33] Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., & Gao, G. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic acids research, gkw982. (PlantTFDB4.0)
[34] Hu, H., Miao, Y. R., Jia, L. H., Yu, Q. Y., Zhang, Q., & Guo, A. Y. (2018). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic acids research, 47(D1), D33-D38. (AnimalTFDB 3.0)