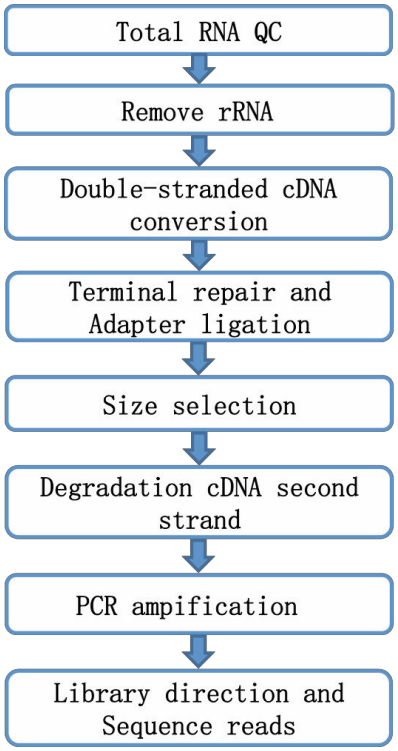
LncRNA-seq Analysis Demo Report

May 1, 2016

Contents

1 Experimental Procedures

From RNA sample preparation to data acquisition, each step such as sample detection, library construction and sequencing, can affect the quality and quantity of data. High quality data is essential for accurate and confidential analysis. In order to ensure the quality and reliability of the sequencing data, every step of data production is under rigid control. The workflow is as follows:



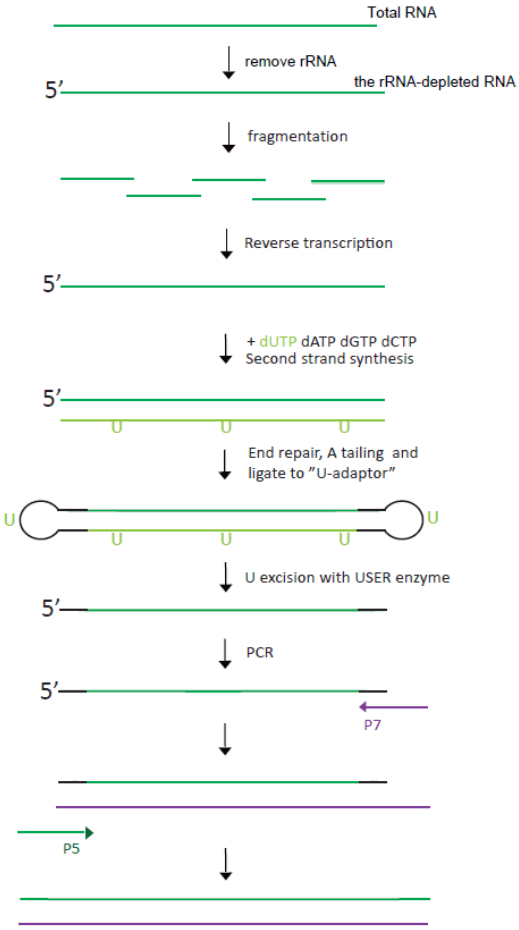
1.1 Quality check of total RNAs

There are four methods for quality check (QC) of RNA samples:
(1) Agarose gel electrophoresis: for RNA integrity and potential contamination
(2) Nanodrop: for RNA purity (OD260/OD280)
(3) Qubit: quantify RNA concentration
(4) Agilent 2100: check RNA integrity again

1.2 Library construction

After RNA QC, rRNAs were removed by using epicentre Ribo-Zero™ Kit. The purified RNAs were first fragmented randomly to short fragments of 150-200 bp by addition of fragmentation buffer, then cDNA synthesis followed using random hexamers. After the first strand was synthesized, a custom second-strand synthesis buffer (Illumina), dNTPs (dUTP, dATP, dGTP and dCTP) and DNA polymerase I were added to synthesize the second-strand. Then followed by purification by AMPure XP beads, terminal repair, polyadenylation, sequencing adapter ligation, size selection and degradation of second-strand U-contained cDNA by the USER enzyme.

The strand-specific cDNA library was generated after the final PCR enrichment. The workflow is as follows:



1.3 Library QC

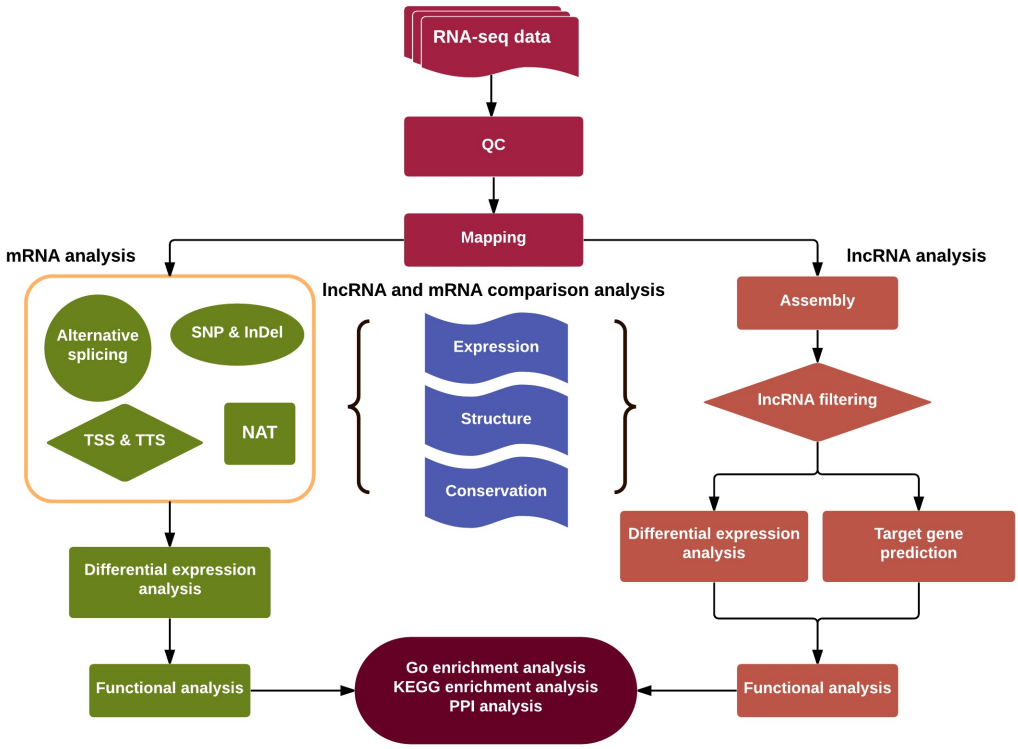
The concentration of library was first quantified by Qubit2.0, then diluted to 1 ng/ul, and the insert size was checked by Agilent 2100 and was further quantified by qPCR (library concentration > 2 nM).

1.4 Sequencing

If the library qualifies, it will be sequenced on an Illumina HiSeq platform according to effective concentration and data volume.

2 Bioinformatics Analysis Procedures

The flowchart below depicts the bioinformatics analysis pipeline we used.



3 Analysis Result

3.1 Raw Data

The original raw image data obtained from high throughput sequencing platforms (e.g. Illumina platform) is transformed to sequenced reads by base calling. The sequenced reads are regarded as raw data or raw reads, which is recorded in FASTQ file (fq) containing sequence information (reads) and corresponding sequencing quality information.

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

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18: ATCACG
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT
+
@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF
```

Line 1 beginning with a '@' character is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence reads. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in

Line 2, and must contain the same number of characters as bases in the sequence.

Table 3.1.1 Illumina sequence identifier details

EAS139	The unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane
2104	Tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	Index sequence

The ASCII value for every character at the fourth line minus 33 will be the corresponding sequencing base quality value at the second line. If the sequencing error rate is recorded by "e" and the base quality for Illumina platform is expressed as Q_{phred}, the equation No.1 as below will be obtained:

Equation 1: $Q_{phred} = -10\log_{10}(e)$

The relationship between sequencing error rate (e) and sequencing base quality value (Q_{phred}) is listed as below (Table 4.2):

Table 3.1.2 Sequencing error rate and corresponding base quality value

Sequencing error rate	Sequencing quality value	Corresponding character
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

3.2 Quality Control

3.2.1 Sequencing Error Rate Examination

For Illumina SBS technology, the distribution of sequencing error rate has two features:

(1) Error rate grows with sequenced reads extension because of the consumption of sequencing reagent. The phenomenon is common in the Illumina high-throughput sequencing platform (Erlich et al., 2008; Jiang et al., 2011).

(2) The reason for the high error rate of the first six bases is that the random hex-primers and RNA template bind incompletely in the process of cDNA synthesis (Jiang et al., 2011).

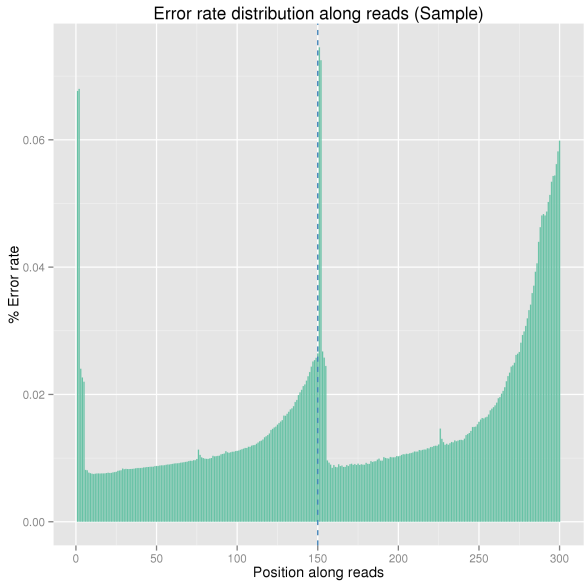


Figure 3.2.1 Sequencing error rate distribution

The x-axis represents position in reads, and the y-axis represents the average error rate of bases of all reads at a position.

3.2.2 Sequencing Data Filtration

Raw sequencing data may contain adapter contaminated and low-quality reads. These sequence artifacts may increase the complexity of downstream analyses, which means that quality control is an essential step. All the downstream analyses will be based on clean reads that pass quality control.

We performed quality control according to the following procedure:
(1) Discard a read pair if either one read contains adapter contamination;
(2) Discard a read pair if more than 10% of bases are uncertain in either one read;
(3) Discard a read pair if the proportion of low quality bases is over 50% in either one read.

RNA-seq Adapter sequences (Oligonucleotide sequences of adapters from TruSeq™ RNA and DNA Sample Prep Kits):
5' Adapter:
5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
3' Adapter:
5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(6-indexs)ATCTCGTATGCCGTCTTCTGCTTG-3′

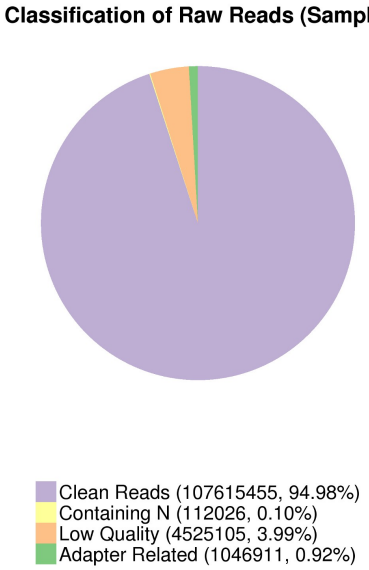


Figure 3.2.2 Raw data filtration result

Note: Reads were discarded in pairs.

(1) Containing N: the number of read pairs with either one read containing uncertain nucleotides more than 10%, and the proportion in raw data.

(2) Low Quality: the number of read pairs with either one read containing low quality (below 5) nucleotides more than 50 percent, and the proportion in raw data.

(3) Adapter related: the number of read pairs filtered out with adapter contamination, and the proportion of filtered read pairs in raw data.

(4) Clean reads: the number of read pairs passed quality control and the proportion in raw data.

3.2.3 Statistics of Sequencing Quality

According to the sequencing feature of Illumina platforms, for paired-end sequencing data we require that Q30 (the percent of bases with phred-scaled quality scores greater than 30) should be above 80%.

Table 3.2.3 Overview of data quality

Sample name	Raw reads	Clean reads	Clean bases(G)	Error rate(%)	Q20(%)	Q30(%)	GC content(%)
Control_1	128375922	122490956	18.37	0.03	97.98	93.76	44.73
Control_2	128375922	122490956	18.37	0.03	96.94	91.33	45.06
Sample_1	113299497	107615455	16.14	0.03	97.83	93.29	47.06
Sample_2	113299497	107615455	16.14	0.03	96.75	91.01	47.06

The details of the table are described below:

(1) Sample name: For PE sequencing, * _1 and * _2 indicate reads on the left and right end, respectively;

(2) Raw reads: Statistics of raw reads, each adjacent four lines contains the information of one read, and the total read number of each file is calculated;

(3) Clean reads(G): Same as raw reads, except that only the filtered reads, which all subsequent analysis is based on, is

calculated;

(4) Clean bases: The product of number and length of sequences, calculated as Giga bases;

(5) Error rate: The error rate of sequencing, calculated based on Equation 1;

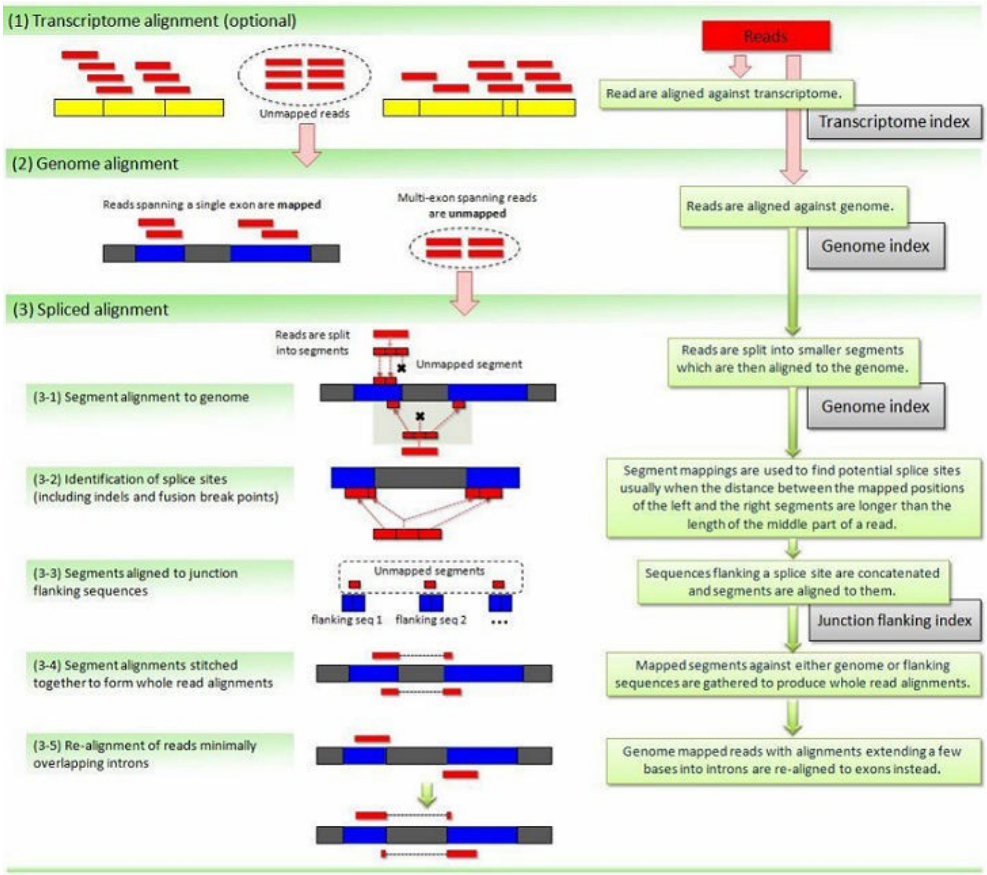
(6) Q20, Q30: The percentage of total number of bases where the Phred score is greater than 20 and 30, respectively;

(7) GC content: The percentage of G and C in all bases;

3.3 Mapping to a Reference Genome

The cleaned reads are aligned to the reference genome with Tophat2 (Kim et al., 2013) and the algorithm of Tophat2 mainly includes three parts:
(1) Map the reads against transcriptome (optional);
(2) Map the full-length reads to the exons;
(3) Map the partial reads to two exons;

The algorithm of TopHat2 is described below (Kim et al., 2013):



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 Statistics of Mapped Reads

Table 3.3.1 Statistics of reads mapped to reference genome

Sample name	Control	Sample
Total reads	103413896	95671374
Total mapped	87319755 (84.44%)	80291861 (83.92%)
Multiple mapped	4739091 (4.58%)	3933783 (4.11%)
Uniquely mapped	82580664 (79.85%)	76358078 (79.81%)
Read-1	41400161 (40.03%)	38289381 (40.02%)
Read-2	41180503 (39.82%)	38068697 (39.79%)
Reads map to '+'	41274699 (39.91%)	38168756 (39.9%)
Reads map to '-'	41305965 (39.94%)	38189322 (39.92%)
Non-splice reads	64293027 (62.17%)	60583242 (63.32%)
Splice reads	18287637 (17.68%)	15774836 (16.49%)

The details of the mapping results are described below:

(1) Total reads: Number of reads after data filtering (clean data);

(2) Total mapped: Number of reads that can be mapped to the genome. Generally, if there is proper reference genome and no contamination during the experimental procedure, the percentage will be higher than 70%;

(3) Multiple mapped: Number of sequences that are mapped to multiple positions in the reference sequences. the percentage of this part of the data is generally less than 10%;

(4) Uniquely mapped: Number of reads that are mapped to the unique position in the reference sequences;

(5) Reads map to '+', Reads map to '-': Number of reads that are mapped to the plus or minus strand, respectively.

(6) Splice reads: Number of reads that are mapped to two exons (also known as the junction reads). Similarly, non-splice reads are those that the full-length reads are mapped to one exon. The percentage of splice reads depends on the length of reads.

3.3.2 Distribution of Reads On Chromosomes

To obtain an overview of the distribution of mapped reads on each chromosome, 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., 2012).

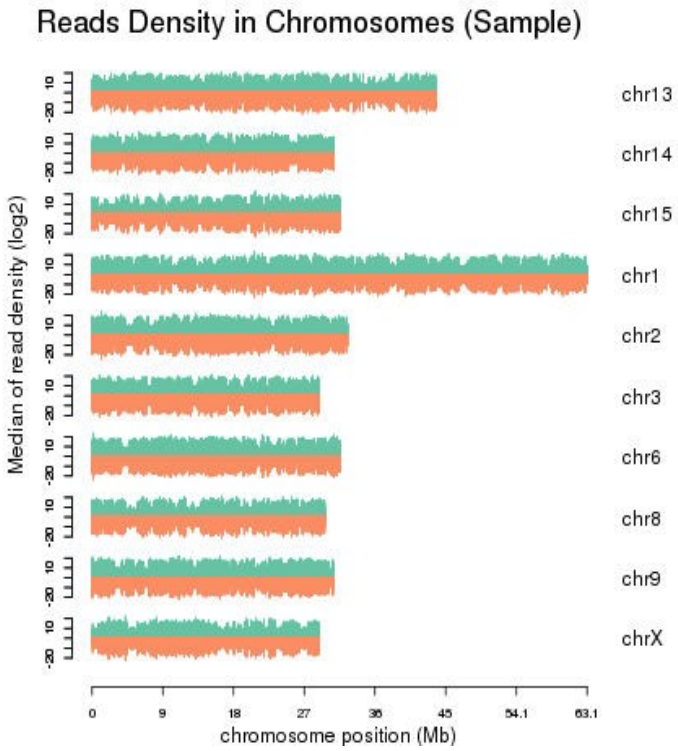


Figure 3.3.2 Distribution of reads on chromosomes

The horizontal axis represents the length of chromosome (Mb), and the vertical axis represents log2(median of read density). Green and red bars represent the plus and minus strands, respectively.

3.3.3 Distribution of Known types of Genes

The coverage of different known gene types in this specie is analysed using the union model by HTSeq. According to the expression quantity, the expressed distribution of various types of genes in sample were made counts and shown in table 3.3.3：

Table 3.3.3 The distribution list of the known types of genes

Sample_name	Sample	Control
mRNA	48722201 (67.41%)	38998933 (58.78%)
misc_RNA	994223 (1.38%)	1074817 (1.62%)
ncRNA	11323217 (15.67%)	11267598 (16.98%)
pseudogene	6464 (0.01%)	8613 (0.01%)
rRNA	16106 (0.02%)	48237 (0.07%)
tRNA	202928 (0.28%)	308863 (0.47%)
Others	11012912 (15.24%)	14635991 (22.06%)

The table above is pictured shown below：

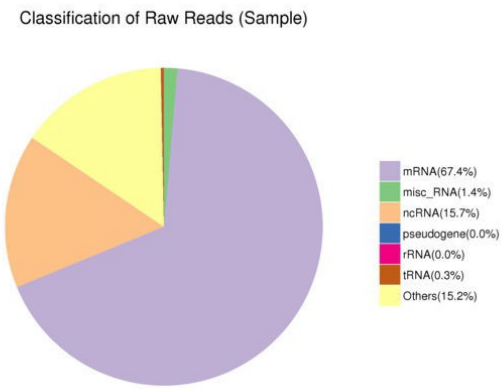


Figure 3.3.3 Distribution of reads in different kinds of genes

3.3.4 Visualization of Aligned Data

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.

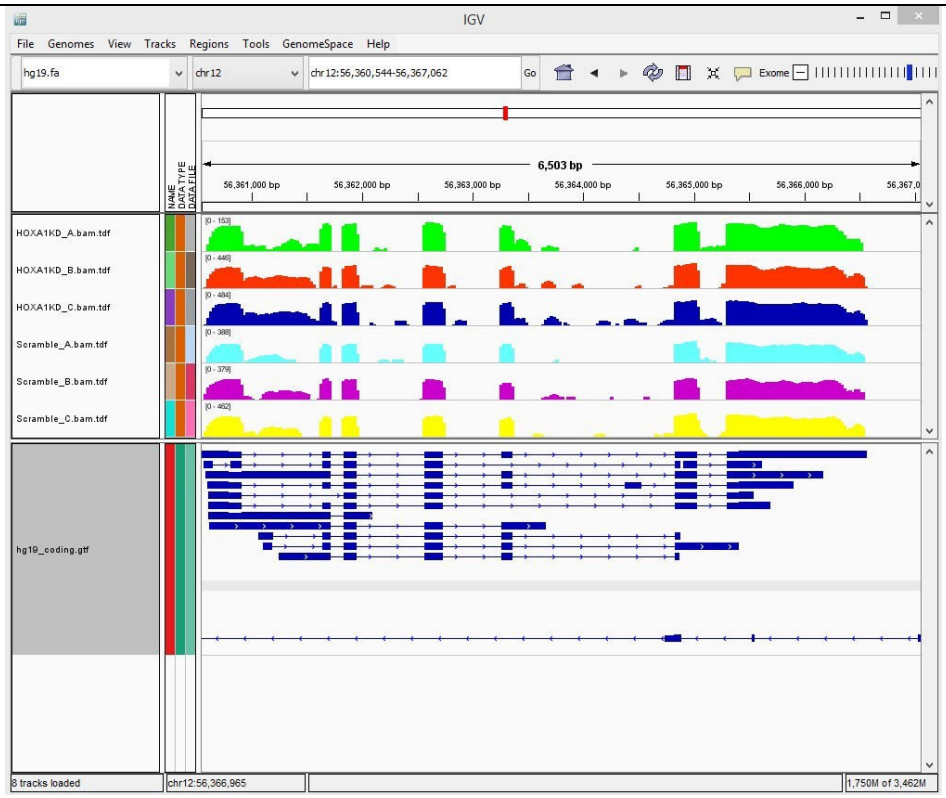


Figure 3.3.4 The interface of the IGV browser

3.4 RNA-Seq Quality Assignment

FPKM (expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced), which considers both the sequencing depth and the gene length, is the most commonly used method for gene expression profiling (Trapnell Cole, et al., 2010), so that the calculated expression levels can be used directly to compare differences in gene expression between samples.

3.4.1 Comparative Analysis of Gene Expression Level

Boxplot and density plot of the FPKMs of all transcripts are used to compare the their expression under different experiments. For samples with replicates, the mean of FPKMs from all replicates is used.

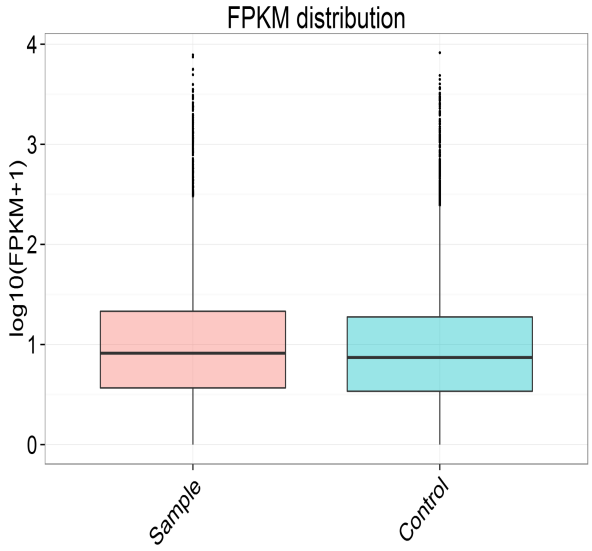


Figure 3.4.1 Comparison of gene expression under different expressions

Note:

(1) boxplot of FPKM values. X an y axis represent the respective sample name and the value of log10(FPKM+1). For each sample, the plot region represents the statistics of maximum, upper quartile, median, lower quartile and minimum, respectively from top to bottom.

(2) FPKM density distribution. X and y axis represent the value of log10(FPKM+1) and the density of genes, respectively.

3.4.2 Correlation Analysis among Samples

Biological replicates are necessary for any biological experiment, including those involving RNA-seq technology (Hansen et al., 2012). Biological replicates in RNA-seq can demonstrate whether the experiment is repeatable. If biological replicates are unavailable, it will be impossible to estimate the level of biological variability in expression for each gene in a study.

The correlation between samples is an important indicator for testing the reliability of the experiment. The closer the correlation coefficient is to 1, the greater the similarity of the samples. ENCODE suggests that the square of the Pearson correlation coefficient should be larger than 0.92, under ideal experimental conditions. In this project, the R^2 should be larger that 0.8.

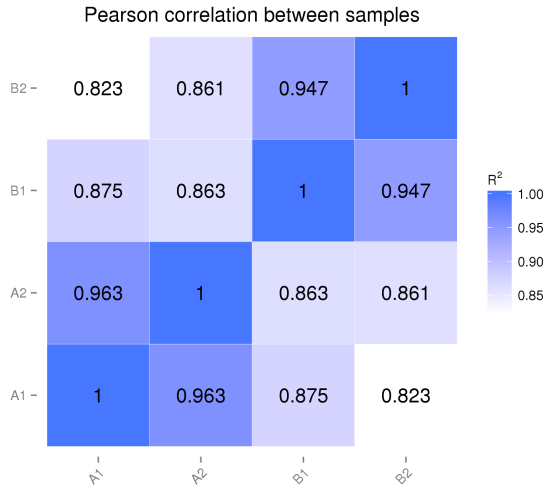


Figure 3.4.2 RNA-Seq correlation

Heat maps of the correlation coefficient between samples are also shown. R^2, the square of the Pearson coefficient correlation coefficient between samples;

3.5 Transcripts Assembly

The Cufflinks software (Trapnell et al., 2010), which uses statistical model, can simultaneously assemble and quantify the expression of isoforms and keep isoform set as small as possible. It can report the maximum likelihood estimate of expression data and use accurate strand information by passing options specific to strand-specific library. The workflow and results of cufflinks assembly are shown below:

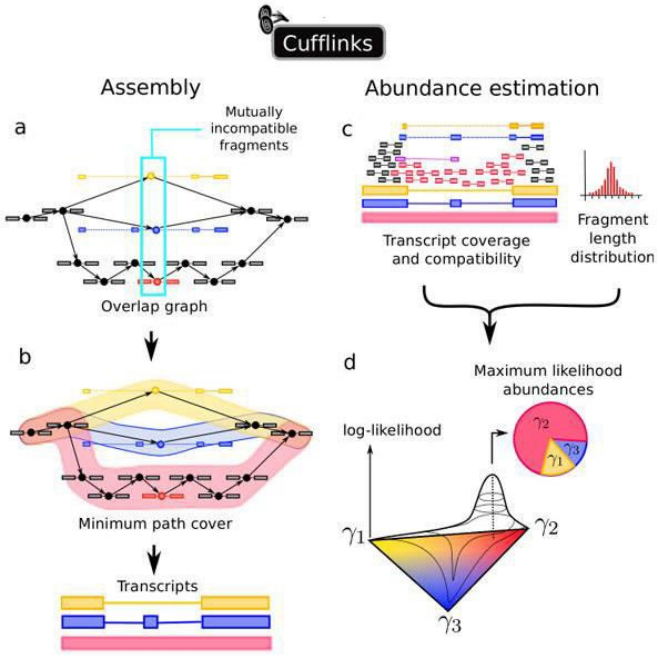


Figure 3.5.1 RNA-Seq correlation

Table 3.5.1 Results of cufflinks assembly (partial)

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
GL000191.1	Cufflinks	transcript	3583	3717	1000	+	.	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "6.1941392125"; frac "0.400000"; conf_lo "0.000000"; conf_hi "0.813112"; cov "101.570893";
GL000191.1	Cufflinks	exon	3583	3717	1000	+	.	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "6.1941392125"; frac "0.400000"; conf_lo "0.000000"; conf_hi "0.813112"; cov "101.570893";
GL000191.1	Cufflinks	transcript	3622	3903	1000	-	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; FPKM "0.1756257804"; frac "0.600000"; conf_lo "0.000000"; conf_hi "0.585419"; cov "2.879894";
GL000191.1	Cufflinks	exon	3622	3903	1000	-	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "1"; FPKM "0.1756257804"; frac "0.600000"; conf_lo "0.000000"; conf_hi "0.585419"; cov "2.879894";
GL000191.1	Cufflinks	transcript	9604	9764	1000	-	.	gene_id "CUFF.3"; transcript_id "CUFF.3.1"; FPKM "1.4539361301"; frac "1.000000"; conf_lo "0.000000"; conf_hi "0.530291"; cov "23.841503";

Note:

(1) Seqname: the name of chromosome or scaffold;

(2) Source: data source, it is "Cufflinks";

(3) Feature: sequence type description, it is "transcript" or "exon";

(4) Start: transcript start position;

(5) End: transcript end position;

(6) Score: score of the assembly;

(7) Strand: transcript strand;

(8) Frame: type of transcript start position, cufflinks does not predict start/end codon, thus it is ".";

(9) Attributes: other descriptions of the sequence, such as gene ID, transcript ID and its quantification;

3.6 Identification of Candidate Long Noncoding RNAs

LncRNA is non-coding transcripts that are longer than 200-nt. Based on their genomic positions, they can be classified to intergenic lncRNAs (lincRNAs), intronic lncRNAs, anti-sense lncRNAs, sense lncRNAs, bidirectional lncRNAs and so on, where lincRNAs account for the largest proportion. While focus on the first three types of lncRNAs, the pipeline with a set of strict filters, as shown below, is used to predict candidate lncRNAs based on their structures and non-coding features.

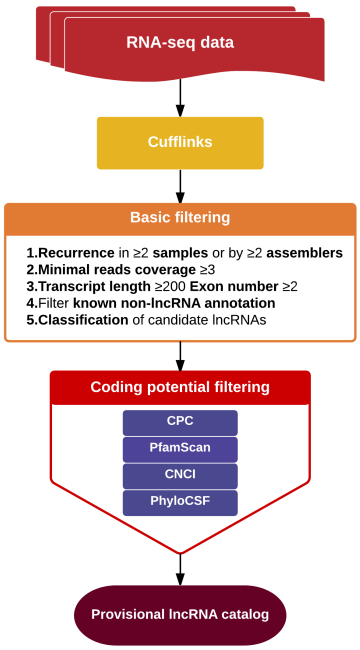


Figure 3.6 Flowchart of lncRNA filtering

3.6.1 Basic filtering

There are five steps for basic filtering:

Step 1: Merge all assembled transcripts by cuffcompare and select transcripts exist in at least two samples;

Step 2: Select transcripts that are longer than 200 bp and have more than 2 exons;

Step 3: Calculate the coverage of each transcript by cufflinks and select transcripts whose coverage $\geqslant$ 3;

Step 4: Compare with known non-lncRNA and non-mRNA transcripts (rRNA, tRNA, snRNA, snoRNA, pre-miRNA, pseudogenes etc.), and filter out the transcripts identical or similar to these ones;

Step 5: Compare with known mRNAs according to the class code of cuffcompare result (http://cufflinks.cbcb.umd.edu/manual.html#class_codes) to get candidate lincRNAs, intronic lncRNAs and anti-sense lncRNAs.

The bar plot below shows the number of transcripts that were filtered out in each step.

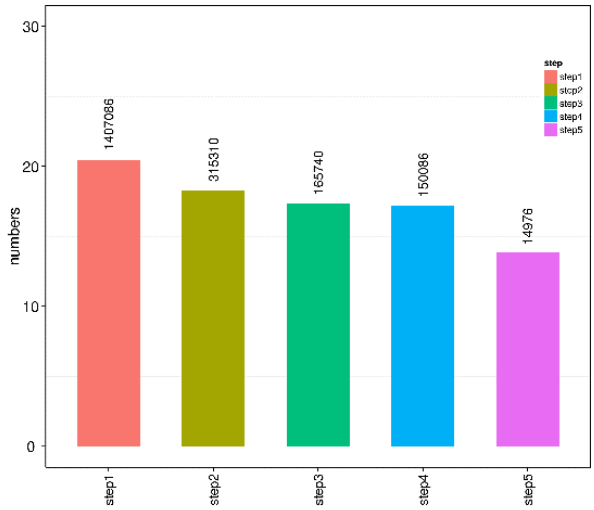


Figure 3.6.1.1 Statistics of lncRNA filtering

Horizontal axis represents the filtering step, and vertical axis represents the number of filtered transcripts in that step.

Table 3.6.1.1 Description of class_code

class_code	meaning
=	Complete match of intron chain
c	Contained
j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
i	A transfrag falling entirely within a reference intron
o	Generic exonic overlap with a reference transcript
p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
u	Unknown, intergenic transcript
x	Exonic overlap with reference on the opposite strand
s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
.	(.tracking file only, indicates multiple classifications)

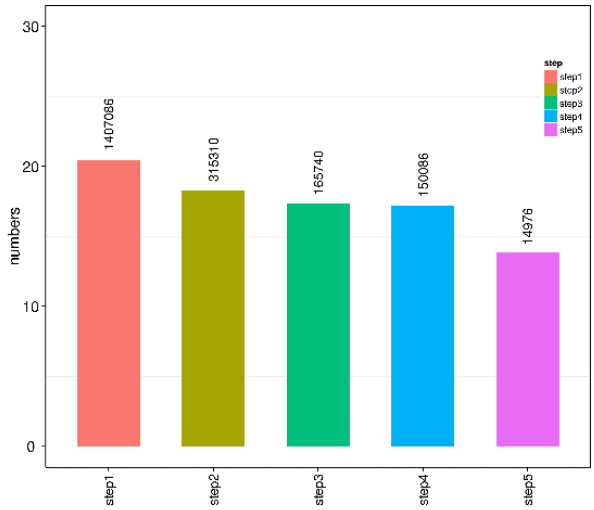


Figure 3.6.1.2 Filtering of lncRNAs based on class_code

Horizontal and vertical axes represent the type of class_code and the number of transcripts, respectively. The class_code "u", "i" and "x" stand for lincRNA, intronic lncRNA and anti-sense lncRNA, respectively.

3.6.2 Coding Potential Filtering

Coding potential is essential to determine if a transcript is a lncRNA, and several popular softwares for coding potential analysis are adopted for coding potential filtering, including CPC, CNCI, Pfam Analysis and PhyloCSF analysis (limited to mammalian only), and the predicted lncRNAs come from the intersection of these methods.

3.6.2.1 CPC Analysis

CPC (Coding Potential Calculator) can calculate coding potential by blastx search against the protein database (The NCBI nr database is used here). Based on the sequence features of the coding frame, the coding potential of the transcript is assessed by support vector machine, and the results are given below.

Table 3.6.2.1 Summary of CPC analysis (partial)

Transcript id	Transcript length	Type	Score
TCONS_00000082	363	coding	1.65903
TCONS_00000117	1901	noncoding	-5.25003
TCONS_00000144	1631	coding	3.30387
TCONS_00000377	489	noncoding	-1.02156
TCONS_00000435	1928	noncoding	-5.06304
TCONS_00000556	981	coding	0.637386

Note:

(1) Transcript id: transcript ID;

(2) Transcript length: Transcript length;

(3) Type: transcript type, either "noncoding" or "coding";

(4) Score: coding potential score. The transcript type is "noncoding" if the score < 0;

3.6.2.2 CNCI Analysis

CNCI (Coding-Non-Coding Index) can distinguish protein-coding and non-coding transcripts from transcript assembly, which is independent of known annotations and can predict the potential of coding or non-coding based on the features of nucleotide triplets (Sun et al., 2013). The results of CNCI are shown below:

Table 3.6.2.2 Summary of CNCI analysis (partial)

Transcript id	Type	Score	Start	End
TCONS_00000082	coding	0.008101845	24	210
TCONS_00000117	noncoding	-0.208177216	1482	1746
TCONS_00000144	coding	0.193568756	375	1374
TCONS_00000377	noncoding	-0.002518776	60	90
TCONS_00000435	noncoding	-0.233758796	0	189
TCONS_00000556	noncoding	-0.174564258	81	141

Note:

(1) Transcript id: transcript ID

(2) Max score: max score of coding potential

(3) Start: ORF start position

(4) End: ORF end position

(5) Protein: protein sequence

3.6.2.3 Pfam Analysis

Pfamscan (Mistry et al., 2007) is used to search protein domains in the pfam HMM database (Bateman et al., 2002) to eliminate sequences matched to known protein domains, and both Pfam-A and Pfam-B databases are used. Pfam-A contains most high quality known protein domains that are manually selected, while Pfam-B covers more domains, which is complementary to Pfam-A. The translated protein sequences are searched against the Pfam-A and Pfam-B databases by hmmscan, and the matched sequences are considered to have coding potential, whereas others are most likely to be non-coding transcripts.

Table 3.6.2.3 Summary of Pfam analysis (partial)

Seq id	Hmm acc	Hmm name	Type	Hmm start	Hmm end	Hmm length	Bit score	E-value
TCONS_00000082-1	PB003422	Pfam-B_3422	Pfam-B	991	1054	1054	40.8	9.20E-11
TCONS_00000117-0	PB008900	Pfam-B_8900	Pfam-B	27	68	131	50.4	2.80E-13
TCONS_00000117-1	PF13900.1	GVQW	Domain	1	48	48	103.1	5.30E-30
TCONS_00000435-0	PF13900.1	GVQW	Domain	1	48	48	101.9	1.30E-29
TCONS_00000435-1	PB008900	Pfam-B_8900	Pfam-B	33	64	131	31.6	1.80E-07
TCONS_00000435-1	PB000655	Pfam-B_655	Pfam-B	72	175	319	61.3	1.30E-16

Note:

(1) Seq id: transcript ID+[0,1,2], transcripts not in the list are "noncoding"

(2) Hmm acc: pfam domain ID

(3) Hmm name: pfam domain name

(4) Type: pfam domain type

(5) Hmm start: start position of pfam domain

(6) Hmm end: end position of pfam domain

(7) Hmm length: length of pfam domain

(8) Bit score: alignment score

(9) E-value: E-value of the alignment, the criteria is: E-value 0.001

3.6.2.4 PhyloCSF Analysis

PhyloCSF (phylogenetic codon substitution frequency) is a tool that calculate the coding potential of transcripts by using genome-wide sequence alignment of multiple organisms. Two main arguments of PhyloCSF are phylogenetic tree and codon matrix (Lin et al., 2011). Based on the genome-wide sequence alignment of multiple organisms, the Codon Substitution Frequency (CSF) is calculated (CSF refers to the frequency of codon substitution in multiple sequence alignment, and the codon substitution ratio of coding and non-coding regions can be used to effectively distinguish coding and non-coding sequences), and the coding potential of the transcripts is scored by combining the distance information from phylogenetic tree of organisms. According to previous studies, different species are found to have different PhyloCSF threshold. Therefore, some known lncRNAs and mRNAs are sampled to calculate the threshold. Since the screening model is designed for mammals, this tool is limited to mammals only.

Table 3.6.2.4 Summary of phyloCSF analysis (partial)

Transcript id	Max score	Start	End	Protein
TCONS_00000082	7.1440	1835	1912	MQQNCVSGLVPVCQLNSSGCSLSDDG
TCONS_00003321	97.1766	1077	1295	MYNADSISAQSKLKEAEKQEEKQIGKSVKQEDRQTPCSPD STANVRIEEKHVRRSSVKKIEKMKEKVCRLPQL
TCONS_00005703	34.5244	342	599	MYSRSQASPSCGGGGGQGGLPRGLGWASLGGVFCEFAAK GLGWVWGGPGVGLGSVLVSKASLTFASQITGAFPLDNSAL RPTGSGF
TCONS_00001396	43.6306	390	725	MGCPGAGTGNPWDQPRLSLPFLAGVELALLHRSPAKGRK MASGGLGLVLKAFCPQGVAGAPVLPQQEAIWGQQCPLGA GASGPGV EEFGKCWNGCLVCPCSFSVTLLPTNSS
TCONS_00004972	61.0833	86066	86257	MDLLVLSQGHQTNTLDIIHIHKEALTKVMESRQHVAEGKT QVQKKVQRLMTSESQEQDFFGHFG

Note:

(1) Transcript id: transcript ID

(2) Max score: max score of coding potential

(3) Start: ORF start position

(4) End: ORF end position

(5) Protein: protein sequence

3.6.2.5 Venn Diagram of Coding Potential Analysis

The noncoding transcripts identified by CPC, CNCI, Pfam and PhyloCSF were summarized and shown as a venn diagram below, and the intersection of these results is considered to be the final lncRNA data set for further analysis.

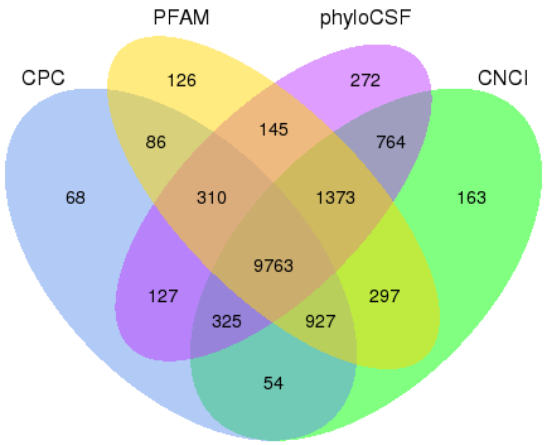


Figure 3.6.2.5 Venn diagram of results from four tools mentioned above

Number in each circle and overlap represent the respective total and shared number of noncoding transcripts predicted by the software.

3.7 lncRNA Expression Analysis

The expression of the filtered lncRNAs was analyzed by cuffdiff (http://cufflinks.cbcb.umd.edu/manual.html#cuffdiff), and the results are shown below:

Table 3.7 FPKM of lncRNA in each sample (partial)

transcript_id	Sample	Control
TCONS_00004163	0.64279	0.639313
TCONS_00072426	0.594465	1.58591
TCONS_00002202	0.593291	0.474413
TCONS_00046915	1.83413	1.90649
TCONS_00050940	1.00281	1.64808
TCONS_00098287	0.518418	0.131542

3.8 lncRNA Target Prediction

As lncRNAs function mainly in *cis* or *trans* manner on their protein-coding target genes, lncRNA target prediction consists the following two sections.

cis-acting target prediction

The *cis*-acting target prediction assumes that the function of lncRNA is related to adjacent protein coding genes. Therefore, coding genes that are 100 kb upstream or downstream of lncRNA are considered to be target genes.

Table 3.8.1 cis-acting target gene prediction results

lncRNA_geneid	mRNA_geneid
XLOC_000150	55160
XLOC_000223	51538
XLOC_000223	2170
XLOC_000223	347735
XLOC_000795	127018
XLOC_000828	5664

Note:

(1) lncRNA_geneid: lncRNA gene ID

(2) mRNA_geneid: cis-acting target gene of this lncRNA

3.9 Functional Enrichment Analysis of lncRNA Target Genes

The GO enrichment analysis for *cis*-acting and *trans*-acting target genes were conducted, and only the *cis*-acting target genes are shown in the report.

3.9.1 GO Enrichment of LncRNA Target Genes

Gene Ontology (GO, http://www.geneontology.org/), as the standard classification system of gene function, can elucidate the functions of lncRNA targets that are differentially expressed. The GOseq R package (Young et al, 2010), which is based on Wallenius non-central hyper-geometric distribution, is used for gene ontology analysis. The Wallenius distribution, compared to hyper-geometric distribution, has the feature that the probability of sampling from a population is different from sampling from another one by assessing the bias of gene length, which can calculate the probability of GO term enrichment more accurately.

3.9.1.1 GO Enrichment of LncRNA Target Genes

The results of GO enrichment of lncRNA target genes are shown below:

Table 3.9.1.1 GO enrichment of lncRNA target genes

GO_accession	Description	Term_type	Over_represented_pValue	padj	fg	bg
GO:0004827	proline-tRNA ligase activity	molecular_function	0.0022518	1	1	6
GO:0006433	prolyl-tRNA	biological_process	0.0022518	1	1	6

	aminoacylation					
GO:0004000	adenosine deaminase activity	molecular_function	0.003709	1	1	6
GO:0000103	sulfate assimilation	biological_process	0.0061562	1	1	6
GO:0004020	adenylylsulfate kinase activity	molecular_function	0.0061562	1	1	6
GO:0019239	deaminase activity	molecular_function	0.0064888	1	1	6

Note:

(1) GO_accession: the unique id in Gene Ontology database

(2) Description: function description in gene ontology

(3) Term_type: type of the GO term (one of cellular_component, biological_process or molecular_function)

(4) Over_represented_pValue: statistical significance on enrichment

(5) padj: adjusted p-value. Normally, padj < 0.05 means the gene is enriched in that term

(6) fg: the number of lncRNA target genes related to the GO term

(7) bg: the number of lncRNA target genes that have GO annotation.

3.9.1.2 DAG of GO-enriched LncRNA Target Genes

The Directed Acyclic Graph (DAG) is used to visualize the GO enrichment, where branches represent inclusion of the two GO terms, and the scope of the term definitions becomes smaller and smaller from top to bottom. Normally, the top 10 results from GO enrichment are selected as main nodes in directed acyclic graph, where the associated terms are also represented and the depth of colors indicates enrichment level. DAGs for biological process, molecular function and cellular component are shown respectively.

The DAGs of GO enrichment of lncRNA target genes are shown below:

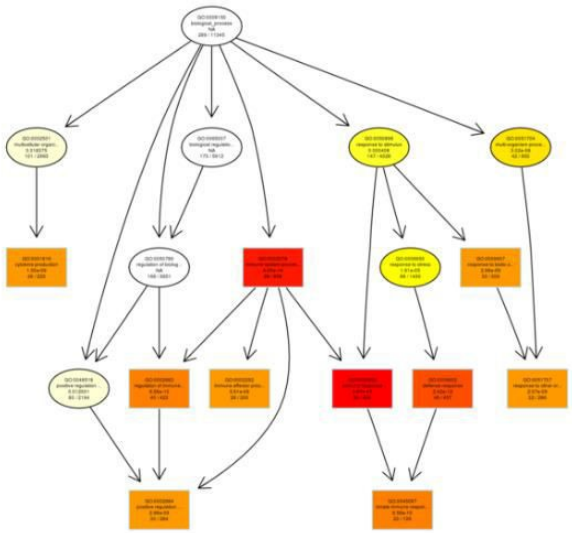


Figure 3.9.1.2 DAGs of GO enrichment of lncRNA target genes

Node represents GO term, and box represents the top 10 terms of GO enrichment. Deeper color indicates higher enrichment and vice versa. The GO term and the padj value of enrichment are shown in each node.

3.9.1.3 Bar plot of GO-enriched LncRNA Target Genes

The top 30 enriched GO terms in biological process, cellular component and molecular function are shown in the bar plot below. If there are less than 30 GO terms, all of them are shown in the plot.

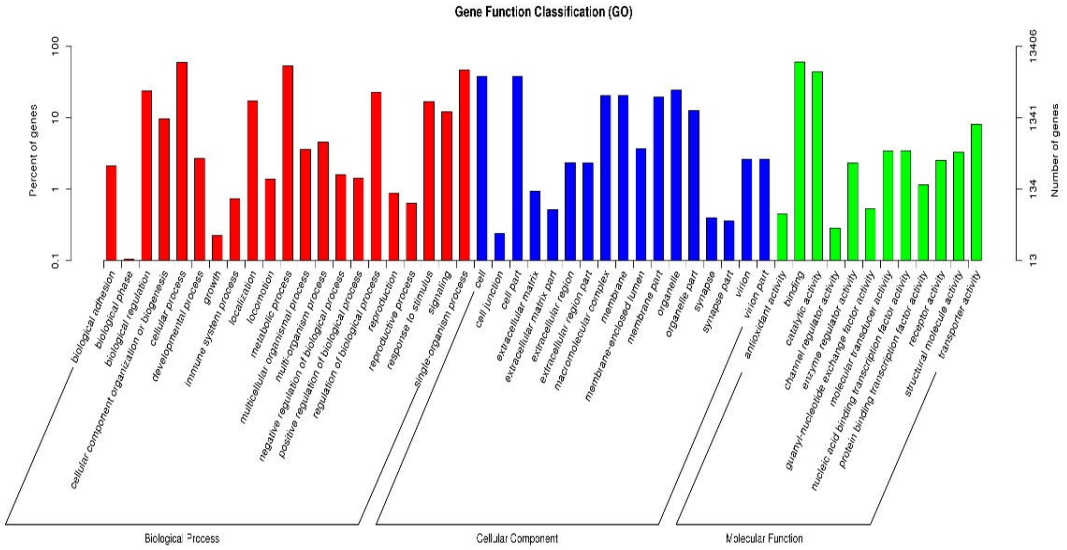


Figure 3.9.1.3 Bar plot of GO enrichment of lncRNA target genes

Node represents GO term, and box represents the top 10 terms of GO enrichment. Deeper color indicates higher enrichment and vice versa. The GO term and the padj value of enrichment are shown in each node.

3.9.1.4 Clustering of GO-enriched lncRNA Target Genes

Clustering of genes based on GO term enrichment is essential for studying the differences of lncRNA target gene expression among samples, where it is easy to find important genes that are differentially expressed. The vertical clustering is useful to determine correlation of samples based on gene expression levels, while the horizontal clustering is useful for finding some classes that have similar function and expression. In this analysis, the differences of gene expression from the top 30 significant GO terms are shown in the result. If the number is less than 30, all of them are used.

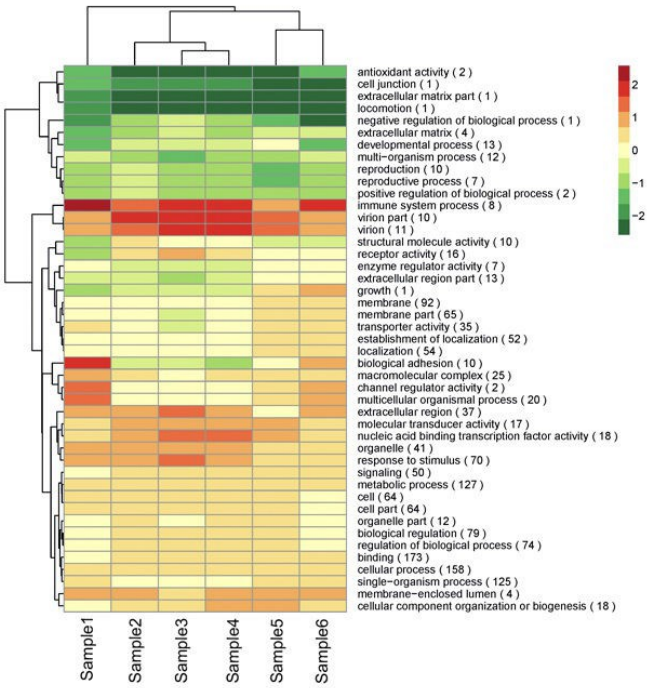


Figure 3.9.1.4 Clustering of enriched GO terms

The union of all terms on level 3 are used for clustering, and the expression level of all genes in each term is calculated. Terms in red and green colors represent high and low expression of genes in the corresponding terms, respectively, and the number in parenthesis after the term indicates the number of corresponding lncRNA target genes.

3.9.2 KEGG Enrichment of LncRNA Target Genes

The KEGG enrichment analysis for *cis*-acting and *trans*-acting target genes were conducted, and only the *cis*-acting target genes are shown in the report.

3.9.2.1 KEGG Enrichment of LncRNA Target Genes

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

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

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

List of enriched KEGG terms:

Table 3.9.2.1 KEGG enrichment of lncRNA target genes

#Term	Id	fg	bg	P-Value	padj
Metabolism of xenobiotics by cytochrome P450	hsa00980	24	74	0.000823954	0.136732588
Steroid hormone biosynthesis	hsa00140	20	57	0.000994419	0.136732588
Transcriptional misregulation in cancer	hsa05202	43	179	0.002238153	0.205164005
Drug metabolism - cytochrome P450	hsa00982	20	68	0.005389137	0.274839746
Ascorbate and aldarate metabolism	hsa00053	11	27	0.005467068	0.274839746
Retinol metabolism	hsa00830	19	64	0.006060405	0.274839746

Note:

(1) #Term: Description of the KEGG pathway

(2) Id: unique pathway ID in the KEGG database

(3) fg: number of lncRNA target genes in the pathway

(4) bg: number of genes in the pathway

(5) P-value: statistical significance of the enrichment

(6) padj: adjusted p-value. Normally, padj < 0.05 means the term is enriched

3.9.2.2 Scatter Plot of KEGG Enrichment of LncRNA Target Genes

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 lncRNA target genes counts to this pathway in the annotated genes counts. The more the Rich factor is, the higher is the degree of enrichment. Qvalue is the adjusted p-value after multiple hypothesis testing, and its range is [0,1]. The more the qvalue is close to zero, the more significant is the enrichment. Top 20 most significant enriched pathways are chosen in KEGG scatter plot, and if the enriched pathways counts is less than 20, then put all of them into the plot. KEGG enrichment scatter diagram is as follows.

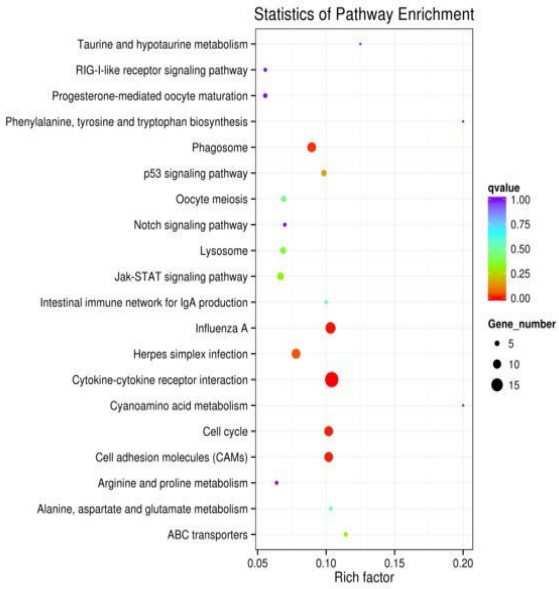


Figure 3.9.2.2 Scatter plot of enriched KEGG pathways of lncRNA target genes

Vertical coordinates represent pathway name, and horizontal coordinates represent Rich factor. The size and color of point represent the number of lnRNA target genes in the pathway and the range of different Q value, respectively.

3.9.2.3 Enriched KEGG Pathway of LncRNA Target Genes

The results of Enriched KEGG pathway of lncRNA target genes are shown below. For convenience of viewing the distribution of lncRNA target genes in pathways, those genes were added to the figures, and they can be viewed as described below: open the folder results/mRNA_Enrichment/KEGGEnrichment, where each html file contains different comparison of samples. Open one file and the pathways can be viewed by clicking on it. The KO node with red box indicates differential lncRNA target genes, and the mouse hovering on the KO node will popup the details of differential genes. All the operations described above can be done offline, and if the network connection is available, clicking on each node will open the associated webpage of KO node from the official KEGG database.

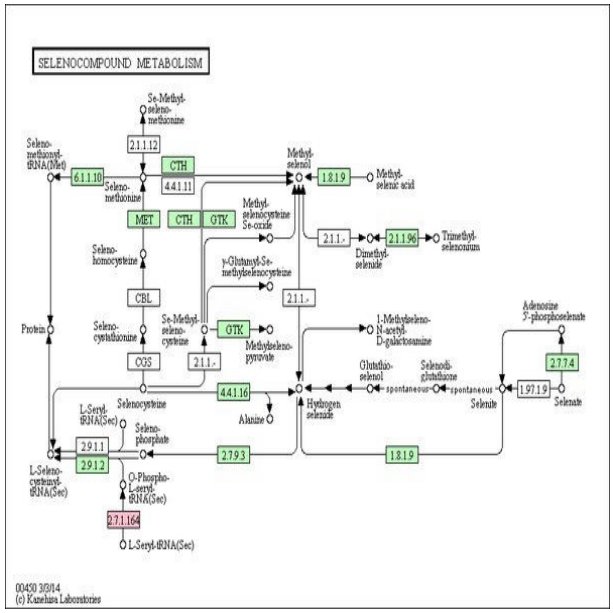


Figure 3.9.2.3 Enriched KEGG pathways of lncRNA target genes

3.9.2.4 Clustering of KEGG-enriched lncRNA Target Genes

Clustering based on the enriched KEGG pathway is essential for studying the differences of lncRNA target gene expression among samples, where it is easy to find some important differential expression in some pathways. The vertical clustering is useful to determine correlation of samples based on gene expression levels, while the horizontal clustering is useful for finding some classes that have similar function and expression. In this analysis, the differences of gene expression from the top 30 significant KO terms are shown in the result. If there are less than 30 significant KO terms, all of them are used.

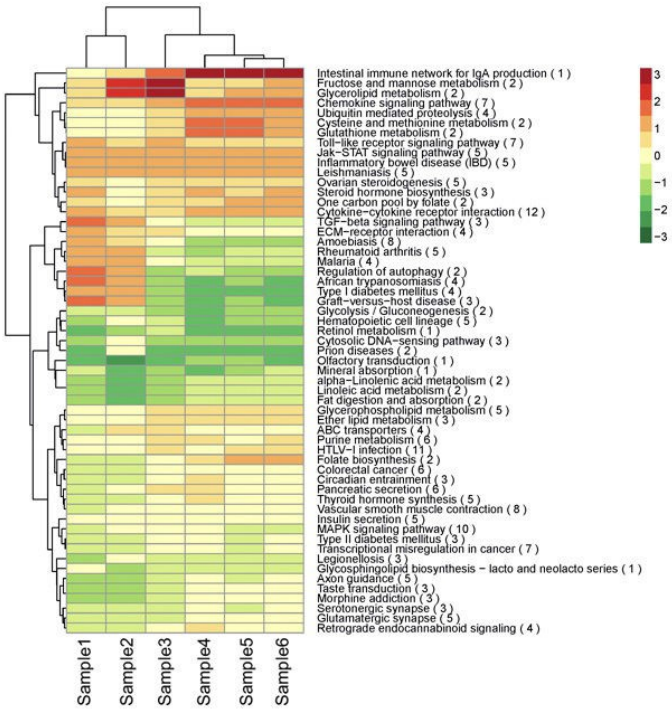


Figure 3.9.2.4 Clustering of KEGG enrichment

The union of all pathways are used for clustering, and the expression level of all genes in each pathway is calculated. Pathways in red and green colors represent high and low expression of genes in the corresponding pathways, respectively, and the number in parenthesis after the pathway indicates the number of corresponding lncRNA target genes.

3.10 lncRNA conservation analysis

3.10.1 Sequence conservation analysis

Sequence conservation of lncRNA is generally lower than that of mRNA, and the phyloP (http://compgen.bscb.cornell.edu/phast/) is used to score the conservation of mRNA and lncRNA. The cumulative distribution of conservation scores is shown below:

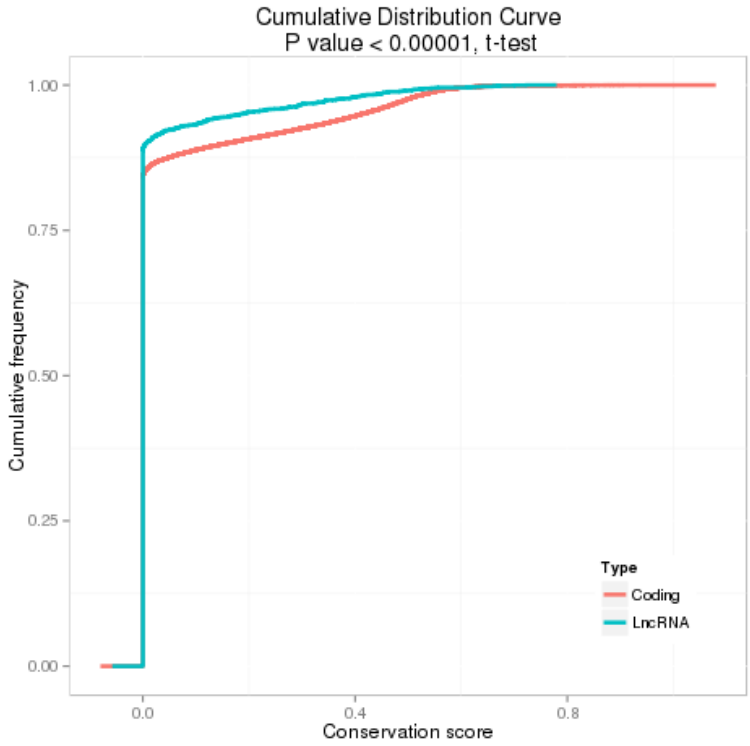


Figure 3.10.1 Cumulative distribution of conservation scores of lncRNA and mRNA

3.10.2 Site conservation analysis

Site conservation of lncRNA sequences exists among various species, and the positions of lncRNA in different species can be visualized by the UCSC browser.The site conservation of lncRNA is given below:

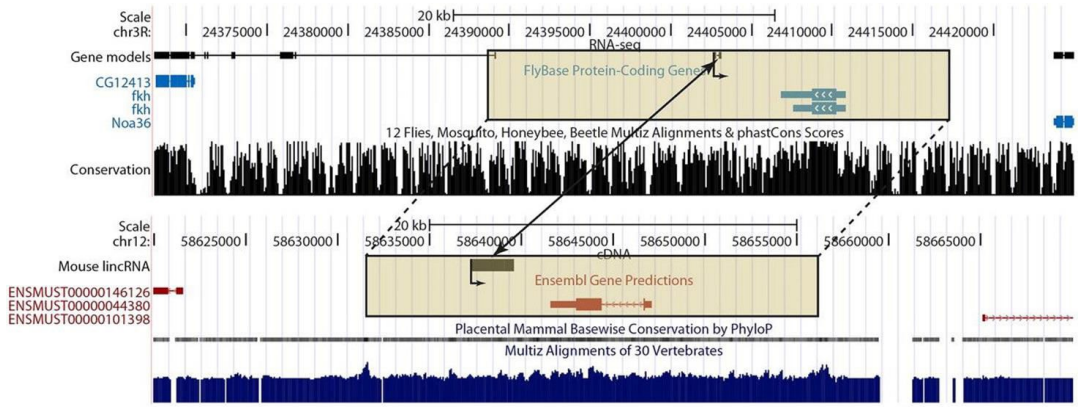
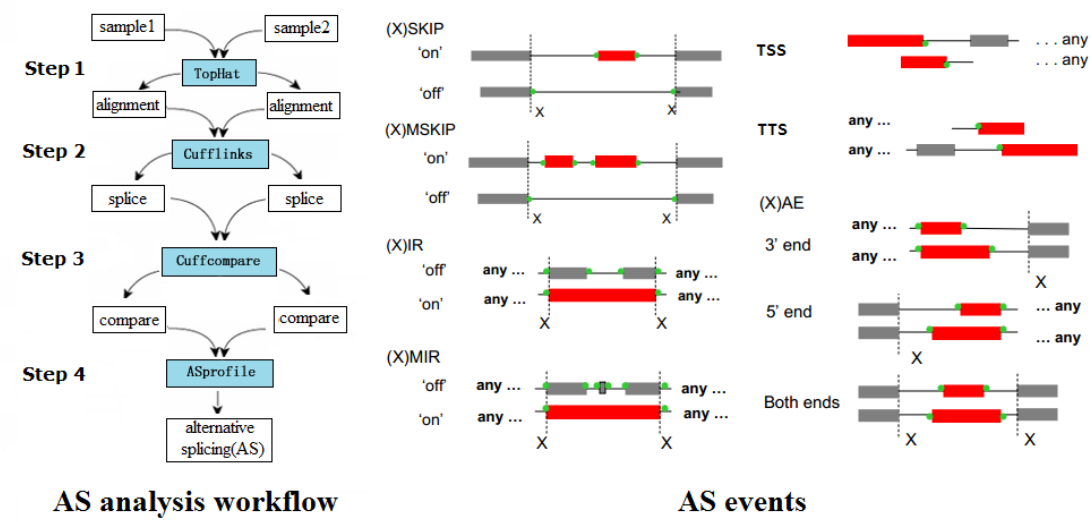


Figure 3.10.2 Conservation of lncRNAs among species

3.11 Alternative splicing (AS) analysis

The ASprofile (Florea et al., 2013) is used to classify and quantify the AS events predicted by cufflinks (Trapnell et al., 2010). The workflow of ASprofile are shown below:



AS analysis workflow **AS events**

The 12 classes alternative splicing events are described below:

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

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

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

(4) XSKIP: Approximate SKIP (XSKIP_ON,XSKIP_OFF pair)

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

(6) XMSKIP: Approximate MSKIP (XMSKIP_ON,XMSKIP_OFF pair)

(7) IR: Intron retention (IR_ON, IR_OFF pair)

(8) XIR: Approximate IR (XIR_ON, XIR_OFF pair)

(9) MIR: Multi-IR (MIR_ON, MIR_OFF pair)

(10) XMIR: Approximate MIR (XMIR_ON, XMIR_OFF pair)

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

(12) XAE: Approximate AE

3.11.1 Classification and quantification of AS events

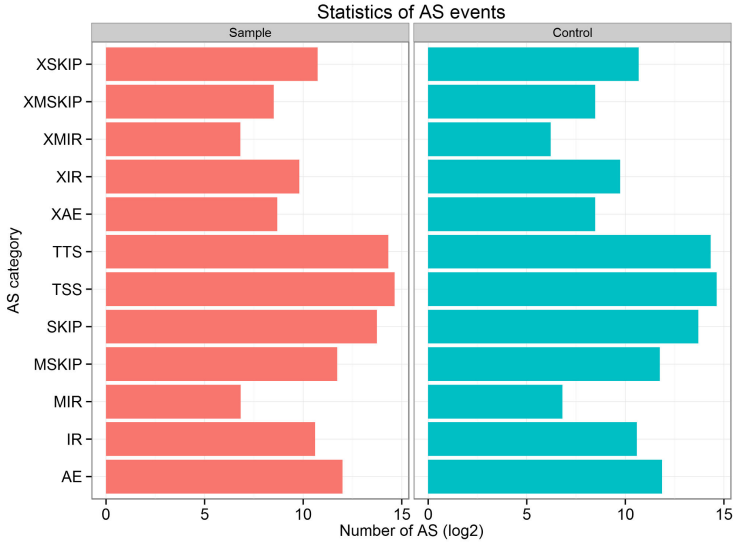


Figure 3.11.1 Classification and quantification of AS events

The vertical axis represents the abbreviations of the AS event, and horizontal axis represents the number of the AS event.

Different samples are distinguished by different sub-figures and colors.

3.11.2 Statistics of types and expression of AS events

Table 3.11.2 Types and expression of AS events

event_id	event_type	gene_id	chrom	event_start	event_end	event_pattern	strand	fpkm	ref_id
1000001	TSS	100127946	chr11	69830650	69830710	69830710	+	0	XM_001717040.2
1000002	TTS	100127946	chr11	69866516	69866574	69866516	+	0	XM_001717040.2
1000003	TSS	100129216	chr11	71589499	71589556	71589556	+	0	NM_001242853.1
1000004	TTS	100129216	chr11	71595453	71595607	71595453	+	0	NM_001242853.1

Note:

(1) event_id: AS event ID

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

(3) gene_id: gene ID from the cufflinks assembly

(4) chrom: chromosome ID

(5) event_start: start position of AS event

(6) event_end: end position of AS event

(7) event_signature: characteristics of AS event (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: strand of gene

(9) fpkm: gene expression level of the AS event

(10) ref_id: gene ID in the reference sequence file

3.12 SNP and InDel analysis

Single Nucleotide Polymorphisms (SNP) is a type of genetic maker that refers to the single nucleotide variation in the genome. There are plenty of SNPs with rich polymorphisms. Theoretically, each SNP site has four types of variation, but in fact there are only two types, namely transformation and transversion, the ratio of which is 1:2. SNPs occur most frequently in the CG sequences, and more often C is converted to T, because C is often methylated in CG, and it will change to T after spontaneously deamination. Normally SNP refers to the single nucleotide variation where the frequency of variation is greater than 1%. InDel (insertion and deletion) refers to the insertion and deletion of small fragments, which is relative to the reference genome, and it may contain one or more bases.

The samtools and picard-tools are used to analyze the mapping results, such as sorting the chromosome and removing duplicate reads, and SNP calling and InDel calling is done by GATK2 (A McKenna, 2010). The table shown below are results after filtering, where the columns in the InDel result are the same as those in the SNP result.

Table 3.12.1 SNP results

#CHROM	POS	REF	ALT	GeneID	Control	Sample
chr1	14653	C	T	10246	6,58	16,60
chr1	14677	G	A	100302652	51,45	63,49
chr1	14907	A	G	10551	7,9	9,4
chr1	14930	A	G	100528064	0,11	7,7

Note:

(1) #CHROM: Chromosome/Scaffold ID of SNPs.

(2) POS: Position of SNPs on corresponding chromosome/scaffold.

(3) REF: Reference genotype.

(4) ALT: SNP genotype (Alternative genotype).

(5) Gene_id: Gene ID from reference GTF file.

(6) other coloums: genotype of each sample at this site (the number represents the reads number supporting the site. In detail, the number before and after comma represents the reads number supporting REF and ALT, respectively.)

3.13 mRNA expression analysis

3.13.1 Quantification of mRNA expression

The expression of mRNAs and lncRNAs is assessed by cuffdiff

(http://cufflinks.cbcb.umd.edu/manual.html#cuffdiff), and the results are shown below:

Table 3.13.1 FPKMs of mRNA from each sample

transcript_id	ST_2	ST_3	SN_2	SN_3
ENST00000618881	0	0	0	0
ENST00000618882	0	1.1894	0	0
ENST00000618887	0	0	2.75718	1.08655
ENST00000496116	1.83384	1.76824	1.36016	0.987781
ENST00000496117	1.25275	0	0	0

3.13.2 Differential expression of mRNAs

Statistically, differential expression analysis of lncRNAs and mRNAs has no bias on molecular type. If the sample has biological replicates, the differential expression is analyzed by cuffdiff, and edgeR is used otherwise.

Table 13.2 Results of differential expression analysis

Gene Id	ST_2	SN_2	log2FoldChange	pval	p-adjusted
ENST00000618881	0	0	0	1	1
ENST00000618882	0	0	0	1	1
ENST00000618887	0	2.75718	-inf	0.0409747	0.502693
ENST00000496116	1.83384	1.36016	0.431093	0.676565	0.99999
ENST00000496117	1.25275	0	inf	0.0637675	0.502693
ENST00000496114	0	0.167968	-inf	0.127744	1

Note:

(1) Gene Id: gene ID

(2) ST_2: mean of FPKMs in sample 1

(3) SN_2: mean of FPKMs in sample 2

(4) log2FoldChange: log2(Sample1/Sample2)

(5) pvalue(pval): p-value

(6) qvalue(p-adjusted): adjusted p-value. Lower qvalue indicates more significant differential expression

3.14 Functional enrichment of differential mRNAs

The differential genes generated by cuffdiff are used for mRNA enrichment analysis.

3.14.1 GO Enrichment of Differential mRNAs

3.14.1.1 GO enrichment of differential mRNAs

Table 3.14.1.1 GO enrichment of differential mRNAs

GO_accession	Description	Term_type	Over_represented_pValue	padj	fg	bg
GO:0005515	protein binding	molecular_function	2.47E-27	1.04E-23	1083	3739
GO:0005488	binding	molecular_function	3.17E-21	6.66E-18	2268	3739
GO:0016787	hydrolase activity	molecular_function	8.85E-12	1.24E-08	673	3739

GO:0031012	extracellular matrix	cellular_component	7.14E-09	7.50E-06	80	3739
GO:0006810	transport	biological_process	4.63E-08	3.24E-05	543	3739
GO:0051234	establishment of localization	biological_process	4.63E-08	3.24E-05	543	3739

Note:

(1) GO_accession: the unique id in Gene Ontology database

(2) Description: function description in gene ontology

(3) Term_type: type of the GO term (one of cellular_component, biological_process or molecular_function)

(4) Over_represented_pValue: statistical significance on enrichment

(5) padj: adjusted p-value. Normally, padj < 0.05 means the gene is enriched in that term

(6) fg: the number of differential genes related to the GO term

(7) bg: the number of differential genes that have GO annotation.

3.14.1.2 DAG of GO-enriched Differential mRNAs

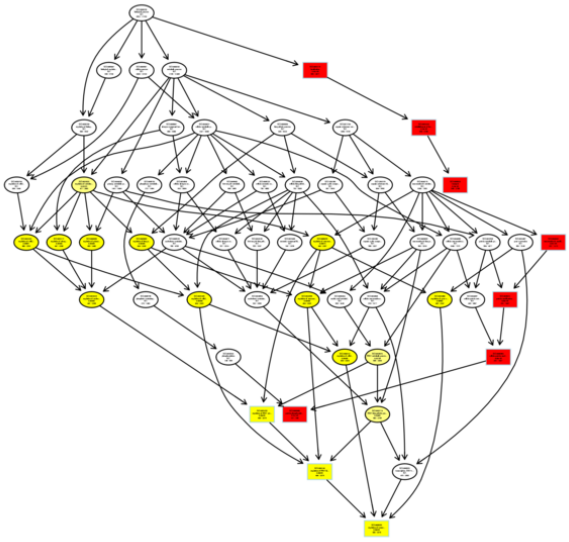


Figure 3.14.1.2 DAGs of GO enrichment

Node represents GO term, and box represents the top 10 terms of GO enrichment. Deeper color indicates higher enrichment and vice versa. The GO term and the padj value of enrichment are shown in each node.

3.14.1.3 Bar plot of GO-enriched Differential mRNAs

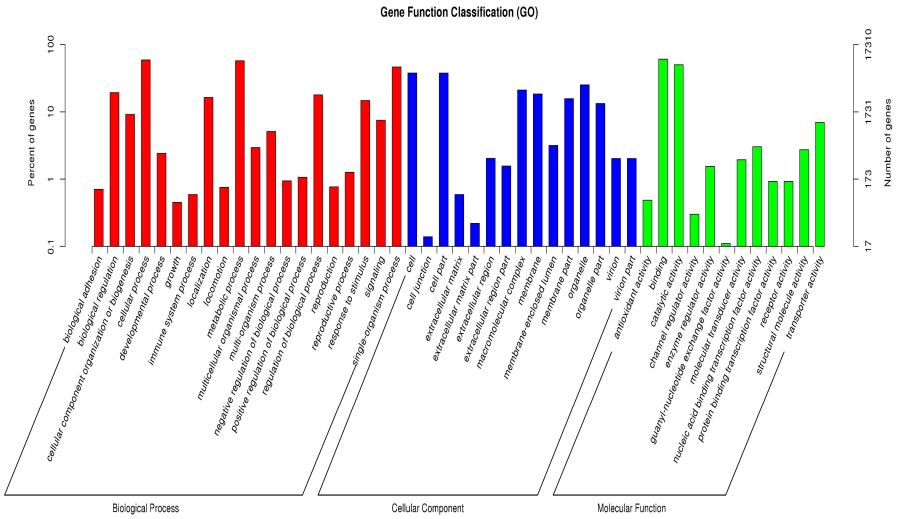


Figure 3.14.1.3 Bar plot of GO enrichment

Node represents GO term, and box represents the top 10 terms of GO enrichment. Deeper color indicates higher enrichment and vice versa. The GO term and the padj value of enrichment are shown in each node.

3.14.1.4 Clustering of GO-enriched Differential mRNAs

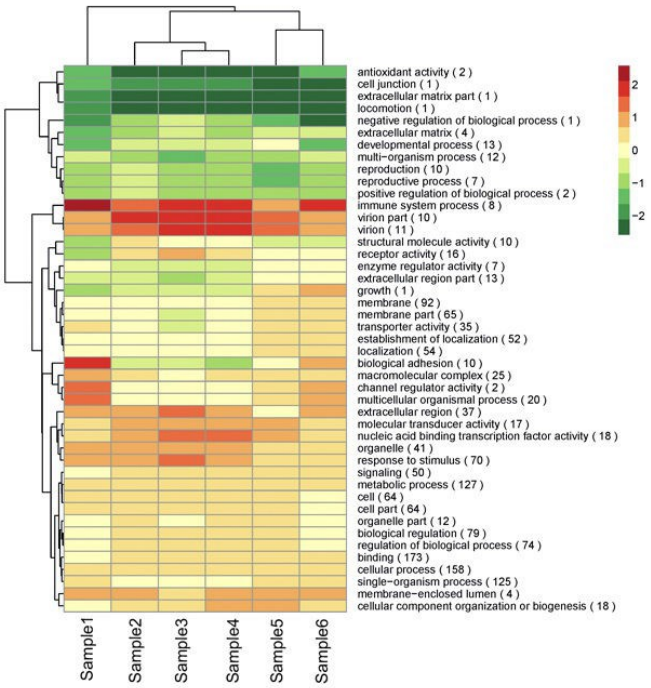


Figure 3.14.1.4 Clustering of enriched GO terms

The union of all terms on level 3 are used for clustering, and the expression level of all genes in each term is calculated. Terms in red and green colors represent high and low expression of genes in the corresponding terms, respectively, and the number in

parenthesis after the term indicates the number of corresponding differential genes.

3.14.2 KEGG Enrichment of Differential mRNAs

3.14.2.1 Summary of KEGG Enrichment of Differential mRNAs

Table 3.14.2.1 KEGG enrichment of differential mRNAs

#Term	Id	fg	bg	P-Value	padj
Focal adhesion	hsa04510	92	207	0.003657421	0.77563385
Pathways in cancer	hsa05200	134	327	0.005540242	0.77563385

Note:

(1) #Term: Description of the KEGG pathway

(2) Id: unique pathway ID in the KEGG database

(3) fg: number of differential genes in the pathway

(4) bg: number of genes in the pathway

(5) P-value: statistical significance of the enrichment

(6) padj: adjusted p-value. Normally, padj < 0.05 means the term is enriched

3.14.2.2 Scatter Plot of KEGG Enrichment of Differential mRNAs

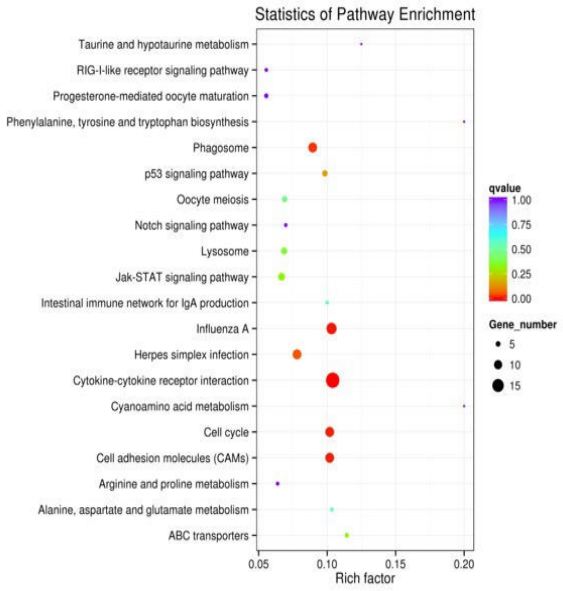


Figure 3.14.2.2 Scatter plot of enriched KEGG pathways of differential mRNAs

Vertical coordinates represent pathway name, and horizontal coordinates represent Rich factor. The size and color of point represent the number of differential genes in the pathway and the range of different Q value, respectively.

3.14.2.3 Enriched KEGG Pathway of Differential mRNAs

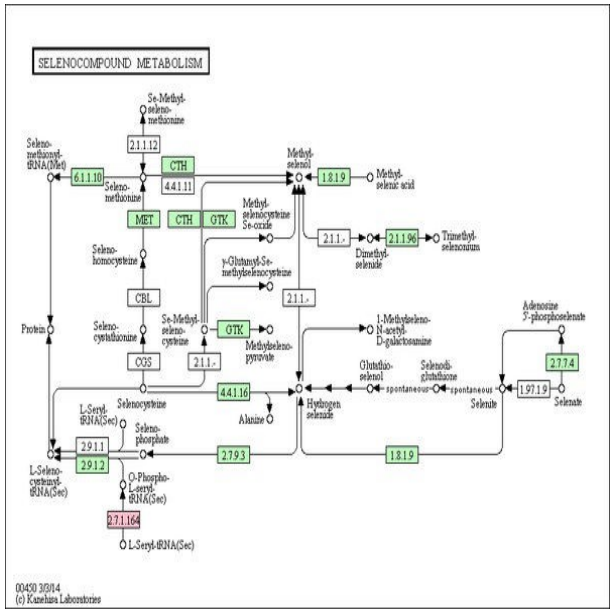


Figure 3.14.2.3 Enriched KEGG pathways of differential mRNAs

3.14.2.4 Clustering of KEGG-enriched Differential mRNAs

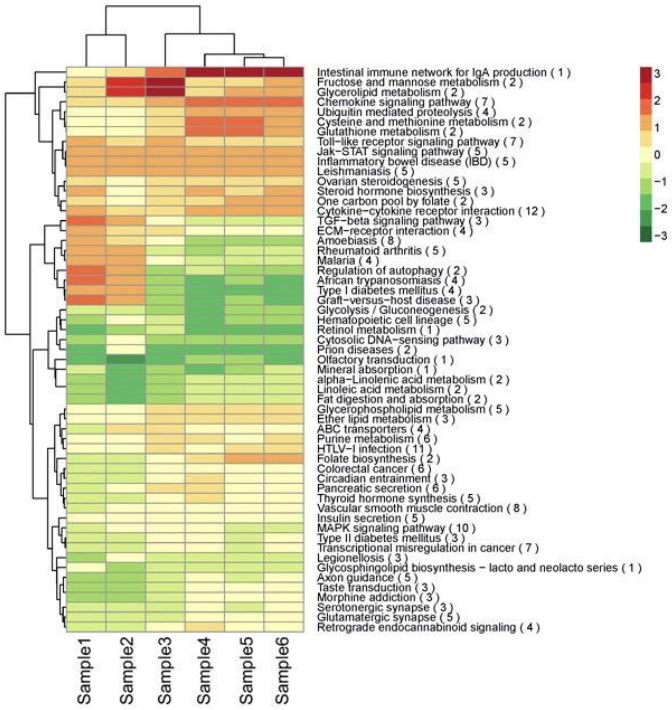


Figure 3.14.2.4 Clustering of KEGG enrichment

The union of all pathways are used for clustering, and the expression level of all genes in each pathway is calculated. Pathways in red and green colors represent high and low expression of genes in the corresponding pathways, respectively, and the number in parenthesis after the pathway indicates the number of corresponding differential genes.

3.15 Network analysis of protein-protein interactions of differential mRNAs

The STRING protein-protein interaction database (http://string-db.org/) is used to construct the interaction network. If the organism exists in the database, the target gene set (such as differentially expressed gene list), are retrieved directly for network construction. Otherwise, the target gene set is blastx searched (Evalue set to 1e-10) against the close species or model organisms in the string database, and the results are used network construction.

The interaction network data file are provided and can be imported to the Cytoscape for editing. Users can summarize and edit the graph according to the topological attributes of some networks. For example, the size of node is in proportion to its degree, that is, the more the edges connected to it, the larger the node and its degree, indicating that these nodes may be the core nodes in the network. The color of node is related to its clustering coefficient. The color gradients from green to red means the corresponding clustering coefficient changes from low to high. The clustering coefficient represents the connectivity of the node and its adjacent nodes, and a higher clustering coefficient means the connectivity is better. According to the purpose and need of the research, users can also customize the graph by adjusting the position and color of the node and annotating the expression levels and so on. It should be noted that the blastx alignment can not ensure good accuracy. This part of analysis, which may assist the user to find some important transcripts, is supplied for reference purpose only. The demonstration of interaction network generated by Cytoscape is shown below:

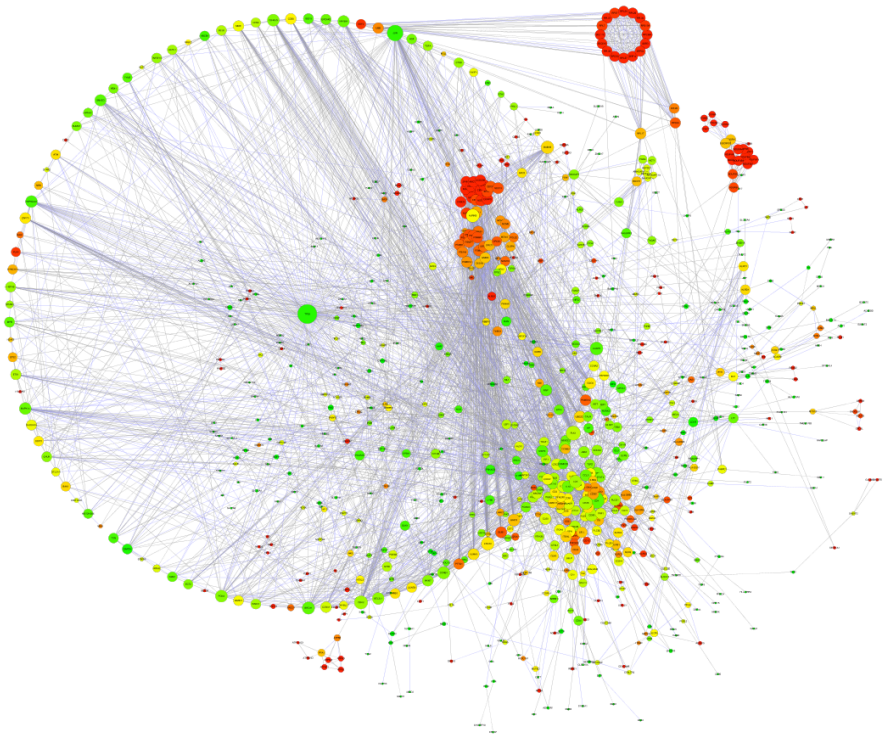


Figure 3.15 Demonstration of interaction network generated by Cytoscape

3.16 Comparison of expression levels of lncRNAs and mRNAs

3.16.1 Comparison of expression levels of lncRNAs and mRNAs

The mean of expression levels of lncRNAs and mRNAs are used and log10-transformed ($\log_{10}$(FPKM+1)) for use in the violin plot.

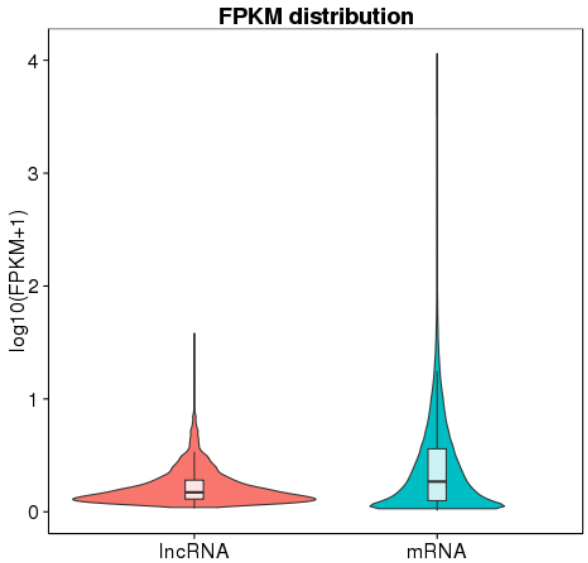


Figure 3.16.1 Violin plot of expression levels of lncRNAs and mRNAs

Horizontal axis represents the molecular type, and vertical axis represents $\log_{10}$(FPKM+1). The width of violin indicates the number of transcripts under current expression level.

3.16.2 Expression analysis of differential lncRNAs and mRNAs

The expression of differential transcripts or genes is visualized by volcano plot. For samples with replicates, the threshold is qvalue < 0.05, otherwise the threshold is qvalue < 0.05 and $|\log_2\text{FoldChange}| > 1$.

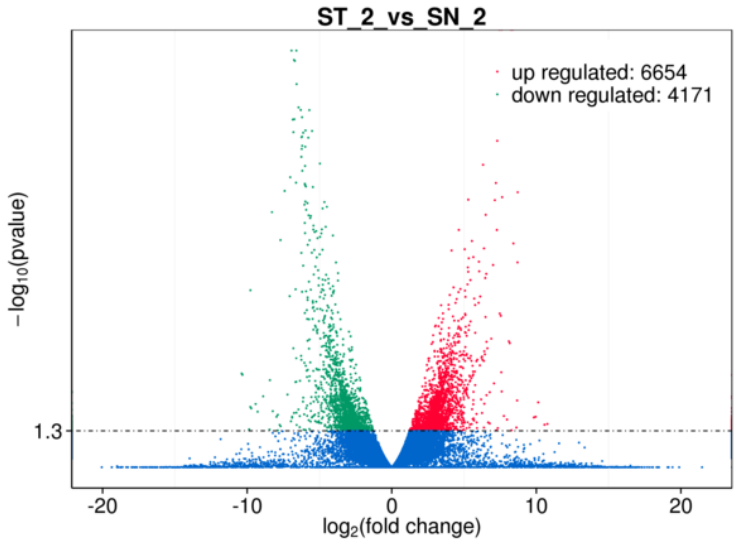


Figure 3.16.2 Volcano plot of differential transcripts

The differential expression with statistical significance are represented by red (up-regulated mRNAs), green (down-regulated mRNAs), yellow (up-regulated lncRNAs) and brown (down-regulated lncRNAs) points, respectively. Horizontal axis represents the fold change of transcripts in different samples, and vertical axis represents the statistical significance of differential expression.

3.16.3 Distribution of lncRNAs and mRNAs in chromosomes

Genes are usually regularly distributed in chromosomes, and those that have similar functions may cluster in the same chromosome. Meanwhile, The adjacent genes usually have similar functions, or involved in the same cell type or metabolic pathway, and they are more possibly regulated by each other, compared to genes in long distance. Therefore, for differential expression studies, the distribution of genes and adjacent genes in the chromosome may be important, which is the key for selection of differential genes. In addition, higher density of differential genes within a region on the chromosome can help us to find interested genes.

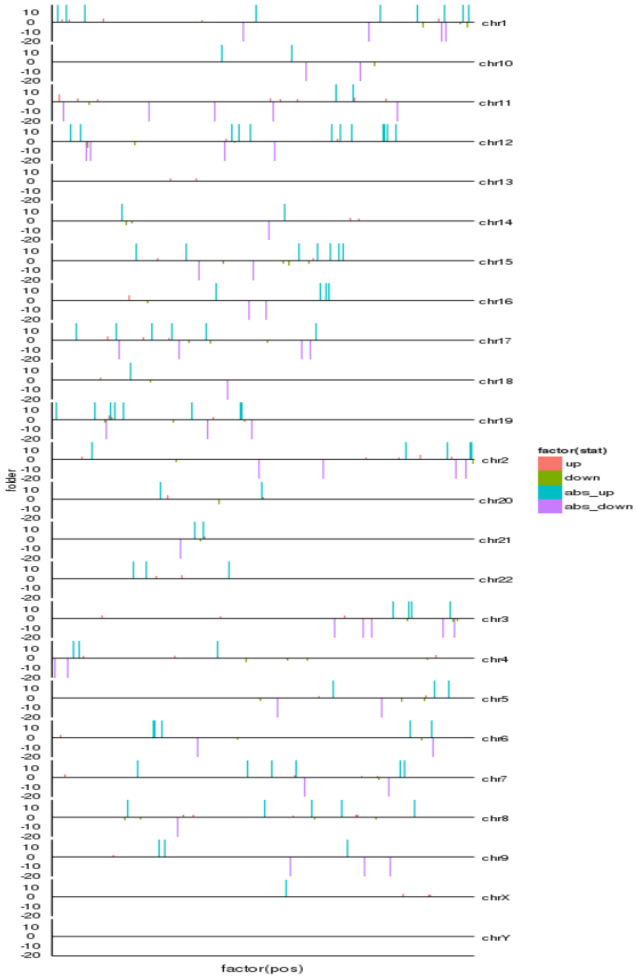


Figure 3.16.3 Distribution of differential genes on chromosomes

The differential genes are screened based on FPKM values from different samples (The threshold is qvalue < 0.05).

3.16.4 Clustering of differential lncRNAs and mRNAs

The clustering analysis is used to assess the expression of transcripts under different experimental conditions. The functions of novel transcripts or the unknown functions of known transcripts can be identified by clustering of genes with the same or similar expression, since these transcripts may have similar functions, or involved in the same metabolic pathway or cellular component. The FPKMs of transcripts are used for hierarchical clustering, where different color indicates different grouping. The ones within the same group have similar expression, which may have similar function or involved in the same biological process.

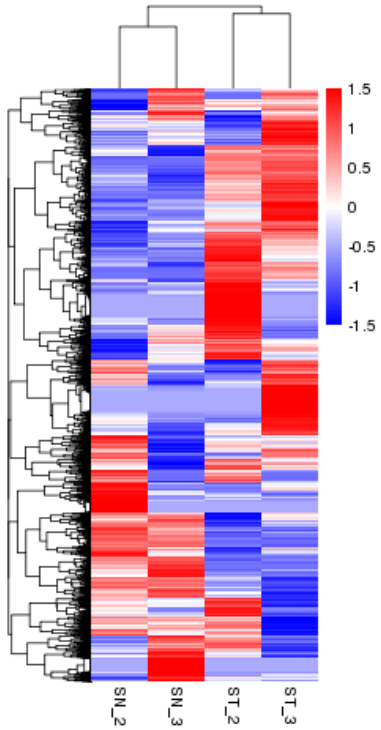


Figure 3.16.4 Clustering of differentially expressed transcripts

Hierarchical clustering based on FPKMs, where log10(FPKM+1) is used for clustering. Red color represents genes with higher expression, while blue color represents genes with lower expression.

3.16.5 Venn diagram of differential expression

When there are 2-5 samples, the comparison of each group can be visualized as venn diagram, which is intuitive to explore unique and common transcripts from each sample.

Figure 16.5 Venn diagram of differentialy expressed genes

The sum of the number in each big circle represents the total number of differential transcripts in the comparison, and the number in the overlap region represents the number of shared transcripts.

3.17 Comparison of structures of lncRNAs and mRNAs

To study the difference of the lncRNAs and mRNAs molecules and whether the predicted lncRNAs consist with the annotated lncRNAs, the structures of lncRNAs and mRNAs are compared based on the length of the transcript and the number of exons and ORFs.

3.17.1 Length comparison of lncRNAs and mRNAs

The result of length comparison is shown below:

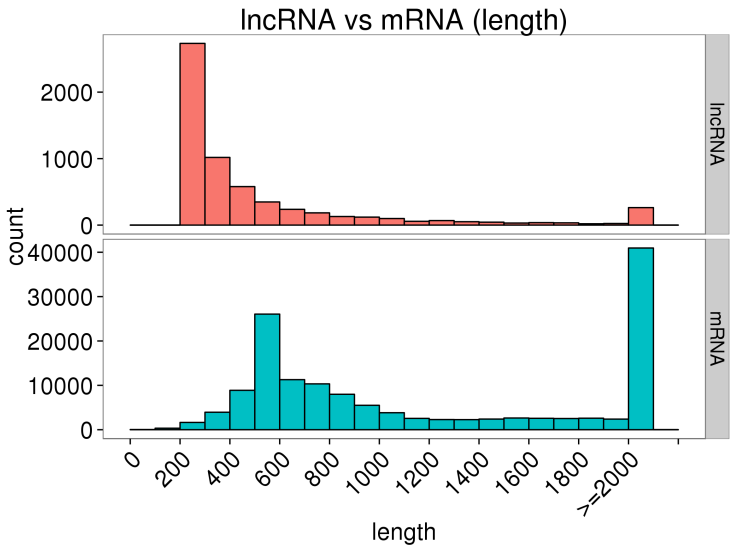


Figure 17.1 Length comparison of lncRNAs and mRNAs

The figures in the top and bottom are the length distribution of lncRNAs and mRNAs, respectively. Horizontal axis represents the length of transcripts, and vertical axis represents the number of transcripts for each length.

3.17.2 Comparison of exon numbers of lncRNAs and mRNAs

The result of the comparison is shown below:

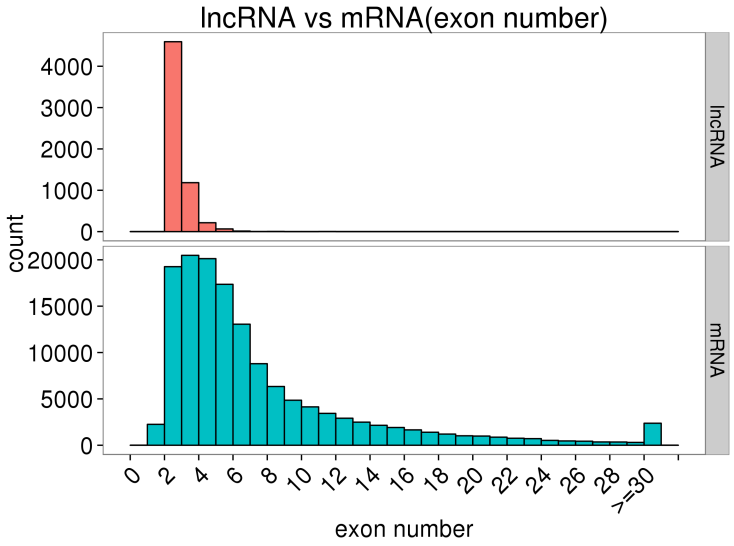


Figure 3.17.2 Comparison of exon numbers of lncRNAs and mRNAs

The figures in the top and bottom are the distribution of exon numbers of lncRNAs and mRNAs, respectively. Horizontal axis represents the number of exons, and vertical axis represents the number of transcripts for each exon number.

3.17.3 Comparison of ORF length of lncRNAs and mRNAs

The ORFs of known genes are retrieved based on the gene annotations, and the ORFs of lncRNAs are predicted by estscan and translated to protein sequences. The length distribution of ORFs is shown below:

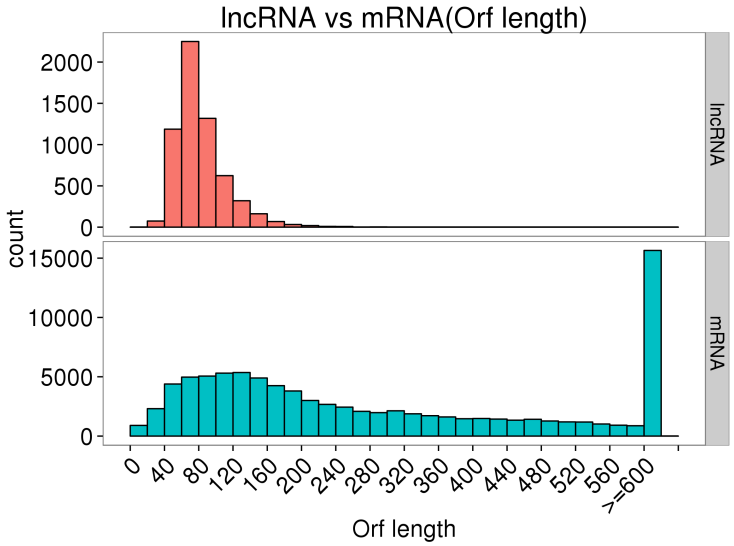


Figure 3.17.3 Comparison of ORF length of lncRNAs and mRNAs

The figures in the top and bottom are the distribution of ORF length of lncRNAs and mRNAs, respectively. Horizontal axis represents the length of ORFs, and vertical axis represents the number of transcripts for each ORF length.

3.18 LncRNA-mRNA interaction network

LncRNAs and mRNAs can be associated by the targeting relation, and mRNAs can be associated by protein-protein interactions, then the lncRNA-mRNA-protein interaction network can be created. The differential lncRNAs and the targeted *cis*- or *trans*-acting mRNAs are associated, and mRNAs are associated by using the STRING database (http://string-db.org/) (See the section "Network analysis of protein-protein interactions of differential mRNAs" above for details).

The data files of lncRNA-mRNA and mRNA-mRNA interaction networks are provided and can be imported to the Cytoscape for editing. Users can summarize and edit the graph according to the topological attributes of some networks. According to the purpose and need of the research, users can also customize the graph by adjusting the position and color of the node and annotating the expression levels and so on. It should be noted that the blastx alignment can not ensure good accuracy. This part of analysis, which may assist the user to find some important genes, is supplied for reference purpose only. The demonstration of interaction network generated by Cytoscape is shown below:

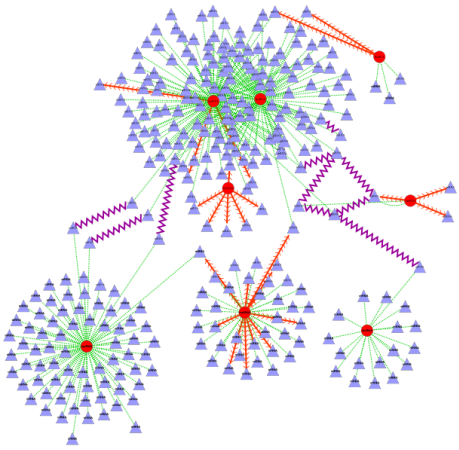


Figure 3.18 Demonstration of interaction network generated by Cytoscape

LncRNAs and target genes are shown in circle and triangle, respectively. The solid line in red color represents the interaction of lncRNAs and *cis*-acting targets, the dashed line in green color represents the interaction of lncRNAs and *trans*-acting targets, and the solid line in purple color represents the mRNA-mRNA interaction. When there are not *trans*-acting targets, the dashed line in green color is not available accordingly.

4 References

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

Bateman A. et al (2002). The Pfam protein families database[J]. Nucleic acids

research, 30(1): 276-280. (Pfam)
Cock P.J.A. et al (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research 38, 1767-1771.
Erlich Y.et al (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing.Nature Methods,5,679-682.
Florea, L. et al (2013). Thousands of exon skipping events differentiate splicing patterns in sixteen human tissues. F1000Research, 2:188. (ASprofile)
Guttman M. et al (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature biotechnology. (scripture)
Hansen. et al (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13.2: 204-216.
Jiang L.C. et al (2011). Synthetic spike-in standards for RNA-seq experiments. Genome research 21, 1543-1551.
Kanehisa, M., M. Araki, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic acids research. (KEGG)
Kim, D.G. et al (2012). TopHat2: Parallel mapping of transcriptomes to detect InDels, gene fusions, and more. (TopHat2)
Kong, L. et al (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 36: W345-349. (CPC)
Langfelder P. et al (2008). WGCNA: an R package for weighted correlation network analysis. (WGCNA)
Langmead B. et al (2012). Fast gapped-read alignment with Bowtie 2. Nature methods. (Bowtie 2)
Langmead B. et al (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. (Bowtie)
Lin M.F. et al (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 27(13): i275-i282. (phyloCSF)
Mao, X. et al (1995). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics. (KOBAS)
Marquez Y. et al (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res. 22, 1184–1195.
McKenna A. et al (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. (GATK)
Mistry J. et al (2007). Predicting active site residue annotations in the Pfam database. BMC bioinformatics, 8(1): 298. (pfamscan)
Mortazavi A. et al (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods.
Siepel A. et al (2005). Phylogenetic hidden Markov models. In R. Nielsen, ed., Statistical Methods in Molecular Evolution, pp. 325-351, Springer, New York.

Siepel A. et al (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034-1050

Sun L. et al (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic acids research, 41(17): e166-e166. (CNCI)

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

Trapnell C. 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 (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. (TopHat)

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

Yan L.Y. et al (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol.

Young M.D. et al (2010).Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology. (GOseq)