Med RNA-Seq Analysis Report

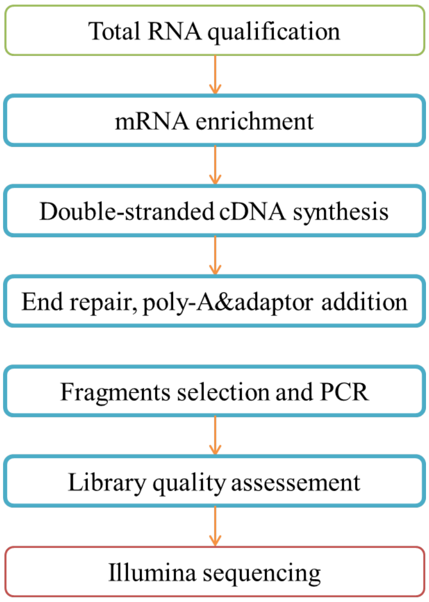
Demo Report

May 1, 2016

Contents

1 The Workflow of Library Construction and Sequencing

As all bioinformatics analysis results depends on the quality of raw data, we do our best to make sure that each step of data production has been checked carefully. The workflow is as follows:


Total RNA qualification
mRNA enrichment
Double-stranded cDNA synthesis
End repair, poly-A&adaptor addition
Fragments selection and PCR
Library quality assessement
Illumina sequencing

1.1 Total RNA quantification and qualification

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

(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.

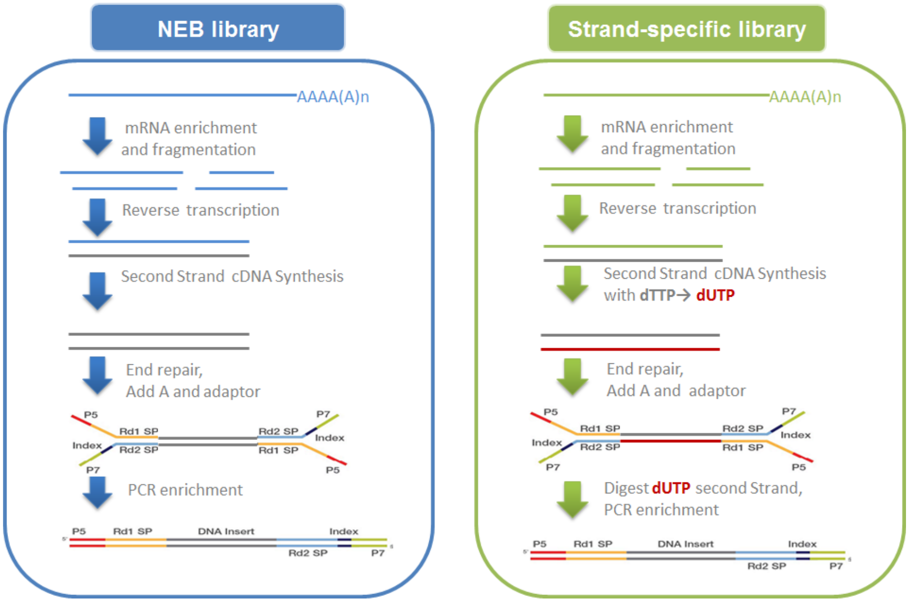
1.2 Library construction and quality assessment

1.2.1 Library construction

Briefly, mRNA from Eukaryote organisms is purified from total RNA using poly-T oligo-attached magnetic beads (For prokaryotes, mRNA was purified through the removal of rRNA by using kit). The mRNA is first fragmented randomly by addition of fragmentation buffer. According to demand of customer, we have ENB library and strand specific library to choose.

ENB library: Then first strand cDNA is synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H-). Second strand cDNA synthesis is subsequently performed using DNA Polymerase I and RNase H. Double-stranded cDNA is purified using AMPure XP beads. Remaining overhangs of the purified double-stranded cDNA are converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3’ ends of DNA fragments, NEBNext Adaptor with hairpin loop structure is ligated to prepare for hybridization (1). In order to select cDNA fragments of preferentially 150~200 bp in length, the library fragments are purified with AMPure XP system (Beckman Coulter, Beverly, USA). Finally, the final library is gotten by PCR amplification and purification of PCR products by AMPure XP beads.

Strand specific library: The construct method of strand specific library is similar to that of ENB library expect that when synthesizing the second strand cDNA, dTTP is replayed by dUTP. After overhangs of the purified double-stranded cDNA are converted into blunt ends, adenylation of 3’ ends of DNA fragments, NEBNext Adaptor with hairpin loop structure is ligated to prepare for hybridization, the second strand cDNA is digested by USER enzyme. The following step is identical with ENB library construction. Below is the workflow chart:


NEB library
AAAA(A)n
mRNA enrichment and fragmentation
Reverse transcription
Second Strand cDNA Synthesis
End repair, Add A and adaptor
P5
Rd1 SP
Rd2 SP
P7
Index
PCR enrichment
DNA Insert
Strand-specific library
AAAA(A)n
mRNA enrichment and fragmentation
Reverse transcription
Second Strand cDNA Synthesis with dTTP→ dUTP
End repair, Add A and adaptor
Digest dUTP second Strand, PCR enrichment
P5
Rd1 SP
DNA Insert
Index
Rd2 SP
P7

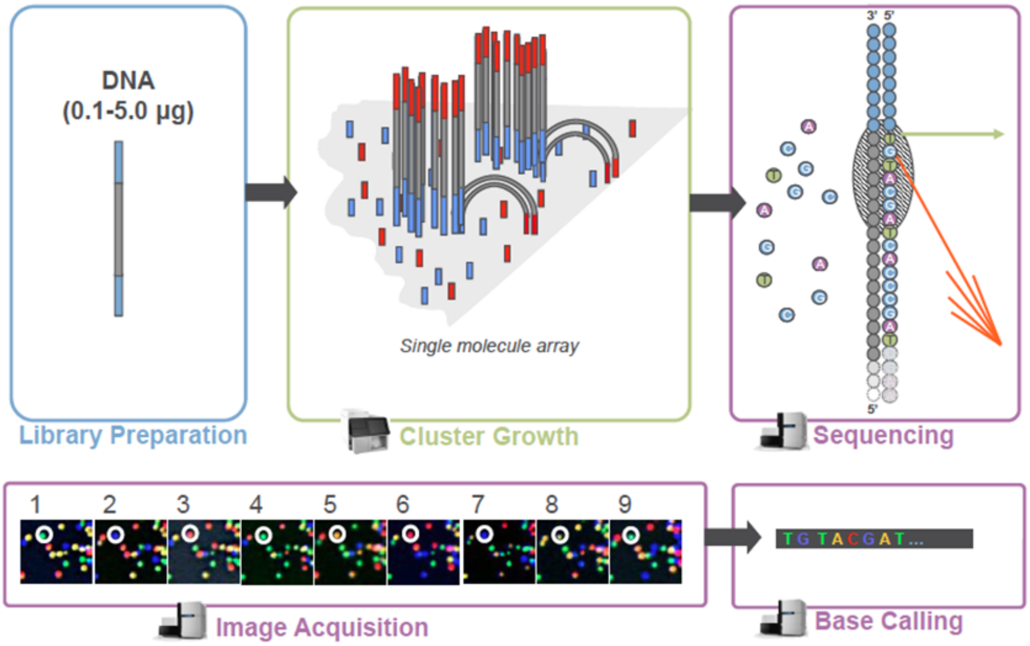
(1) Adaptor:P5/P7 is PCR primers and those primers are complementary to sequences on flow cell; Rd1/Rd2 SP are read1/read2 sequencing primers; Index is used for identifying different libraries

1.2.2 Library quality assessment

After library construction, diluting library to 1.5ng/ul with the preliminary quantitative result by Qubit2.0 and detecting the insert size by Agilent 2100. Q-PCR is used to accurately quantify the library effective concentration (> 2nM), in order to ensure the library quality.

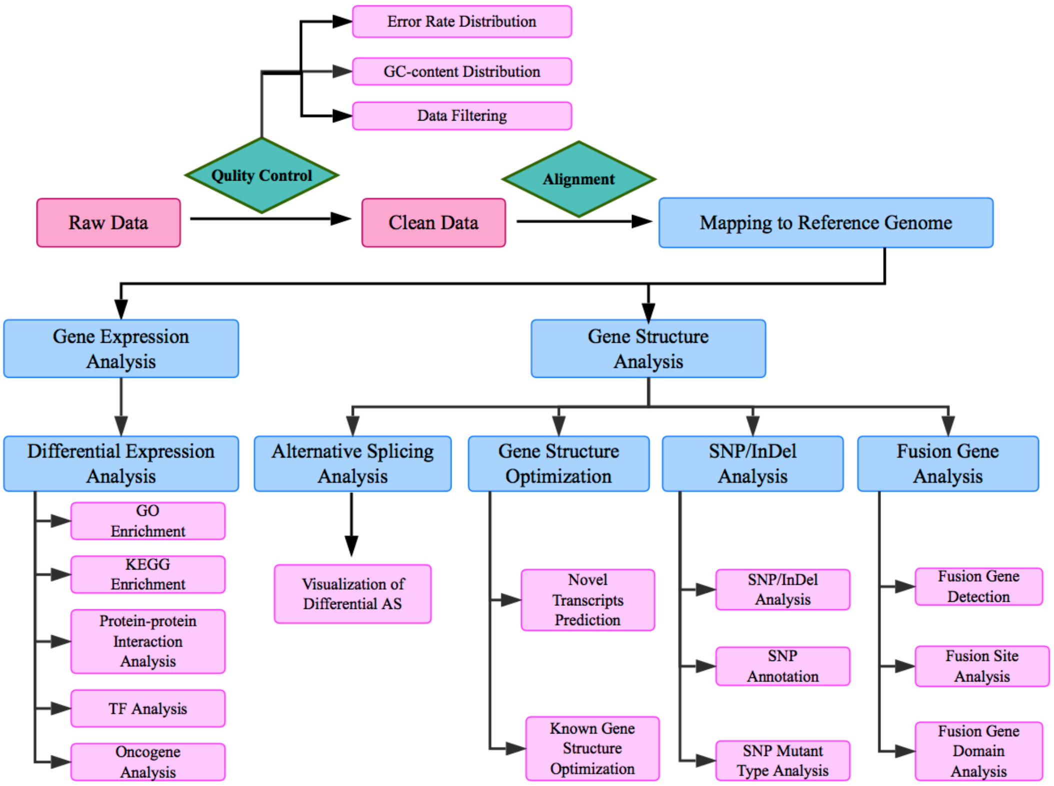
1.3 Sequencing

Libraries are fed into HiSeq/MiSeq machines after pooling according to activity and expected data volume if library quality is up to standard.The sequencing overview is as follows:



2 Project Results

The workflow of bioinformatics analysis



2.1 Raw data processing and quality assessment

2.1.1 The description of raw data

Raw image data file from high-throughput sequencing (like Illumina HiSeq^{TM}) was transformed to Sequenced Reads (called Raw Data or Raw Reads) by CASAVA base recognition (Base Calling). Raw Data is stored in FASTQ(fq) format files, which contain reads sequence and corresponding base quality. FASTQ format is described by four lines:

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT

NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT

+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHFAE0ECFFD/AEHH
```

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence letters. 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 symbols as letters in the sequence.

The details of Sequencing identifier of Illumina are as follows:

(1) HWI-ST1276:71 HWI-ST1276, Instrument - unique identifier of the sequencer; 71, run number - Run number on instrument

(2) C1162ACXX:1:1101:1208:2458 means the coordinate of read on C1162ACXX (Flowcell ID) flowcell, line 1, 1101 tile is(x=1208, y=2458)

(3) 1:N:0:CGATGT the first number is 1 or 2, 1 means single reads or the first read of paired ends, 2 means the second of paired ends; the second letter means whether reads is adjusted(Y means yes, N means no); the third number represent the number of Control Bits in sequence; six bases on the fourth place is Illumina index sequence.

2.1.2 Error rate distribution

"e" represents sequencing error rate, "Qphred" represents base quality values of Illumina HiSeq^{TM}, the Qphred=-10log10(e). Base Quality and Phred score relationship in illumina Casava 1.8 as follows:

Phred score	ASCII code	error rate	correct rate	Q-sorce
10	+ (10+33)	1/10	90%	Q10
20	5 (20+33)	1/100	99%	Q20
30	? (30+33)	1/1000	99.90%	Q30
40	I (40+33)	1/10000	99.99%	Q40

For RNA-seq technology, sequencing error rate distribution has two features, see Fig 1:

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

(2) The first six bases have a relatively high error rate. The reason is incomplete binding of the random hex-primers and RNA template in cDNA synthesis (Jiang et al.). In general, a single base error rate should be lower than 1%.

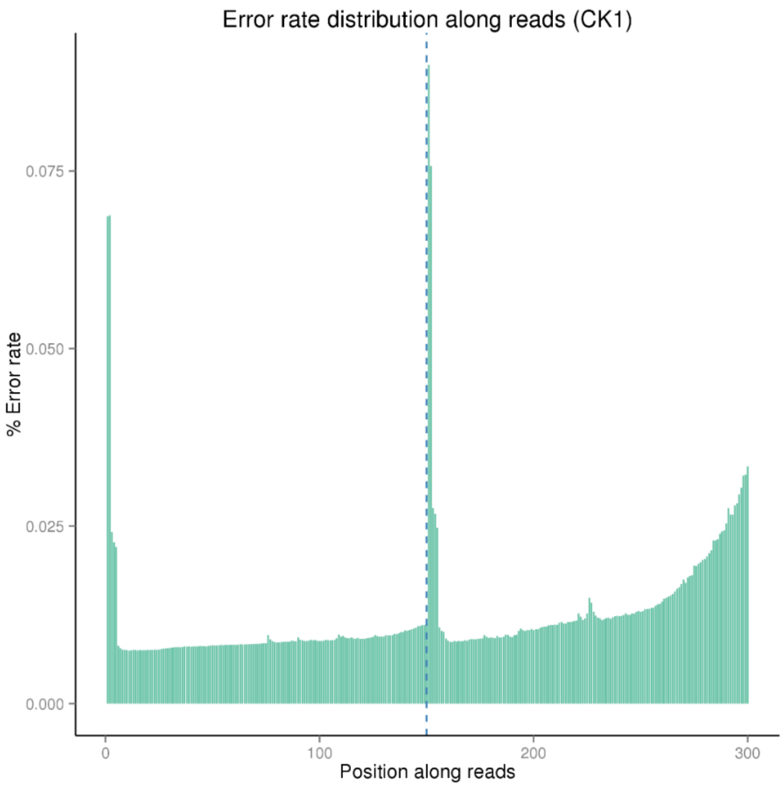


Fig 1 Error rate distribution

Horizontal axis for reads position, vertical axis for single base error rate

2.1.3 GC-content

GC-content is used to detect potential AT/GC separation.

For RNA-seq, in view of the random interrupted and the principle of G/C, A/T content are respectively equal, G and C, A and T should be equal and content is stable on the whole through the whole sequencing process for non-stranded library (If the library is strand-specific, it may occur AT separation or GC separation). A large variation of sequencing error in the first 6-7 bases is allowed considering the usage of random primer in library construction. It's normal that the first few bases have certain preference in existing high-throughput sequencing technology, for the usage of 6bp random primer in library construction. Fig 2

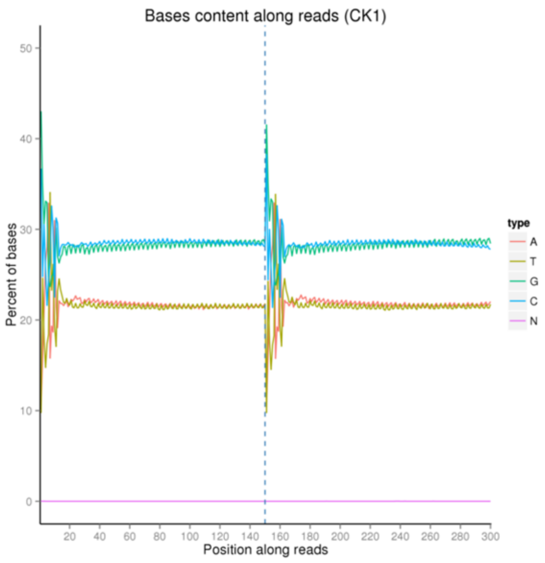


Fig 2 GC-content

Horizontal axis for reads position, vertical axis for single base percentage. Different color for different base type

First 125bp is the GC-content of reads 1 in PE reads, Remaining 125bp is the GC-content of reads 2 in PE reads.

2.1.4 Data filtering

The Sequenced Reads/raw reads often contain low quality reads or reads with adaptors Fig 3, which will affect the analysis quality. To avoid this, it's necessary to filter the raw reads and get the clean reads. Raw reads filtering as follows:

(1) Remove reads containing adaptors;

(2) Remove reads containing N > 10% (N represents base that could not be determined);

(3) The Qscore (Quality value) of over 50% bases of the read is <= 5.

RNA-seq Adapter information:

NEBNext$^{®}$ Ultra^{TM} RNA Library Prep Kit

RNA 5' Adapter(5’Adaptor)

5’AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC TTCCGATCT

RNA 3' Adapter (3’Adaptor,The underlined 6bp bases is Index)

5’GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATG CCGTCTTCTGCTTG

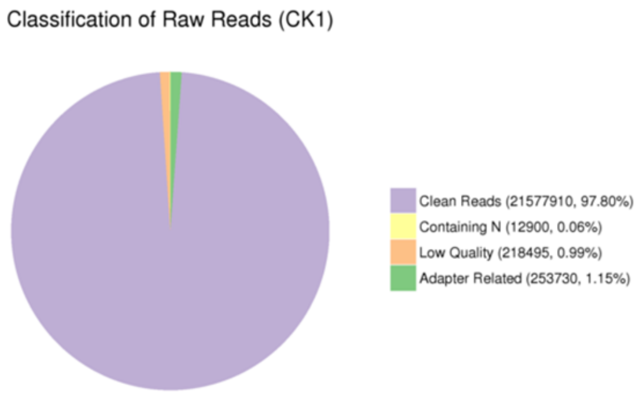


Fig 3 Composition of raw data

Different color for different components

(1) Adaptor related: (reads containing adapter) / (total raw reads).

(2) Containing N: (reads with more than 10% N) / (total raw reads).

(3) Low quality: (reads of low quality) / (total raw reads).

(4) Clean reads: (clean reads) / (total raw reads).

2.1.5 Data quality summary

Table 1 Data quality summary

Sample name	Raw reads	Clean reads	Clean bases	Error rate(%)	Q20(%)	Q30(%)	GC content(%)
CK1	20546961	20521798	1.03G	0.01	97.58	94.89	48.35
CK2	22247752	22217540	1.11G	0.01	97.64	95.03	48.88
CK3	23088290	23059051	1.15G	0.01	97.58	94.87	48.96
treat1	25477684	25445679	1.27G	0.01	97.58	94.88	49.44
treat2	23154667	23121864	1.16G	0.01	97.55	94.82	49.94
treat3	21918524	21892392	1.09G	0.01	97.71	95.19	49.31

(1) Sample name: sample names.

(2) Raw reads: Four rows as a unit to calculate the sequence number of each raw data file.

(3) Clean reads: Calculated as Raw Reads, statistics object is cleand data file. The subsequent analyzes are all based on clean reads.

(4) Clean bases: (Number of sequences) * (sequence length), use G for unit.

(5) Error rate: base error rate.

(6,7) Q20, Q30: (Base number of Phred value > 20(> 30)) / (Total base number).

(8) GC content: (G&C base number) / (Total base number).

2.2 Mapping reads to reference genome

The human genome and its annotation is more and more comprehensive since the beginning of the human genome project. It provides us with great convenience to analyze transcriptome. Ensemble is a joint scientific project between the European Bioinformatics Institute and the Welcome Trust Sanger Institute, which was launched in 1999 in response to the imminent completion of the Human Genome Project. After 10 years in existence, Ensemble's aim remains to provide a centralized resource for geneticists, molecular biologists and other researchers studying the genomes of our own species and other vertebrates and model organisms. Ensemble is one of several well-known genome browsers for the retrieval of genomic information. We choose human genome sequence and annotation from Ensemble project as reference.

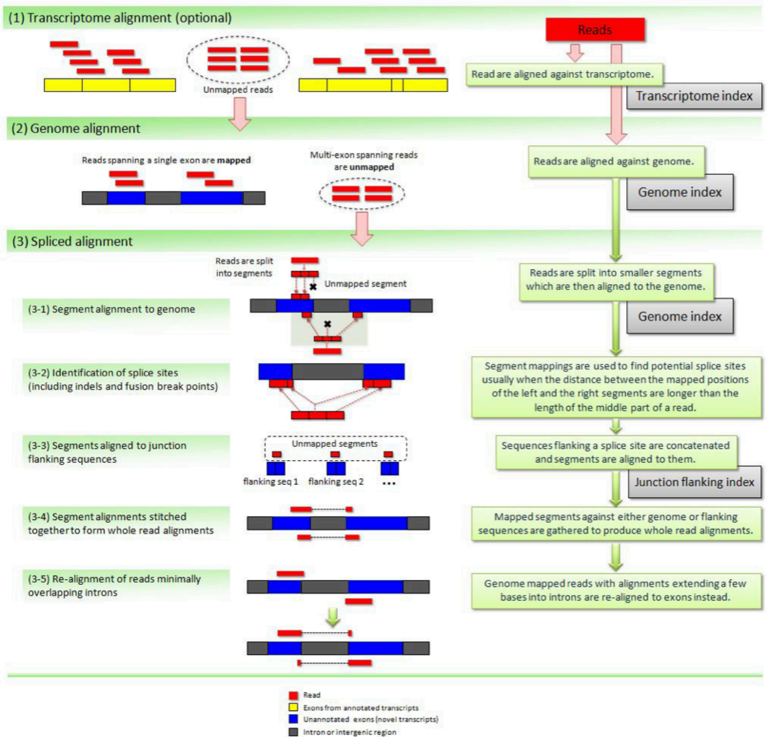
The first step of mapping is to align reads to the reference genome. We use TopHat2 (Kim et al, 2013) to accomplish the alignment. TopHat2 mapping can be divided into three steps as follow:

(1) Align reads to transcriptome (optional).

(2) Align the whole read to exon,   disallowing large gaps.

(3) Unmapped reads are split into shorter segments and aligned independently. The genomic regions surrounding the mapped read segments are then searched for possible spliced connections.

The following figure show the alignment process (Kim et al, 2013):


(1) Transcriptome alignment (optional)
Reads
Unmapped reads
Read are aligned against transcriptome.
Transcriptome index
(2) Genome alignment
Reads spanning a single exon are mapped
Multi-exon spanning reads are unmapped
Reads are aligned against genome.
Genome index
(3) Spliced alignment
Reads are split into segments
Unmapped segment
Reads are split into smaller segments which are then aligned to the genome.
(3-1) Segment alignment to genome
Genome index
(3-2) Identification of splice sites (including indels and fusion break points)
Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.
(3-3) Segments aligned to junction flanking sequences
Unmapped segments
flanking seq 1
flanking seq 2
...
Sequences flanking a splice site are concatenated and segments are aligned to them.
Junction flanking index
(3-4) Segment alignments stitched together to form whole read alignments
Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.
(3-5) Re-alignment of reads minimally overlapping introns
Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.
Read
Exons from annotated transcripts
Unannotated exons (novel transcripts)
Intron or intergenic region

2.2.1 Mapping result

Table 2 The summary of mapping result

Sample_name	CK1	CK2	CK3	treat1	treat2	treat3
Total reads	20521798	22217540	23059051	25445679	23121864	21892392
Total mapped	19964936 (97.29%)	21480583 (96.68%)	22438267 (97.31%)	24630579 (96.8%)	22292052 (96.41%)	21222468 (96.94%)
Multiple mapped	1670231 (8.14%)	1878172 (8.45%)	1796125 (7.79%)	1919307 (7.54%)	1544381 (6.68%)	1628363 (7.44%)
Uniquely mapped	18294705 (89.15%)	19602411 (88.23%)	20642142 (89.52%)	22711272 (89.25%)	20747671 (89.73%)	19594105 (89.5%)
Non-splice reads	14740624 (71.83%)	16476604 (74.16%)	16555159 (71.79%)	18190504 (71.49%)	17192225 (74.35%)	16277545 (74.35%)
Splice reads	3554081 (17.32%)	3125807 (14.07%)	4086983 (17.72%)	4520768 (17.77%)	3555446 (15.38%)	3316560 (15.15%)

(1) Total reads: total clean reads;

(2) Total mapped: the ratio of mapped reads with total reads(the ratio should higher than 70%);

(3) Multiple mapped: the ratio of multiple mapped reads with total reads(the ratio should lower than 10%);

(4) Uniquely mapped the ratio of uniquely mapped reads with total reads;

(5) Reads map to '+', Reads map to '-': the ratio of Reads map to '+' strand, Reads map to '-' strand with total reads;

(6) Splice reads: also called Junction reads, the ratio of mapped splice reads with total reads;

2.2.2 Reads distribution on genome

Reads will be mapped to exon, intron and intergenic.

Because of complete annotation of human genome, the ratio of reads mapped to exon is the highest. Reads mapped to intron may due to the remaining pre-mRNA or alternative splicing event. Reads Mapped to intergenic reads may due to the incomplete gene annotation. In addition, some ncRNA with polyA tail will affect the mapping result.

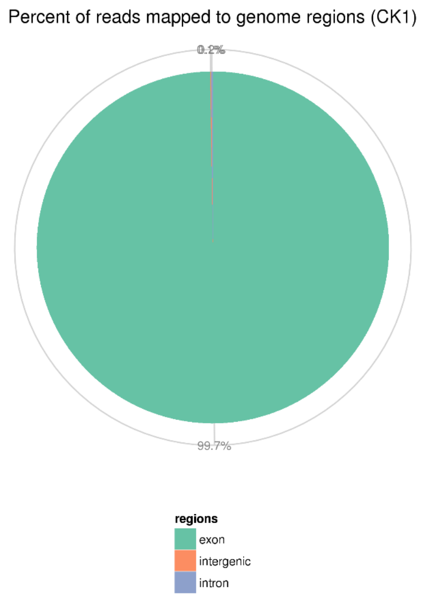


Fig 4 Reads distribution on reference genome

2.2.3 Distribution of mapped reads on chromosome

After getting the information about distribution of mapped reads on genome area, we will focus the distribution of mapped reads on chromosome. On one hand we can get insight into how many regions on chromosome is covered, on the other hand we can get the information about transcription activity on chromosome.

We calculate mapped reads density on every chromosome, then plot based on the mapped reads density. The plot method are as follows:

(1) 1K bp as a unit when window size is set, counting number of reads that mapped to every base in window;

(2) Mid-value is chosen to represent the number of reads that mapped to each base in window; the longer chromosome is, the more reads will be calculated (Marquez et al.);

(3) $\log_2$ (mid-value) is used to plot.

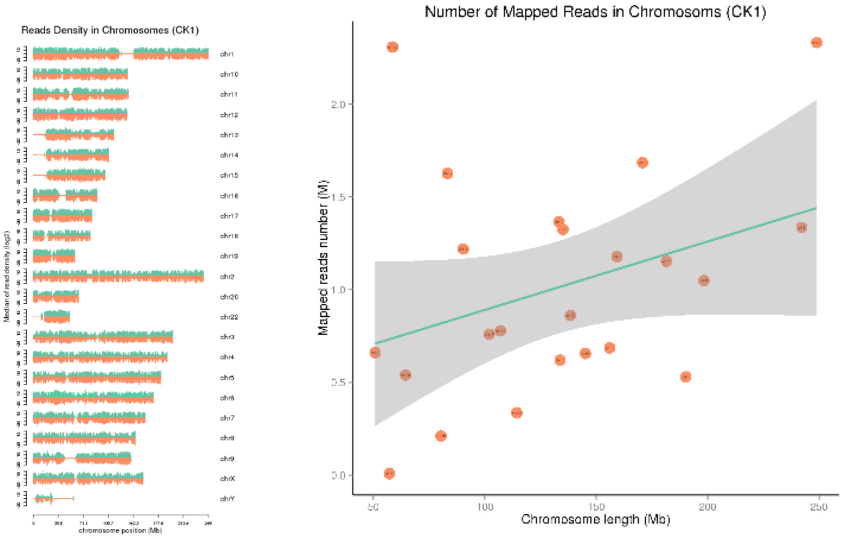


Fig 5 Distribution of mapped reads on chromosomes

Two figures for every sample; Fig 5.1: X axis represents the length of the chromosomes (in Mb), Y axis indicates the log2(mid-value). Green represents sense stands, red represents antisense strands.

Fig 5.2: X axis represents length of the chromosomes; Y axis indicates number of mapped reads on each chromosome. Grey region indicates the 95% confidence interval.

3.2.4 Visualization of mapping results

We provide mapping result in the format of BAM together with reference genome and annotation file. BAM file can be visualized by using IGV (Integrative Genomics Viewer). The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations. By using IGV we can (1) visualize reads position on genome, including reads distribution on chromosomes, the distribution of exon, intron, junction, intergenic and so on at different scale (2) visualize reads density in different area at different scale (3) visualize annotation of gene and its isoform (4) visualize other annotation information(IGVQuickStart.pdf).

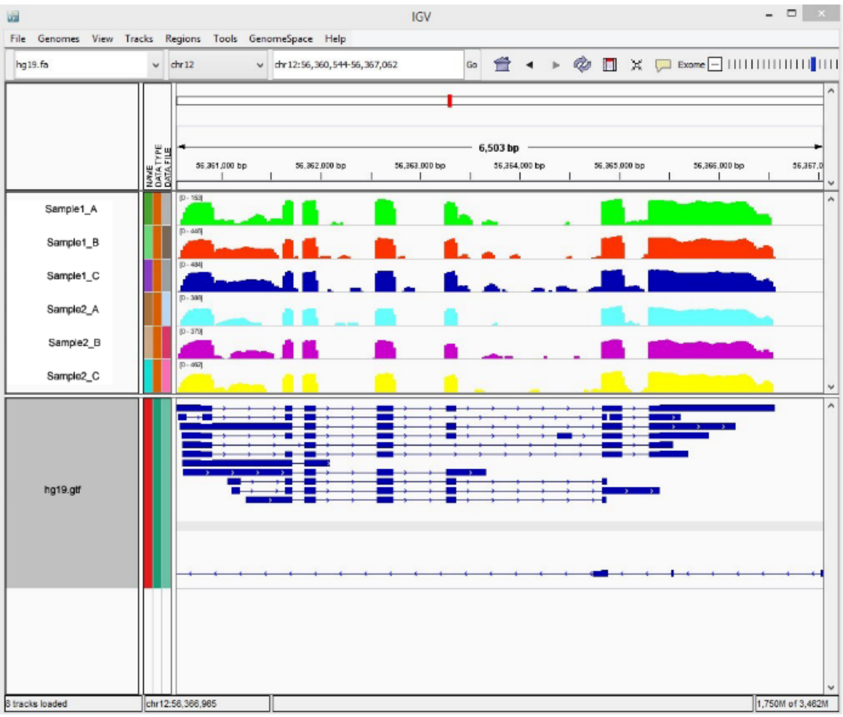


Fig 6 Interface of IGV

2.3 Gene expression level analysis

Gene expression level analysis is the core task in RNA-seq experiment. By calculating number of mapped reads, we can know about gene expression level.

2.3.1 Gene expression level analysis

In RNA-seq experiment, we use read counts to estimate gene expression level. FPKM (Fragments Per Kilo bases per Million reads) is the commonest method of estimating gene expression levels, which takes the effects of both sequencing depth and gene length on counting of fragments into consideration. (Mortazavi et al. 2008)

Table 3 Summary of gene expression level analysis(FPKM)

Gene_id	CK1	CK2	CK3	treat1	treat2	treat3
ENSG00000239119	1.98234332	12.23884406	0	7.117899187	0	45.60708762
ENSG00000160307	2.11441291	0	0.371801617	0.451431653	6.455887693	0.136913804
ENSG00000118513	2.522100797	2.030818637	11.69005409	4.763925501	1.576701966	1.688621284
ENSG00000233608	1.419661629	1.385197963	0.516487207	1.097433848	2.353037031	1.854386039

(1) Gene_ID: Ensemble gene ID

(2) Other columns: Gene expression level of different samples represented by FPKM

2.3.2 Gene expression level distribution

Gene expression level distribution among different samples can be displayed by box plots.

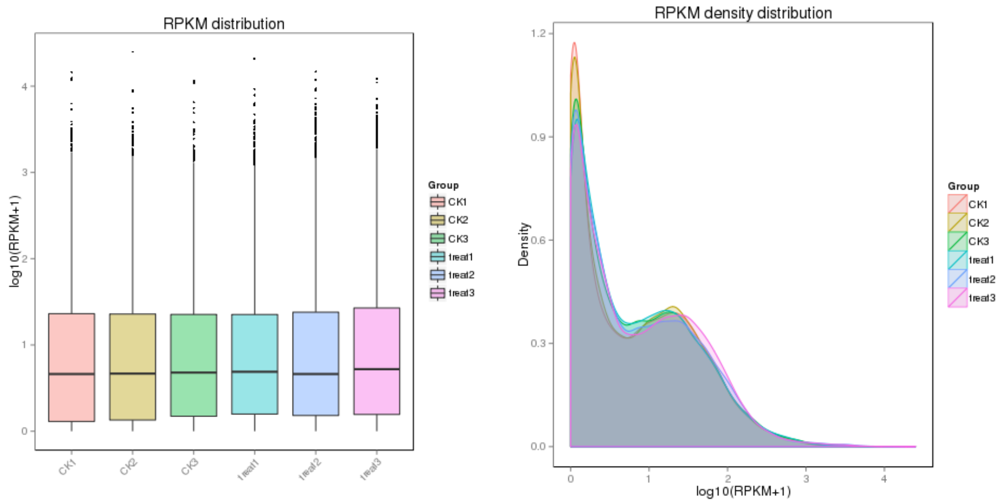


Fig 7 Gene expression level distribution among different samples

X axis represents the name of sample, Y axis indicates the log10(FPKM+1), parameters of box plots are indicated, including maximum, upper quartile, mid-value, lower quartile and minimum.

2.3.3 Correlation analysis

The correlation between samples is an important evaluating indicator to test the reliability of the experiment. The closer the correlation coefficient is to 1, the higher the samples' similarity is.

We can easily get the correlation coefficient among different groups by using correlation coefficient matrix. Low correlation coefficient among groups means those groups are different. High correlation coefficient among intra-group means the design of replicates is logical. correlation coefficient matrix is show in Fig 8.

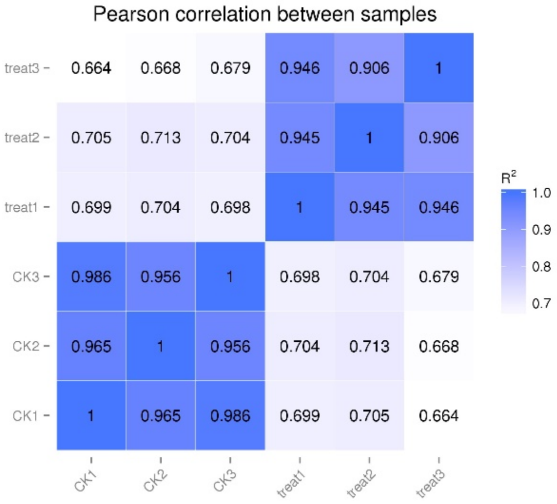


Fig 8 Correlation coefficient matrix

Correlation coefficient matrix among different samples

2.4 Differential expression analysis

Readcount obtained from Gene Expression Analysis is used to do differential expression analysis.

2.4.1 Result of differential expression analysis

For the samples with biological replicates, differential expression analysis of two conditions/groups was performed using the DESeq R package (Anders et al., 2010). It provides statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. So, if the readcount of the i-th gene in j-th sample is K_{ij}, there is: $K_{ij} \sim NB(\mu_{ij},\sigma_{ij}^2)$ And the resulting P values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate.

$$K_{ij} \sim NB(\mu_{ij},\sigma_{ij}^2)$$

For the samples without biological replicates, readcount was adjusted by TMM, then differential expression analysis was performed by using the DEGseq R package.

Table 4 Result of differential expression analysis

Gene Id	CK	treat	log2FoldChange	pval	p-adjusted
ENSG00000000971	67.52166303	9.636614695	2.8088	1.52E-13	1.30E-11
ENSG00000003436	77.76179996	20.84844525	1.8991	1.14E-10	6.64E-09
ENSG00000004142	149.0292613	305.406558	-1.0351	6.18E-10	3.29E-08
ENSG00000004478	216.7455112	572.6836262	-1.4017	9.75E-30	2.38E-27

(1) Gene_ID: Ensemble gene ID

(2) Readcount_Sample1: Adjusted readcount of sample1

(3) Readcount_Sample2: Adjusted readcount of sample2

(4) log2FoldChange: log2(Sample1/Sample2)

(5) Pvalue (pval): Pvalue in statistical tests.

(6) Qvalue (padj): Corrected pvalue. The smaller the p-adjusted, the more significant differentially expressed genes.

2.4.2 Differential expression genes identification

Volcano diagram can visually show the whole distribution of differential expression genes. For the samples with biological replicates, the threshold of differential expression genes is: padj < 0.05. For the samples without biological replicates, the threshold of differential expression genes is: |log2(FoldChange)| > 1 and qvalue < 0.005.

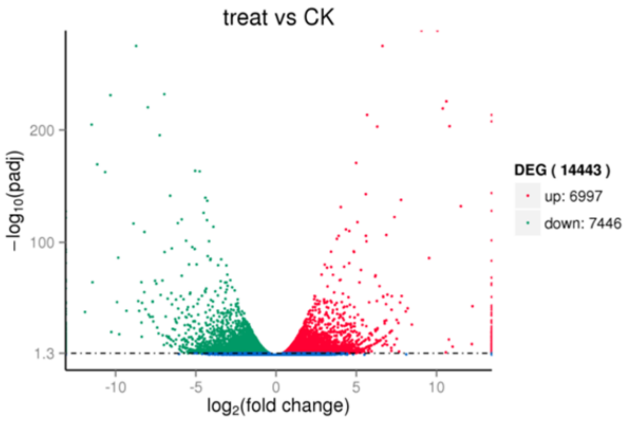


Fig 9 Volcano diagrams of differential expression genes

Horizontal axis for the fold change of genes in different samples. Vertical axis for statistically significant degree of changes in gene expression levels, the smaller the corrected pvalue, the bigger -log10 (corrected pvalue), the more significant the difference. The point represents gene, blue dots indicate no significant difference in genes, red dots indicate upregulated differential expression genes, green dots indicate downregulated differential expression genes.

2.4.3 Venn diagram of differential expression genes

When the comparison group number is from 2 to 5, Venn diagram can be drawn to illustrate the number of common and unique differential expression genes among comparison groups.

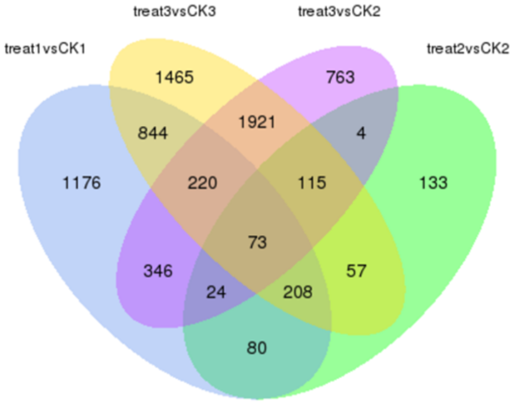


Fig 10 Venn diagram of differential expression genes

The sum of numbers in each big circle is the total number of differential expression genes in this compare group, and the overlapping part is the number of common differential expression genes among comparison groups.

2.4.4 Cluster Analysis of differential expression genes

Cluster Analysis of differential expression genes was used to estimate expression pattern of differential expression genes under different experimental conditions. Through clustering genes in similar expression pattern, unknown functions of transcripts were recognized, with the reason that same kind of transcripts have similar functions or participate in the same metabolic processes or cellular pathways. Hierarchical clustering analysis was carried out with the $\log_{10}$(FPKM+1) of union differential expression genes of all comparison groups under different experimental conditions.

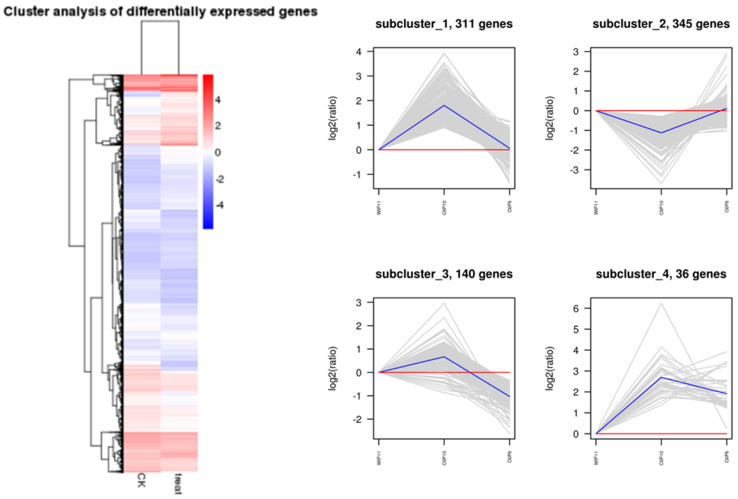


Fig 11 Hierarchical clustering heatmap of differential expression genes

Red represents high expression genes, Blue represents low expression genes. Color descending from red to blue, indicated log10(FPKM+1) from large to small.

2.5 Function analysis of differential expression genes

2.5.1 GO enrichment analysis

Gene Ontology（GO, http://www.geneontology.org/） is an international, standardized classification system described the gene function. After filtered the differentially expressed genes based on experimental targets, we can research the distribution of differentially expressed genes in Gene Ontology to illustrate functional differences between differentially expressed genes. The software of GO enrichment analysis we use is GOseq (Young et al., 2010), based on Wallenius non-central hyper-geometric distribution. Standard methods give biased results on RNA-seq data due to over-detection of differential expression for long and highly expressed transcripts. GOseq was design to compensate for the effect of selection bias in the statistical test of a category's enrichment among differentially expressed genes.

Table 5 GO enrichment analysis results

GO accession	Description	Term type	Over represented p-Value	Corrected p-Value	DEG item	DEG list
GO:0005509	calcium ion binding	molecular_function	1.13E-05	0.04737	29	436
GO:0001558	regulation of cell growth	biological_process	0.0005136	0.38229	4	436
GO:0005520	insulin-like growth factor binding	molecular_function	0.0005136	0.38229	4	436
GO:0008191	metalloendopeptidase inhibitor activity	molecular_function	0.00060024	0.38229	3	436

(1) GO accession: Unique ID in Gene Ontology database

(2) Description: Description of Gene Ontology function

(3) Term type: The classification of Gene Ontology function

(3) Over_represented_pValue: Statistics significance level of GO enrichment analysis

(4) Corrected_pValue: Corrected statistics significance level of GO enrichment analysis,commonly, Corrected_pValue < 0.05 means the GO term is significant enriched

(5) DEG_item: The number of differential expression genes that related to the GO term

(6) DEG_list: The number of differential expression genes

The GO enrichment analysis results are displayed through histogram and Directed Acyclic Graph (DAG). Histogram displays the number of genes that are significantly enriched in each GO term at the biological process, cellular component, molecular function level. The most 30 significantly enriched GO term will be displayed, if not, all the GO term will be displayed. DAG is another way to display the GO Enrichment analysis results. Branches mean hierarchical relationship, and the function ranges become more and more specified from top to bottom. In general, the top 10 results of GO enrichment analysis are chosen as main nodes (showed by box) in directed acyclic graph, and related GO Terms are shown together by hierarchical relationship, the enrichment degree is illustrated by color shades, the darker the shades, the higher the enrichment degree. In our project, DAG of biological process, molecular function and cellular component are drawn respectively.

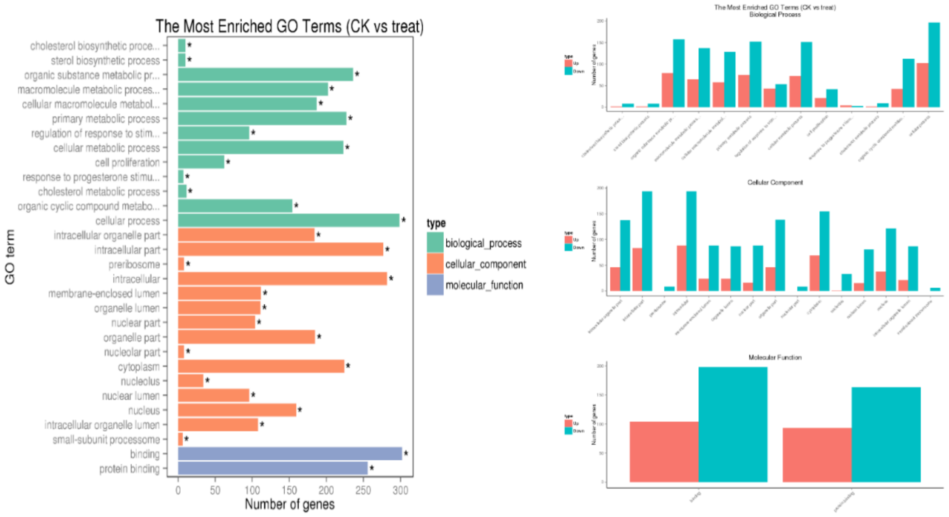


Fig 12 Histogram and DAG of the top enrichment GO term

Fig 12.1: Vertical axis for GO term, horizontal axis for the number of the differential expression genes. the biological process, cellular component, molecular function are signed by different color.”*”means the GO term is significant enriched. Fig 12.2: every node represent one GO term, boxes represent the top 10 enrichment GO term. the enrichment degree is illustrated by color shades, the darker the shades, the higher the enrichment degree. Every node represent the name of the GO term and the Pvalue of enrichment analysis.

2.5.2 KEGG pathway enrichment analysis

We can quickly identify the pathway that differential genes are involved by KEGG Pathway Enrichment Analysis. KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.kegg.jp/) is an important database resource for pathway analysis(Kanehisa,2008). KEGG Pathway as a unit in enriched pathways analysis, significantly enriched pathways are identified by using hypergeometric test.

Table 6 The list of differential genes Enriched pathways

Term	Database	ID	Sample number	Background number	P-Value	Corrected P-Value
Axon guidance	KEGG PATHWAY	hsa04360	17	127	0.000715691	0.171050058
ECM-receptor interaction	KEGG PATHWAY	hsa04512	12	87	0.003322011	0.312197538
Protein digestion and absorption	KEGG PATHWAY	hsa04974	12	89	0.003918798	0.312197538
Focal adhesion	KEGG PATHWAY	hsa04510	18	207	0.029012063	0.999999976

(1) Term: The description of KEGG pathway

(2) Database: Database used in analysis

(3) ID：Database used in analysis

(4) Sample number：The number of differential expression genes in certain pathway

(5) Background number：The number of genes in certain pathway

(6) P-value：Significant levels of Enriched pathways analysis

(7) Corrected P-value：Corrected significant levels. Commonly, Corrected P-value<0.05 means enriched pathways are statistically significantly

KEGG enriched pathway results can be demonstrated by using Scatterplot. In figure 13, KEGG enriched pathway results are evaluated by Rich factor, Qvalue and the number of differential expression genes involved in those pathway. Rich factor refers to the ratio of the number of differentially expressed genes in the pathway and the number of all genes annotated in the pathway. The bigger Rich factor is, the more significant enrichment degree is. Qvalue is the result after multiple hypothesis testing of pvalue, its range is [0,1], and the closer to zero it is, the more significant the enrichment is. Here we display the top 20 significantly enriched in the figure, if the enriched pathway is less than 20, the whole will be show. In addition, we also provide web interface to illustrate differential expression genes in metabolic pathway.

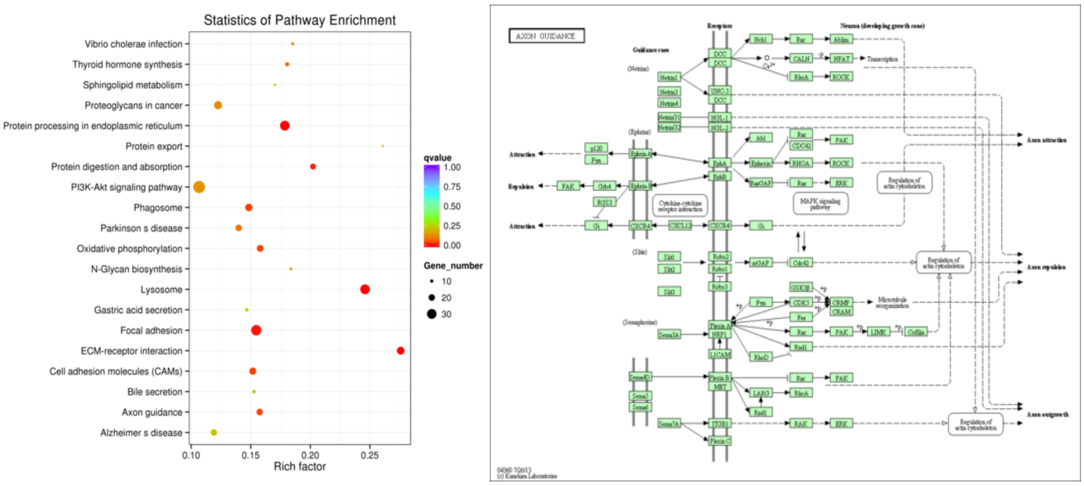


Fig 13 KEGG enrichment scatterplots and metabolic pathway

Fig 13.1: Vertical axis for pathway term, horizontal axis for Rich factor, the size of the spot represents the number of the differential expression genes. The color of the spot represent the range of the Qvalue; Fig 13.2: in KEGG metabolic pathway, red nodes represent the upregulate genes, green nodes represent down regulate genes, yellow nodes represent both up and down regulate genes. When the mouse hovers over the node, more detailed information of differential expression genes, like log2(Fold change) (number in parentheses), will show in the Pop-up box, coloring ditto.

2.5.3 Protein protein interaction analysis

We construct the protein-protein interaction network for differential expression gene by searching STRING protein interaction database (http://string-db.org/).

We provide protein-protein interaction network file, this network file can be imported into Cytoscape software, which can be visualized and edited. The central organizing metaphor of Cytoscape is a network graph, with molecular species represented as nodes and intermolecular interactions represented as links, that is, edges, between nodes • Customize network data display using powerful visual styles. • View a superposition of gene expression ratios and p-values on the network. Expression data can be mapped to node color, label, border thickness, or border color, etc. according to user-configurable colors and visualization schemes. • Layout networks in two dimensions. A variety of layout algorithms are available, including cyclic and spring-embedded layouts. • Zoom in/out and pan for browsing the network. • Use the network manager to easily organize multiple networks. And this structure can be saved in a session file. • Use the bird's eye view to easily navigate large networks. • Easily navigate large networks (100,000+ nodes and edges) by efficient rendering engine.

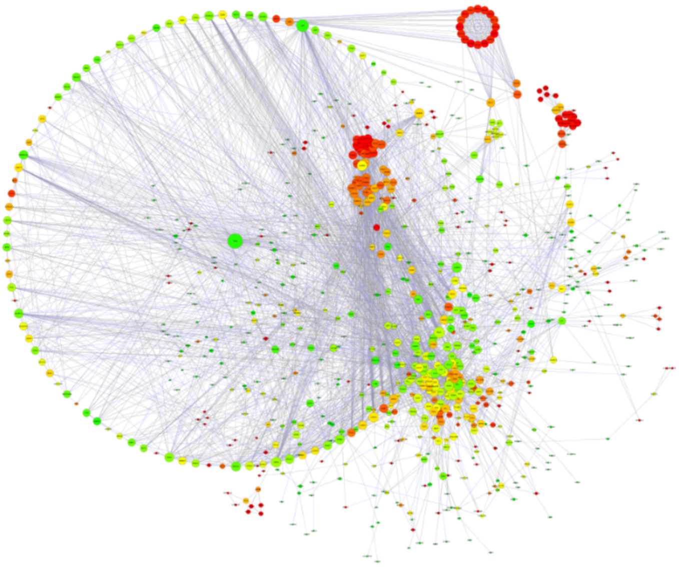


Fig 14 Visualization of Cytoscape

2.5.4 Function annotation of transcription factor

Gene expression regulation at transcriptional level is the critical step of gene expression. Transcription factors band to the upstream sequences of the gene, then activate or inhibit gene expression. TFCat is a curated catalog of mouse and human transcription factors (TF) based on a reliable core collection of annotations obtained by expert review of the scientific literature. Annotated genes are assigned to a functional category and confidence level. We use the differential expression TF to search the TFCat, TFCat provides the annotation of the TF and the corresponding reference (PubMed ID).

Table 7 Results of TF analysis

Gene_ID	Gene_Symbol	treat	CK	log2FoldChange	Description	Function	Evidence	PUBMED_ID
ENSG00000006704	GTF2IRD1	42.01614	17.66971	1.2497	general transcription factor II I repeat domain-containing 1	DNA Binding; Transactivation	Strong	14645227
ENSG00000019549	SNAI2	15.44245	49.29195	-1.6744	snail homolog 2 (Drosophila)	DNA Binding	Strong	10518215
ENSG00000073282	TP63	1.510145	74.02954	-5.6153	transformation related protein 63	Transactivation	Strong	9774969
ENSG00000090447	TFAP4	28.22998	84.84078	-1.5875	transcription factor AP4	DNA Binding; Transactivation	Not Selected	2833704

(1) Gene_ID： Ensembl gene ID

(2) Gene_Symbol: Gene name

(3,4) Sample： Read count of the TF

(5) log2FoldChange： log2 fold change between different sample

(6) Description： The description of gene

(7) Function： The function of TF

(8) Evidence： The evaluation of evidence

(9) PUBMED ID： Reference ID in PubMed

2.5.5 Function annotation of oncogene

Proto-oncogene is normal gene that involve in cell development, cell division and cell differentiation. It can turn into oncogene when gene sequence is changed. Usually, the expression of some specific oncogene will be upregulated in tumor or malignant cells lines. To investigate the differential expression of oncogene in different sample will be helpful for exploring the mechanism of disease development. COSMIC is an online database of somatically acquired mutations found in human cancer. COSMIC, an acronym of Catalogue of Somatic Mutations In Cancer, curates data from papers in the scientific literature and large scale experimental screens from the Cancer Genome Project at the Sanger Institute. The database is freely available without restriction via its website. By searching the database using differential expression genes, we can identify oncogene and its annotation.

Table 8 The result of oncogene analysis

Gene_ID	Gene_Symbol	treat	CK	log2FoldChange	Description	Tumor Type(Somatic)	Tumour Types(Germline)
ENSG00000019582	CD74	10.52230335	412.031453	-5.2912	"CD74 molecule, major histocompatibility complex, class II invariant chain"	NSCLC	N.A.
ENSG00000039068	CDH1	0	18.90004505	-5.1404	"cadherin 1, type 1, E-cadherin (epithelial) (ECAD)"	"lobular breast, gastric"	gastric
ENSG00000051108	HERPUD1	81.52349377	40.07751934	1.0244	"homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1"	prostate	N.A.
ENSG00000092820	EZR	745.3054634	1492.763253	-1.0021	ezrin	NSCLC	N.A.

(1) Gene_ID： Ensembl gene ID

(2) Gene_Symbol： Gene name

(3,4) Sample： The expression level of oncogene

(5) log2FoldChange： log2 Fold Change between samples

(6) Description： The description of gene

(7) Tumor Type (Somatic)：The type of Somatic cancer

(8) Tumor Type (Germline)：The type of germline cancer

2.6 Alternative splicing analysis

Alternative splicing (AS) is a regulated process during gene expression that results in a single gene coding for multiple proteins. In this process, particular exons of a gene may be included within or excluded from the final, processed messenger RNA (mRNA) produced from that gene. Consequently, the proteins translated from alternatively spliced mRNAs will contain differences in their amino acid sequence and, often, in their biological functions. Notably, alternative splicing allows the human genome to direct the synthesis of many more proteins than would be expected from its 20,000 protein-coding genes.

MISO (Mixture of Isoforms) is a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data, and identifies differentially regulated isoforms or exons across samples. By modeling the generative process by which reads are produced from isoforms in RNA-Seq, the MISO model uses Bayesian inference to compute the probability that a read originated from a particular isoform. MISO uses the inferred assignment of reads to isoforms to quantitate the abundances of the underlying set of alternative mRNA isoforms. Confidence intervals over estimates can be obtained, which quantify the reliability of the estimates. MISO estimates of isoform expression by using Ψ values (Ψ values, for “Percent Spliced In” or “Percent Spliced Isoform”) and differential isoform expression. Confidence intervals for expression estimates and quantitative measures of differential expression were represent by “Bayes factors”.

2.6.1 Quantification of alternative splicing event

In the most common type of alternative splicing in mammals, an exon is included or excluded from the mature mRNA; ‘percentage spliced in’ (PSI or Ψ) denotes the fraction of mRNAs that represent the inclusion isoform. Reads aligning to the alternative exon or to its junctions with adjacent constitutive exons provide support for the inclusion isoform, whereas reads aligning to the junction between the adjacent constitutive exons support the exclusion isoform; the relative read density of these two sets forms the standard estimate of Ψ. This estimate ignores reads that align to the bodies of the flanking constitutive exons, which could have derived from either isoform. Nevertheless, these constitutive reads contain latent information about the splicing of the alternative exon, as higher expression of the exclusion isoform will generally increase the density of reads in the flanking exons relative to the alternative exon, and lower expression of the exclusion isoform will decrease this ratio of densities. MISO captures this, as well as the information in the lengths of library inserts in paired-end data, by recasting the analysis of isoforms as a Bayesian inference problem. Generative process for MISO model. White, alternatively spliced exon; gray and black, flanking constitutive exons. RNAseq reads aligning to the alternative exon body (white) or to splice junctions involving this exon support the inclusive isoform, whereas reads joining the two constitutive exons (blackgray exon junction) support the exclusive isoform. Reads aligning to the constitutive exons are common to both isoforms.

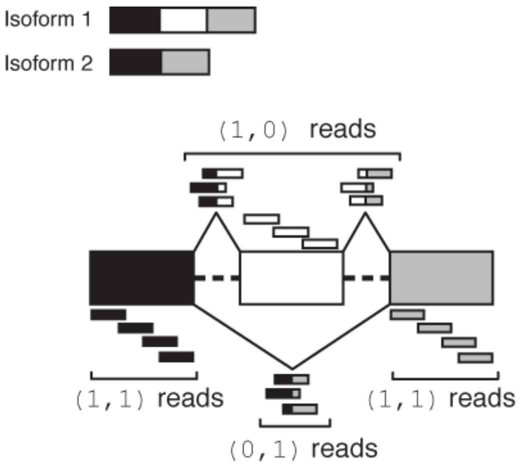


Table 9 Result of quantification AS event by MISO

event_name	ensg_id	gsymbol	chrom	strand	miso_posterior_mean	ci_low	ci_high	counts	assigned_counts
chr18:66377258:66377381:-@chr18:66368982:66369027:-@chr18:66367642:66367722:-	ENSG00000166479	TMX3	chr18	-	0.67	0.34	0.96	(0,0):117,(1,0):4	'0:4'
chr18:56001050:56001124:+@chr18:56010138:56010335:+@chr18:56016769:56016846:+	ENSG00000049759	NEDD4L	chr18	+	0.45	0.03	0.92	(0,0):403,(1,0):22	'0:22'
chr18:46783380:46783474:-@chr18:46735921:46736085:-@chr18:46690055:46690157:-	ENSG00000141627	DYM	chr18	-	0.16	0.01	0.53	(0,0):145,(0,1):1	'0:0,1:1'
chr18:43795777:43796561:+@chr18:43819971:43820153:+@chr18:43833663:43833786:+	ENSG00000152242	C18orf25	chr18	+	0.16	0.08	0.25	(0,0):91,(0,1):17,(1,0):11,(1,1):171	'0:31,1:168

1) Event_name: AS event name (exon name structure "chromosome number: start site: stop site: sense and antsense strand” @ is linker);

(2) Gene_ID: Ensemble gene ID;

(3) gsymbol: Gene symbol;

(4) Chrom: Chromosome that AS event happen;

(5) Strand: Strand direction;

(6) MISO_posterior_mean: Estimation of Ψ;

(7,8) ci_low, ci_high: Confidence interval of Psi;

(9) Counts: Counts: Reads number for supporting isoform. “(0,1):X” means reads number for supporting isoform2 is X, “(1,0):Y”means reads number for supporting isoform1 is Y, (1,1):Z means reads number for supporting isoform1 and isoform2 is Z.”(0,0):W” means reads number for not being taken account of is W;

(10) Assigned_counts: Reads number of every isoform;

2.6.2 Detection of differentially expressed isoforms

Differential splicing of alternative exons entails a difference in Ψ values, ΔΨ, and can be evaluated statistically using the Bayes factor (BF), which quantifies the odds of differential regulation occurring. MISO is used to calculate the posterior probability distributions of Ψ and ΔΨ for the two samples. The latter distribution is used to calculate the BF, defined as the ratio of the posterior probability of the alternative hypothesis, $\Delta\Psi \neq 0$, to that of the null hypothesis, $\Delta\Psi = 0$ (Online Methods); thus, higher values of the BF indicate increased confidence in differential regulation. Using MISO to get PSI of AS event among different samples, we define differential event based on (ΔΨ>=0.2， BF>=10)

Table 10 Result of differential analysis of AS event

event_name	Gene_ID	gsymbol	CK	treat	diff	bayes_factor
chr18:32870236:32870355:+@chr18:32870974:32871196:+@chr18:32885940:32890730:+	NA	NA	0.56	0.81	-0.25	246004272
chr3:169490853:169491250:+@chr3:169491819:169491885:+@chr3:169492053:169492349:+	ENSG00000085274	MYNN	0.3	0.55	-0.25	1093589985
chr3:15100876:15100982:-@chr3:15094798:15094974:-@chr3:15090019:15094140:-	NA	NA	0.52	0.11	0.41	1971.93
chr3:49507565:49507866:+@chr3:49524687:49524848:+@chr3:49526059:49526178:+	ENSG00000173402	DAG1	0.66	0.23	0.43	27.01

(1) Event_name: AS event name;

(2) Gene_ID: Ensemble gene ID;

(3) Gsymbol: gene symbol;

(4,5) Sample_posterior_mean: Estimation of Ψ;

(6) Diff: Difference value of Ψ;

(7) Bayes factor: Indicates the degree of difference, the greater the value the higher the degree of difference;

2.6.3 Visualization of differential analysis result

Using sashimi_plot to visualize differential analysis result. sashimi_plot Characteristics of Sashimi_plot as follows:

(1) The RNA-Seq read densities along exons are shown as histograms, color-coded by the sample. The RNA-Seq densities are aligned to the isoforms drawn at the bottom of the plot, which are automatically read from the GFF annotation of the events given as input.

(2) Junction reads are visualized as arcs connecting the pair of exons that the junction borders. The thickness of the arc is in proportion to the number of junction reads present in the sample, but the actual number of junction reads can be optionally plotted too (as in the main example.)

(3) MISO expression estimates are (optionally) shown on the right, including the full posterior distribution (as black histograms) over Ψ, with the Ψ estimate drawn as a thick red line and lower and upper 95% confidence intervals plotted as dotted grey lines. The actual value of Ψ along with the value of each confidence interval bound is shown to the right of the histograms.

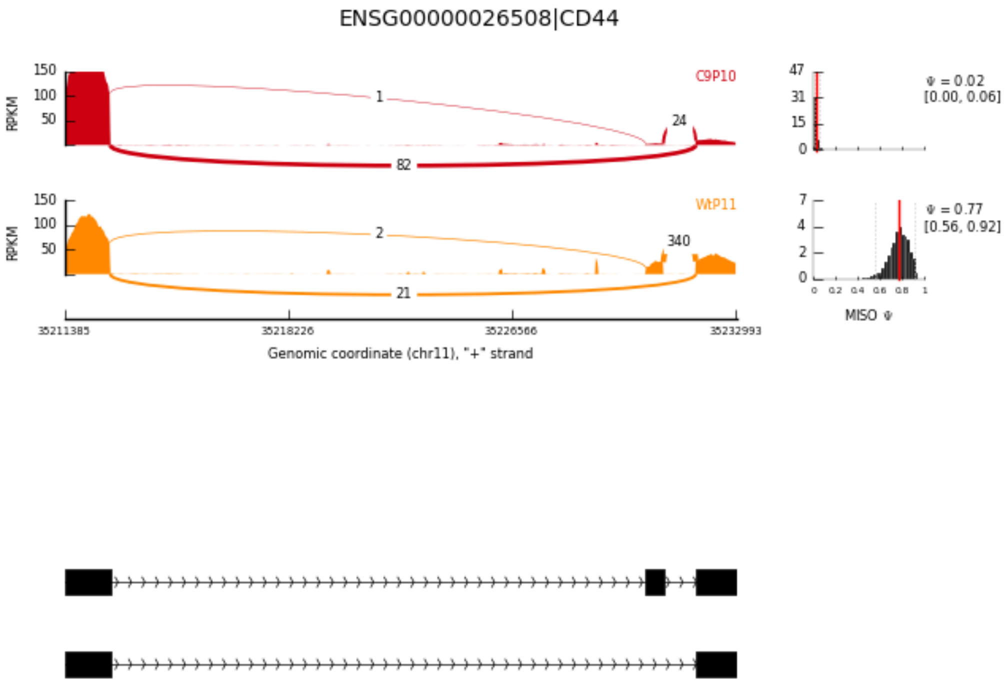


Fig 15 Visualization of differential analysis result

2.7 SNP/InDel analysis

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

Analysis tools, such as Samtools and Picard, are used to sort the reads according to the genome coordinates, followed by screening out repeated reads. Finally, GATK3 is used to carry out SNP calling and INDEL calling. After filtering, results such as those shown in the following table are obtained, in which INDEL and SNPs share the same columns.

2.7.1 SNP detection

Table 11 The result of SNP analysis (Just show one sample)

Chrom	POS	ID	REF	ALT	QUAL	FILTER	TYPE	DP	Allele Depth	genotype
chr1	14653	rs375086259	C	T	2130.92	PASS	SNP	114	78,36	C/T
chr1	14677	rs201327123	G	A	980.92	PASS	SNP	143	117,26	G/A
chr1	19918	.	G	C	31.78	PASS	SNP	13	10,3	G/C
chr1	20144	rs143346096	G	A	442.92	PASS	SNP	40	31,9	G/A

（1）Chrom：Chromosome ID of SNPs.

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

（3）ID：The ID of this variation in dbSNP.

（4）REF: Reference genotype.

（5）ALT：Alternative genotype.

（6）QUAL：Quality value of SNP variation.

（7）FILTER：Whether through quality control.

（8）TYPE：variation type(SNP or INDEL).

（9）DP：The sequence depth of the site (representing the number of reads to support the site).

（10）Allele Depth：The number of reads supporting either the reference genotype or SNP genotype.

（11）genotype: The genotype of the locus.

2.7.2 SNP annotation result

We use ANNOVAR(Wang K et al.) to annotate SNP site, which includes annotation information on dbSNP, the 1000-genome project and other published databases. Annotation contains the variation's position, type, etc.

Table 12 The annotation result of SNPs

Chr	Start	End	Ref	Alt	cytoBand	Func.refGene	Gene.refGene	GeneDetail .refGene	ExonicFunc. refGene	AAChange.r efGene	snp138	cosmic70	gwasCatalog
chr1	14653	14653	C	T	1p36.33	ncRNA_exonic	ENSG0000022 7232	.	.	.	rs62635297	.	.
chr1	14677	14677	G	A	1p36.33	ncRNA_exonic	ENSG0000022 7232	.	.	.	rs112391680	.	.
chr1	19918	19918	G	C	1p36.33	ncRNA_intronic	ENSG0000022 7232	.	.	.	.	.	.
chr1	20144	20144	G	A	1p36.33	ncRNA_intronic	ENSG0000022 7232	.	.	.	rs77331792	.	.

(1) Chr：Chromosome ID of SNPs.

(2) Start：The start site of variation on chromosome.

(3) End：The end site of variation on chromosome.

(4) REF：Genotype on reference sequence.

(5) ALT：Alternative genotype.

(6) CytoBand:snp: Chromosome band.

(7) Func.refGene: Tells whether the variant hit exons or hit intergenic regions, or hit introns, or hit a non-coding RNA genes. (exonic, splicing, UTR5, UTR3, intronic, ncRNA_exonic, ncRNA_intronic, ncRNA_UTR3, ncRNA_UTR5, ncRNA _splicing, upstream, downstream, intergenic).

(8) Gene.refGene : Gene symbol.

(9) GeneDetail.refGene: Description of variations in UTR, splicing, ncRNA, splicing or intergenic.

(10) ExonicFunc.refGene: The amino acid changes as a result of the exonic variant.(synonymous_SNV, missense_SNV, stopgain_SNV, stopgloss_SNV or unknown).

(11) AAChange.refGene: When 'Func' equals 'exonic' or 'exonic;splicing', this value gives the change of amino acid in each related transcript. For example, AIM1L:NM_001039775:exon2:c.C2768T:p.P923L, AIM1L is gene name containing this variation; NM_001039775 is ID of transcript; exon2 means the variation is on the second exon of the transcript; c.C2768T means the 2, 768 base on cDNA is changed from C to T due to this variation; p.P923L means the 923 amino acid on protein is changed from Pro to Leu due to this variation.

(12) snp138: The ID of this variation in dbSNP(version 138).

(13) cosmic70: COSMIC annotation information(version 70).

(14) gwasCatalog: Tells whether this variation has been identified by published Genome-Wide Association Studies (GWAS), collected in the Catalog of Published Genome-Wide Association Studies at the National Human Genome Research Institute (NHGRI). It lists the diseases related to this variation. '.' means this variation has not been reported by published GWAS study.

2.7.3 InDel detection

InDel refers to the insertion or deletion of nucleotides in sample genome. Normally, the length of InDel varies from 1bp to 50bp and the InDel occurrence would be reduced rapidly by increasing InDel length whose theoretical distributions correspond to Power-law. In mammal genomes, a single base InDel is the most frequent and deletion occurs more than insertion, which means there is a deletion preference in genome. According to researches, InDel contributes more than the replacement to the evolution of the genome. Below is a statistical analysis of the results for InDel:

Table 13 The result of InDel analysis (Just show one sample)

Chrom	POS	ID	REF	ALT	QUAL	FILTER	TYPE	DP	Allele Depth	genotype
chr1	761957	rs59038458	A	AT	1542.88	PASS	INDEL	19	7,12	A/AT
chr1	778302	rs112119688	C	CCT	1076.13	PASS	INDEL	19	7,12	C/CCT
chr1	789513	rs200509509	GA	G	922.88	PASS	INDEL	49	37,12	GA/G
chr1	878516	rs201043644	GA	G	375.15	PASS	INDEL	46	38,8	GA/G

（1）Chrom： Chromosome ID of InDels.

（2）POS： Position of InDels on corresponding chromosome.

（3）ID： The ID of this variation in dbSNP.

（4）REF: Reference genotype.

（5）ALT： Alternative genotype.

（6）QUAL： Quality value of InDel variation.

（7）FILTER： Whether through quality control.

（8）TYPE: variation type (SNP or InDel).

（9）DP: The sequence depth of the site (representing the number of reads to support the site).

（10）Allele Depth: The number of reads supporting either the reference genotype or InDel genotype.

（11）genotype: The genotype of the locus.

2.7.4 InDel annotation result

We use ANNOVAR(Wang K et al.) to annotate InDel site, which includes annotation information on dbSNP, the 1000-genome project and other published databases. Annotation contains the variation's position, type, etc.

Table 14 The annotation result of InDels

Chr	Start	End	Ref	Alt	cytoBand	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene	AAChange.refGene	snp138	cosmic70	gwasCatalog
chr1	761957	761957	-	T	1p36.33	ncRNA_exonic	ENSG00000225880	.	.	.	rs59038458	.	.
chr1	778302	778302	-	CT	1p36.33	ncRNA_intronic	ENSG00000228794	.	.	.	rs112119688	.	.
chr1	789514	789514	A	-	1p36.33	ncRNA_exonic	ENSG00000228794	.	.	.	rs200509509	.	.
chr1	878517	878517	A	-	1p36.33	intronic	ENSG00000187634	.	.	.	rs201043644	.	.

(1) Chr: Chromosome ID of SNPs.

(2) Start: The start site of variation on chromosome.

(3) End: The end site of variation on chromosome.

(4) REF: Genotype on reference sequence.

(5) ALT: Alternative genotype.

(6) CytoBand: snp: Chromosome band.

(7) Func.refGene: Tells whether the variant hit exons or hit intergenic regions, or hit introns, or hit a non-coding RNA genes. (exonic, splicing, UTR5, UTR3, intronic, ncRNA_exonic, ncRNA_intronic, ncRNA_UTR3, ncRNA_UTR5, ncRNA _splicing, upstream, downstream, intergenic).

(8) Gene.refGene: Gene symbol.

(9) GeneDetail.refGene: Description of variations in UTR, splicing, ncRNA, splicing or intergenic.

(10) ExonicFunc.refGene: The amino acid changes as a result of the exonic variant.(synonymous_SNV, missense_SNV, stopgain_SNV, stopgloss_SNV or unknown).

(11) AAChange.refGene: When 'Func' equals 'exonic' or 'exonic;splicing', this value gives the change of amino acid in each related transcript. For example, AIM1L:NM_001039775:exon2:c.C2768T:p.P923L, AIM1L is gene name containing this variation; NM_001039775 is ID of transcript; exon2 means the variation is on the second exon of the transcript; c.C2768T means the 2, 768 base on cDNA is changed from C to T due to this variation; p.P923L means the 923 amino acid on protein is changed from Pro to Leu due to this variation.

(12) snp138: The ID of this variation in dbSNP(version 138).

(13) cosmic70: COSMIC annotation information(version 70).

(14) gwasCatalog: Tells whether this variation has been identified by published Genome-Wide Association Studies (GWAS), collected in the Catalog of Published Genome-Wide Association Studies at the National Human Genome Research Institute (NHGRI). It lists the diseases related to this variation. '.' means this variation has not been reported by published GWAS study.

2.7.5 Sequence analysis of gene frameshift mutations

InDel in coding sequence or splicing site may affect the protein translation. Frameshift mutation (non-trebling InDel) may change the entire ORF, so that frameshift mutation has greater impact on gene function, compared to non-frameshift mutation.

```
>line203 ENST00000399794 WILDTYPE
```

```
MVEADRPGKLFIGGLNTETNEKALETVFGKYGRIVEVLLIKDRETNKSRGFAFVTFE
SPADAKDAARDMNGKSLDGKAIKVEQATKPSFERGRHGPPPPP
RSRGPPRGFGAGRGGSGGTRGPPSRGGHMDDGGYSMNFNMSSSRGPLPVKRGPPPRS
GGPSPKRSAPSGLVRSSSGMGGRAPLSRGRDSYGGPPRREPLP
SRRDVYLSPRDDGYSTKDSYSSRDYPSSRDTRDYAPPPRDYTYRDYGHSSSRDDYPS
RGYGDRDGYGRDRDYSDHPSGGSYRDSYESYGNSRSAPLTRGP
PPSYGGSSRYDDYSSSRDGYGGSRDSYSSSRSDLYSSCDRVGRQERGLPPSVERGYP
SSRDSYSSSSRGAPRGAGPGGSRSDRGGGRSRY*
```

2.8 Fusion gene analysis

A fusion gene is a hybrid gene formed from two previously separate genes. It can occur as a result of: translocation, interstitial deletion, or chromosomal inversion. The first fusion gene was discovered in cancer cells in hematological system. BCR-ABL gene fusion in chronic myelogenous leukemia (CML)is one of the most famous gene fusion event. BCR-ABL fusion gene is the target of imatinib in CML therapy. High throughput RNA-Seq has facilitated the research on fusion gene. As more and more fusion genes were discovered, researchers found that except cancer cells in hematological system, fusion genes also can be found in other tumor cells. Until now hundreds of fusion genes have been identified playing important rules during tumorigenesis. To identify fusion gene would facilitate the research on disease especially cancer, also important for Clinical molecular typing and new tumor drug designing.

For human and mouse data, we use SOAPfuse and Tophat fusion to identify fusion genes in tumor samples, then visualize results. SOAPfuse identify fusion transcripts from paired-end RNA-Seq data. SOAPfuse applies an improved partial exhaustion algorithm to construct a library of fusion junction sequences, which can be used to efficiently identify fusion events, and employs a series of filters to nominate high-confidence fusion transcripts. Compared with other released tools, SOAPfuse achieves higher detection efficiency and consumed less computing resources. It also can predict fusion point and visualize results.

2.8.1 Fusion gene detection

Fusion genes were detected in tumor sample by SOAPfuse (Tophat fusion). The number of fusion genes and their fusion partner were counted. Fusion partner can fuse various of version genes Because of different fusion point. The details of fusion genes statistical result can be found in analysis report. Here only show fusion genes in one single sample.

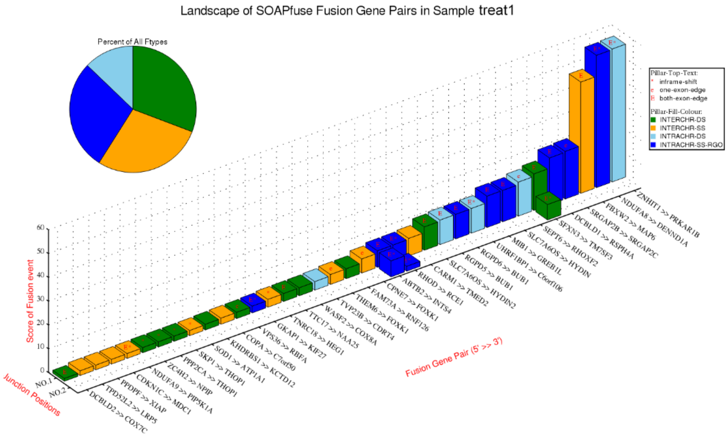


Fig 16 Result of fusion gene in one single sample show in 3D

x-axis show the position of different fusion point ; y-axis show the two fusion partner; z-axis show the score of the gene fusion event, this score is in proportion to reads counts. The different fill color represents different gene fusion event (Classification according to whether fusion partner come from the same gene and DNA sense or antsense). '*' means the downstream fusion partner is frameshift mutation, 'e' means that only one fusion point at the flank of exon, 'E' means both of the two fusion point are at the flank of exon.

2.8.2 Fusion point analysis

Using SOAPfuse to draw the fusion point and fusion process of one single fusion gene.

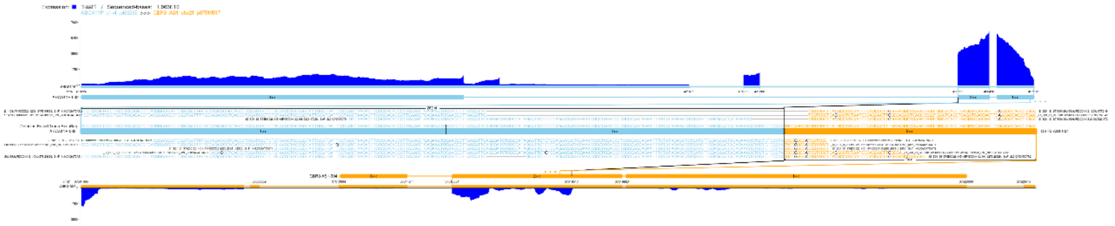


Fig 17 The fusion point and fusion process of one single fusion gene

The figure illustrate the information of fusion point on transcript and genome, and the corresponding reads. Sky blue and red are corresponding to the two gene partners respectively. Dark blue on upper and lower edge represents the distribution of reads align to the gene, the middle part represents the sequence information of fusion gene and reads align to fusion gene.

2.8.3 Fusion gene domain analysis

Fusion gene may yield new gene production with new or alternative function. That alternative production may cause cancer. In addition, some proto-oncogenes may fuse with a promoter, then activate the proto-oncogene. The later one is commonly found in lymphoma. Fusion gene domain analysis and function prediction will be helpful for us to explore the mechanism of tumorigenesis, development and tumor metastasis. More detail result can be found in analysis report. Domain analysis by domain searching in pfam database, the results are show below:

Table 15 The annotation of fusion gene in pfam database

Fusion Gene_ID	Alignment start	Alignment end	Pfam acc	Pfam name	Pfam description	Pfam length	Bit score	E-value
ABCA11P-002/275/CBR3-AS1-004/3148_1	1	40	PF01352	KRAB	KRAB box	87	87.4	2.90E-25
ANGEL1-001/1493/GLTSCR2-006/392_1	132	174	PF03372	Exo_endo_phos	Endonuclease /Exonuclease/ phosphatase family	66	36.2	5.80E-09
ANGEL1-001/1493/GLTSCR2-006/392_2	1	22	PF13900	GVQW	Putative binding domain	66	91.6	2.10E-26
ANGEL1-001/1493/GLTSCR2-006/392_2	24	48	PF13900	GVQW	Putative binding domain	66	91.6	2.10E-26

(1) Fusion Gene_ID： The gene ID of fusion gene identify by SOAPfuse (including fusion gene and the corresponding position)

(2) Alignment start： The start site of aligned domain

(3) Alignment end：The end site of aligned domain

(4) Pfam acc： Aligned domain ID in Pfam database

(5) Pfam name： Aligned domain name in Pfam database

(6) Pfam description: The description of aligned domain

(7) Pfam length： The length of aligned domain

(8) Bit score： The score of alignment

(9) E-value: Alignment value

2.8.4 The distribution of fusion gene on genome

The Circos displays the distribution of all the fusion genes identified in this project.

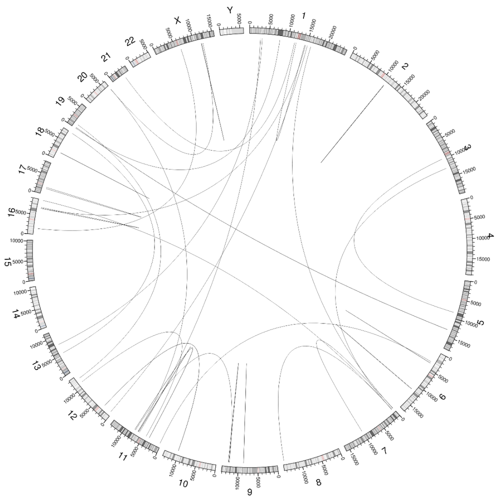


Fig 18 Circos displays all the fusion genes found in all the samples.

In the figure, the circle is constituted by 24 chromosomes, each line in the circle represents a fusion gene. The end of each line represent the fusion point.

3 Appendix

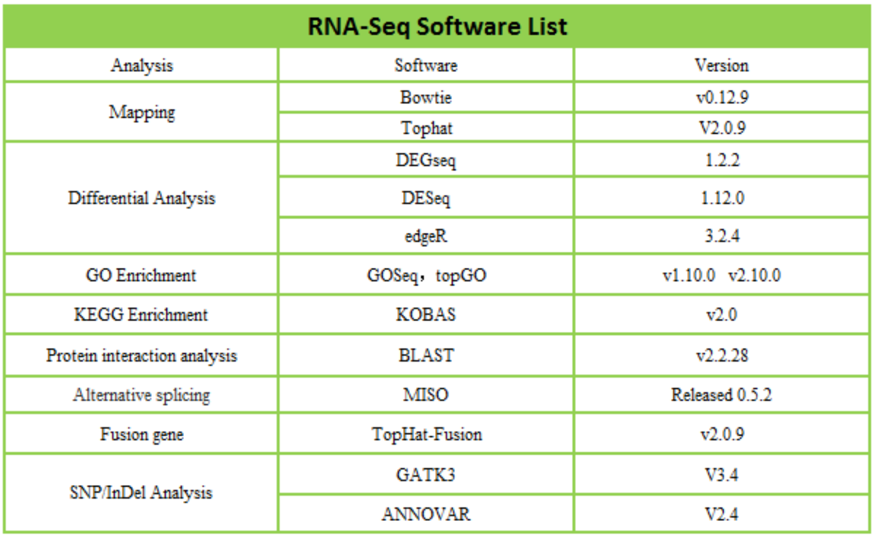
3.1 File catalog

We provide results file catalog, results can be viewed by catalog, clicking catalog list, the corresponding file will be viewed (notice: making sure report file and results file are in the same folder.

```
├── 1. QC：Data Quality Control
│   ├── 1.1.   OriginalData: Raw Data（fastq format）
│   ├── 1.2.   ErrorRate: Error Rate
│   ├── 1.3.   GC：GC Content Distribution
│   ├── 1.4.   ReadsClassification:  Data Filtering
│   └── 1.5.   DataTable: Data Quality Control Summary (raw, clean, Q20, Q30,
GC etc)
├── 2. Mapping: Mapping to a Reference Genome
│   ├── 2.1.   MapStat: Overview of Mapping Status
│   ├── 2.2.   MapReg: Mapped Regions in Reference Genome (exons, introns, or
intergenic regions)
│   ├── 2.3.   ChrDen:Distribution of Mapped Reads in Chromosomes
│   └── 2.4.   IGV: Visualization of Mapping Status of Reads using IGV
├── 3. GeneExprQuatification:   Expression Quantification
├── 4. DiffExprAnalysis: Gene Expression Difference Analysis
│   ├── 4.1 Correlation
│   ├── 4.2.   DEGsList: List of Differentially Expressed Genes (all,up regulated,
               down -regulated )
│   ├── 4.3.   DEGs Filter：Volcano plot
│   ├── 4.4.   Venn Diagram:   The Venn Diagram
│   │   └── treat1vsCK1_treat2vsCK2_treat3vsCK3
│   └── 4.5.   DEGcluster: Cluster Analysis of Gene Expression Differences
│       └── Subcluster
├── 5.DEG_GOEnrichment
│   ├── 5.1.DEG_GOList
│   ├── 5.2.DAG
│   └── 5.3.BAR
├── 6.DEG_KEGGEnrichment
│   ├── 6.1.DEG_KEGGList
│   ├── 6.2.DEG_KEGGScat
│   └── 6.3.DEG_KEGGPath
│       ├── ALL
│         ├── DOWN
│         └── UP
├── 7. DEG_PPI:   Protein-Protein Interaction Network Analysis
├── 8. Function Annotation
│    ├── 8.1.    TF Annotation:   The Transcription Factor Analysis Results
│    └── 8.2.    cancer Annotation:    Function Annotation of Oncogene
├── 9. MISO:   Alternative Splicing Analysis
│    ├── 9.1event_analysis
│    ├── 9.2different_analysis
│    └── 9.3sashimi_plot
│         ├── treat1_vs_CK1
│         ├── treat2_vs_CK2
│         └── treat3_vs_CK3
├── 10.      SNP:   SNP & InDel
└── 11.      Gene Fusion:   Fusion Gene Analysis
     ├── 11.1FusionList
     ├── 11.2FusonStats
     ├── 11.3FusionFigures
     └── 11.4FusionCircos
```

3.2 Software catalog

A list of software used in our pipelines is presented.



RNA-Seq Software List		
Analysis	Software	Version
Mapping	Bowtie	v0.12.9
	Tophat	V2.0.9
Differential Analysis	DEGseq	1.2.2
	DESeq	1.12.0
	edgeR	3.2.4
GO Enrichment	GOSeq，topGO	v1.10.0 v2.10.0
KEGG Enrichment	KOBAS	v2.0
Protein interaction analysis	BLAST	v2.2.28
Alternative splicing	MISO	Released 0.5.2
Fusion gene	TopHat-Fusion	v2.0.9
SNP/InDel Analysis	GATK3	V3.4
	ANNOVAR	V2.4

4 Reference

1.Cock P J A, Fields C J, Goto N, 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. (FASTQ)

2.Erlich Y, Mitra PP, delaBastide M, et al. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing.Nat Methods. 2008 Aug;5(8):679-82.(sequencing error rate distribution)

3.Jiang L, Schlesinger F, Davis CA, et al. (2011). Synthetic spike-in standards for RNA-seq experiments.Genome Res. 2011 Sep;21(9):1543-51. (sequencing error rate distribution)

4.Altschul S F, Madden T L, Schäffer A A, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402. (BLAST)

5.Finn R D, Tate J, Mistry J, et al. (2008). The Pfam protein families database. Nucleic Acids Res 36, D281–D288. (Pfam)

6.Yarden Katz, Eric T. Wang, Edoardo M, et al. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods 7, 1009-1015 (MISO)

7.Götz S, García-Gómez J M, Terol J, et al. (2008).High-throughput functional annotation and data mining with the Blast2GO suite.Nucleic Acids Research 36, 3420-3435. (BLAST2go)

8.Chepelev I, Wei G, Tang Q, et al. (2009). Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic acids research 37, e106-e106. (SNP)

9.Cingolani, P., et. al. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in Genetics 3.(SnpSift)

10.McKenna A, Hanna M, Banks E, et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. (GATK)

11.Trapnell C, Williams B A, Pertea G, et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech 28, 511–515. (FPKM)

12.Dillies M A, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis[J]. Briefings in bioinformatics, 2013, 14(6): 671-683. (normalization methods)

13.Anders S, Huber W. (2010).Differential expression analysis for sequence count data.Genome Biology,doi:10.1186/gb-2010-11-10-r106. (DESeq)

14.Wang L, Feng Z, Wang X, et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-138. (DEGseq)

15.Young M D, Wakefield M J, Smyth G K, et al. (2010).Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology, doi:10.1186/gb-2010-11-2-r14. (GOseq)

16.Kanehisa M, Araki M, Goto S, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids research 36:D480-D484. (KEGG)

17.Mao X, Cai T, Olyarchuk J G, et al. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.Bioinformatics 21, 3787–3793. (KOBAS)

18.Shannon P, Markiel A, Ozier O, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498-2504. (Cytoscape)

19.Fulton DL1, Sundararajan S, Badis G, et al. (2009). TFCat: the curated catalog of mouse and human transcription factors. Genome Biol. 10(3):R29.(TFCat)

20.Jia W, Qiu K, He M, et al. (2013): SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome biology 14:R12.(SOAPfuse)