BGISEQRNArefDemo [BGISEQ Transcriptome Resequencing Report]

2017/9/18

Table of Contents

Results

1 Abstract

In our project, we sequenced 6 samples on BGISEQ-500 Platform in total and generated about 2.34 Gb per sample. The average genome mapping rate is 96.30% and the average gene mapping rate is 70.78%. 22,597 genes were identified in which 22,127 of them are known genes and 528 of them are novel genes. 10,604 novel transcipts were identified in which 8,969 of them are previously unknown splicing event for known genes, 528 of them are novel coding transcripts without any known features, and the remaining 1,107 are long noncoding RNA.

2 Sequencing Reads Filtering

The sequencing reads which containing low-quality, adaptor-polluted and high content of unknown base (N) reads, should be processed to be removed before downstream analyses. The original data performance is shown in Figure1. Clean reads quality metrics are shown as Table1. The distribution of base quality is shown as Figure2.

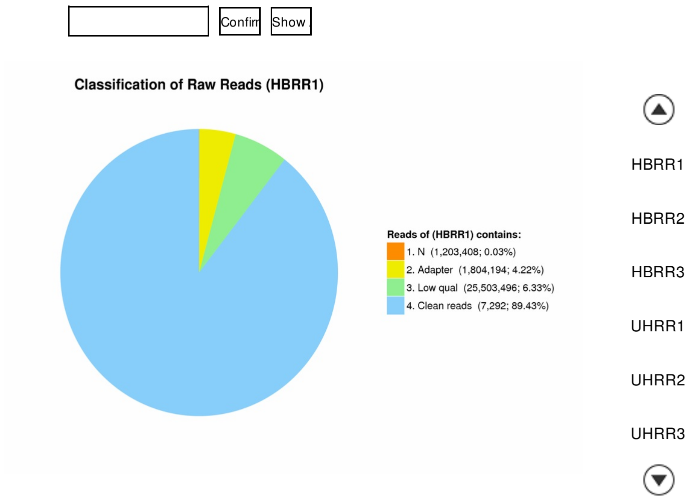


Figure 1 Raw data filter composition chart.

N: The total amount of reads which contain more than 5% unknown N base; the N reads ratio; Adaptor: The total amount of reads which contain adaptors; the adaptor ratio; Low quality: More than 20% of bases in the total read have quality score lower than 15; low quality reads ratio; Clean reads: Reads filtered with N reads, reads have adaptors and low quality reads; clean reads ratio.

Table 1 Clean reads quality metrics (Download)

Sample	Total Raw Reads(Mb)	Total Clean Reads(Mb)	Total Clean Bases(Gb)	Clean Reads Q20(%)	Clean Reads Q30(%)	Clean Reads Ratio(%)
HBRR1	28.52	25.50	2.30	98.99	95.28	89.43
HBRR2	28.52	26.08	2.35	99.02	95.39	91.46
HBRR3	28.52	26.22	2.36	98.94	95.11	91.95
UHRR1	28.52	25.98	2.34	98.78	94.48	91.09
UHRR2	28.52	26.12	2.35	98.78	94.45	91.59
UHRR3	28.52	25.82	2.32	98.72	94.27	90.54

Samples: Sample names

Total Raw Reads(Mb): The reads amount before filtering, Unit: Mb

Total Clean Reads(Mb): The reads amount after filtering, Unit: Mb

Total Clean Bases(Gb): The total base amount after filtering, Unit: Gb

Clean Reads Q20(%): The Q20 value for the clean reads

Clean Reads Q30(%): The Q30 value for the clean reads

Clean Reads Ratio(%): The ratio of the amount of clean reads

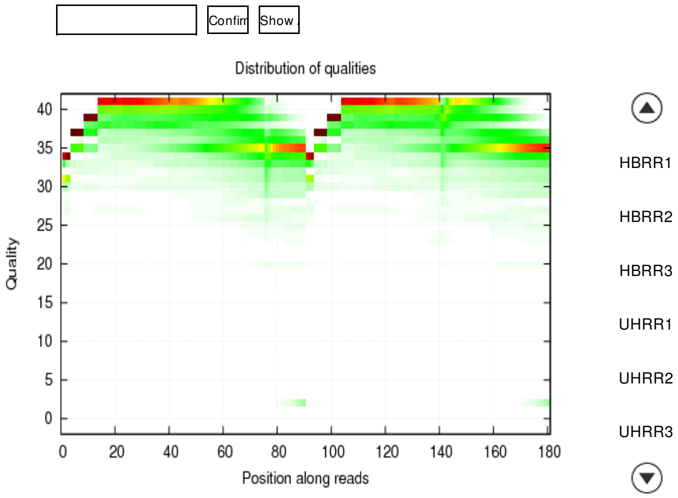


Figure 2 Distribution of base quality on clean reads.

X axis represents base positions along reads. Y axis represents base quality value. Each dot in the image represents the number of total bases with certain quality value of the corresponding base along reads. Darker dot color means greater base number. If the proportion of the bases with low quality (<20) is very low, that means the sequencing quality of this lane is good.

3 Genome Mapping

After reads filtering, we map clean reads to reference genome using HISAT [2]. On average 96.30% reads are mapped, and the uniformity of the mapping result for each sample suggests that the samples are comparable. The mapping details are shown as Table2.

Table 2 Summary of Genome Mapping (Download)

Sample	Total CleanReads	Total MappingRatio	Uniquely MappingRatio
HBRR1	25,503,496	95.94%	82.36%
HBRR2	26,083,620	95.92%	82.60%
HBRR3	26,221,456	95.94%	81.99%
UHRR1	25,978,082	96.55%	81.40%
UHRR2	26,122,186	96.84%	81.29%
UHRR3	25,821,926	96.63%	81.68%

Sample: Sample name

Total CleanReads: The amount of clean reads

Total MappingRatio: The percentage of mapped reads

Uniquely MappingRatio: The percentage of reads that map to only one location of reference

At the same time, we provide the bam files for the genome mapping result. IGV(Integrative Genomics Viewer) tool can be used to review the mapping result. IGV supports importing multiple samples for comparison and show the distribution of reads in the exon, intron, UTR, intergenic areas based on the annotation result. Figure3 is an example of IGV display. In addition, Figure4 shows an example of sashimi-plot, which can plot reads densities along exons and junctions for multiple samples. IGV Genomic Data Browsing Method please refer to the directory IGV/IGV_readme.pdf.

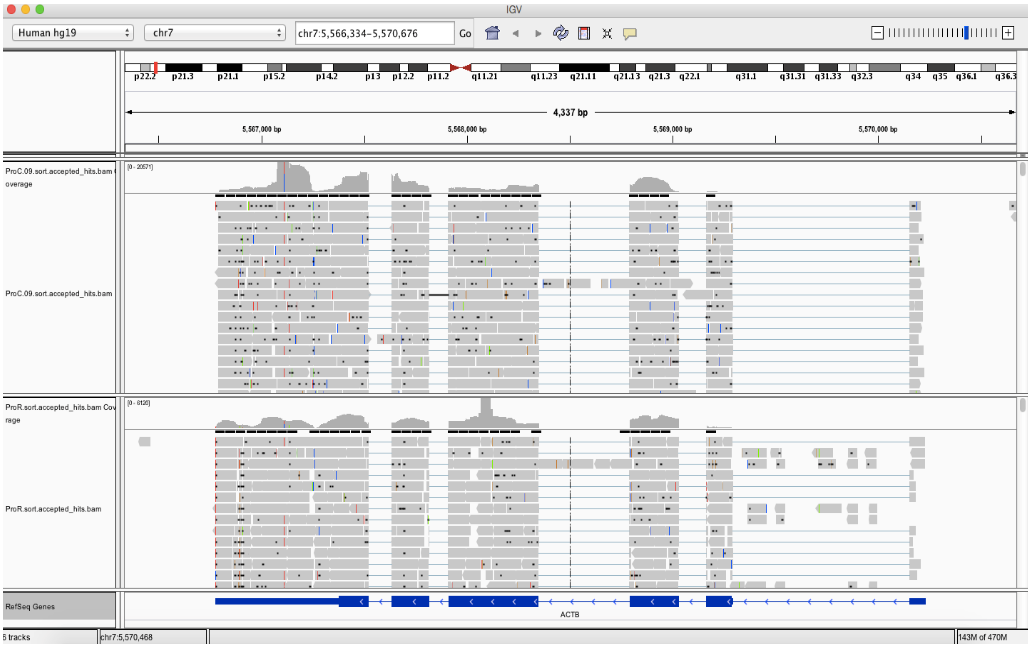

IGV
Human hg19
chr7
chr7:5,566,334-5,570,676
Go
p22.2 p21.3 p21.1 p15.2 p14.2 p13 p12.2 p11.2 q11.21 q11.23 q21.11 q21.13 q21.3 q22.1 q31.1 q31.31 q31.33 q32.3 q34 q35 q36.1 q36.3
4,337 bp
5,567,000 bp
5,568,000 bp
5,569,000 bp
5,570,000 bp
ProC.09.sort.accepted_hits.bam Coverage
[0 - 20571]
ProC.09.sort.accepted_hits.bam
ProR.sort.accepted_hits.bam Coverage
[0 - 6120]
ProR.sort.accepted_hits.bam
RefSeq Genes
ACTB
6 tracks
chr7:5,570,468
143M of 470M

Figure 3 IGV comparison results.

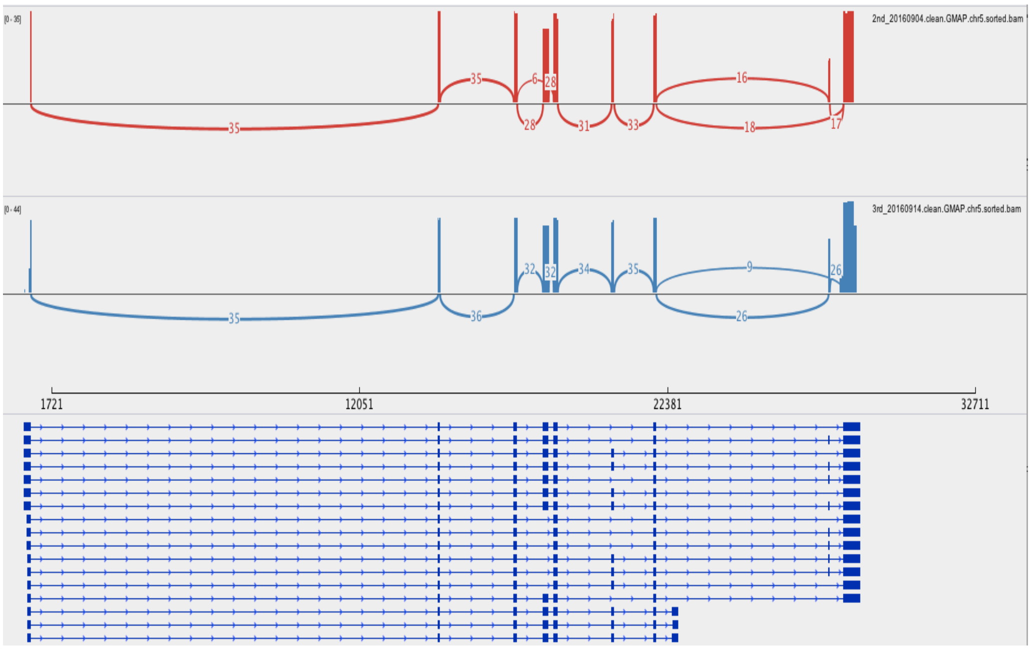


Figure 4 Sashimi plots to screen differentially spliced exons along genomic regions.

4 Novel Transcripts Prediction

After genome mapping, we use StringTie[3] to reconstruct transcripts, and with genome annotation information we identify novel transcripts by using Cuffcompare(a tool of Cufflinks [4]) and predict the coding ability of those new transcripts using CPC [5]. In total, we identify 10,604 novel transcripts, the detailed information is shown as Table3.

Table 3 Summary of Novel Transcripts. (Download)

Total_Novel_Transcript	Coding_Transcript	Noncoding_Transcript	NovelIsoform	NovelGene
10,604	9,497	1,107	8,969	528

Total_Novel_Transcript: The amount of predicted novel transcripts

Coding_Transcript: The amount of predicted coding transcripts

Noncoding_Transcript: The amount of predicted noncoding transcripts

NovelIsoform: The amount of predicted coding transcripts that previously unknown splicing event for a known gene

NovelGene: The amount of predicted coding transcripts that previously unknown

5 SNP and INDEL Detection

After genome mapping, we use GATK\[7\]to call SNP and INDEL variant for each sample. Final results are stored in VCF format. The SNP summary is shown as Table4, and Figure5. We also generate a friendly-interfaced SNP summary in EXCEL format shown as Table19. Then, we statistic the location of SNP and INDEL , shown as Figure6 and Figure7.

Table 4 SNP variant type summary. (Download)

Sample	A-G	C-T	Transition	A-C	A-T	C-G	G-T	Transversion	Total
HBRR1	25,728	25,344	51,072	4,246	2,791	5,875	4,356	17,268	68,340
HBRR2	26,466	26,009	52,475	4,512	2,910	6,129	4,492	18,043	70,518
HBRR3	24,569	24,069	48,638	4,162	2,669	5,646	4,247	16,724	65,362
UHRR1	24,580	24,216	48,796	4,289	2,846	6,018	4,401	17,554	66,350
UHRR2	24,330	23,682	48,012	4,368	2,950	5,975	4,384	17,677	65,689
UHRR3	25,940	25,331	51,271	4,499	3,049	6,173	4,536	18,257	69,528

Sample: Sample name

A-G: The amount of A-G variant type

C-T: The amount of C-T variant type

Transition: The amount of A-G and C-T variant type

A-C: The amount of A-C variant type

A-T: The amount of A-T variant type

C-G: The amount of C-G variant type

G-T: The amount of G-T variant type

Transversion: The amount of A-C, A-T, C-G and G-T variant type

Total: The amount of all variant type

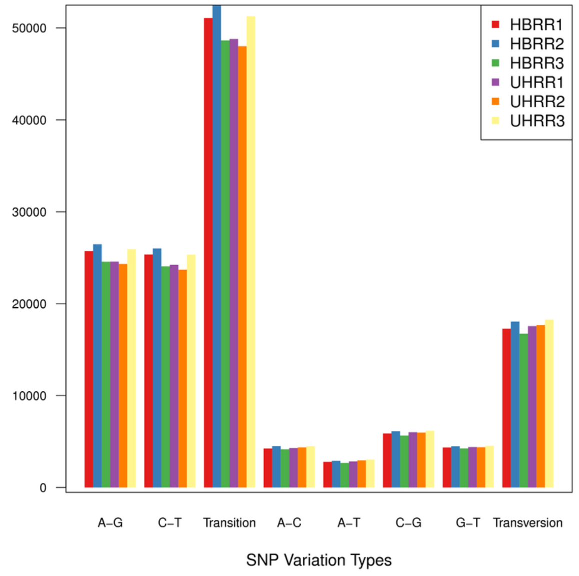


Figure 5 SNP variant type distribution.

X axis represents the type of SNP. Y axis represents the number of SNP.

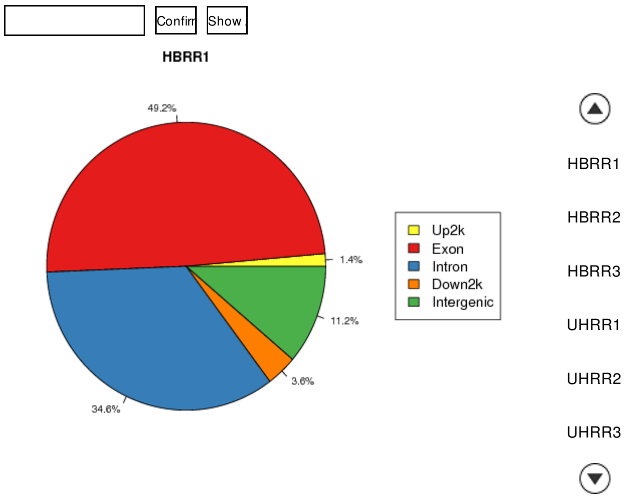


Figure 6 Distribution of SNP location.

Up2k means upstream 2,000 bp area of a gene. Down2k means downstream 2,000 bp area of a gene.

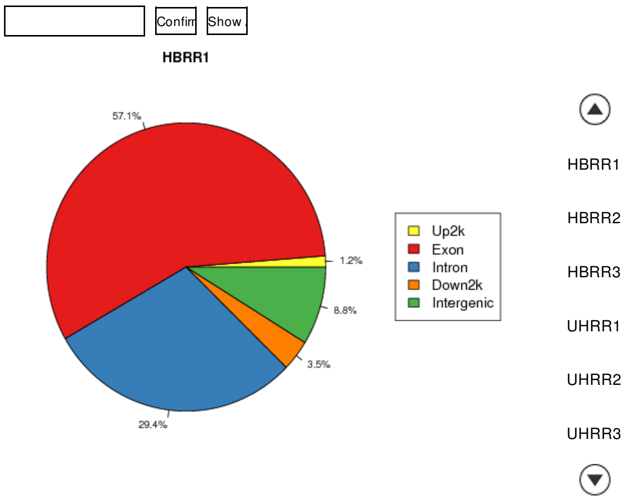


Figure 7 Distribution of INDEL location.

Up2k means upstream 2,000 bp area of a gene. Down2k means downstream 2,000 bp area of a gene.

The VCF format SNP and INDEL result of each sample are shown as tables below(see VCF format in help page VCF format).

Table 5 SNP list of HBRR1: (Download)

Table 6 SNP list of HBRR2: (Download)

Table 7 SNP list of HBRR3: (Download)

Table 8 SNP list of UHRR1: (Download)

Table 9 SNP list of UHRR2: (Download)

Table 10 SNP list of UHRR3: (Download)

Table 11 Summary of SNP result in VCF format: (Download)

Table 12 INDEL list of HBRR1: (Download)

Table 13 INDEL list of HBRR2: (Download)

Table 14 INDEL list of HBRR3: (Download)

Table 15 INDEL list of UHRR1: (Download)

Table 16 INDEL list of UHRR2: (Download)

Table 17 INDEL list of UHRR3: (Download)

Table 18 Summary of INDEL results in VCF format: (Download)

Table 19 Summary of SNP results in excel: (Download)

6 Gene Fusion Detection

After reads filtering, we use SOAPfuse[8] to detect gene fusion for each sample, shown as Figure8.

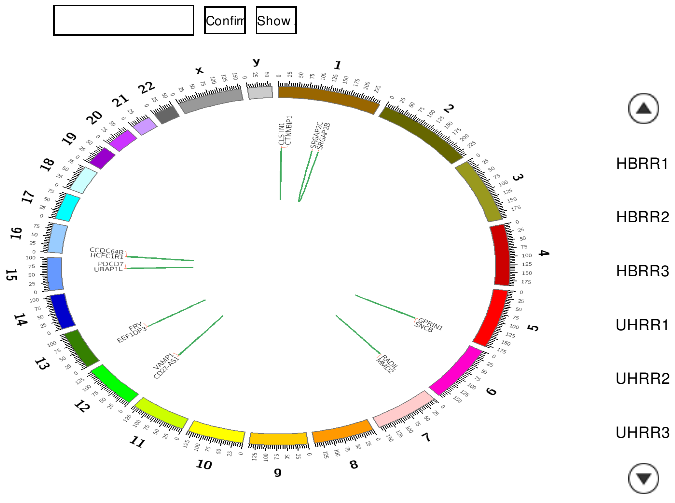


Figure 8 Circos diagram for gene fusion visualization.

The outer circle represents the chromosome information, the lines represent the fusion genes (Red lines represent the gene fusion which occurs between chromosomes, green lines represent the gene fusion which occurs within chromosomes, gene fusion analysis only for human samples.)

SOAPfuse software can show the gene fusion very clearly. SOAPfuse demo result for example as shown in Figure9. For details, see the result directory 'Transcriptome_Resequencing_Report/BGI_result/3.Structure/GeneFusion'.

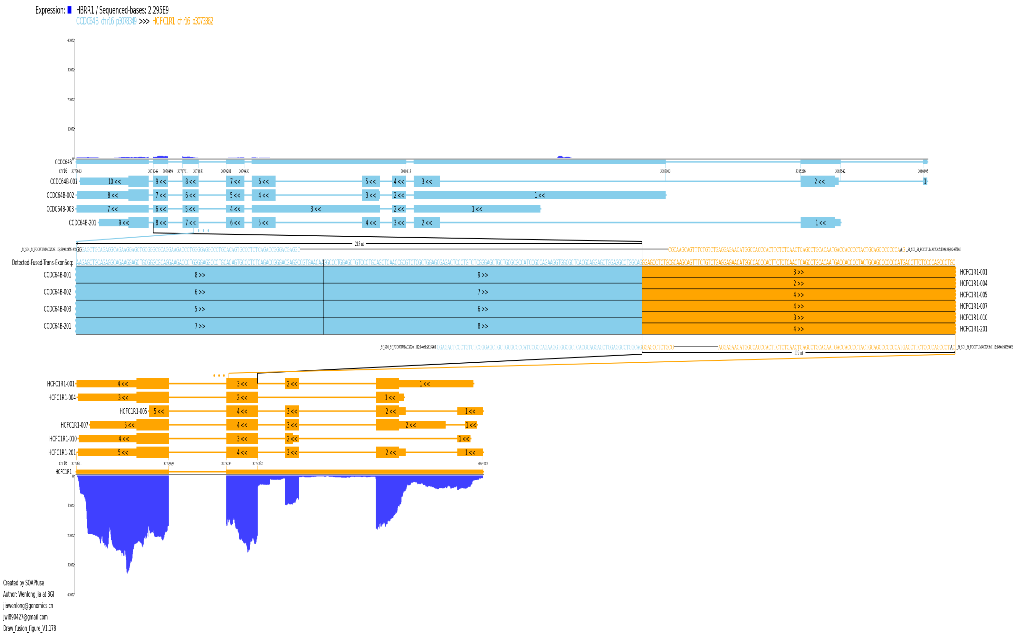


Figure 9 SOAPfuse demo result.

Gene fusion details for each sample is shown as tables below (see Gene fusion format on help page Gene fusion format instruction):

Table 20 Gene fusion list of HBRR1: (Download)

Table 21 Gene fusion list of HBRR2: (Download)

Table 22 Gene fusion list of HBRR3: (Download)

Table 23 Gene fusion list of UHRR1: (Download)

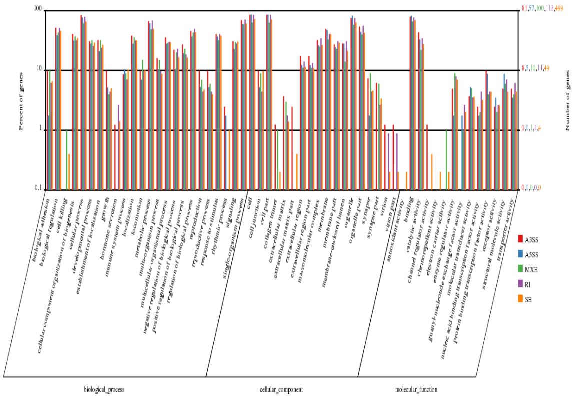
Table 24 Gene fusion list of UHRR2: (Download)

Table 25 Gene fusion list of UHRR3: (Download)

7 Differentially Splicing Gene Detection

After genome mapping, we use rMATS \[9\]to detect differentially splicing gene (DSG) between samples. DSGs are regulated by alternative splicing (AS), which allows the production of a variety of different isofroms from one gene only. Changes in relative abundance of isoforms, regardless of the expression change, indicate a splicing-related mechanism. We detect five types of AS events, including Skipped Exon (SE), Alternative 5' Splicing Site (A5SS), Alternative 3' Splicing Site (A3SS), Mutually exclusive exons (MXE) and Retained Intron (RI). The Gene Ontology classification is shown as Figure10 and the summary of gene splicing is shown in Figure11.

□ Confirm Show



HBRR-VS-UHRR

HBRR1-VS-UHRR1

HBRR2-VS-UHRR2

Figure 10 Gene Ontology classification of DSGs.

X axis represents the Gene Ontoloty functions. Y axis represents the number of DSGs. Different columns represent different samples.

The results of each provided compare plan are shown as tables below (see DSG format in help page Differentially Splicing Gene format).

Table 26 A3SS regulated DSG list of HBRR-VS-UHRR: (Download)

Table 27 A5SS regulated DSG list of HBRR-VS-UHRR: (Download)

Table 28 MXE regulated DSG list of HBRR-VS-UHRR: (Download)

Table 29 RI regulated DSG list of HBRR-VS-UHRR: (Download)

Table 30 SE regulated DSG list of HBRR-VS-UHRR: (Download)

Table 31 A3SS regulated DSG list of HBRR1-VS-UHRR1: (Download)

Table 32 A5SS regulated DSG list of HBRR1-VS-UHRR1: (Download)

Table 33 MXE regulated DSG list of HBRR1-VS-UHRR1: (Download)

Table 34 RI regulated DSG list of HBRR1-VS-UHRR1: (Download)

Table 35 SE regulated DSG list of HBRR1-VS-UHRR1: (Download)

Table 36 A3SS regulated DSG list of HBRR2-VS-UHRR2: (Download)

Table 37 A5SS regulated DSG list of HBRR2-VS-UHRR2: (Download)

Table 38 MXE regulated DSG list of HBRR2-VS-UHRR2: (Download)

Table 39 RI regulated DSG list of HBRR2-VS-UHRR2: (Download)

Table 40 SE regulated DSG list of HBRR2-VS-UHRR2: (Download)

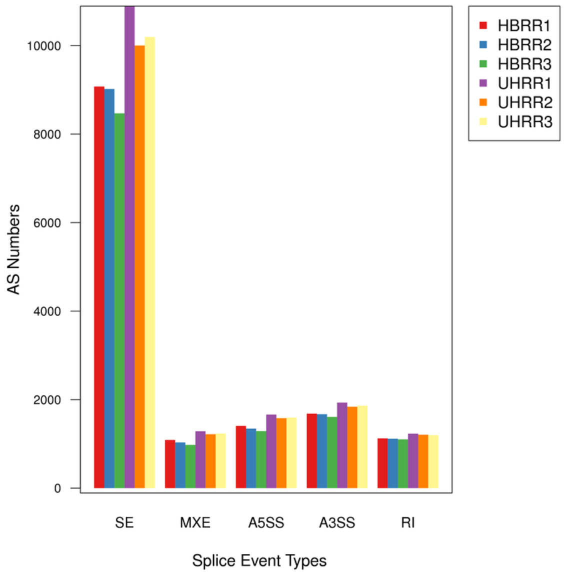


Figure 11 Statistic of Splicing.

X axis means the type of splicing. Y axis means the amount. Different columns represent different splicing events.

8 Gene Expression Analysis

8.1 Gene Mapping and Expression

After novel transcript detection, we merge novel coding transcripts with reference transcripts to get complete reference, then we map clean reads to it using Bowtie2\[10\], then calculate gene expression level for each sample with RSEM \[11\]. The gene mapping ratio is shown as Table41. And the number of genes and transcripts of each sample is shown as Table42.

Table 41 Summary of gene mapping ratio. (Download)

Sample	Total CleanReads	Total MappingRatio	Uniquely MappingRatio
HBRR1	25,503,496	67.01	58.46
HBRR2	26,083,620	66.27	57.82
HBRR3	26,221,456	65.87	57.43
UHRR1	25,978,082	76.13	64.63
UHRR2	26,122,186	75.17	63.32
UHRR3	25,821,926	74.22	62.90

Sample: Sample name

Total CleanReads: The amount of Clean reads

Total MappingRatio: The percentage of mapped reads (%)

Uniquely MappingRatio: The percentage of uniquely mapped reads (%)

Table 42 Genes and Transcripts statistics (Download)

Sample	Total GeneNumber	Known GeneNumber	Novel GeneNumber	Total TranscriptNumber	Known TranscriptNumber	Novel TranscriptNumber
HBRR1	19,758	19,385	373	35,736	30,127	5,609
HBRR2	19,813	19,430	383	35,825	30,179	5,646
HBRR3	19,676	19,306	370	35,414	29,818	5,596
UHRR1	20,113	19,721	392	36,738	31,290	5,448
UHRR2	20,059	19,663	396	36,276	30,827	5,449
UHRR3	20,022	19,622	400	36,415	30,958	5,457

Sample: Sample name

Total GeneNumber: The amount of all genes

Known GeneNumber: The amount of known genes

Novel GeneNumber: The amount of novel genes

Total TranscriptNumber: The amount of all transcripts

Known TranscriptNumber: The amount of known transcripts

Novel TranscriptNumber: The amount of novel transcripts

The expressed gene list summary is shown in Table49.

Table 43 Expressed gene list of HBRR1: (Download)

Table 44 Expressed gene list of HBRR2: (Download)

Table 45 Expressed gene list of HBRR3: (Download)

Table 46 Expressed gene list of UHRR1: (Download)

Table 47 Expressed gene list of UHRR2: (Download)

Table 48 Expressed gene list of UHRR3: (Download)

Table 49 The result of gene expression. (Download)

gene_id	transcript_id(s)	length	expected_count	FPKM
1	BGI_novel_T005039,NM_130786	1,559.21	38.37	2.99
10	NM_000015	1,317.00	1.00	0.09
100	BGI_novel_T005809,NM_000022	1,566.00	36.00	2.79
1000	NM_001308176,NM_001792	4,067.92	461.00	13.45

gene_id: The gene ID

transcript id: The transcript ID

length: Gene length

expected count: The reads amount which mapped to the gene

FPKM: The gene FPKM

8.2 Reads Coverage and Distribution Analysis of Transcripts

We calculate the reads coverage and the reads distribution of each detected transcript, shown as Figure12and Figure13, respectively.

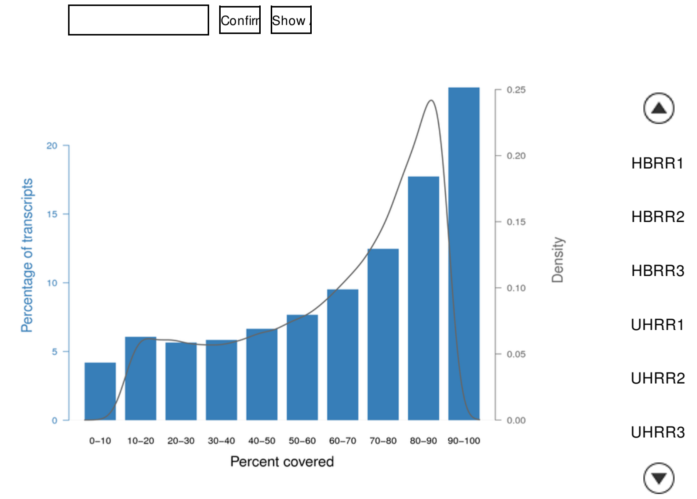


Figure 12 Reads coverage on transcripts.

X axis represents the reads coverage. Y axis on left side represents the percentage of transcripts. Y axis on right side represents the density of transcripts.

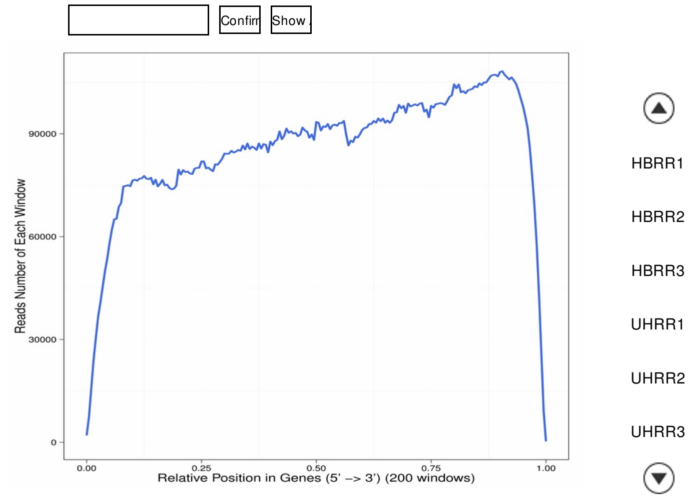


Figure 13 Reads distribution on transcripts.

X axis represents the position along transcripts. Y axis represents the number of reads.

8.3 Correlation Between Samples

After that, we calculate pearson correlation between all samples, shown as Figure14. Hierarchical clustering between all samples is also performed, shown as Figure15.

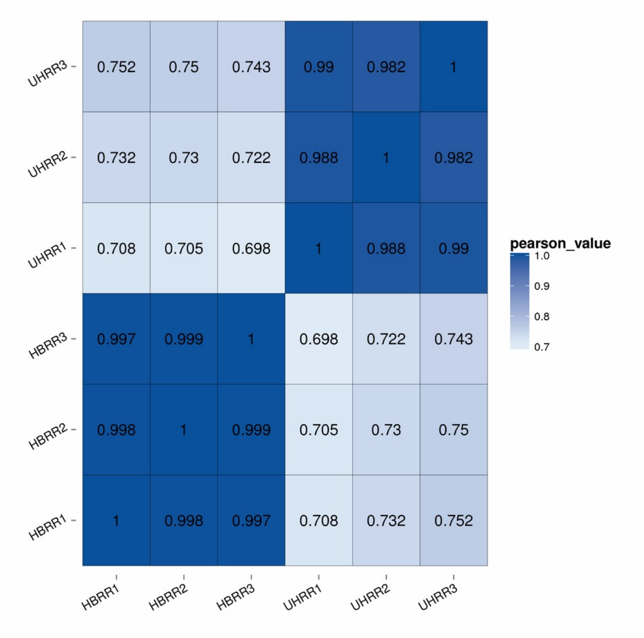


Figure 14 Heatmap of Pearson correlation between samples.

Both X and Y axis represent each sample. Coloring indicate Pearson correlation (high: blue; low: white).

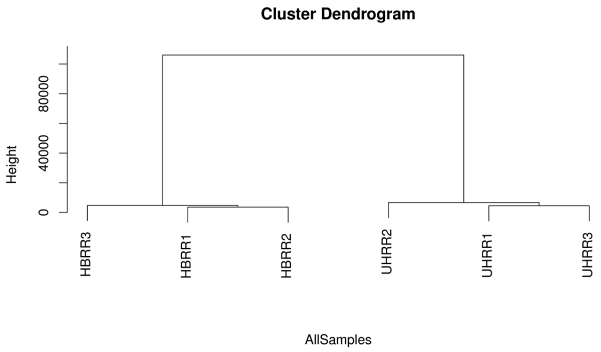


Figure 15 Hierarchical clustering between samples.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation). We perform PCA analysis based on the PCA plan provided by our customer, shown as Figure16.

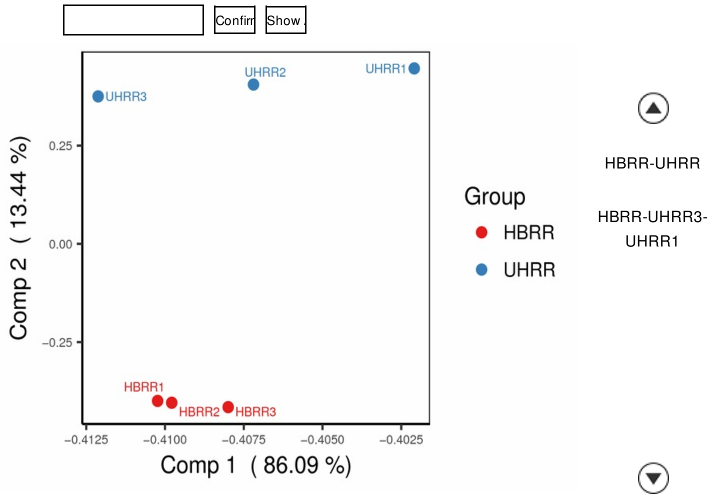


Figure 16 PCA analysis.

X axis represents the contributor rate of first component. Y axis represents the contributor rate of second component. Points represent each sample. The samples in one group shows the same color.

8.4 The Distribution of Gene Expression

Based on the expression information, we preform box plot to show the distribution of the gene expression level of each sample, besides we can observe the dispersion of the distribution, as shown in Figure17. The density map can show the change of gene abundance and reflect the concentration of gene expression in the sample interval, as show in Figure18.

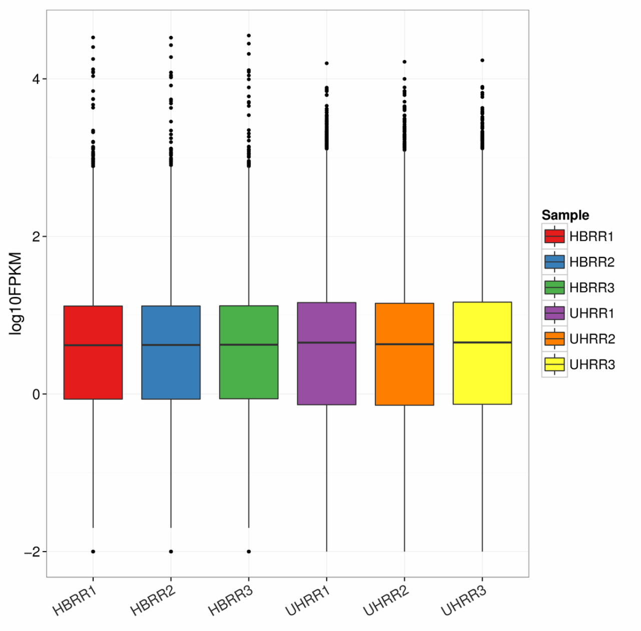


Figure 17 Gene expression Box-plot.

X axis represents the sample name. Y axis represents the log10FPKM value.

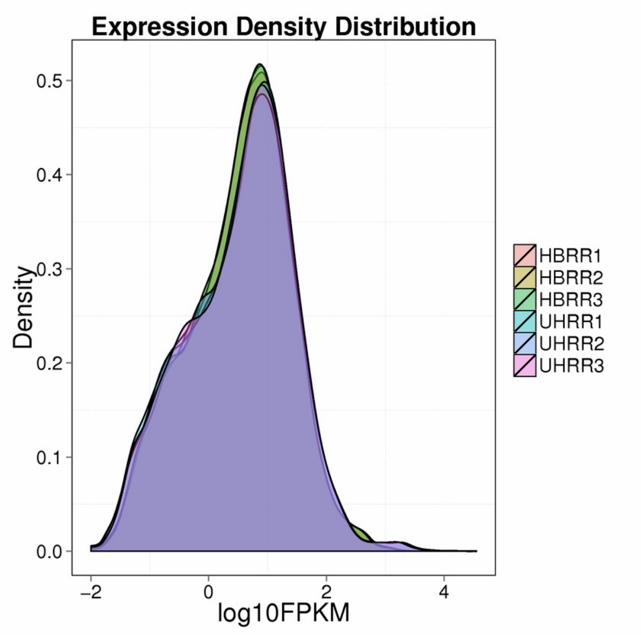


Figure 18 Gene expression density map.

X axis represents the log10FPKM value. Y axis represents the gene density.

To show the gene amount under different FPKM value, we calculate the gene amount under three different FPKM ranges（FPKM <= 1、FPKM 1~10、FPKM >= 10）, shown as Figure19.

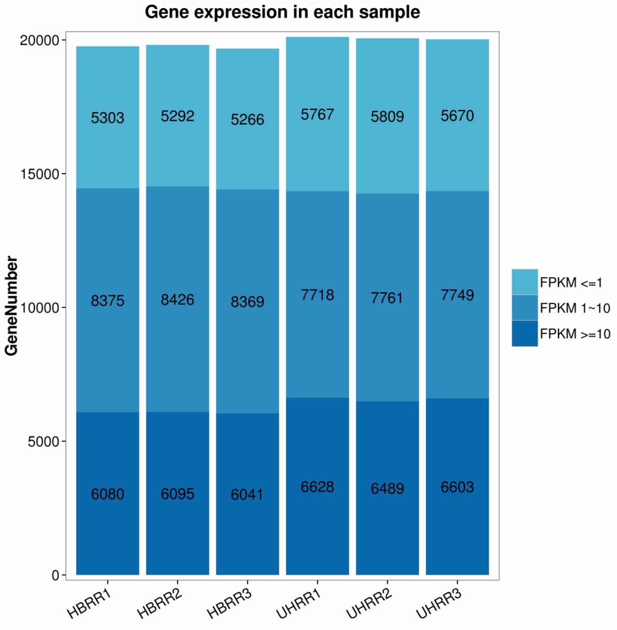


Figure 19 Gene expression distribution.

X axis represents the sample name. Y axis represents the gene amount. The dark color means the high expression level which FPKM value >= 10, while the light color means the low expression level which FPKM value <= 1.

To show the amount of novel genes and known genes, we calculate the gene amount (shown as Figure20) and analyse the expression ratio (shown as Figure21) respectively.

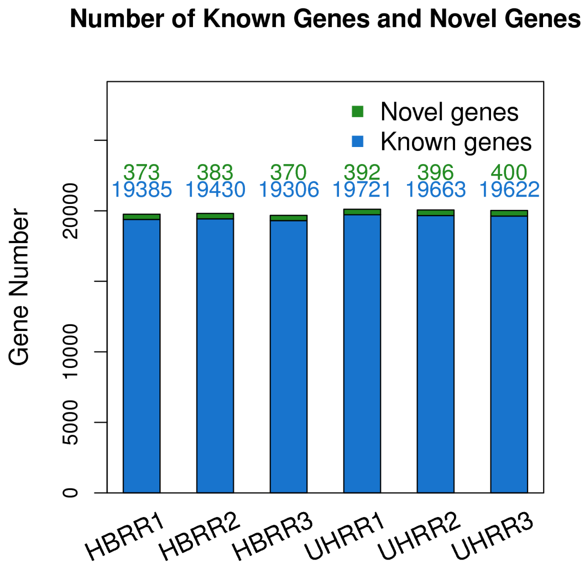


Figure 20 Statistics of novel genes and known genes.

X axis represents the sample name. Y axis represent the gene amount.

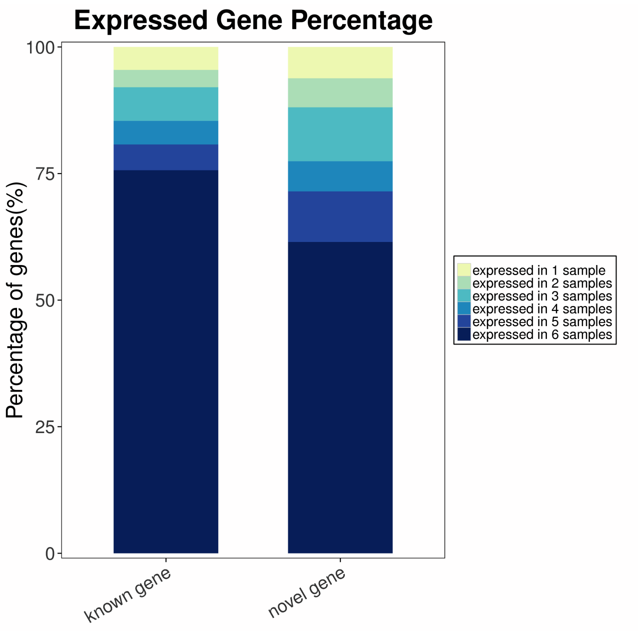


Figure 21 Expression distribution of novel and known genes.

X axis represents the gene type. Y axis represents the percentage of genes expressed in different samples.

8.5 Gene Expression Analysis Between Samples

We also use venn diagram to display expressed gene between samples, shown as Figure22.

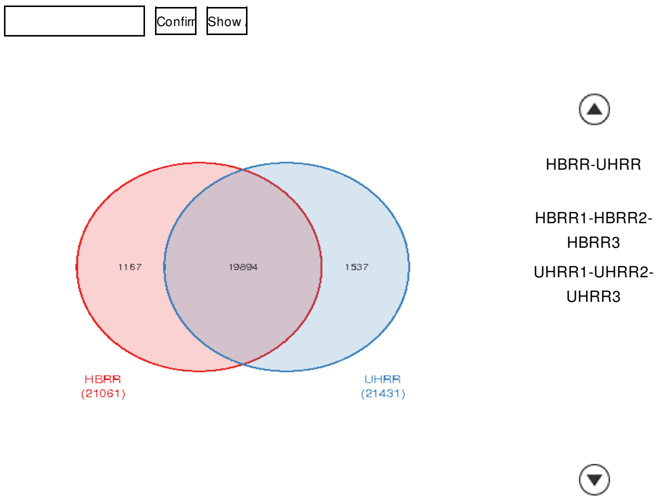


Figure 22 Venn diagram analysis.

9 Circos Diagram

Based on the SNP , INDEL , gene expression and gene fusion(Only for human samples) result, we use Circos\[12\]to perform the analysis, shown as Figure23.

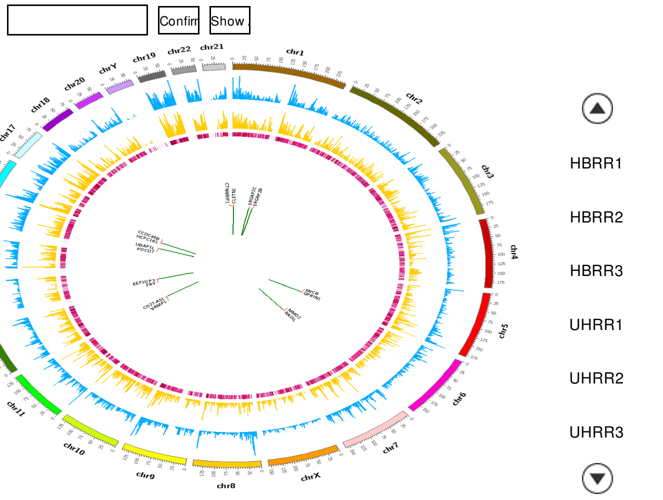


Figure 23 Circos diagram.

From the outside circle to the inner circle, the first circle represents chromosome, the second, third and forth circle represents the SNP amount, INDEL amount and FPKM value, respectively. The connection lines in the inner circle represent the gene fusion.(Red lines represent the gene fusion occurs between chromosomes, green lines represent the gene fusion which occurs within chromosomes, gene fusion analysis only for human samples.)

10 Time Series Analysis

In different time stages, some of the genes may have similar expression pattern. According to the gene expression information, those genes can be clustered into time related clusters. Those genes with the same gene expression pattern will be in the same gene cluster. It has been mentioned in some papers that time series analysis can be used to identify the tissue specific genes. The gene cluster results are shown in Figure24. Details refer to : BGI_result/4.Quantify/GeneExpression/Clustering_Mfuzz.

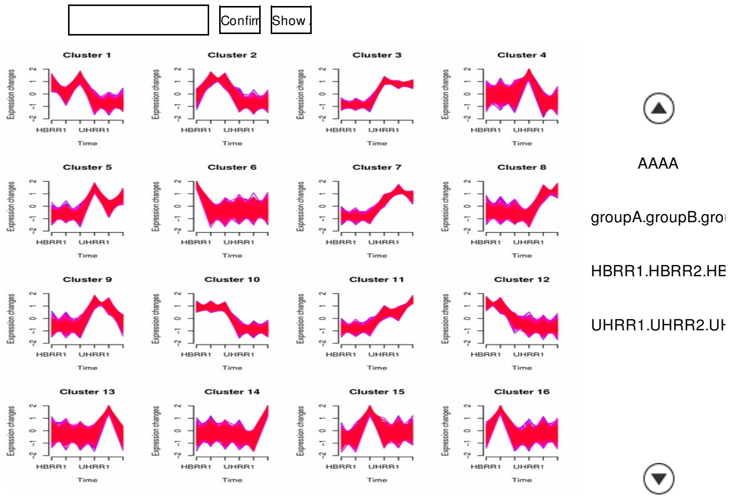


Figure 24 Time series analysis Mfuzz result.

X axis represents the timeline, Y axis represents the normalized gene expression level.

The genes in each cluster are shown in the below table:

Table 50 Distribution of genes in clusters (Download)

gene_id	cluster	HBRR1_fpkm	HBRR2_fpkm	HBRR3_fpkm	...
10005	1	21.42	18.90	20.42	...
10008	1	0.43	0.46	0.65	...
100128946	1	0.30	0.20	0.32	...
100129396	1	1.17	0.67	0.91	...

gene_id: Gene ID
cluster: Cluster ID
Sample(Group)1_fpkm: The FPKM value for sample(group)1
Sample(Group)2_fpkm: The FPKM value for sample(group)2
Sample(Group)3_fpkm: The FPKM value for sample(group)3
...: symbol, GO, Pathway and NR annotation

Table 51 Gene distribution in each clusters of AAAA: (Download)
Table 52 Gene distribution in each clusters of groupA.groupB.groupC: (Download)
Table 53 Gene distribution in each clusters of HBRR1.HBRR2.HBRR3: (Download)
Table 54 Gene distribution in each clusters of UHRR1.UHRR2.UHRR3: (Download)

Membership represents the value between 0~1, it can be used to evaluate if the genes follow the change trends of the clusters. If the membership value goes to 1, it means the gene follow the change trend of the clusters.

Table 55 The membership value of the genes for AAAA: (Download)
Table 56 The membership value of the genes for groupA.groupB.groupC: (Download)
Table 57 The membership value of the genes for HBRR1.HBRR2.HBRR3: (Download)
Table 58 The membership value of the genes for UHRR1.UHRR2.UHRR3: (Download)

11 Differentially Expressed Gene Detection

Based on the gene expression level, we can identify the DEG (Differentially expression genes) between samples or groups. We use DEGseq,DEseq2,EBseq,NOIseq and PossionDis algorithms to detect the DEGs, results shown as below:

Table 59 DEG result for HBRR2-VS-UHRR2.PossionDis_Method (Download)

GeneID	HBRR2-Expression	UHRR2-Expression	log2FoldChange(UHRR2/HBRR2)	FDR	Pvalue
BGI_novel_G000242	0.01	6,962.88	19.41	0.00e+00	0.00e+00
BGI_novel_G000239	0.01	998.30	16.61	0.00e+00	0.00e+00
336	0.01	811.92	16.31	0.00e+00	0.00e+00
3046	0.01	765.82	16.22	0.00e+00	0.00e+00

GeneID: Gene ID

Sample1-Expression: Gene expression in sample1

Sample2-Expression: Gene expression in sample2

log2FoldChange(sample2/sample1): The log2 value of ratio of Sample1-Expression to Sample2-Expression

FDR: False discovery rate

Pvalue: p-value

Table 60 DEG list for comparison HBRR-VS-UHRR.DEseq2_Method: (Download)

Table 61 DEG list for comparison HBRR-VS-UHRR.EBseq_Method: (Download)

Table 62 DEG list for comparison HBRR-VS-UHRR.NOIseq_Method: (Download)

Table 63 DEG list for comparison HBRR1-VS-UHRR1.DEGseq_Method: (Download)

Table 64 DEG list for comparison HBRR1-VS-UHRR1.PossionDis_Method: (Download)

Table 65 DEG list for comparison HBRR2-VS-UHRR2.DEGseq_Method: (Download)

Table 66 DEG list for comparison HBRR2-VS-UHRR2.PossionDis_Method: (Download)

Summary of DEGs is shown in Figure25. We use MA plot, Volcano plot, Scatter plot and Heatmap plot to show the distributions of DEGs in Figure26, Figure27, Figure28 and Figure29 respectively.

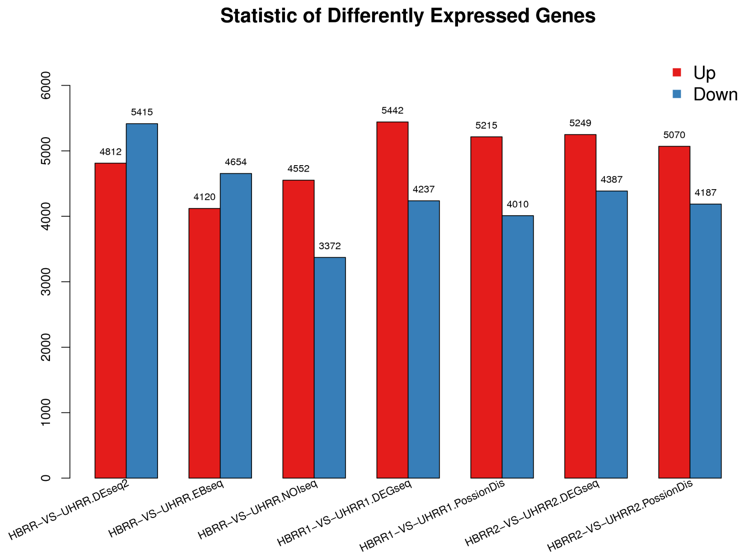


Figure 25 Summary of DEGs.

X axis represents comparison method between each group. Y axis represents DEG numbers. Red color represents up-regulated DEGs. Blue color represents down-regulated DEGs.

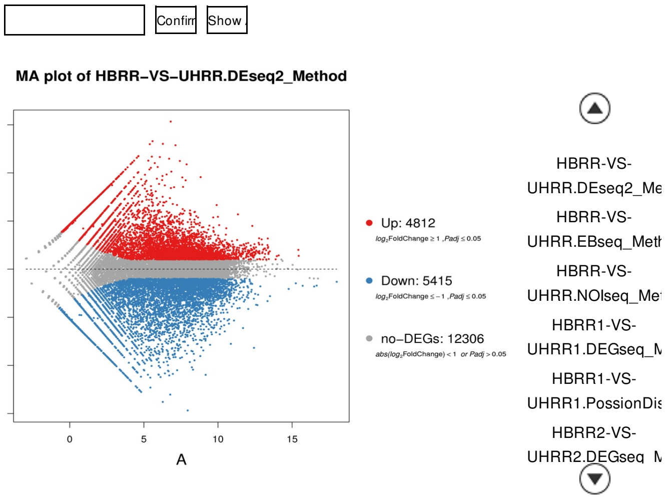


Figure 26 MA plot of DEGs.

X axis represents value A (log2 transformed mean expression level). Y axis represents value M (log2 transformed fold change). Red dots represent up-regulated DEGs. Blue dots represent down-regulated DEGs. Gray points represent non-DEGs.

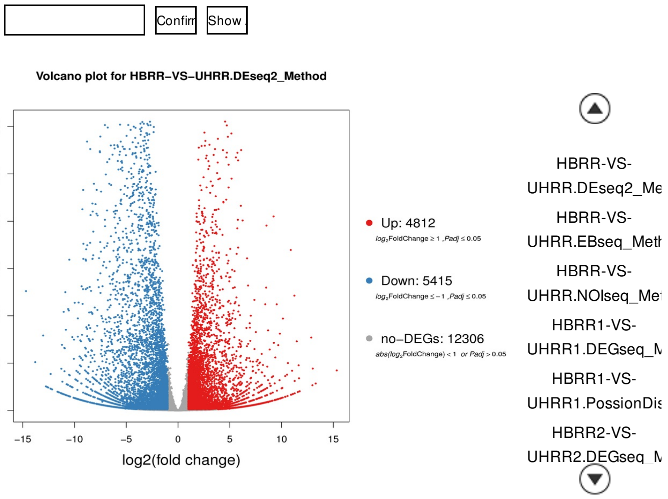


Figure 27 Volcano plot of DEGs.

X axis represents log2 transformed fold change. Y axis represents -log10 transformed significance. Red points represent up-regulated DEGs. Blue points represent down-regulated DEGs. Gray points represent non-DEGs.

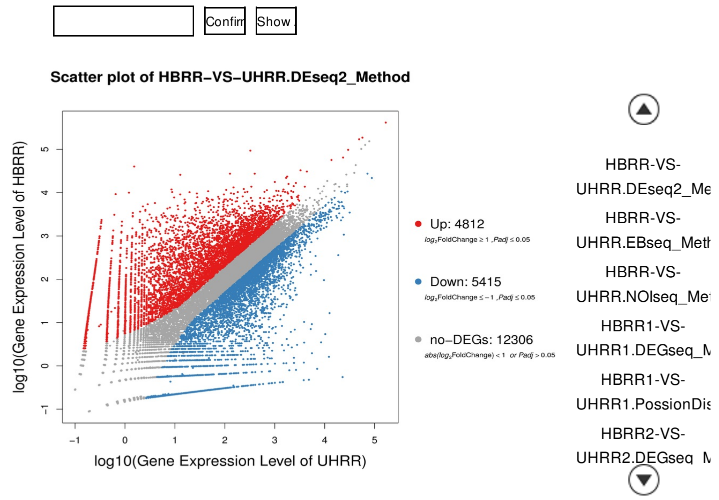

Confirm
Show
Scatter plot of HBRR−VS−UHRR.DEseq2_Method
log10(Gene Expression Level of HBRR)
log10(Gene Expression Level of UHRR)
−1
0
1
2
3
4
5
Up: 4812
log_2FoldChange ≥ 1 ,Padj ≤ 0.05
Down: 5415
log_2FoldChange ≤ −1 ,Padj ≤ 0.05
no−DEGs: 12306
abs(log_2FoldChange) < 1 or Padj > 0.05
HBRR-VS-UHRR.DEseq2_Me
HBRR-VS-UHRR.EBseq_Meth
HBRR-VS-UHRR.NOIseq_Me
HBRR1-VS-UHRR1.DEGseq_M
HBRR1-VS-UHRR1.PossionDis
HBRR2-VS-UHRR2.DEGseq_M

Figure 28 Scatter plot of DEGs.

X Y axis represents log10 transformed gene expression level, red color represents the up-regulated genes, blue color represents the down-regulated genes, gray color represents the non-DEGs.

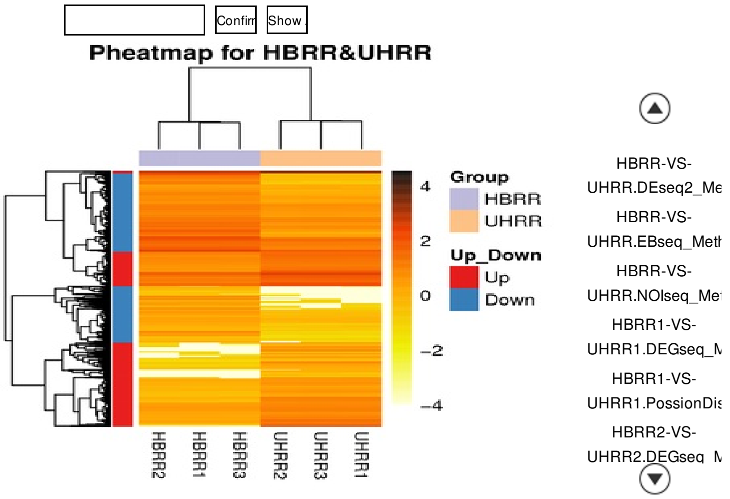


Figure 29 Heatmap of DEGs.

X axis represents the sample. Y axis represents the DEGs. The color represents the log10 transformed gene expression level. (The dark color means the high expression level while the light color means the low expression level.)

12 Venn Diagram of DEG

We perform DEGs by Venn diagrams, as shown in Figure30.

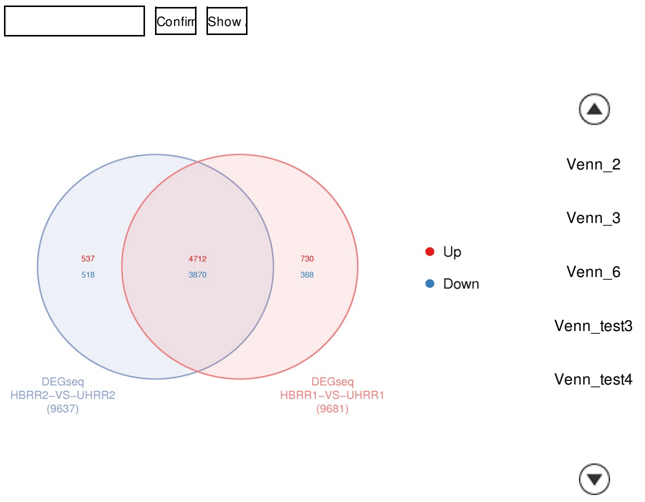


Figure 30 Venn diagram of DEGs.

The red number represents the up-regulated gene amount, blue number represents the down-regulated gene amount.

13 Clustering Analysis of DEG

We perform hierarchical clustering for DEGs ,shown as Figure31.

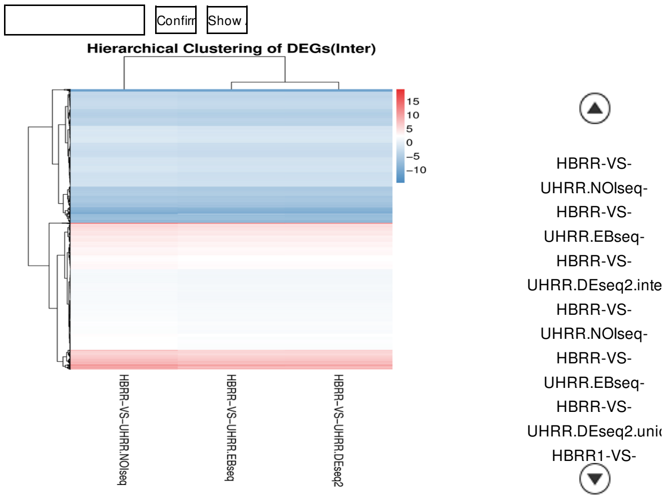


Figure 31 Heatmap of hierarchical clustering of DEGs.

X axis represents each comparing sample. Y axis represents DEGs. Coloring indicates the log2 transformed fold change (high:red, low:blue).

The ordered DEG lists after hierarchical clustering are shown as tables below (The file name suffix '.inter' or '.union' indicate the file contents are intersection or union of DEGs among different comparison groups. See Cluster list format in help page Cluster list format):

Table67 Clustering DEGs list of HBRR-VS-UHRR.NOIseq-HBRR-VS-UHRR.EBseq-HBRR-VS-UHRR.DEseq2.inter: (Download)

Table68 Clustering DEGs list of HBRR-VS-UHRR.NOIseq-HBRR-VS-UHRR.EBseq-HBRR-VS-UHRR.DEseq2.union: (Download)

Table69 Clustering DEGs list of HBRR1-VS-UHRR1.PossionDis-HBRR2-VS-UHRR2.PossionDis-HBRR1-VS-UHRR1.DEGseq-HBRR2-VS-UHRR2.DEGseq-HBRR-VS-UHRR.NOIseq-HBRR-VS-

UHRR.EBseq-HBRR-VS-UHRR.DEseq2.inter: (Download)

Table 70 Clustering DEGs list of HBRR1-VS-UHRR1.PossionDis-HBRR2-VS-UHRR2.PossionDis-HBRR1-VS-UHRR1.DEGseq-HBRR2-VS-UHRR2.DEGseq-HBRR-VS-UHRR.NOIseq-HBRR-VS-UHRR.EBseq-HBRR-VS-UHRR.DEseq2.union: (Download)

Table 71 Clustering DEGs list of HBRR1-VS-UHRR1.PossionDis-HBRR2-VS-UHRR2.PossionDis.inter: (Download)

Table 72 Clustering DEGs list of HBRR1-VS-UHRR1.PossionDis-HBRR2-VS-UHRR2.PossionDis.union: (Download)

Table 73 Clustering analysis format instruction. (Download)

Field	Description
Gene	Gene ID
A-VS-B	log2FoldChange of A-VS-B
C-VS-D	log2FoldChange of C-VS-D
...	...
...	Gene symbol, GO, Kegg and NR annotation

14 Gene Ontology Analysis of DEG

With DEGs, we perform Gene Ontology (GO) classification and functional enrichment. GO has three ontologies: molecular biological function, cellular component and biological process. We would perform functional enrichment respectively. The GO classification results are shown as Figure32. The GO enrichment results are shown as Figure33. The GO classification of up-regulated and down-regulated genes are shown as Figure34. (Click help page How to read DEG GO enrichment analysis resultto know how to read the GO analysis result)

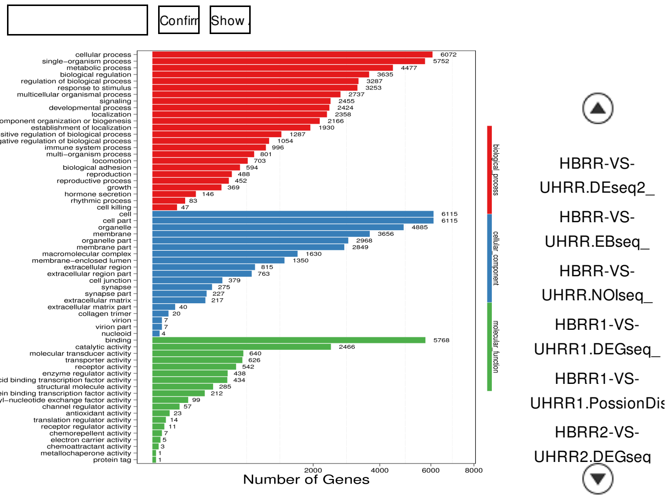


Figure 32 GO classification of DEGs.

X axis represents number of DEG. Y axis represents GO term.

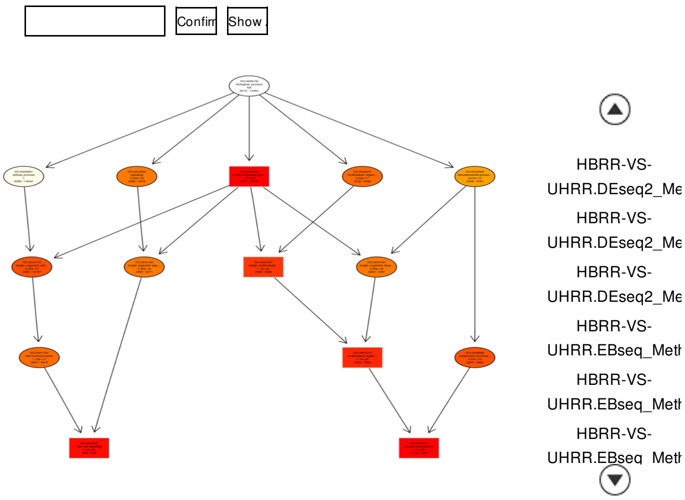


Figure 33 GO functional enrichment of DEGs.

We use DAG (Directed Acyclic Graph) to show the GO enrichment result. Each node shows the name of the GO term and the p-value. The darker (red) the color is, the lower p-value which indicates the more significant enrichment.

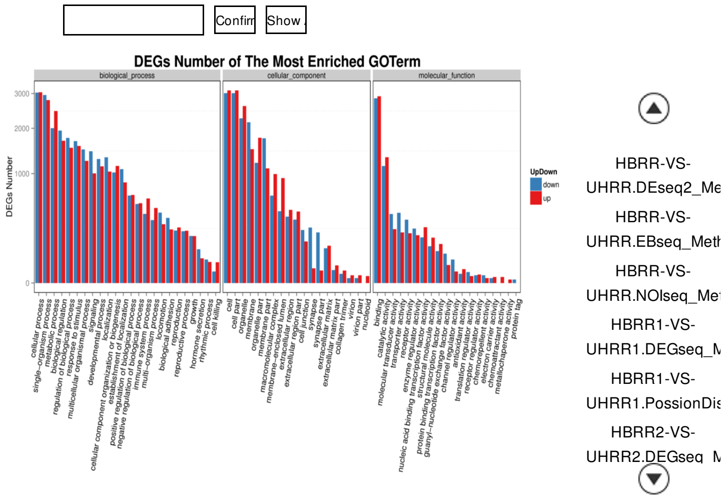


Figure 34 GO classification of up-regulated and down-regulated genes.

X axis represents GO term. Y axis represents the amount of up/down-regulated genes.

15 Pathway Analysis of DEG

With DEGs, we perform KEGG pathway classification and functional enrichment. The pathway classification results are shown as Figure35, and the pathway functional enrichment results are shown as Figure36. The pathway functional enrichment result for up/down regulation genes are shown as Figure37 (click help page How to read DEG pathway enrichment analysis result to know how to read the pathway analysis result).

Table 74 Pathway functional enrichment results. (Download)

#Pathway	HBRR2-VS-UHRR2.PossionDis_Method (8224)	All-Unigene (20106)	Pvalue	Qvalue	Pathway ID
Axon guidance	321	551	6.778957e-17	2.054024e-14	ko04360
Cell adhesion molecules (CAMs)	416	752	3.366312e-16	5.099963e-14	ko04514
ECM-receptor interaction	236	422	2.358674e-10	2.382261e-08	ko04512
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	254	461	3.626528e-10	2.747095e-08	ko05412

#Pathway: The name of the pathway

group_Method: The number of DEGs which annotated to specific pathway

All-Gene: The number of genes which annotated to specific pathway

Pvalue: p-value

Qvalue: corrected p-value

Pathway ID: The ID of the pathway

For details, refer to the help page \title{kegg_help}.

Table 75 Pathway functional enrichment results of HBRR-VS-UHRR.DEseq2Pathway (Download)

Table 76 Pathway functional enrichment results of HBRR-VS-UHRR.EBseqPathway (Download)

Table 77 Pathway functional enrichment results of HBRR-VS-UHRR.NOIseqPathway (Download)

Table 78 Pathway functional enrichment results of HBRR1-VS-UHRR1.DEGseqPathway (Download)

Table 79 Pathway functional enrichment results of HBRR1-VS-UHRR1.PossionDisPathway (Download)

Table 80 Pathway functional enrichment results of HBRR2-VS-UHRR2.DEGseqPathway (Download)

Table 81 Pathway functional enrichment results of HBRR2-VS-UHRR2.PossionDisPathway (Download)

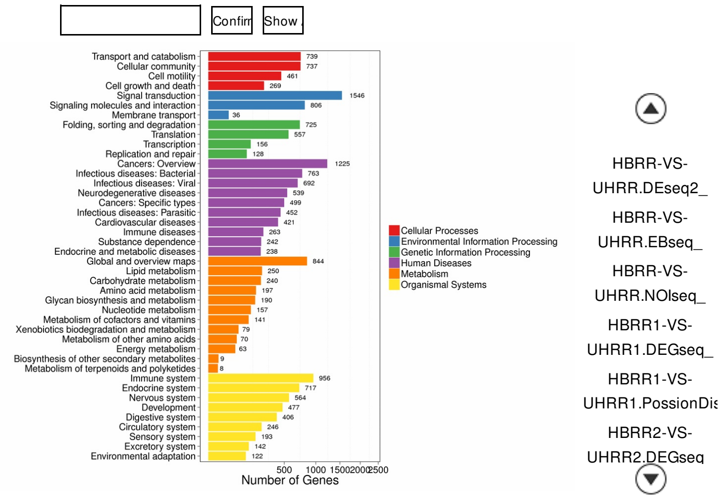


Figure 35 Pathway classification of DEGs.

X axis represents number of DEG. Y axis represents functional classification of KEGG. There are seven branches for KEGG pathways: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Disease (For animals only), Metabolism, Organismal Systems and Drug Development.

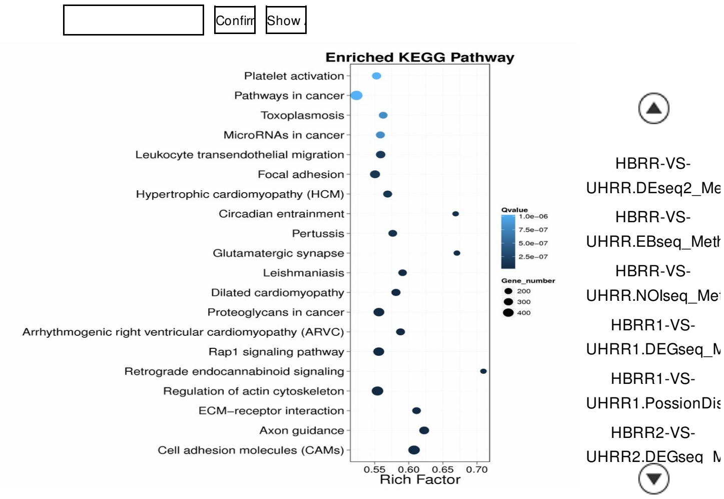


Figure 36 Pathway functional enrichment of DEGs.

X axis represents enrichment factor. Y axis represents pathway name. The color indicates the q-value (high: white, low: blue), the lower q-value indicates the more significant enrichment. Point size indicates DEG number (The bigger dots refer to larger amount). Rich Factor refers to the value of enrichment factor, which is the quotient of foreground value (the number of DEGs) and background value (total Gene amount). The larger the value, the more significant enrichment.

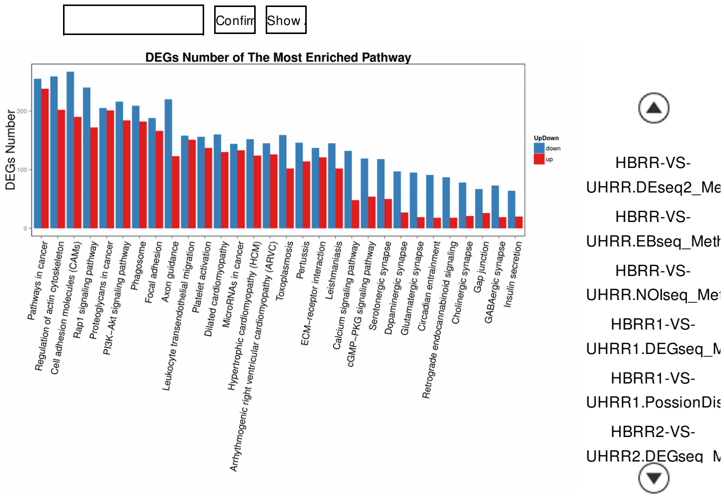


Figure 37 Pathway functional enrichment result for up/down regulation genes.

X axis represents the terms of Pathway. Y axis represents the number of up/down regulation genes.

16 Transcription Factor Prediction

We predict the DEGs that encode Transcription Factor (TF) in plant and animal research. (Click here TF to know how to read TF result). At the same time, we also conduct DEGs classification on TF family, shown as Figure38. We also conduct clustering analysis of the expression of TF , shown as Figure39.

Table 82 TF coding DEGs list (Download)

GeneID	Transcripts	TF_family
4904	BGI_novel_T000313	CSD
4904	NR_132737	CSD
4904	NM_004559	CSD
7545	NM_003412	zf-C2H2

GeneID: Gene ID

Transcripts: Transcript ID

TF_family: Transcription factor family

For details, see the help page TF.

Table 83 TF coding DEGs list of HBRR-VS-UHRR.DEseq2_Method: (Download)

Table 84 TF coding DEGs list of HBRR-VS-UHRR.EBseq_Method: (Download)

Table 85 TF coding DEGs list of HBRR-VS-UHRR.NOIseq_Method: (Download)

Table 86 TF coding DEGs list of HBRR1-VS-UHRR1.DEGseq_Method: (Download)

Table 87 TF coding DEGs list of HBRR1-VS-UHRR1.PossionDis_Method: (Download)

Table 88 TF coding DEGs list of HBRR2-VS-UHRR2.DEGseq_Method: (Download)

Table 89 TF coding DEGs list of HBRR2-VS-UHRR2.PossionDis_Method: (Download)

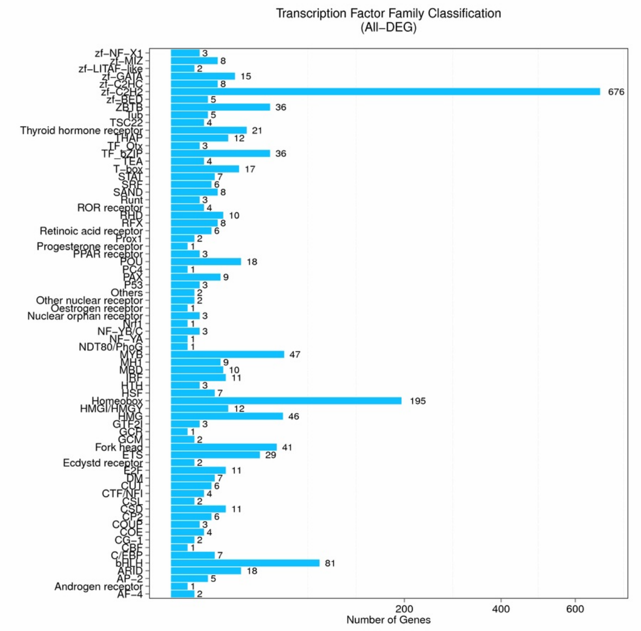


Figure 38 DEGs classification on TF family.

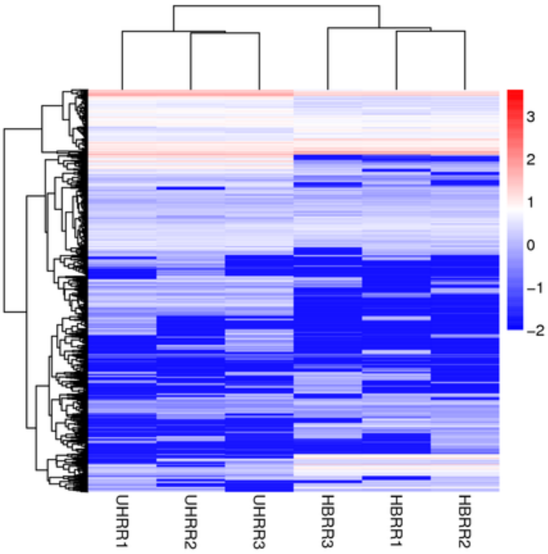


Figure 39 Expression heatmap of TF coding DEGs.

17 Protein-Protein Interaction Networks of DEG

We use STRING\[21\]database to analyze the protein and protein interaction and construct the interaction networks of DEGs. We select the top 100 interaction networks to draw the picture, see Figure40. We also provide the input files for Cytoscape network analysis directly. Cytoscape is a software platform for visualizing complex networks and integrating these with any type of attribute data.

Table 90 Protein-protein interaction result of DEGs （Download)

gene1	gene2	protein_cluster1	protein_cluster2	score
55971	6405	9606.ENSP00000005260	9606.ENSP00000002829	167
1856	10083	9606.ENSP00000005340	9606.ENSP00000005226	199
51087	56603	9606.ENSP00000007699	9606.ENSP00000001146	179
57172	10368	9606.ENSP00000009105	9606.ENSP00000005284	250

gene1: Interaction gene 1

gene2: Interaction gene 2

protein_cluster1: Protein encoded by Gene1 in the STRING database

protein_cluster2: Protein encoded by Gene2 in the STRING database

score: The larger the value, the more reliable result

Table 91 Protein-protein interaction result of HBRR-VS-UHRR.DEseq2_Method: （Download)

Table 92 Protein-protein interaction result of HBRR-VS-UHRR.EBseq_Method: （Download)

Table 93 Protein-protein interaction result of HBRR-VS-UHRR.NOIseq_Method: （Download)

Table 94 Protein-protein interaction result of HBRR1-VS-UHRR1.DEGseq_Method: （Download)

Table 95 Protein-protein interaction result of HBRR1-VS-UHRR1.PossionDis_Method: （Download)

Table 96 Protein-protein interaction result of HBRR2-VS-UHRR2.DEGseq_Method: （Download)

Table 97 Protein-protein interaction result of HBRR2-VS-UHRR2.PossionDis_Method: （Download)

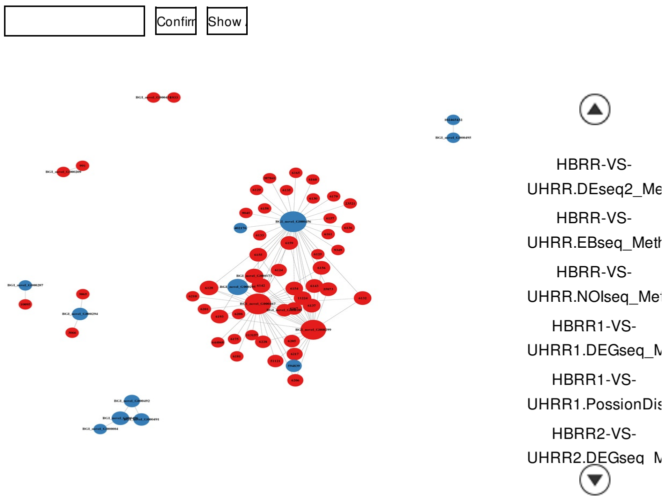


Figure 40 Protein-protein interaction network.

The red dots refer to up-regulated genes, while the blue dots refer to down-regulated genes. The size of the circle indicates the number of interactions.

18 Fungal Pathogenic Gene Prediction

Based on the PHI database, we perform the fungal pathogenic gene prediction for all the DEGs, result shown as below:

Table 98 Result of Fungal pathogenic gene prediction (Download)

GeneID	Transcript	Coverage	Identity	Function
10382	NM_001289127	51.59	80.8	...
10382	NM_001289123	51.22	80.8	...
10382	NM_001289129	52.69	80.8	...
10399	NM_006098	83.20	73.4	...

GeneID: Gene ID

Transcript: Transcript ID

Coverage: Transcript coverage

Identity: Transcript identity

Function: Accession number in PHI | Accession number in UniProt | Protein name | Taxonomy ID in NCBI | Species name | Protein function (More details, please see the help page PHI)

Table 99 DEG list of fungal pathogenic gene prediction result of HBRR-VS-UHRR.DEseq2_Method: (Download)

Table 100 DEG list of fungal pathogenic gene prediction result of HBRR-VS-UHRR.EBseq_Method: (Download)

Table 101 DEG list of fungal pathogenic gene prediction result of HBRR-VS-UHRR.NOIseq_Method: (Download)

Table 102 DEG list of fungal pathogenic gene prediction result of HBRR1-VS-UHRR1.DEGseq_Method: (Download)

Table 103 DEG list of fungal pathogenic gene prediction result of HBRR1-VS-UHRR1.PossionDis_Method: (Download)

Table 104 DEG list of fungal pathogenic gene prediction result of HBRR2-VS-UHRR2.DEGseq_Method: (Download)

Table 105 DEG list of fungal pathogenic gene prediction result of HBRR2-VS-UHRR2.PossionDis_Method:

（Download)

19 Plant Disease Resistance Gene Detection

Based on the PRG database, we perform the plant disease resistance gene analysis for all the DEGs:

Table 106 Result of Plant disease resistance gene prediction （Download)

GeneID	Name	Type	Species	Class
6130	Lus10020500	Putative_R-Genes,_predicted_from_Pythozome	Linum usitatissimum	RLP
3838	orange1.1g037562m	Putative_R-Genes,_predicted_from_Pythozome	Citrus sinensis	NL
3838	orange1.1g037562m	Putative_R-Genes,_predicted_from_Pythozome	Citrus sinensis	NL
5521	MDP0000254848	Putative_R-Genes,_predicted_from_Pythozome	Malus x domestica	RLK-GNK2

GeneID: Gene ID

Name: The name of the disease resistance genes

Type: The type of the plant resistance genes

Species: Species name

Class: The type of domain for the disease resistance genes (More details, check the help page PRG)

Table 107 DEG list of plant disease resistance gene prediction result of HBRR-VS-UHRR.DEseq2_Method: （Download)

Table 108 DEG list of plant disease resistance gene prediction result of HBRR-VS-UHRR.EBseq_Method: （Download)

Table 109 DEG list of plant disease resistance gene prediction result of HBRR-VS-UHRR.NOIseq_Method: （Download)

Table 110 DEG list of plant disease resistance gene prediction result of HBRR1-VS-UHRR1.DEGseq_Method: （Download)

Table 111 DEG list of plant disease resistance gene prediction result of HBRR1-VS-UHRR1.PossionDis_Method: （Download)

Table 112 DEG list of plant disease resistance gene prediction result of HBRR2-VS-UHRR2.DEGseq_Method: （Download)

Table 113 DEG list of plant disease resistance gene prediction result of HBRR2-VS-UHRR2.PossionDis_Method: （Download)

● Methods

1 Transcriptome Resequencing Study Process

1.1 Experiment Workflow

Total RNA sample QC

Using Agilent 2100 Bioanalyzer (Agilent RNA 6000 Nano Kit) to do the total RNA sample QC: RNA concentration, RIN value, 28S/18S and the fragment length distribution. For plant and fungi samples, we use NanoDrop^{TM}to identify the purity of the RNA samples.

Library construction

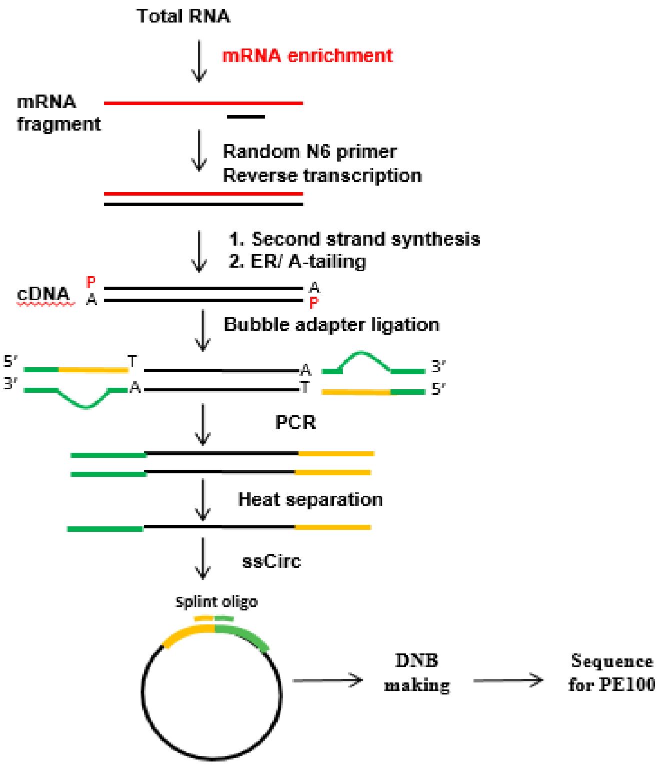


Figure 1 Transcriptome experimental workflow.

The first step in the workflow involves purifying the poly-A containing mRNA molecules using poly-T oligo-attached magnetic beads. Following purification, the mRNA is fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments are copied into first strand cDNA using reverse transcriptase and random primers. This is followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. These cDNA fragments then have the addition of a single 'A' base and subsequent ligation of the adapter. The products are then purified and enriched with PCR amplification. We then quantified the PCR yield by Qubit and pooled samples together to make a single strand DNA circle (ssDNA circle), which gave the final library. DNA nanoballs (DNBs) were generated with the ssDNA circle by rolling circle replication (RCR) to enlarge the fluorescent signals at the sequencing process. The DNBs were loaded into the patterned nanoarrays and pair-end reads of 100 bp were read through on the BGISEQ-500 platform for the following data analysis study. For this step, the BGISEQ-500 platform combines the DNA nanoball-based nanoarrays and stepwise sequencing using Combinational Probe-Anchor Synthesis Sequencing Method.

1.2 Bioinformatics Workflow

Firstly, we filter the low quality reads (More than 20% of the bases qualities are lower than 10), reads with adaptors and reads with unknown bases (N bases more than 5%) to get the clean reads. Then we map those clean reads onto reference genome, followed with novel gene prediction, SNP & INDEL calling and gene splicing detection. Finally, we identify DEGs (differentially expressed genes) between samples and do clustering analysis and functional annotations. The analysis pipeline is shown in Figure2.

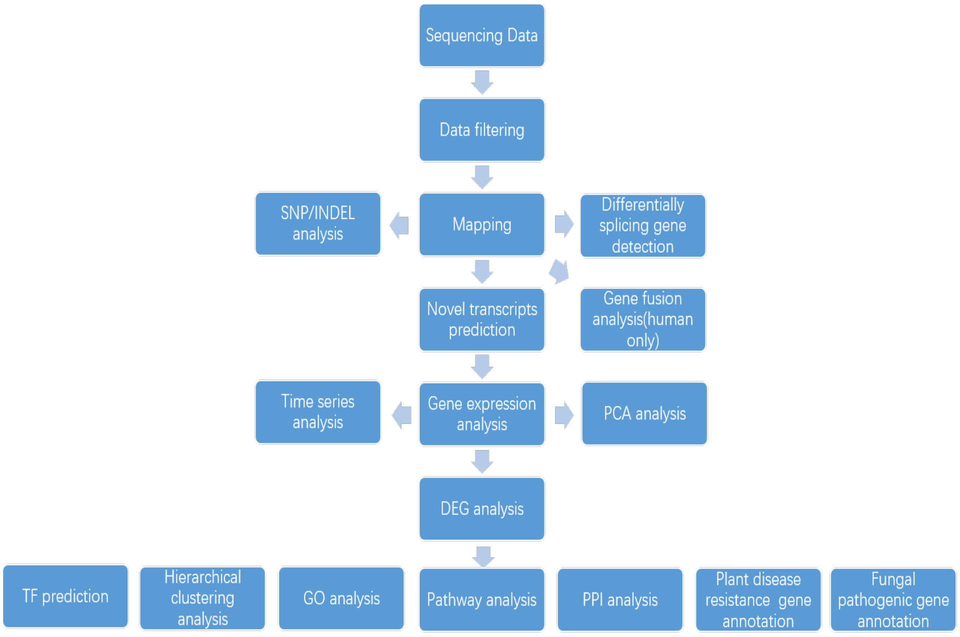

Sequencing Data
Data filtering
SNP/INDEL analysis
Mapping
Differentially splicing gene detection
Novel transcripts prediction
Gene fusion analysis(human only)
Time series analysis
Gene expression analysis
PCA analysis
DEG analysis
TF prediction
Hierarchical clustering analysis
GO analysis
Pathway analysis
PPI analysis
Plant disease resistance gene annotation
Fungal pathogenic gene annotation

Figure 2 Transcriptome Resequencing analysis pipeline.

2 Sequencing Reads Filtering

We use internal software SOAPnuke to filter reads, followed as:

1) Remove reads with adaptors;

2) Remove reads in which unknown bases(N) are more than 5%;

3) Remove low quality reads (we define the low quality read as the percentage of base which quality is lesser than 15 is greater than 20% in a read).

After filtering, the remaining reads are called "Clean Reads" and stored in FASTQ format[1](see FASTQ Format in help page FASTQ Format).

Software information:

```
SOAPnuke:
version: v1.5.2
parameters: -l 15 -q 0.2 -n 0.05 -i
website: https://github.com/BGI-flexlab/SOAPnuke
```

3 Genome Mapping

We use HISAT (Hierarchical Indexing for Spliced Alignment of Transcripts) to do the mapping step. For HISAT which is much faster, sensitive and high accuracy analysis software. The mapping method is shown as Figure3[2].

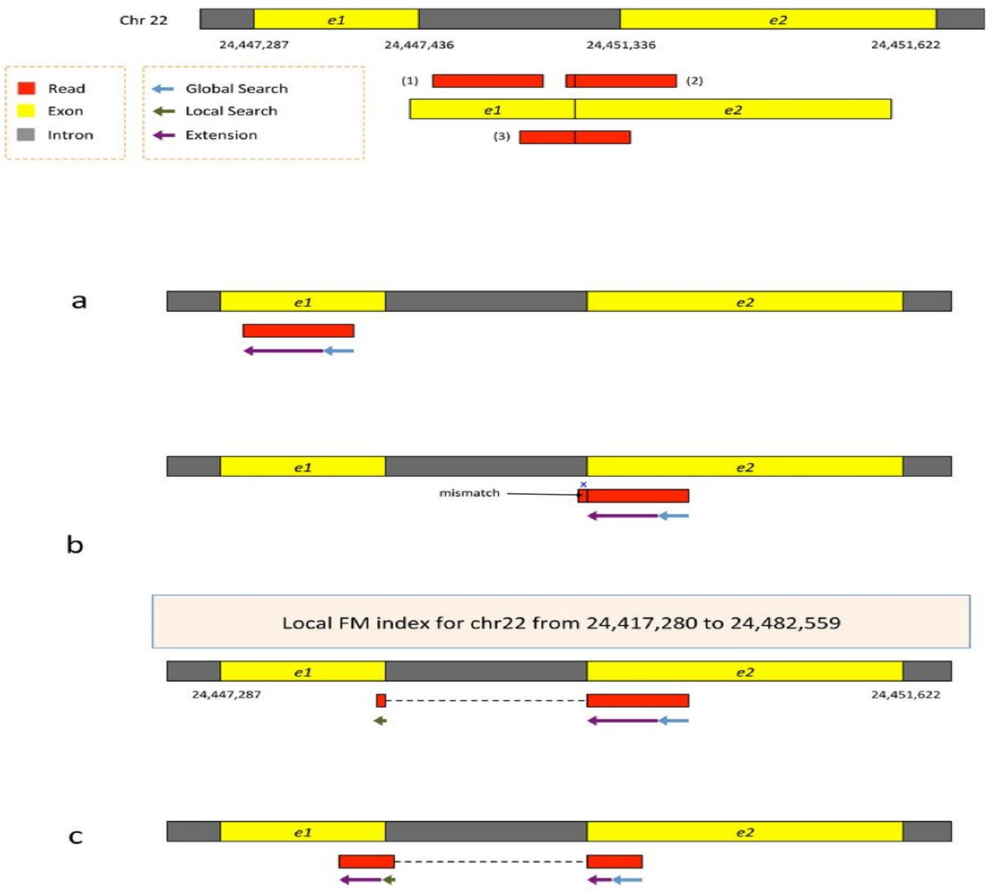


Figure 3 HISAT mapping demo show.

Software information:

```
HISAT2:
Version: v2.0.4
Parameters: --phred64 --sensitive --no-discordant --no-mixed -I 1 -X 1000
Website: http://www.ccb.jhu.edu/software/hisat
```

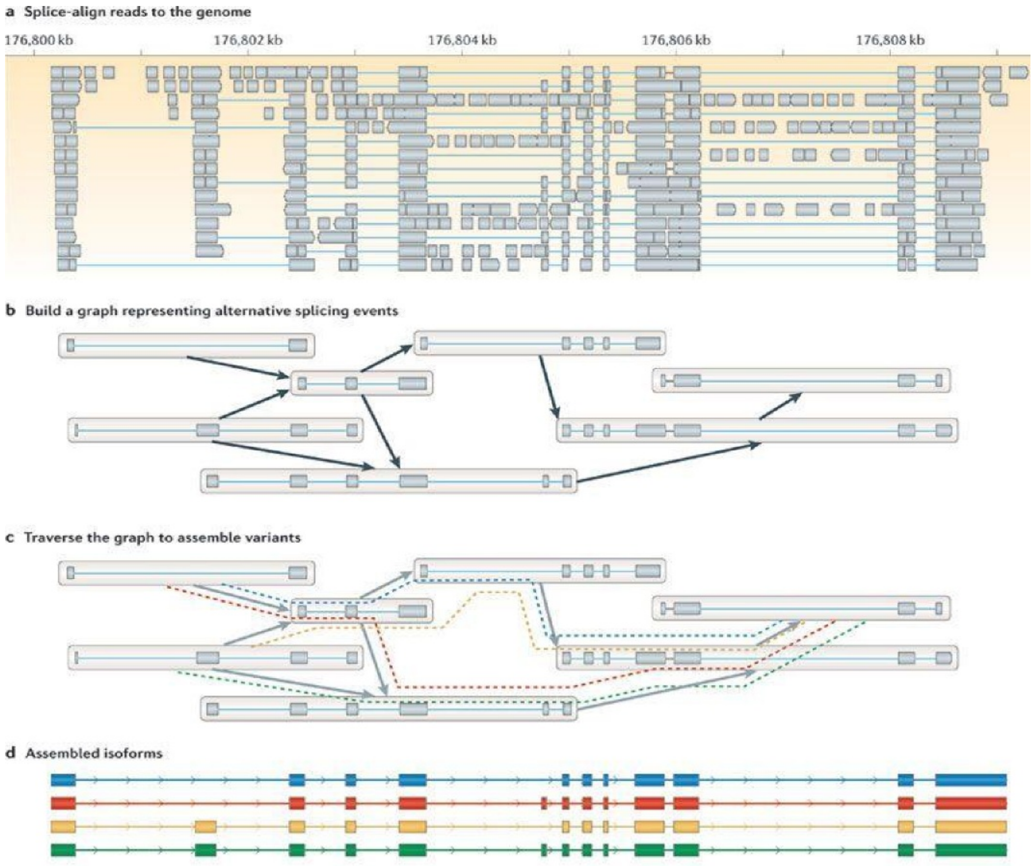
4 Novel Transcript Prediction

We use StringTie[3] to reconstruct transcripts, and use Cuffcompare（Cufflinks [4] tools） to compare reconstructed transcripts to reference annotation, after that, we select 'u','i','o','j' class code types as novel transcripts, class code type details is shown as Table1. And then, we use CPC [5] to predict coding potential of novel transcripts, then we merge coding novel transcripts with reference transcripts to get a complete reference, and downstream analysis will base on this reference. StringTie is an much faster and accurate software for transcriptome assembly, compared to Cufflinks software [2].The pipeline for transcriptome assembly based on reference please see Figure4[6].

Table 1 Explanation of class code. （Download）

Class_Code	Explanation
u	Unknown, intergenic transcript.
i	A transfrag falling entirely within a reference intron.
o	Generic exonic overlap with a reference transcript.
j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript.

Please refer to Cufflinks Websitefor class code details.



a Splice-align reads to the genome

b Build a graph representing alternative splicing events

c Traverse the graph to assemble variants

d Assembled isoforms

Nature Reviews | Genetics

Figure 4 Transcriptome assembly based on reference.

Software information:

```
StringTie:
Version: v1.0.4
Parameters: -f 0.3 -j 3 -c 5 -g 100 -s 10000 -p 8
Website: http://ccb.jhu.edu/software/stringtie
Cufflinks:
Version: v2.2.1
Parameters: -p 12
Website: http://cole-trapnell-lab.github.io/cufflinks
CPC:
Version: v0.9-r2
Parameters: Default
```

```
Website: http://cpc.cbi.pku.edu.cn
```

5 SNP and INDEL Detection

With genome mapping result, we use GATK\[7\]to call SNP and INDEL for each sample. After filtering out the unreliable sites, we get the final SNP and INDEL in VCF format.

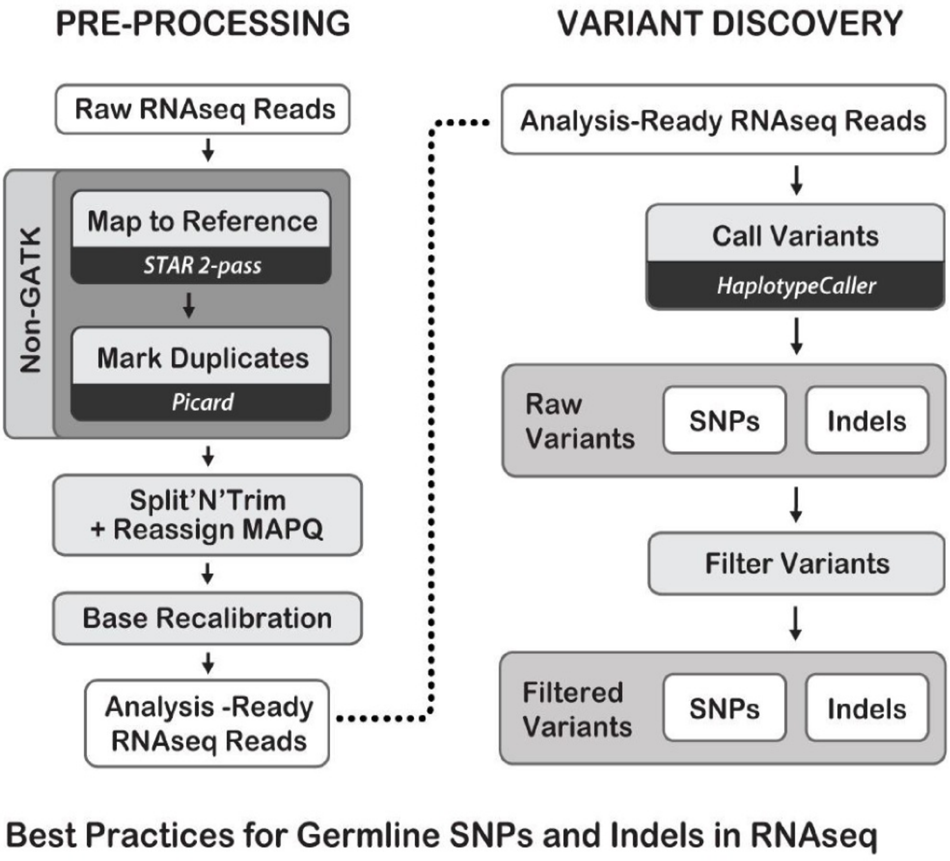


Figure 5 Pipeline for calling SNP and INDEL in RNAseq.

6 Gene Fusion Detection

A fusion gene is a hybrid gene formed from two previously separate genes as result of translocation, interstitial deletion or chromosomal inversion. Figure6demonstrates these three types of how fusion gene forms. Often, fusion genes are oncogenes that cause cancer, and most of them are found from hematological cancers, sarcomas and prostate cancer. Oncogenic fusion genes may lead to a gene product with a new or different function from the two fusion partners. Alternatively, a proto-oncogene is fused to a strong promoter, and thereby the oncogenic function is set to function by an upregulation caused by the strong promoter of the upstream fusion partner.

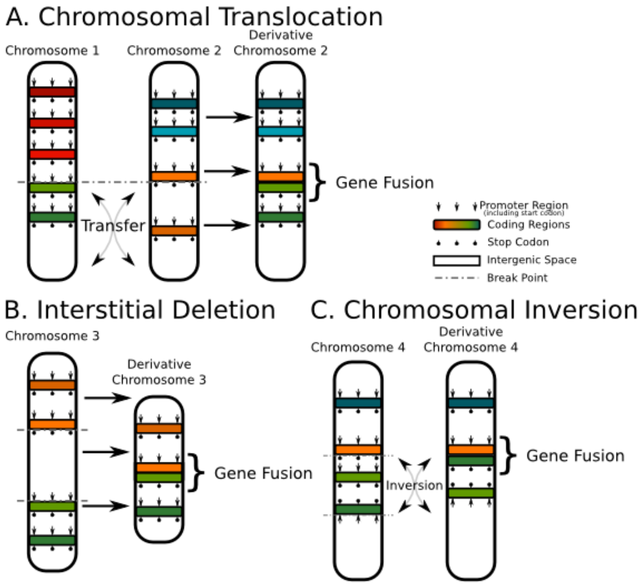


Figure 6 Gene fusion types.

A)chromosomal translocation B)Interstitial Deletion C)Chromosomal Inversion.

Presence of certain chromosomal aberrations and their resulting fusion genes is commonly used within cancer diagnostics for precise diagnosis. Chromosome banding analysis, fluorescence In Situ hybridization (FISH), and reverse transcription polymerase chain reaction (RT- PCR) are common methods employed at diagnostic laboratories. These methods all have their distinct shortcomings due to the very complex nature of cancer genomes. Recent developments such as high throughput sequencing and custom DNA microarrays bear promise of introduction of more efficient methods.

SOAPfuse[8]is a novel tool to identify fusion transcripts from paired-end RNA-Seq data. The tool 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 is much more accurate, sensitive and efficient for fusion discovery and consumed less computing resources. Furthermore, SOAPfuse provides predicted junction sequences of fusion transcripts and schematic diagrams of fusion events, which greatly improve the efficiency of fusions detection and strongly promote disease research, especially cancer research. These advantages mentioned above is significant for clinical molecular typing and new anti-tumor drug development. Figure7 is the workflow of SOAPfuse.

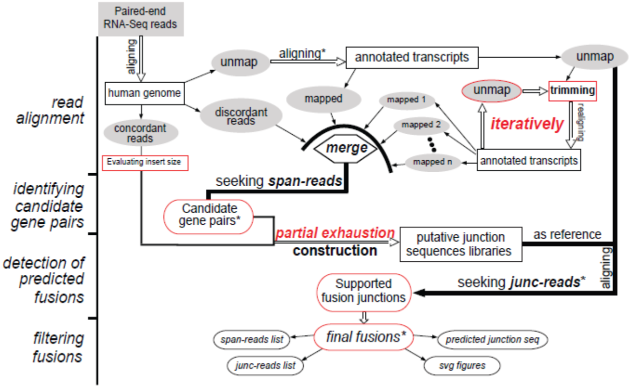


Figure 7 SOAPfuse workflow.

Software information:

```
SOAPfuse:
Version: v1.18
Parameters: -fs 1 -tp 9 -fm
Website: http://soap.genomics.org.cn/soapfuse.html
```

7 Differentially Splicing Gene Detection

It is important to distinguish differential isoform relative abundance, from differential isoform expression. Changes in relative abundance of isoforms, regardless of the expression change, indicate a splicing-related mechanism. On the other hand, there can be measurable changes in the expression of isoforms across samples, without necessarily changing the relative abundance, which possibly indicates a transcription-related mechanism. We use rMATS\[9\] to detect differentially splicing gene(that is differential isoform relative abundance between samples), a computational tool to detect differential alternative splicing events from RNA-Seq data, it calculates the inclusion isoform and skipping isofrom, shown as Figure8. The statistical model of MATS calculates the P-value and false discovery rate (FDR) that the difference in the isoform ratio of a gene between two conditions, in our project, gene that with FDR <= 0.05 is defined as significant differentially splicing gene (DSG).

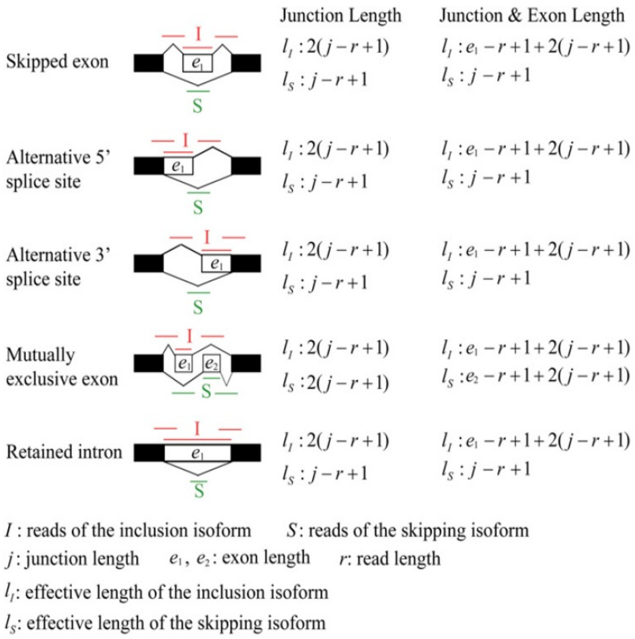

Junction Length
Junction & Exon Length
Skipped exon
I
S
e_1
$l_I : 2(j-r+1)$
$l_S : j-r+1$
$l_I : e_1 - r + 1 + 2(j-r+1)$
$l_S : j-r+1$
Alternative 5' splice site
$l_I : 2(j-r+1)$
$l_S : j-r+1$
$l_I : e_1 - r + 1 + 2(j-r+1)$
$l_S : j-r+1$
Alternative 3' splice site
$l_I : 2(j-r+1)$
$l_S : j-r+1$
$l_I : e_1 - r + 1 + 2(j-r+1)$
$l_S : j-r+1$
Mutually exclusive exon
e_2
$l_I : 2(j-r+1)$
$l_S : 2(j-r+1)$
$l_I : e_1 - r + 1 + 2(j-r+1)$
$l_S : e_2 - r + 1 + 2(j-r+1)$
Retained intron
$l_I : 2(j-r+1)$
$l_S : j-r+1$
$l_I : e_1 - r + 1 + 2(j-r+1)$
$l_S : j-r+1$
I : reads of the inclusion isoform
S: reads of the skipping isoform
j: junction length
e_1, e_2: exon length
r: read length
l_I: effective length of the inclusion isoform
l_S: effective length of the skipping isoform

Figure 8 Relative abundance calculation of differential isoforms.

Software information:

```
rMATS:
Version: v3.0.9
Parameters: -analysis U -t paired -a 8
Website: http://rnaseq-mats.sourceforge.net
```

8 Gene Expression Analysis

We mapped clean reads to reference using Bowtie2\[10\], and then calculate gene expression level with RSEM \[11\]. RSEM is a software package for estimating gene and isoform expression levels from RNA-Seq data. Then, wecalculate pearson correlation between all samples using cor, perform hierarchical clustering between all samples using hclust, perform PCA analysis with all samples using princomp, and draw the diagrams with ggplot2 with fuctions of R.

Software information:

```
 Bowtie2 :
Version: v2.2.5
Parameters: -q --phred64 --sensitive --dpad 0 --gbar 99999999 --mp 1,1 --np 1
--score-min L,0,-0.1 -I 1 -X 1000 --no-mixed --no-discordant  -p 1 -k 200
Website: http://bowtie-bio.sourceforge.net/ Bowtie2 /index.shtml
 RSEM :
Version: v1.2.12
Parameters: default
Website: http://deweylab.biostat.wisc.edu/ RSEM
```

9 Circos Diagram

Circos is a software package for visualizing data and information\[12\]. We visualize SNP , INDEL , gene expression and gene fusion (Only for human samples) result based on Circos diagram.

Software information:

```
Circos:
Version: v0.69
Website: http://www.circos.ca
```

10 Time Series Analysis

Clustering analysis is a common method for gene expression analysis. There are two types of clustering: Hard clustering and soft clustering. Vast majority of clustering algorithms applied produce hard partitions of the data, i.e. each gene or protein is assigned exactly to one cluster. Hard clustering is favourable if clusters are well separated. However, this is generally not the case for gene expression time-course data, where gene/protein clusters frequently overlap. Additionally, hard clustering algorithms are often highly sensitive to noise. We use Mfuzz[13]analysis software to do the soft cluster which is more suitable for gene expression data. It is more noise robust and a priori pre-filtering of genes/proteins can be avoided. Moreover, it can also carry out more targeted search for regulatory elements.

Software information:

```
Mfuzz:
Version: v2.34.0
Parameters: -c 12 -m 1.25
Website: http://mfuzz.sysbiolab.eu
```

11 DEG Detection

We detect DEGs with DEGseq, DEseq2, EBseq, NOIseq 和 PossionDis as requested.DEGseq is based on the poisson distribution, performed as described at Wang L, Feng Z, Wang X, et al.\[14\] DEseq2 is based on the negative binomial distribution, performed as described at Michael I, et al.\[15\] EBseq is based on mpirical Bayesian model, performed as described at Leng N, et al.\[16\]NOIseq is based on noisy distribution model, performed as described at Tarazona S, et al.\[17\] PossionDis is based on the poisson distribution, performed as described at Audic S, et al.\[18\]

Software information:

```
DEGseq:
Parameters: Fold Change >= 4.00 and Adjusted Pvalue <= 0.001
DEseq2:
Parameters: Fold Change >= 2.00 and Adjusted Pvalue <= 0.05
EBseq:
Parameters: Fold Change >= 2.00 and Posterior Probability of being Equivalent
Expression(PPEE) <= 0.05
NOIseq:
```

```
Parameters: Fold Change >= 2.00 and Probability >= 0.8
PossionDis:
Parameters: Fold Change >= 2.00 and  FDR  <= 0.001
```

12 Hierarchical Clustering Analysis of DEG

We perform hierarchical clustering for DEGs using pheatmap, a function of R. For cluster more than two groups, we perform the intersection and union DEGs between them, respectively.

13 Gene Ontology Analysis of DEG

With the GO annotation result, we classify DEGs according to official classification, and we also perfrom GO functional enrichment using phyper, a function of R. The pvalue calculating formula in hypergeometric test is:

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

See wiki for details https://en.wikipedia.org/wiki/Hypergeometric_distribution。

Then we calculate false discovery rate（FDR ）for each pvalue, in general, the terms which FDR not larger than 0.01 are defined as significant enriched.

14 Pathway Analysis of DEG

With the KEGG annotation result, we classify DEGs according to official classification, and we also perform pathway functional enrichment using phyper, a function of R. The pvalue calculating formula in hypergeometric test is:

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

See wiki for details https://en.wikipedia.org/wiki/Hypergeometric_distribution.

Then we calculate false discovery rate（FDR ）for each pvalue, in general, the terms which FDR not larger than 0.01 are defined as significant enriched.

15 Transcription Factor Prediction of DEG

We use getorf to find ORF of each DEG . For plants, we align ORF to TF domains (from PlntfDB) using hmmsearch\[19\]. For animals, we align ORF to animal TF database (AnimalTFDB) using DIAMOND\[20\].

Software information:

```
PlntfDB:
Version: v23.0
Website: http://plntfdb.bio.uni-potsdam.de/v3.0/
```

```
AnimalTFDB:
Version: v2.0
Website: http://www.bioguo.org/AnimalTFDB/
getorf:
Version: EMBOSS:6.5.7.0
Parameters: -minsize 150
Website: http://www.bioinformatics.nl/cgi-bin/emboss/help/getorf
hmmseach:
Version: v3.0
Parameters: default
Website: http://hmmer.org
DIAMOND:
Version: v0.8.31
Parameters: --more-sensitive --evalue 1e-5
Website:https://github.com/bbuchfink/diamond
```

16 PPI Analysis of DEG

We use DIAMOND\[20\]to map the DEGs to the STRING\[21\]database to obtain the interaction between DEG -encoded proteins using homology with known proteins. We select the top 100 interaction networks to draw the picture, for the entire interaction result we provide an input file that can be imported directly into Cytoscape for network analysis. Cytoscape is a software for complex network analysis and visualization. For more information, refer to the official documentation.

```
STRING:
Version: v10
Website: http://string-db.org/
DIAMOND:
Version: v0.8.31
Parameters(Running): --evalue  1e-5 --outfmt 6 --max-target-seqs 1 --more-
sensitive
Parameters(Selecting): query coverage >= 50%、identity >= 40%
Website: https://github.com/bbuchfink/diamond
```

17 Fungal Pathogenic Gene Prediction

We use BLAST\[22\]or DIAMOND\[20\]to map the DEGs to the PHI-base\[23\]database to detect the fungal pathogenic genes based on the query coverage and identity requirement\[24\].

```
PHI-base:
Version: v4.1
Website: http://www.phi-base.org/
BLAST+:
Version: v2.5.0
Parameters(Running): -evalue 1e-5 -outfmt 6 -max_target_seqs 1
Parameters(Selecting): query coverage >= 50%、identity >= 40%
Website: https://blast.ncbi.nlm.nih.gov/
DIAMOND:
```

```
Version: v0.8.31
Parameters(Running): --evalue  1e-5 --outfmt 6 --max-target-seqs 1 --more-
sensitive
Parameters(Selecting): query coverage >= 50%、identity >= 40%
Website: https://github.com/bbuchfink/diamond
```

18 Plant Disease Resistant Gene Prediction

We use BLAST\[22\]or DIAMOND \[20\]to map the DEGs to the PRGdb\[25\]database to detect the plant disease resistant genes based on the query coverage and identity requirement\[26\].

```
PRGdb:
Version: v2.0
Website: http://prgdb.crg.eu/
blast+:
Version: v2.5.0
Parameters(Running): -evalue 1e-5 -outfmt 6 -max_target_seqs 1
Parameters(Selecting): query coverage >= 50%, identity >= 40%
Website: https://blast.ncbi.nlm.nih.gov/
DIAMOND:
Version: v0.8.31
Parameters(Running): --evalue  1e-5 --outfmt 6 --max-target-seqs 1 --more-
sensitive
Parameters(Selecting): query coverage >= 50%, identity >= 40%
Website: https://github.com/bbuchfink/diamond
```

● Help

1 FASTQ Format

The original image data is transferred into sequence data via base calling , which is defined as raw data or raw reads and saved as FASTQ file. Those FASTQ files are the original data provided for users, including detailed read sequences and the read quality information. In each FASTQ file, every read is described by four lines, listed as follows:

```
@CL100012105L2C001R003_48/1
CAGCCAGCCAGTGGCAGTGCGAGGTGGAGGAGGCAAACAAGTGTAATCGTTTATACATACCCACAGGTGTTAAAA
AGTAATCGAAGTACGAAGAGGAACA
+
FGFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFGFGFFEFFFGFFFFFFFFFFFFGF:FFFFCFFFFFGFGG
FFEGFFFFFFFFFFFFFFF>FFEFF
```

The first line starts with an "@", /1 means read1, /2 means read2.

The second line is the reads sequence.

The third line starts with an "+".

The forth line is sequencing quality value, in which each letter corresponds to the base in line 2.

For details , Pls check the website: http://en.wikipedia.org/wiki/FASTQ_format .

2 Relationship between sequencing error rate and sequencing quality value

The base quality is equal to ASCII value of the character in line 4 minus 64 (we call the quality system is Phred+64), e.g. the ASCII value of c is 99, then its base quality value is 35. Table1 demonstrates the relationship between sequencing error rate and the sequencing quality value. Specifically, if the sequencing error rate is denoted as E and base quality value is denoted as Q, the relationship is as following formula:

$$\mathbf{SQ} = -\mathbf{10} \times (\mathbf{log}\frac{\mathbf{E}}{\mathbf{1} - \mathbf{E}})/(\mathbf{log10})$$

$$\mathbf{E} = \frac{\mathbf{Y}}{\mathbf{1} + \mathbf{Y}}$$

$$\mathbf{Y} = \frac{\mathbf{SQ}}{e^{-10 \times log10}}$$

Table 1 Relationship between sequencing error rate and sequencing quality value. (Download)

Sequencing Error Rate(%)	Sequencing Quality Value	Character(Phred+64)	Character(Phred+33)
1.00	20	T	5
0.10	30	^	?
0.01	40	h	I

```
Note: The quality value system of BGISEQ-500 is Phred+33. To considerate the
compatibility of the subsequent analysis software, we will convert the
quality system from Phred+33 to Phred+64.
```

3 VCF format

Variant Call Format (VCF) is a flexible and extendable format for variation data such as single nucleotide polymorphism (SNP), insertions/deletions (INDEL), copy number variants and structural variants. See details at UCSC website http://genome.ucsc.edu/FAQ/FAQformat.html#format10.1

Table 2 VCF instruction. (Download)

Item	Description
CHROM	Name of the chromosome (or contig, scaffold, etc.)
POS	Position in chromosome
ID	If the variants existed in dbSNP database, then the ID is the dbSNP ID.
REF	The base in the reference genome
ALT	The variant base
QUAL	Quality score
FILTER	Filter details
INFO	Other information

4 Gene fusion format instruction

Gene fusion list of each sample is stored in tab-seperated text file. See http://soap.genomics.org.cn/soapfuse.html Output Files for detail.

Table 3 Gene fusion format instruction. (Download)

Item	Description
up_gene	5' fusion gene
up_chr	5' fusion gene chromosome
up_strand	5' fusion gene strand
up_Genome_pos	5' fusion gene position along chromosome
up_loc	5' fusion gene location (E means fused-at-exon edge, M means fused-at-exon region, I means fused-at-intron)
dw_gene	3' fusion gene
dw_chr	3' fusion gene chromosome
dw_strand	3' fusion gene strand
dw_Genome_pos	3' fusion gene position along chromosome
dw_loc	3' fusion gene location (E means fused-at-exon edge, M means fused-at-exon region, I means fused-at-intron)
Span_reads_num	Number of span-reads
Junc_reads_num	Number of junction reads
Fusion_Type	The type of fusion
down_fusion_part_frame-shift_or_not	whether down stream fusion partner is frame-shift or in-frame-shift

5 Differentially Splicing Gene format

Differentially Splicing Gene(DSG) result format is described in Table4.

Table 4 Differentially Splicing Gene format instruction (Download)

Field	Description	Notes
GeneID	gene identity	-
Chr	chromosome	-
Strand	strand	-
Control-IC	inclusion junction counts for Control sample, replicates are separated by comma	-

	comma	
Control-SC	skipping junction counts for Control sample, replicates are separated by comma	-
Treat-IC	inclusion junction counts for Treat sample, replicates are separated by comma	-
Treat-SC	skipping junction counts for Treat sample, replicates are separated by comma	-
Pvalue	statistical significance	-
FDR	false discovery ratio	-
longExonStart	the long exon start position on chromosome	for A3SS and A5SS event
longExonEnd	the long exon end position on chromosome	for A3SS and A5SS event
shortExonStart	the short exon start position on chromosome	for A3SS and A5SS event
shortExonEnd	the short exon end position on chromosome	for A3SS and A5SS event
flankingExonStart	the flanking exon start position on chromosome	for A3SS and A5SS event
flankingExonEnd	the flanking exon end position on chromosome	for A3SS and A5SS event
1stExonStart	the first exon start position on chromosome	for MXE event
1stExonEnd	the first exon end position on chromosome	for MXE event
2ndExonStart	the secend exon start position on chromosome	for MXE event
2ndExonEnd	the secend exon end position on chromosome	for MXE event
riExonStart	the intron-retained exon start position on chromosome	for RI event
riExonEnd	the intron-retained exon end position on chromosome	for RI event
skipExonStart	the skipped exon start position on chromosome	for SE event
skipExonEnd	the skipped exon end position on chromosome	for SE event
upstreamExonStart	the upstream exon start position on chromosome	for RI and SE event
upstreamExonEnd	the upstream exon end position on chromosome	for RI and SE event
downstreamExonStart	the downstream exon start position on chromosome	for RI and SE event
downstreamExonEnd	the downstream exon end position on chromosome	for RI and SE event
LongExonTranscripts	the transcripts that contain long exon, separated by comma	for A3SS and A5SS event
ShortExonTranscripts	the transcripts that contain short exon, separated by comma	for A3SS and A5SS event
1stExonTranscripts	the transcripts that contain first exon, separated by comma	for MXE event
2ndExonTranscripts	the transcripts that contain secend exon, separated by comma	for MXE event
RetainTranscripts	the transcripts that contain intron-retained exon, separated by comma	for RI event
AbandonTranscripts	the transcripts that exclude intron-retained exon, separated by comma	for RI event
InclusionTranscripts	the transcripts that include certain exon, seperated by comma	for SE event
SkippingTranscripts	the transcripts that exclude certain exon, seperated by comma	for SE event

6 DEG list format

The result of differentially expressed genes format is described in Table5.

Table 5 Format description of DEGs. （Download)

Field	Description
GeneID	Gene ID
Length	Gene length
Sample1-Expression	Gene expression of control sample(s)
Sample2-Expression	Gene expression of treat sample(s)
log2FoldChange(Sample2/Sample1)	log2 transformed fold change between control and treat samples
Pvalue	Statistic of pvalue(PossionDis or DEseq2 method used)
FDR	Statistic of false discovery rate(PossinoDis method used)
Padj	Statistic of adjusted pvalue(DEseq2 method used)
PPEE	Statistic of posterior probability of being equivalent expression(EBseq method used)
Probability	Statistic of probability of being DEG(NOIseq method used)
Up/Down-Regulation(Sample2/Sample1)	Flags indicate up-regulated DEG(Up) or down-regulated DEG(Down) or non-DEG(*)
...	Gene symbol, GO, KEGG and NR annotation

7 Cluster list format

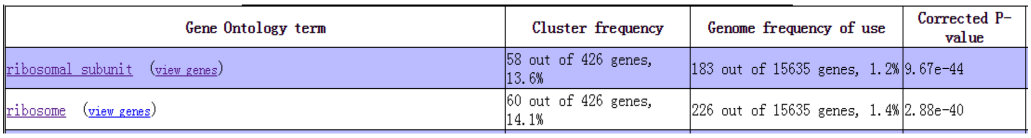
The format of cluster list is described as Table6.

Table 6 Format description of DEGs clustering list. （Download)

Field	Description
Gene	Gene ID
A-VS-B	log2FoldChange of A-VS-B
C-VS-D	log2FoldChange of C-VS-D
...	...
...	Gene symbol, GO, Kegg and NR annotation

8 How to read DEG GO enrichment analysis result

Make sure that the computer has installed java and use IE browser to open GOView.html. The left navigation includes three types of GO terms for each control-treatment pairwise (C: cellular component, P: biological process, F: molecular function). Click one of them, the enriched GO terms result will be listed as Figure2.

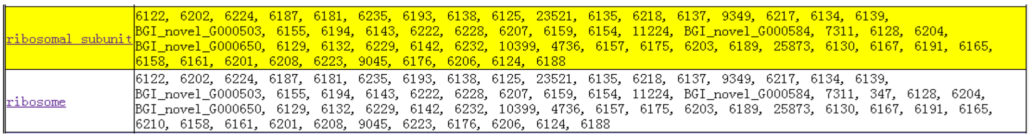


Gene Ontology term	Cluster frequency	Genome frequency of use	Corrected P-value
ribosomal subunit (view genes)	58 out of 426 genes, 13.6%	183 out of 15635 genes, 1.2%	9.67e-44
ribosome (view genes)	60 out of 426 genes, 14.1%	226 out of 15635 genes, 1.4%	2.88e-40

Figure 2 Significantly enriched GO terms in DEGs.

Column 1 is GO term name. Column 2 is the ratio of DEGs enriched to this GO term. Column 3 is the ratio of genes enriched to this GO term in background database. Column 4 is Corrected P-value which indicates the degree of enrichment and the smaller Corrected P-value, the more significantly DEGs enriched to this GO term. The result list has been sorted by Corrected P-value.

Click the term name 'ribosomal subunit' in Figure2, you can go to http://amigo.geneontology.org/amigo for more information when the computer is Internet-connected. Click 'view genes' in Figure2, you can get gene IDs that enriched to this GO term as Figure3.



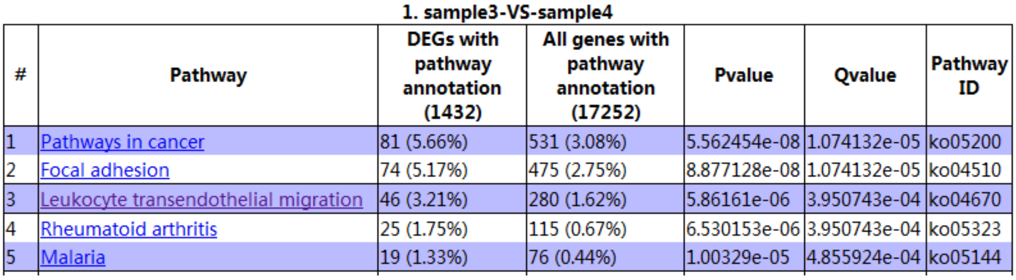
ribosomal subunit	6122, 6202, 6224, 6187, 6181, 6235, 6193, 6138, 6125, 23521, 6135, 6218, 6137, 9349, 6217, 6134, 6139, BGI_novel_G000503, 6155, 6194, 6143, 6222, 6228, 6207, 6159, 6154, 11224, BGI_novel_G000584, 7311, 6128, 6204, BGI_novel_G000650, 6129, 6132, 6229, 6142, 6232, 10399, 4736, 6157, 6175, 6203, 6189, 25873, 6130, 6167, 6191, 6165, 6158, 6161, 6201, 6208, 6223, 9045, 6176, 6206, 6124, 6188
ribosome	6122, 6202, 6224, 6187, 6181, 6235, 6193, 6138, 6125, 23521, 6135, 6218, 6137, 9349, 6217, 6134, 6139, BGI_novel_G000503, 6155, 6194, 6143, 6222, 6228, 6207, 6159, 6154, 11224, BGI_novel_G000584, 7311, 347, 6128, 6204, BGI_novel_G000650, 6129, 6132, 6229, 6142, 6232, 10399, 4736, 6157, 6175, 6203, 6189, 25873, 6130, 6167, 6191, 6165, 6210, 6158, 6161, 6201, 6208, 9045, 6223, 6176, 6206, 6124, 6188

Figure 3 Gene ID list related to GO terms.

In the example, the following DEGs were annotated to the term 'ribosomal subunit': 6122, 6202, 6224, 6187, 6181, 6235, 6193, 6138, 6125, 23521, 6135, 6218, 6137, 9349, 6217, 6134, 6139, BGI_novel_G000503, 6155, 6194, 6143, 6222, 6228, 6207, 6159, 6154, 11224, BGI_novel_G000584, 7311, 6128, 6204, BGI_novel_G000650, 6129, 6132, 6229, 6142, 6232, 10399, 4736, 6157, 6175, 6203, 6189, 25873, 6130, 6167, 6191, 6165, 6158, 6161, 6201, 6208, 6223, 9045, 6176, 6206, 6124, 6188.

9 How to read DEG pathway enrichment analysis result

Open html report for pathway enrichment result and the enriched KEGG pathways will be listed as Figure4.



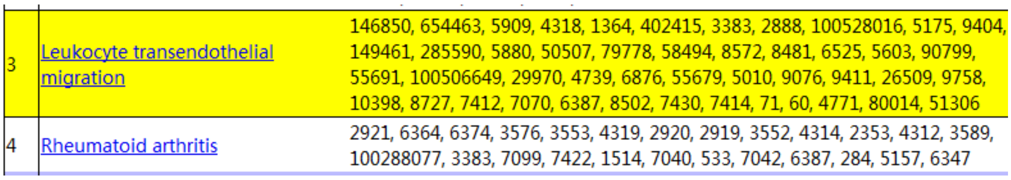
1. sample3-VS-sample4

#	Pathway	DEGs with pathway annotation (1432)	All genes with pathway annotation (17252)	Pvalue	Qvalue	Pathway ID
1	Pathways in cancer	81 (5.66%)	531 (3.08%)	5.562454e-08	1.074132e-05	ko05200
2	Focal adhesion	74 (5.17%)	475 (2.75%)	8.877128e-08	1.074132e-05	ko04510
3	Leukocyte transendothelial migration	46 (3.21%)	280 (1.62%)	5.86161e-06	3.950743e-04	ko04670
4	Rheumatoid arthritis	25 (1.75%)	115 (0.67%)	6.530153e-06	3.950743e-04	ko05323
5	Malaria	19 (1.33%)	76 (0.44%)	1.00329e-05	4.855924e-04	ko05144

Figure 4 Pathway enrichment analysis of DEGs.

Column 1 is ordinal number. Column 2 is pathway name. Column 3 is the ratio of DEGs enriched to this pathway. Column 4 is the ratio of genes enriched to this pathway in background database. Pvalue and Qvalue are both values that indicate the degree of enrichment and Qvalue is corrected Pvalue. The smaller they are, the more significantly DEGs enriched to this pathway. The result list has been sorted by Qvalue. The last column pathway ID is corresponding to pathway name.

Click pathway name 'Leukocyte transendothelial migration' in Figure4, you can get gene IDs that enriched to it as Figure5.

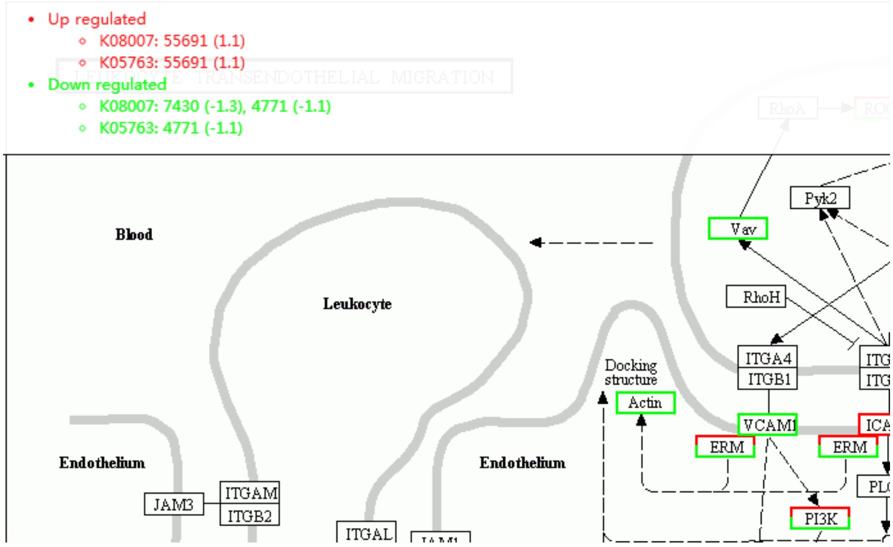


3	Leukocyte transendothelial migration	146850, 654463, 5909, 4318, 1364, 402415, 3383, 2888, 100528016, 5175, 9404, 149461, 285590, 5880, 50507, 79778, 58494, 8572, 8481, 6525, 5603, 90799, 55691, 100506649, 29970, 4739, 6876, 55679, 5010, 9076, 9411, 26509, 9758, 10398, 8727, 7412, 7070, 6387, 8502, 7430, 7414, 71, 60, 4771, 80014, 51306
4	Rheumatoid arthritis	2921, 6364, 6374, 3576, 3553, 4319, 2920, 2919, 3552, 4314, 2353, 4312, 3589, 100288077, 3383, 7099, 7422, 1514, 7040, 533, 7042, 6387, 284, 5157, 6347

Figure 5 Gene ID list related to pathway.

There are 46 DEGs enriched to the pathway 'Leukocyte transendothelial migration'.

Furthermore, detecting the most significant pathways, the enrichment analysis of DEG pathway significance, allows us to see detailed pathway information in KEGG database. For example, clicking the hyperlink on 'Leukocyte transendothelial migration' in Figure5 will get detailed information as shown in Figure6.



- Up regulated
  - K08007: 55691 (1.1)
  - K05763: 55691 (1.1)
- Down regulated
  - K08007: 7430 (-1.3), 4771 (-1.1)
  - K05763: 4771 (-1.1)

Figure 6 An example of KEGG pathway of 'Leukocyte transendothelial migration'.

Up-regulated genes are marked with red borders and down-regulated genes with green borders. Non-change genes are marked with black borders. When mouse hover on border with red or green, the related DEGs appear on the top left. Clicking gene name in the figure, the page will redirect to KEGG website if the computer is Internet-connected.

10 TF

In molecular biology and genetics, a transcription factor (sometimes called a sequence-specific DNA-binding factor) is a protein that binds to specific DNA sequences, thereby controlling the rate of transcription of genetic information from DNA to messenger RNA. Transcription factors perform this function alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes. See wiki for detail https://en.wikipedia.org/wiki/Transcription_factor.

Table 7 TF format description （Download)

Item	Description
GeneID	Gene ID
Length	Gene length
Sample1-Expression	Expression level of sample1
Sample2-Expression	Expression level of sample2
log2FoldChange(Sample2/Sample1)	log2 transformed fold change between sample1 and sample2
up/down	Up regulation / down regulation
Transcripts	Transcripts
TF_family	TF family
Included_domain	Included domain (for plant)
Excluded_domain Excluded domain (for plant)	
PlntfDB_link/AnimalDB_link	Linkage for the database
...	Gene symbol, GO, Kegg and NR annotation

11 PHI

The protein function description details are shown as below:

loss of pathogenicity: The deletion mutant of this gene can’t cause the disease, indicating that the gene determines the pathogen pathogenicity.

reduced virulence: The pathogenicity of this gene deletion mutant is reduced, indicating that the gene determines the pathogenicity of the pathogen.

increased virulence (Hypervirulence): This gene determines the pathogenicity of pathogens.

effector (plant avirulence determinant): Currently a plant pathogen specific term which was previously known as an avirulence gene. An effector gene is required for the direct or indirect recognition of a pathogen only in resistant host genotypes which possess the corresponding disease resistance gene. Positive recognition leads to activation of plant defences and the pathogen fails to cause disease. Note some effector genes are required to cause disease on susceptible hosts but most are not.

enhanced antagonism: The gene affects the interaction of pathogens and plants, and the deletion mutants of the gene can cause plant disease.

unaffected pathogenicity: The gene does not result to the pathogenicity of pathogens.

lethal: The gene is necessary for the survival of the strain.

chemistry target: resistance to chemical: The mutant of the gene lead to the resistance to the chemistry.

chemistry target: sensitivity to chemical: The mutant of the gene lead to the sensitivity to the chemistry.

PHI format description see as Table8。

Table 8 PHI format description （Download)

Item	Description
GeneID	Gene ID
Length	Gene length
control-Expression	Gene expression level of control
treatment-Expression	Gene expression level of treatment
log2FoldChange(treatment/control)	log2 transformed fold change between control and treatment
Up/Down-Regulation	Up regulation / down regulation
Transcript	Transcript ID
Coverage	Transcript coverage
Identity	Transcript identity
Function	Accession number in PHI | Accession number in UniProt | Protein name | Taxonomy ID in NCBI | Species name | Protein function

12 PRG

The type of domain of disease resistant genes in PRG database.

CNL: The combination of coiled-coil domain, nucleotide binding site and leucine-rich repeat, which is short for CC-NB-LRR.

TNL: The combination of Toll-interleukin receptor-like domain, nucleotide binding site and leucine-rich repeat, which is short for TIR-NB-LRR.

RLP: The combination of receptor serine-threonine kinase-like domain and extracellular leucine-rich repeat, which is short for ser/thr-LRR.

RLK: The combination of kinase domain and extracellular leucine-rich repeat, which is short for Kin-LRR.

Others: Other types.

For more information pls view the website: http://prgdb.crg.eu/wiki/Category:Classes 。

PRG format description see as Table9。

Table 9 PRG format description （Download)

Item	Descritption
GeneID	Gene ID
Length	Gene length
control-Expression	Gene expression level of control
treatment-Expression	Gene expression level of treatment
log2FoldChange(treatment/control)	log2 transformed fold change between control and treatment
Up/Down-Regulation	Up regulation / down regulation
Transcript	Transcript ID
Coverage	Transcript coverage
Identity	Transcript identity
PRGID	PRG ID
Name	The name of the disease resistance genes
Type	The type of the disease resistance genes
Species	Species name
Class	The type of domain for the disease resistance genes
GenBank_ID	GenBank ID
GenBank_Locus	GenBank locus ID
Description	Gene description

● FAQs

1. How to upload the RNA data onto NCBI?

See guidance on website: https://www.ncbi.nlm.nih.gov/guide/howto/submit-sequence-data/

2. What is Q20 and Q30? How about the quality of my data?

Q scores are defined as a property that is logarithmically related to the base calling error probabilities $(P)^2$, $Q = -10\log_{10}P$. If Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times. This means that the base call accuracy (i.e., the probability of a correct base call) is 99.9%. If base call accuracy of 99% (Q20) will have an incorrect base call probability of 1 in 100. Q20 represents the percentage of the bases which Q score are more than 20. Q30 represents the percentage of the bases which Q score are more than 30. For example, if we sequenced 1Gb data in total, the Q score of 0.9Gb bases is no less than 20 , then we can say the Q20 is 90%. Normally, for NGS sequencing, Q20≥90%, Q30≥80% can be defined as good quality data.

3. What's the differentce between FPKM and RPKM?

RPKM stands for Reads Per Kilobase of transcript per Million mapped reads. In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. RPKM=(1000000*C)/(N*L/1000), C represents the amount of reads which mapped to the specific transcripts, N represents the amount of reads which mapped to any transcripts. L represents the base amount of the specific transcripts.

FPKM stands for Fragments Per Kilobase of transcript per Million mapped reads. FPKM=(1000000*C)/(N*L/1000), C represents the amount of fragment which mapped to the specific transcripts, N represents the amount of fragment which mapped to any transcripts. L represents the base amount of the specific transcripts.

In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. Paired-end RNA-Seq experiments produce two reads per fragment, but that doesn't necessarily mean that both reads will be mappable. For example, the second read is of poor quality. If we were to count reads rather than fragments, we might double-count some fragments but not others, leading to a skewed expression value. Thus, FPKM is calculated by counting fragments, not reads.

4. Once I get the sequencing data, how can I open those files? How can I do the analysis for selected the interested genes?

You can use EditPlus to open the files. Such as water channel protein, you can search for the genes according to the keywords (AQP, aquaporin, etc.). After obtaining the gene ID, you can search for the gene ID in the sequence result file to get the gene sequence.

5. How to select the genes for qPCR validation? How do we evaluate the validation result?

For RNA-Seq, thousands of differentially expressed genes are identified. We need to select the genes for qPCR validation by our own research interests. Normally, we need to select no less than 20 genes which are highly expressed and significantly differentially expressed for validation.

6. Do we need to do the biological replicates when doing the RNA seq? How many replicates per samples is recommended?

Yes, we recommend 3 biological replicates per sample for RNA seq which is more sufficient for bioinformatics analysis.

7. For library construction, why we fragment the RNA instead of cDNA?

Pls refer to paper 'RNA-Seq: a revolutionary tool for Transcriptomics'.

References

[1] Cock, P. J. et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38, 1767-1771 (2010).

[2] Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357-360 (2015).

[3] Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290-295 (2015).

[4] Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562-578 (2012).

[5] Kong, L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 35, W345-W349 (2007).

[6] Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671-682 (2011).

[7] McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303 (2010).

[8] Jia, W. et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol. 14, R12 (2013).

[9] Shen, S. et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl Acad. Sci. USA 111, E5593-E5601 (2014).

[10] Langmead, B. et al. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357-359 (2012).

[11] Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).

[12] Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639-1645 (2009).

[13] Kumar, L. & Futschik, M. E. Mfuzz: a software package for soft clustering of microarray data. Bioinformation 2, 5-7 (2007).

[14] Wang, L. et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-138 (2010).

[15] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

[16] Leng, N. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29, 1035-1043 (2013).

[17] Tarazona, S. et al. Differential expression in RNA-seq: a matter of depth. Genome Res. 21, 2213-2223 (2011).

[18] Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. Genome Res. 7, 986-995 (1997).

[19] Mistry, J. et al. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41, e121 (2013).

[20] Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat. Methods, 12, 59-60 (2015).

[21] von Mering, C. et al. STRING: known and predicted protein-protein as sociations, integrated and transferred across organisms. Nucleic Acids Res. 33, D433-D437 (2005).

[22] Altschul, S. F. et al. Basic local alignment search tool. J. Mol. Biol. 215, 403-410 (1990).

[23] Winnenburg, R. et al. PHI-base: a new database for pathogen host interactions. Nucleic Acids Res. 34, D459-D464 (2006).

[24] Verma, S. et al. Draft genome sequencing and secretome analysis of fungal phytopathogen Ascochyta rabiei provides insight into the necrotrophic effector repertoire. Sci. Rep. 6, 24638 (2016).

[25] Sanseverino, W. et al. PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res. 38, D814-D821 (2010).

[26] Zhou, X. et al. De novo sequencing and analysis of the transcriptome of the wild eggplant species Solanum aculeatissimum in response to Verticillium dahliae. Plant Mol. Biol. Rep. 34, 1193-1203 (2016).

Technical Support E-mail:info@bgitechsolutions.com

Website: www.bgitechsolutions.com