Arigatou Eukaryotic Transcriptome Report

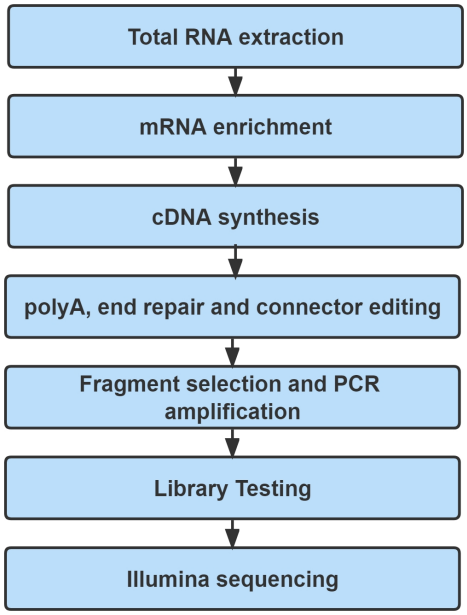
Content

MWXS-XX-XXXX-novo report

Transcriptome broadly refers to the collection of all transcripts in a cell under a certain physiological condition. The object of interest is the sum of all RNAs that can be transcribed in a specific cell under a certain functional state, mainly including mRNA and ncRNA. Transcriptome research serves as a foundation for the study of gene function and structure, and plays an important role in the development of organisms and the incidence of disease. With the development of gene sequencing technology and the reduction of sequencing cost, RNA-seq has become the main method for transcriptome research with the advantages of high throughput, high sensitivity and broad application. In this project, 12 samples were sequenced and a total of 80.04 Gb Clean Data were obtained, with 6 Gb clean data for each sample and 92% or more Q30 bases.

1 Experimental Workflow

The experimental workflow for transcriptome sequencing includes several stages: RNA extraction, RNA quality assessment, library construction, and sequencing. The experimental process is illustrated in the diagram below:



Extraction, Library Construction, and Sequencing Workflow Diagram

1.1 RNA Quality Assessment

High-quality RNA is the foundation for the success of the entire project. To ensure RNA quality, the following methods are used to inspect samples, and only qualified samples proceed to library preparation:

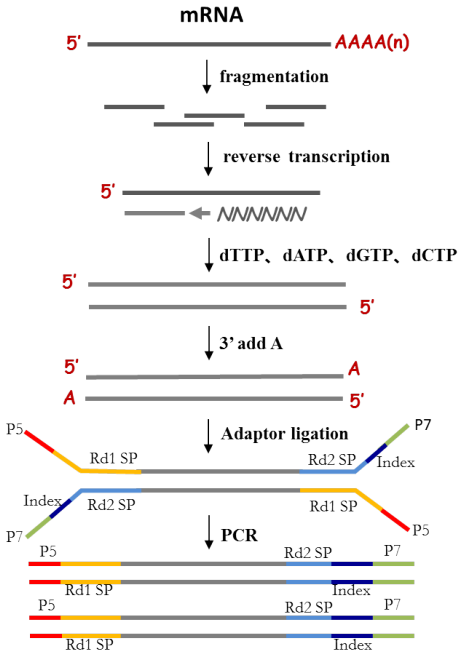
(1) Agarose Gel Electrophoresis: Analyze RNA integrity and check for the presence of DNA contamination.

(2) Qubit 4.0 Fluorometer/MD Enzyme Analyzer: Accurately measure RNA concentration.

(3) Qsep400 Bioanalyzer: Precisely evaluate RNA integrity.

1.2 Library Construction

There are two primary methods for obtaining mRNA: firstly, using the characteristic polyA tails of most mRNAs in eukaryotic organisms, mRNA with polyA tails is enriched using Oligo(dT) magnetic beads. Secondly, ribosomal RNA is removed from total RNA to obtain mRNA. Subsequently, the RNA is fragmented using fragmentation buffer, and the short fragmented RNA serves as a template to synthesize the first-strand cDNA using random hexamers. Then, a buffer, dNTPs (dTTP, dATP, dGTP, and dCTP), and DNA polymerase I are added to synthesize the second-strand cDNA. The purified double-stranded cDNA is further subjected to end repair, A-tailing, and adapter ligation. The cDNA library is size-selected using DNA purification beads and PCR enrichment is performed to obtain the final cDNA library. The experimental process is illustrated in the diagram below:



Library Construction Method Schematic

1.3 Library Quality Check

After library construction is complete, the library's quality is assessed. Sequencing can only proceed when the quality meets the required standards. The quality assessment methods include:

(1) Preliminary quantification using the Qubit dye method, and insert size analysis using a fragment analyzer. Sequencing can proceed only when the insert size meets expectations.

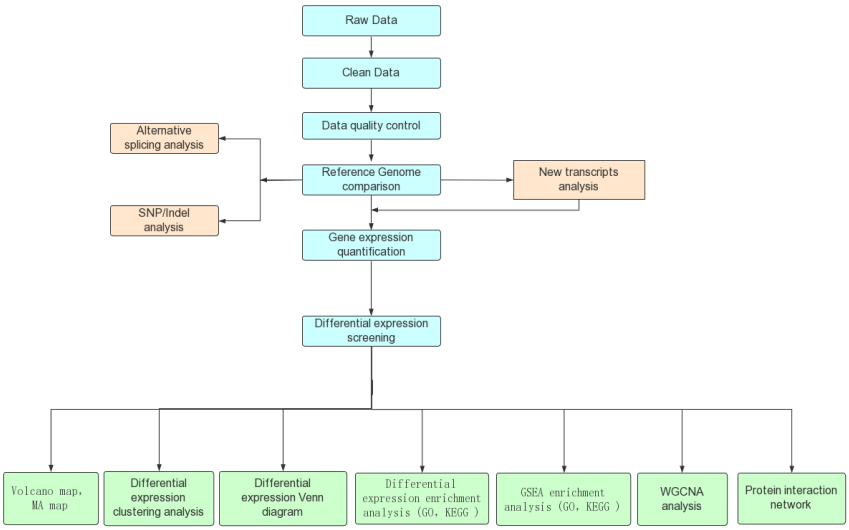
(2) Q-PCR method for accurate quantification of the library's effective concentration (library effective concentration > 2nM), completing library inspection.

1.4 Sequencing

After passing library inspection, different libraries are pooled according to the target amount of data for sequencing on the Illumina platform.

2 Bioinformatics Analysis Process

The output data were filtered to obtain clean data, which were compared with the specified reference genome to obtain mapped data. The mapped data were subjected to differential expression analysis based on gene expression levels across different samples or groups, followed by functional annotation and enrichment analysis of the differentially expressed genes to gain insights into their biological functions. The flow chart for transcriptome bioinformatics analysis is shown below:



Bioinformatics Analysis Process

2.1 Sequencing Data and Quality Control

2.1.1 Description of Sequencing Data

Illumina high-throughput sequencing platform sequences cDNA libraries based on Sequencing by Synthesis (SBS) technology. The image-based sequencing data is then converted by CASAVA base calling into a large amount of high-quality data, called raw data. Raw data are usually provided in fastq format and contain mainly the sequence information of sequenced fragments and respective sequencing quality information. Each read in the fastq file consists of four lines of descriptive information, as follows:

```
@ST-E00600:42:H3JYTALXX:1:1101:1217:1000 1:N:0:TCCGTCTA
GGCCAAAAAGGGGAGTGGGTGGGTAGGGGAGTGCCAGGGCCAAAAAGGGGGAGTGGGTG
+
```

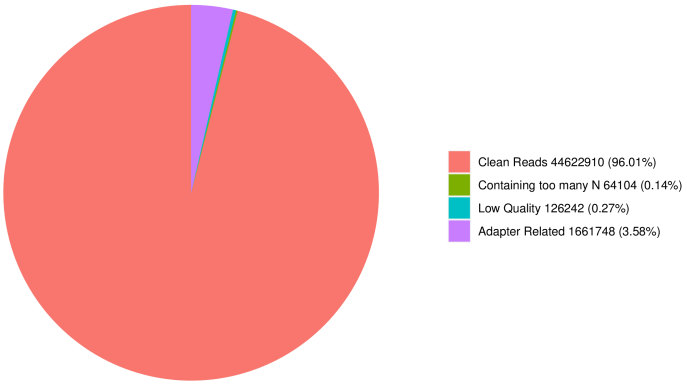
```
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFF
```

The first line of the above file starts with '@', followed by Illumina sequence identifiers and description text; The second line is the base sequence of the sequenced fragment; The third line starts with '+', followed by Illumina sequence identifiers (can also be empty); The fourth line represents the sequencing quality value corresponding to each base of the sequenced fragment, where the ASCII value corresponding to each character in the line minus 33 gives the sequencing quality value of that base.

2.1.2 Sequencing Data Filtering

Before performing data analysis, it is first necessary to ensure that these reads are of high enough quality to ensure the accuracy of subsequent analysis. We used fastp[1] software to perform strict quality control on the data, and the filtering criteria were as follows: （1）Reads with adapters were removed; （2）When the proportion of ambiguous bases (N) in any sequencing read exceeds 10% of the total bases in that read, the paired reads containing this read were removed; （3）When the number of low quality (Q<=20) bases contained in any sequenced read exceeds 50% of the total number of bases in that read, the paired reads containing this read were removed.

Data filtering profile for each sample is summarized in the following figures:



the composition statistics of raw sequencing data

Adapter related: the proportion of reads with adapters; Containing N: the proportion of reads with N bases; Low quality: the proportion of reads with low quality; Clean reads: the proportion of clean reads.

2.1.3 Sequencing Error Rate Distribution

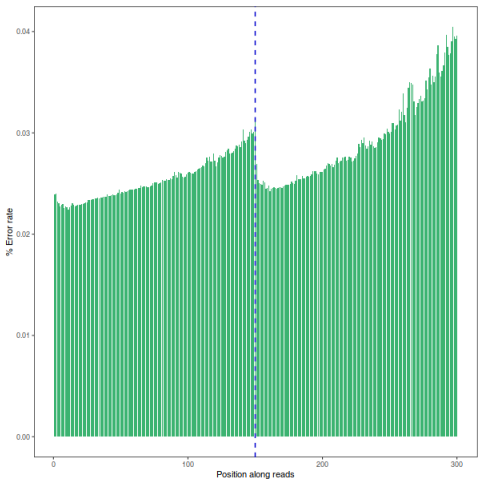
The sequencing process itself may produce errors, and the sequencing error rate distribution check can reflect the quality of the sequencing data. The sequencing quality value of each base in the sequence information is stored in the fastq file. With sequencing error rate expressed as e and Illumina base quality value expressed as Qphred, then: Qphred = -10log10(e). The concise correspondence between Illumina Casava v1.8 base calling and Phred scores is shown in the following table:

Table 1 Concise correspondence between base calling and Phred score

Sequencing error rate	Base mass	phred33 Corresponding characters	phred64 Corresponding characters
5%	13	.	M
1%	20	5	T
0.1%	30	?	^

Higher base quality value indicates a more reliable and accurate base calling. For example, for a base

quality value of Q20, 1 out of 100 bases were called incorrectly, and so on. With current RNAseq sequencing technology, there are two characteristics of sequencing error rate distribution as follows:（1）The sequencing error rate increases as the length of sequenced reads increases. This is caused by the consumption of chemical reagents during sequencing and is a characteristic of Illumina high-throughput sequencing platform;（2）The first 6 bases feature a high sequencing error rate, and this is exactly the length of the random primer required for reverse transcription during RNA-seq library building. The high sequencing error in the first 6 bases is due to the incomplete binding of random primers and RNA templates. This feature is shared by illumina high-throughput sequencing platforms. The distribution of sequencing data error rates for each sample in this project are plotted as follows:

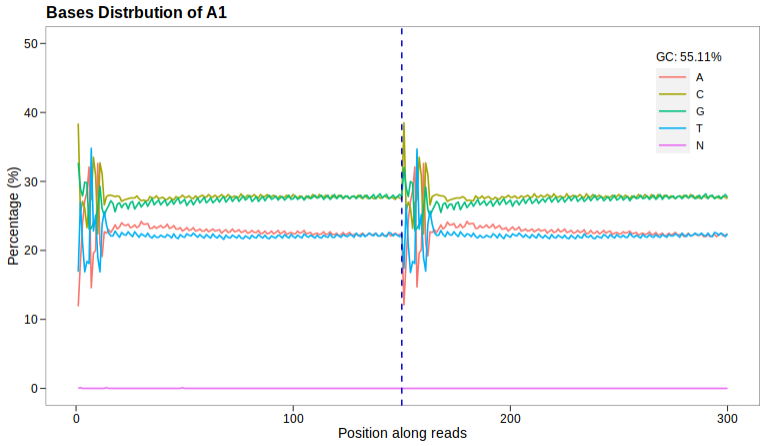


Distribution of Base Error Rate in Reads

The horizontal coordinate indicates the position of the bases in the reads, and the vertical coordinate indicates the single base error rate.

2.1.4 GC Content Distribution

GC content distribution check is used to detect the presence of AT/GC separation. In theory, the GC and AT contents of sequenced reads should be equal at each position and essentially constant throughout the sequencing process as a result of random sequence fragmentation and the principle of double-strand complementarity. However, since reverse transcription uses 6 bp random primers, the first few bases are subject to some preference in nucleotide composition, producing regular fluctuations before stabilizing. The GC content distribution of each sample is illustrated as follows:



GC Content Distribution Plot

The horizontal coordinate indicates the position of the bases in the reads, and the vertical coordinate indicates the proportion of single bases.

2.2 Sequencing Output Statistics

After raw data filtering, and checking for sequencing error rate and GC content distribution, the clean reads used for subsequent analysis, with data summarized in the following table:

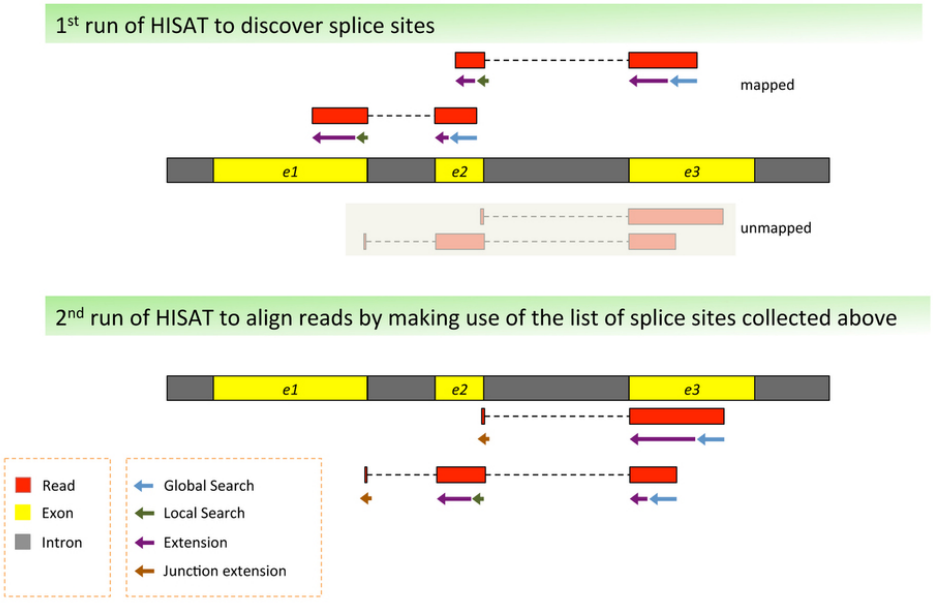
Table 2 Data Output Statistics

Sample	Raw Reads	Clean Reads	Clean Base(G)
A1	46475004	44622910	6.69
A2	49121824	46893456	7.03
A3	44776266	43838320	6.58
B1	44816386	43318554	6.50
B2	43398202	41947518	6.29
B3	44634278	43013954	6.45
C1	45967324	44406072	6.66
C2	47410196	45811112	6.87
C3	45534544	44648410	6.70
D1	46479640	44957230	6.74

- Sample：sample name
- Raw Reads：the number of raw reads
- Clean Reads：the number of clean reads obtained by filtering the raw reads
- Clean Bases：the total number of bases of high-quality reads

2.3 Reference Genome Library

The sequencing fragments were derived from randomly fragmented mRNAs. In order to determine which genes these fragments were transcribed from, the clean reads after quality control were aligned to the reference genome. HISAT2 [2] was used to align the clean reads to the reference genome to obtain information about the location of the reads in the genome or gene, as well as sequence features unique to the sequencing sample. The algorithm of HISAT2 consists of three main parts: （1）Whole alignment of a read to a single exons of the genome; （2）Partial alignment of a read to two exons of the genome; （3）Partial alignment of a read to three or more exons of the genome. In this project, the default parameters of the software were used for sequence alignment, and the algorithm for HISAT2 alignment is illustrated as follows:



Schematic Diagram of Hisat2 Alignment Algorithm

2.3.1 Alignment Efficiency Statistics

Alignment efficiency refers to the percentage of mapped reads out of the total number of clean reads, which represents the most direct indication of transcriptome data utilization. If the reference genome is well-assembled and the sequenced species is closely related to the reference genome, and there is no contamination during the experiment, then the percentage of reads successfully mapped to the genome should be higher than 70% (total mapped).The reference genome used in this project was Oryza_sativa.IRGSP-1.0.52.gff3.gz,download address：ftp://ftp.ensemblgenomes.org/pub/plants/release-52/fasta/oryza_sativa/dna/.Genome Structure Annotation File：Oryza_sativa.IRGSP-1.0.dna.toplevel.fa.gz. The reads alignment profile for each sample in this project is listed in the table below:

Table 3 Alignment Profile Statistics

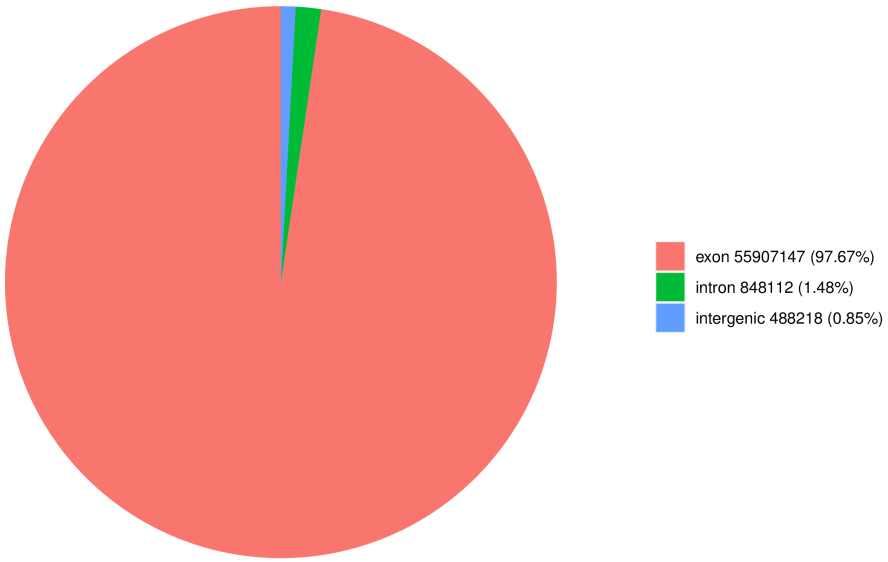
Sample	Total Reads	Reads mapped	Unique mapped
A1	44622910	42234572(94.65%)	41292113(92.54%)
A2	46893456	44813154(95.56%)	43777751(93.36%)
A3	43838320	42175376(96.21%)	41171066(93.92%)
B1	43318554	41221934(95.16%)	40144760(92.67%)
B2	41947518	39921372(95.17%)	38916023(92.77%)
B3	43013954	40858718(94.99%)	39806951(92.54%)
C1	44406072	41972113(94.52%)	41009108(92.35%)
C2	45811112	43586341(95.14%)	42542162(92.86%)
C3	44648410	42140788(94.38%)	41074111(91.99%)
D1	44957230	42558962(94.67%)	41286399(91.83%)

- Sample: sample name
- Total Reads：the total number of clean reads
- Reads mapped：the number of reads mapped to the reference genome
- Unique mapped：the number of reads that are uniquely mapped to the reference genome

2.3.2 Distribution of Mapping Regions

Typically, a large proportion of reads will align to exons, while only a small proportion of reads align to introns and intergenic regions. Reads align to introns may originate from pre-mRNAs or retained introns

from alternative splicing events. Reads align to intergenic regions may originate from ncRNAs or low levels of DNA fragment contamination, or the gene may not be well annotated. The alignment distribution of each sample is illustrated as follows:



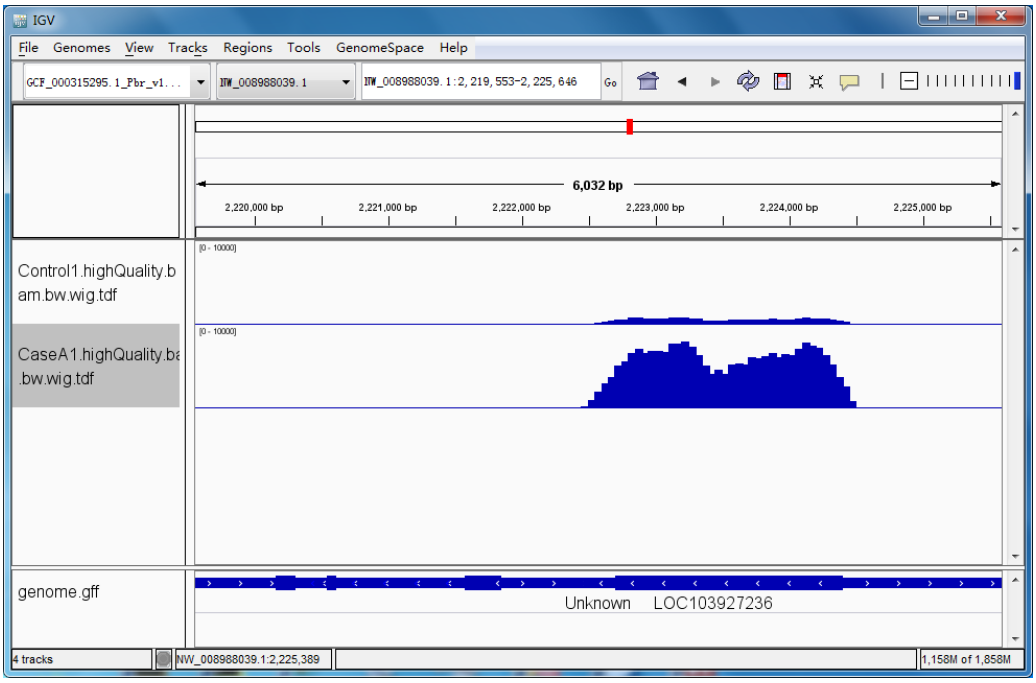
Distribution of Mapping Regions Plot

The number of exon, intron, and intergenic reads were counted and plotted separately.

2.3.3 Visualization of Comparison Results

We converted the bam files obtained by aligning the reads to the reference genome into tdf files that can be recognized by IGV. The distribution of reads on each chromosome and the distribution of functional regions such as exons, introns, and intergenic regions annotated in the genome were then visualized by IGV[3] software, as illustrated below:

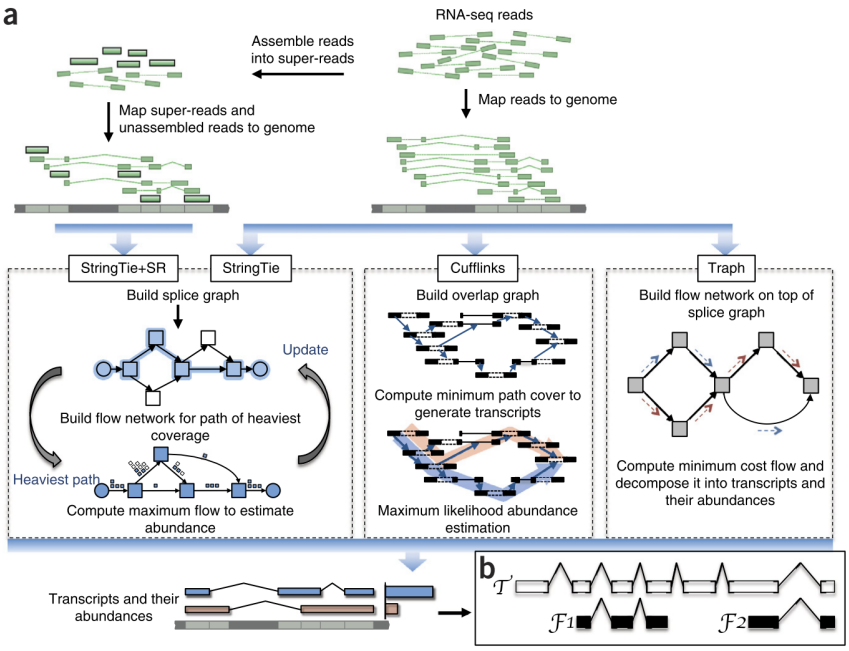
IGV visualization tutorial:src/appendix/IGV_tutorial.pdf



Schematic Diagram of IGV Visualization

2.4 New Gene Analysis

Based on the positional information of reads aligned to the reference genome, StringTie[4] is used to assemble reads into transcripts. StringTie utilizes network flow algorithms and optional de novo assembly to assemble transcripts. Compared to software like Cufflinks, StringTie is capable of assembling more complete and accurate transcripts and also operates at a faster speed. The schematic diagram of its workflow is shown below:



StringTie Schematic diagram

（1）Assemble reads into ‘Super-reads’. If there is an overlap of k-mers between two reads, they are extended until both directions can no longer be extended. Finally, they are assembled into longer sequences, referred to as ‘Super-reads (SR)’.（2）Align the reads to the reference genome. The StringTie + SR method uses a mixture of reads, including super-reads and unassembled reads. （3）Cluster the aligned reads, and for each cluster, construct the corresponding alternative splicing graph. Each alternative splicing represents all possible transcript isoforms of a gene. （4）Identify the path with the highest read coverage from the alternative splicing graph and construct a network flow for that path. （5）Use a network flow algorithm to assign reads to transcripts, maximizing the number of reads covered by transcripts. （6）Remove reads used in step (5) from the splicing graph, iterate through steps (4)-(5) until no more paths can be followed. The assembled transcripts are then compared to the genome’s annotation information using GffCompare to discover new transcripts or genes.

2.4.1 Discovery of New Genes

Extracting information about new transcripts from the comparison results between assembled transcripts and genome annotations, and saving it in GTF (Gene Transfer Format) format. Detailed information about the GTF format can be found at https://genome.ucsc.edu/FAQ/FAQformat.html。

Table 4 Table of New Transcript Structure Information

#ID	source	feature	start	end	score	strand
1	StringTie	transcript	57002	57445	1000	-
1	StringTie	exon	57002	57249	1000	-
1	StringTie	exon	57326	57445	1000	-
1	StringTie	transcript	313402	314287	1000	.
1	StringTie	exon	313402	314287	1000	.
1	StringTie	transcript	406683	407033	1000	.
1	StringTie	exon	406683	407033	1000	.
1	StringTie	transcript	403540	405111	1000	-
1	StringTie	exon	403540	404501	1000	-
1	StringTie	exon	404736	405111	1000	-

- ID：Chromosome number.
- source：Software or database that generated the annotation.
- feature：Annotation type.
- start：Starting coordinate.
- end：Ending coordinate.
- score：Score assessing the reliability of the annotation. A is used when score is missing.
- strand：Positive or negative strand.

2.4.2 Functional Annotation of New Genes

Sequences of new genes are extracted from the genome, and the annotation results are obtained by aligning these new genes with sequences from databases such as KEGG, GO, NR, Swiss-Prot, TrEMBL, and KOG using diamond[5]. The alignment criteria include an E-value threshold of 1e-5. For plant transcription factor prediction, the iTAK[6] software is used, which integrates two databases, PlnTFDB[7] and PlantTFDB [8].For animal transcription factor identification, the animalTFDB[9] database is utilized：

Table 5 Table of New Gene Annotation

gene_id	chr	start	end	strand
novel.5617	8	19091559	19092093	+
novel.5174	7	22735610	22736112	+
novel.3522	4	13805980	13806310	.
novel.905	10	12031111	12031409	.
novel.4744	6	27043167	27044250	.
novel.3543	4	15885745	15885969	.
novel.5004	7	13645708	13649500	+
novel.4505	6	14182904	14183568	+
novel.5040	7	15187545	15191354	-
novel.2707	2	32122510	32123503	-

- gene_id：identifier for New Gene.
- chr：Chromosome name where the new gene is located.
- start：Starting position of the new gene on the chromosome.
- end：Ending position of the new gene on the chromosome.
- strand：Strand (positive or negative) of the new gene on the chromosome.

2.5 Quantification of Gene Expression

The number of fragments of a transcript is related to the amount of sequencing data (or mapped data), transcript length, and transcript expression level. In order for the number of fragments to truly reflect the transcript expression level, it is necessary to normalize the number of mapped reads and transcript length in the sample. FPKM (fragments per kilobase of transcript per million fragments mapped) was used as a measure of transcript or gene expression level, and the FPKM calculation formula is as follows:

$$FPKM = \frac{mapped\ fragments\ of\ transcript}{Total\ Count\ of\ mapped\ fragments\ (Millions) \times Length\ of\ transcript\ (kb)}$$

The mapped fragments of transcript indicates the number of fragments mapped to a transcript, i.e. the number of mapped paired-end reads to a transcript. The total count of mapped fragments (Millions) indicates the total number of fragments mapped to the genome, expressed in 10^6. The length of transcript (kb) means

transcript length in 10^3 bases. The expression levels of some genes (FPKM) are shown in the following table:

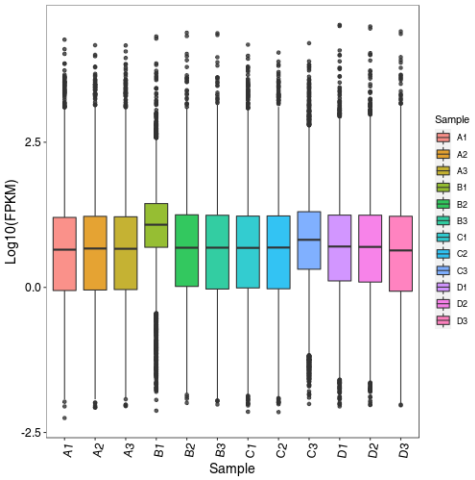
Table 6 Gene Expression Levels (FPKM)

ID	A1	A2	A3	B1	B2	B3
ENSRNA049465704	0.0000	0.0000	0.0000	0.7865	0.0000	0.0000
ENSRNA049465759	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ENSRNA049465818	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ENSRNA049466327	0.0000	0.2935	0.0000	0.0000	0.0000	0.0000
ENSRNA049466562	0.0000	0.0000	0.0000	0.0000	1.5832	0.0000
ENSRNA049466596	0.0000	0.7128	0.0000	0.0000	0.7916	0.7877
ENSRNA049466653	0.0000	0.7417	1.5688	0.0000	0.0000	0.0000
ENSRNA049466739	0.6555	0.0000	0.0000	0.0000	0.0000	0.6893
ENSRNA049467856	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ENSRNA049467888	9.8329	18.0867	10.5535	55.7452	2.7705	7.5820

- ID：Gene numbe
- Second column and after: FPKM expression of all genes in each sample

2.5.1 Overall Distribution of Gene Expression in Samples

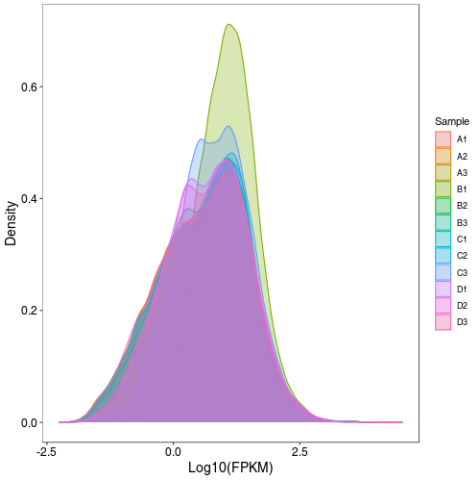
The detection of gene expression using transcriptome data to sensitive. Typically, the FPKM values of protein-coding gene expression levels that can be sequenced span six orders of magnitude from 10^{-2} to 10^{4}.The box plot indicates the dispersion of the distribution of gene expression levels of individual samples, while allowing a visual comparison of the overall gene expression levels of different samples. The FPKM distribution of each sample in this project is illustrated as follows:



Box Plot of Gene Expression

The horizontal coordinate indicates the different samples; the vertical coordinate indicates the logarithmic value of the sample expression (FPKM). This plot measures the expression level of each sample in terms of the overall dispersion of expression.

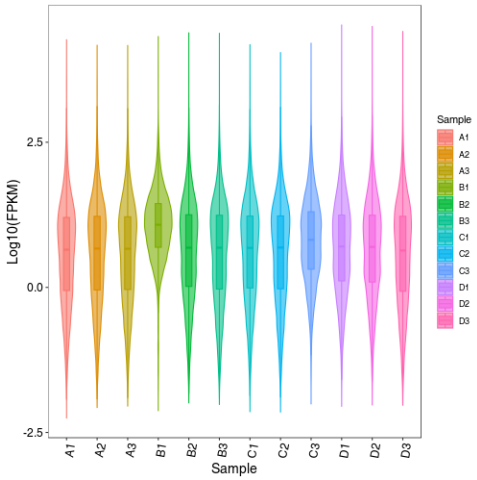
Density plots demonstrate the changes in gene abundance with expression levels in a sample and provide a clear picture of the regions where the bulk of the gene expression levels lie, as shown in the figure below:



Gene Expression Density Distribution

The curves of different colors in the plot represent different samples. For a dot on the curve, its horizontal coordinate indicates the logarithmic value of FPKM of the corresponding sample, and the vertical coordinate of the dot indicates the probability density.

Violin plots are used to show the distribution states as well as the probability densities of multiple sets of data, as shown in the figure below:

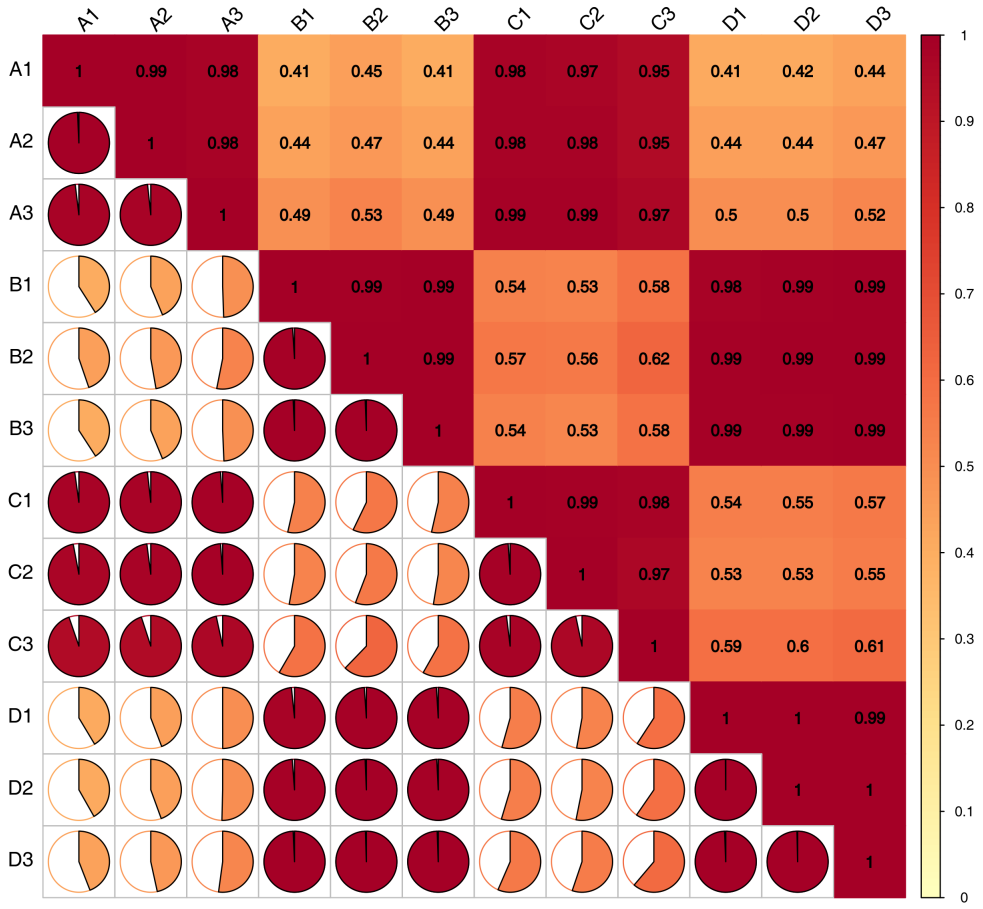


Expression Level Violin Plot

The curves of different colors in the plot represent different samples. The width of each violin plot reflects the number of genes at that expression level.

2.5.2 Sample Correlation Analysis

Gene expression can vary among individuals, with different levels of expression variability among genes. Transcriptome sequencing, qPCR and microarray technologies do not eliminate this variability. In order to find differentially expressed genes of interest, expression differences due to biological variability need to be considered and addressed. A common and effective way to do this is to set up biological replicates in the experimental design. The more consistent the replicate conditions and the greater the number of replicates, the more reliable the search for differentially expressed genes. For projects with biological replicates, assessing the relevance of the biological replicates is important for analyzing transcriptome sequencing data. The correlation of biological replicates not only tests the reproducibility of biological experimental manipulations, but also assesses the reliability of differentially expressed genes and aids in the screening of abnormal samples. Pearson correlation coefficient (expressed as r) is used as an indicator to assess the correlation of biological replicates. The closer the absolute value of r to 1, the stronger the correlation between two replicate samples. The correlation statistics between samples for this project are plotted as shown below:



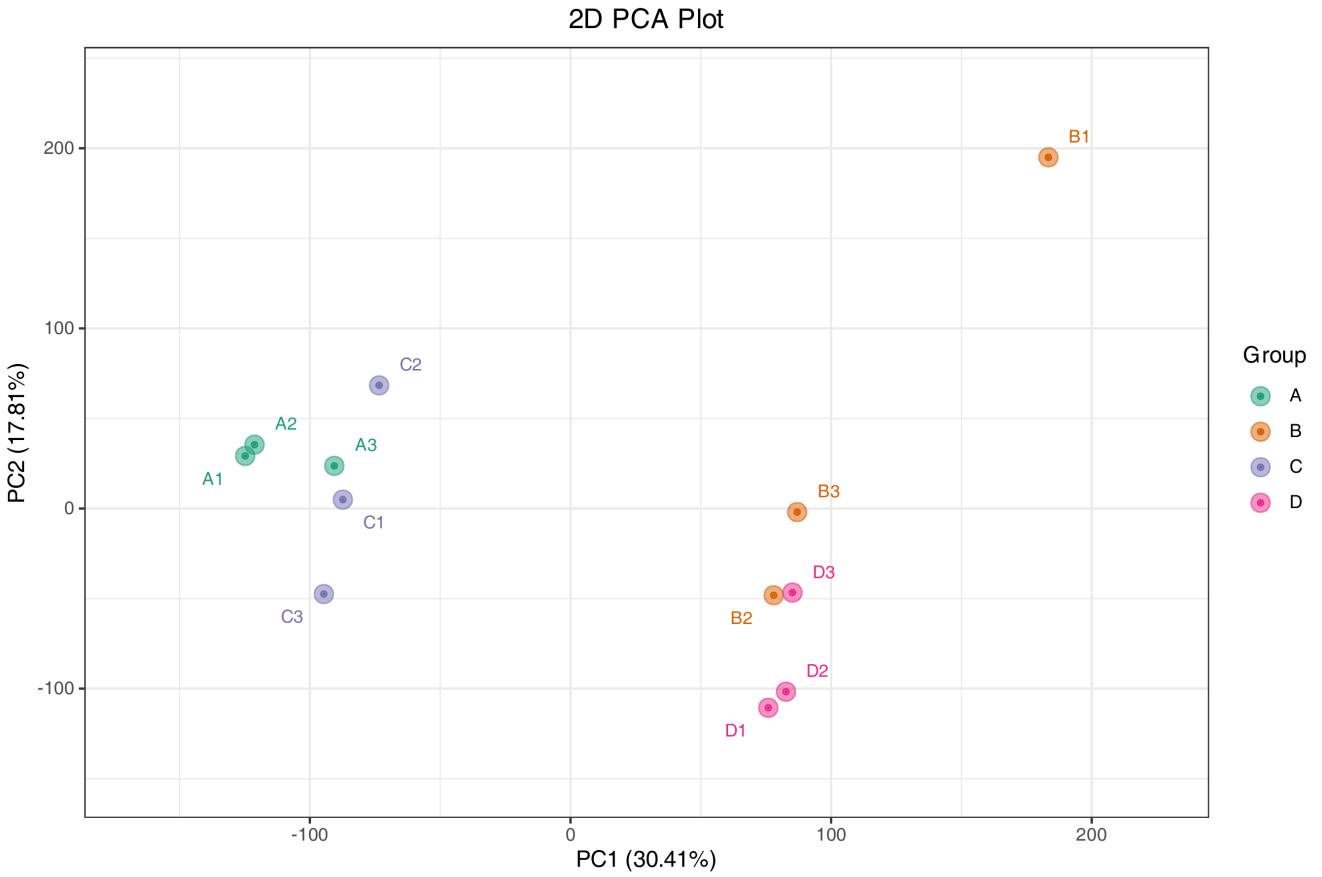
Inter-Sample Correlation Plot

2.5.3 Principal Component Analysis

By using multivariate statistical analysis, high-dimensional and complex data can be simplified and downscaled with maximum retention of the original information, and reliable mathematical models can be established to summarize and conclude the expression characteristics of the research object. Principal component analysis (PCA) is an unsupervised pattern recognition method for statistical analysis of multidimensional data, which converts a set of potentially correlated variables into a set of linearly uncorrelated variables by orthogonal transformation. The converted set of variables are called principal components. This analysis is often used to study how to reveal the internal structure among multiple variables through a few principal components, i.e., to derive a few principal components from the original variables so that they retain as much information as possible about the original variables and are uncorrelated with each other. The usual math-

ematical processing is to make a linear combination of the original multiple indicators as a new composite indicator.

The data processing principle of PCA: the original data is compressed into a number of n principal components to characterize the original data set, PC1 denotes the most significant feature that can describe the multidimensional data matrix, PC2 denotes the most significant feature that can describe the data matrix excluding PC1, and PC3 ……PCn and so on.



PCA Plot of the Sample

2.6 Screening for Differentially Expressed Genes

For samples with biological replicates, differential expression analysis between sample groups was performed using DESeq2[10, 11] to obtain the set of differentially expressed genes between two biological conditions. For samples with no biological replicates, we used edgeR[12]. The input genes are required to

be unstandardized reads count data, rather than standardized data such as RPKM, FPKM.The read counts of genes were implemented using featureCounts[13].After the differential analysis, it is also necessary to perform multiple-hypothesis test correction for the probability of hypothesis testing (P value) by using the Benjamini-Hochberg method to obtain the false discovery rate (FDR).The screening conditions for differential genes: |log2Fold Change| >= 1, and FDR < 0.05。

2.6.1 Raw Reads Counts

We used featureCounts[13] to count the reads on genes for each sample based on the high-quality alignment results, and then combined the gene count results for all samples. Due to the large number of genes, the web report only displays some of the data as shown in the following table:

Table 7 Table of Reads Counts on Genes

ID	A1	A2	A3
ENSRNA049465704	0	0	0
ENSRNA049465759	0	0	0
ENSRNA049465818	0	0	0
ENSRNA049466327	0	1	0
ENSRNA049466562	0	0	0
ENSRNA049466596	0	1	0
ENSRNA049466653	0	1	2
ENSRNA049466739	1	0	0
ENSRNA049467856	0	0	0
ENSRNA049467888	15	29	16

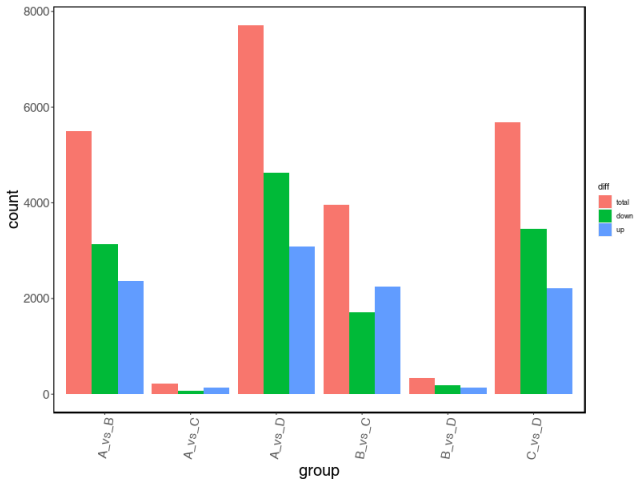
- ID：Gene Number
- The 2 - last column：the Raw readcount data for each sample

2.6.2 Number for Differentially Expressed Genes

After completing the analysis of differentially expressed genes using DESeq2/edgeR, the total number of differentially expressed genes, the number of up-regulated genes, and the number of down-regulated genes in each group were counted as shown in the following table:

Table 8 Table for Differentially Expressed Genes

group	total	down	up
A_vs_B	5496	3132	2364
A_vs_C	213	69	144
A_vs_D	7719	4628	3091
B_vs_C	3961	1708	2253
B_vs_D	331	193	138
C_vs_D	5677	3459	2218



Statistical Plot of Differentially Expressed Genes

2.6.3 Table of Differentially Expressed Genes

The differentially expressed genes calculated for each differential grouping are shown in the table below:

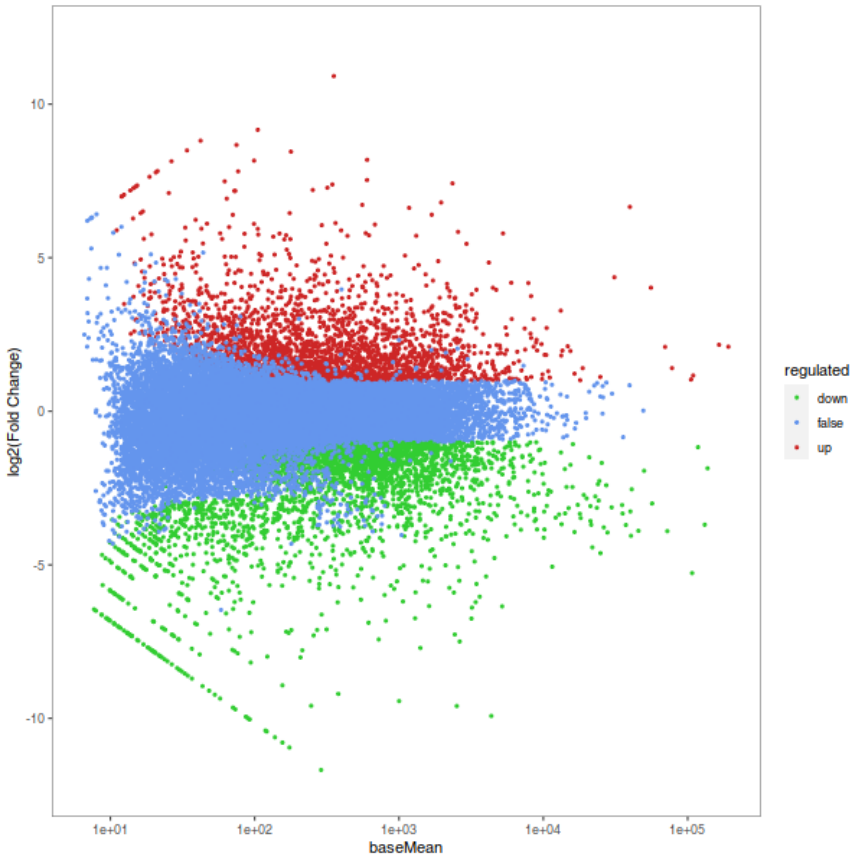
Table 9 List of Differentially Expressed Genes

ID	A1_fpkm	A2_fpkm	A3_fpkm
Os04g0444800	231.4199	191.4418	199.9208
Os08g0189900	4.5910	5.6973	6.6948
Os08g0190100	2.0382	2.3505	2.6723
Os11g0707000	7085.3902	6175.1536	7038.1461
Os04g0127200	8.4983	10.7863	10.9131

- ID：gene number
- middle column: sample expression information

2.6.4 MA Plot of Differentially Expressed Genes

MA plots provide a visual representation of the overall distribution of gene expression levels and folds of difference, as shown below:

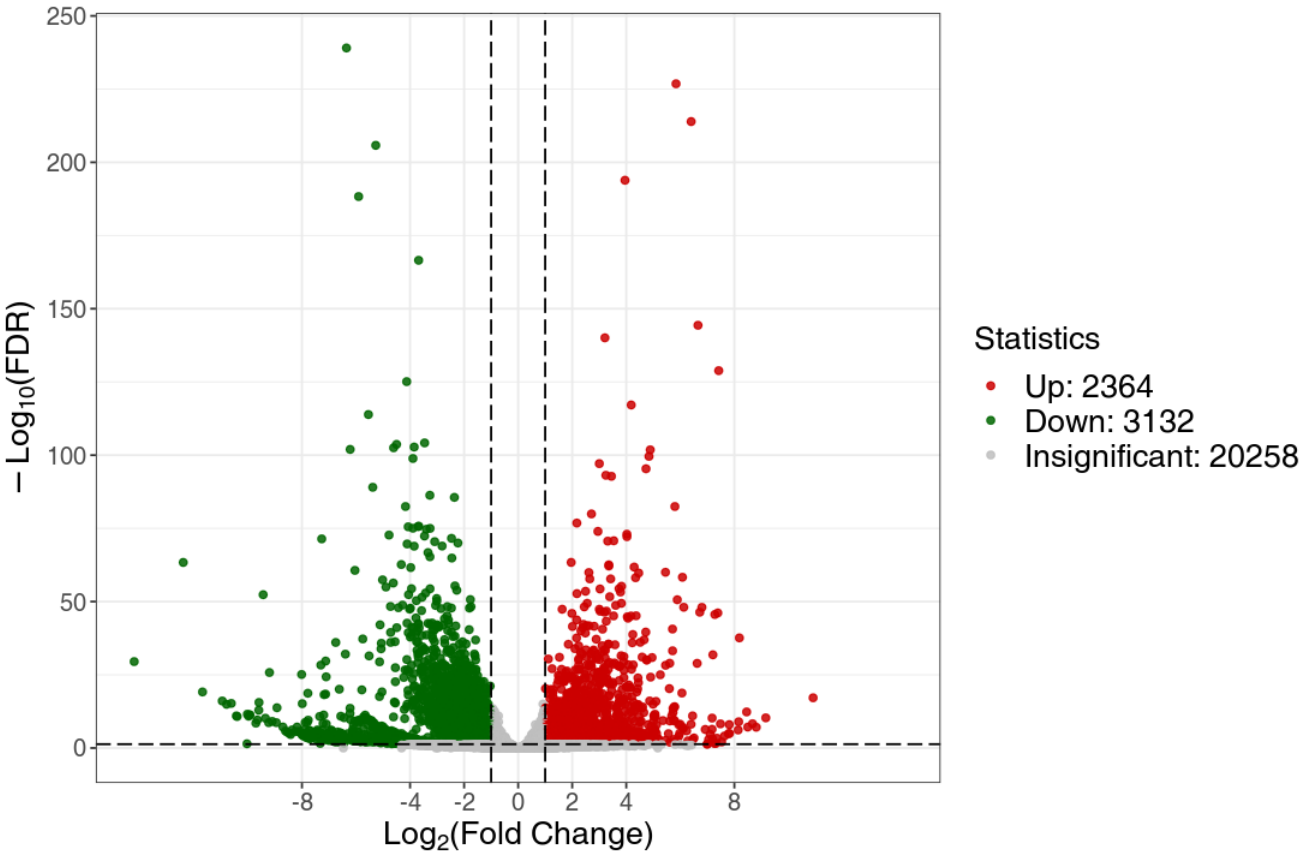


MA Plot of Differentially Expressed Genes

the horizontal coordinate indicates the mean value of gene expression in the two samples; red dots represent up-regulated gene expression, while green dots represent down-regulated gene expression and blue indicates no significant difference in gene expression.

2.6.5 Volcano Plot of Differentially Expressed Genes

Volcano plots provide a visual representation of the overall distribution of differentially expressed genes in the two sets of samples, as illustrated below:

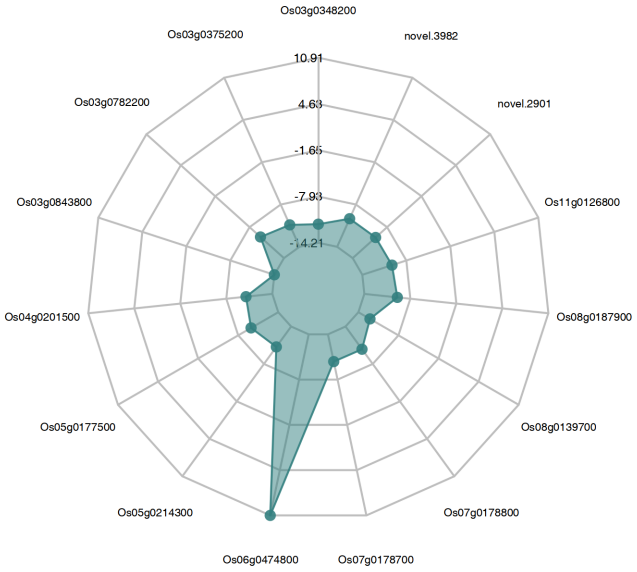


Volcano Plot of Differentially Expressed Genes

The horizontal coordinate indicates the fold change in gene expression and the vertical coordinate indicates the significance level of differentially expressed genes. Red dots represent up-regulated differentially expressed genes, green dots represent down-regulated differentially expressed genes, and gray dots represent non-differentially expressed genes.

2.6.6 Radar Chart of Differentially Expressed Genes

The 15 up-/down-regulated genes with the largest differential multiplicity are presented using radar chart:



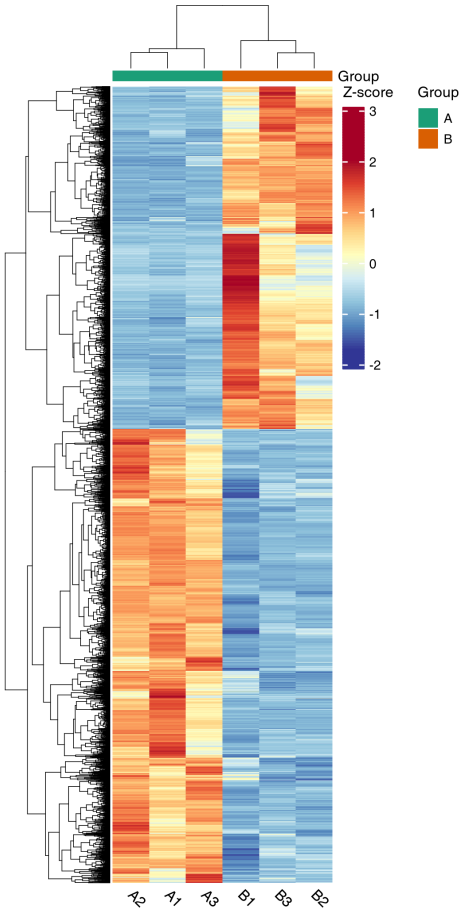
Radar Chart of Differentially Expressed Genes

Each node in the graph represents a gene, and the position of the node corresponds to the size of the log2FoldChange value of that gene.

2.6.7 Gene Expression Cluster Analysis

The clustering analysis is used to determine the expression patterns of differentially expressed genes under different experimental conditions by grouping genes with the same or similar expression patterns into clusters, thus predicting the functions of unknown genes or unknown functions of known genes as genes in the same cluster may have similar functions or be involved in the same metabolic process or cellular pathway together.

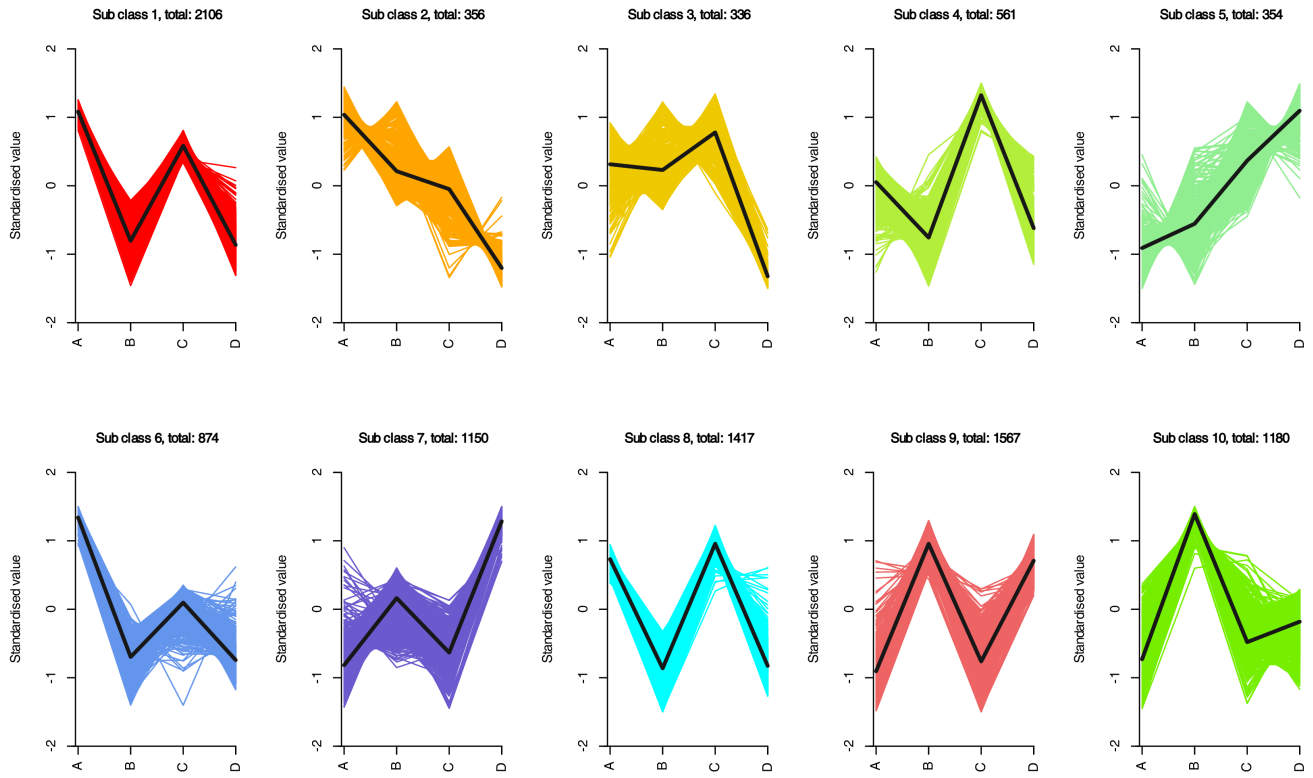
Z-score was used to normalize the differentially expressed genes. Cluster heatmaps for differentially expressed genes across all comparison groups and for each differential grouping are plotted as follows:



Clustering Heatmap of Differentially Expressed Genes

The horizontal coordinates indicate sample names and hierarchical clustering results, and the vertical coordinates indicate differential genes and hierarchical clustering results. Red denotes high expression, whereas blue depicts low expression.

To investigate the expression patterns of genes under different treatment conditions, the FPKM of all differential genes combined were first normalized using the scale function in R language, and then K-means clustering analysis was performed. Genes of the same class exhibit similar trends under different experimental treatments and may have similar functions, as shown below:

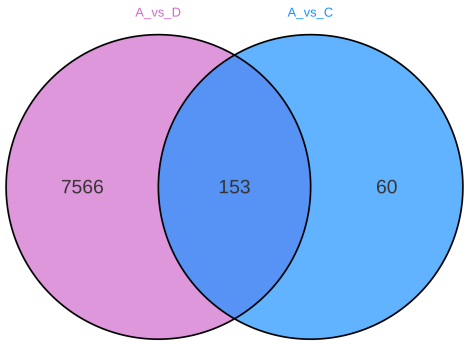


K-Means Cluster Plot

The horizontal coordinate indicates the sample, and the vertical coordinate indicates the normalized expression level.

2.6.8 Venn Diagram of Differentially Expressed Genes

Venn diagrams illustrate the overlap of differentially expressed genes between different comparative combinations. Using the Venn diagram, differentially expressed genes common or unique to certain comparative combinations can be screened, as shown below:



Venn Diagram of Differentially Expressed Genes in Different Groupings

The non-overlapping area of the Venn diagram represents the differentially expressed genes specific to that differential grouping, and the overlapping area represents the differentially expressed genes common to several differential groupings that overlap.

2.7 Analysis of differential gene transcription factors

Transcription factors are a class of proteins that bind to DNA and regulate gene expression. Transcription factors bind to specific DNA sequences, termed promoters, facilitating the transcription of genes by recruiting other regulatory factors and RNA polymerases. Transcription factors can either promote or repress the transcription of genes, and their expression and activity can be regulated by various signals inside and outside the cell. Transcription factors are crucial in biological processes such as cell differentiation, growth, and development. The plantTFDB database contains family classification rules of plant transcription factors, genome-wide transcription factor profiles, rich annotations, transcription factor binding motifs, transcription factor predictions, phylogenetic trees and other related information, involving 165 species. So,it annotates the differential genes through the plantTFDB database and applies FIMO to obtain the target genes corresponding to the transcription factors, ultimately visualizing all the results (Note: only the transcription factor

module annotates with the plantTFDB database, while the other analysis modules all use iTAK for information regarding transcription factor annotations).

2.7.1 Differential transcription factor annotation table

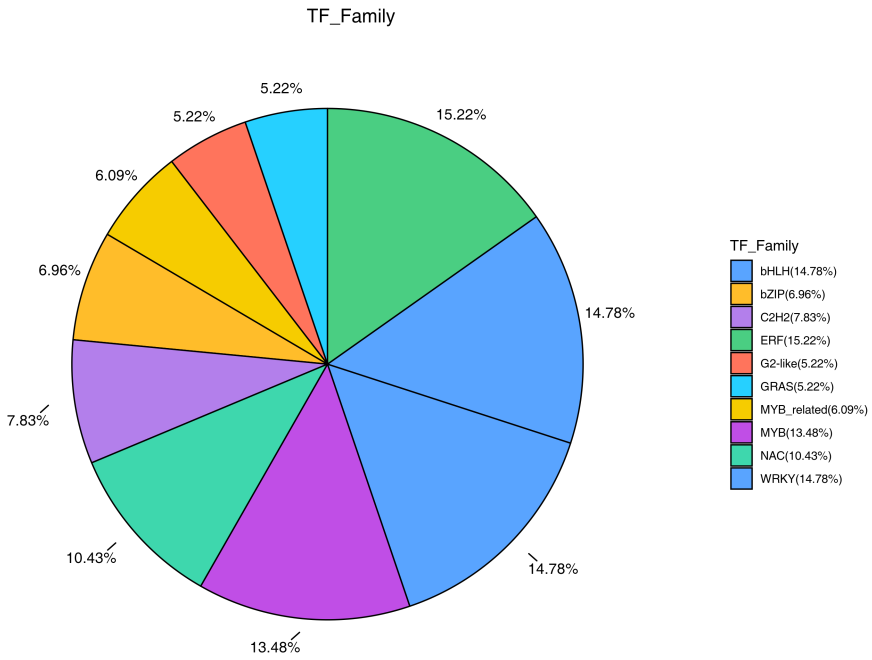
Differential genes were annotated by using the plantTFDB transcription factor database. The detailed annotation information of the differential genes of each differential group is presented, with the results shown in the following table:

Table 10 Differential transcription factor annotation table

ID	TF_Family	A1_fpkm
Os01g0108400	bHLH	21.8192
Os01g0129600	LBD	3.9589
Os01g0140700	RAV	0.3023
Os01g0141000	RAV	17.7261
Os01g0158900	NF-X1	5.1353

1) geneID：ID number of the gene
2) TF_Family：transcription factor gene family
3) *_count：readcount value for each sample ### Pie chart showing the distribution of differential transcription factor gene families

Statistics on differential transcription factor gene families and the number of differential transcription factors in those gene families (only the top ten gene families with the highest number of differential genes are shown) are presented in the figure below:



Pie chart showing the distribution of differential transcription factor gene families

Different gene families are represented by different colors, with the percentage of differential transcription factors in each gene family shown in the figure.

2.7.2 Table summarizing the statistics of differential transcription factors

Differential genes were annotated through the plantTFDB database with differential transcription factor gene families counted, as shown in the following table:

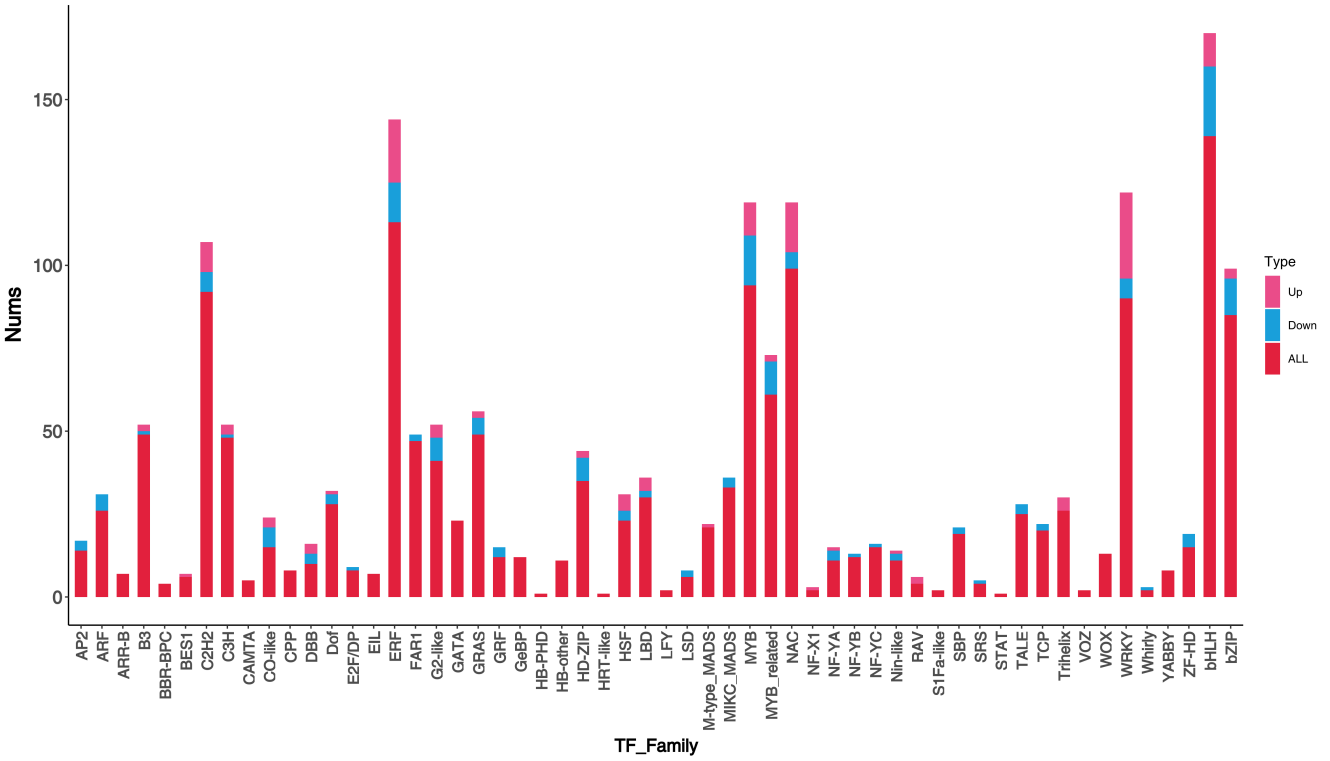
Table 11 Table summarizing the statistics of differential transcription factors

Family	type	num	gene_name
AP2	ALL	14	OsPLT5;OsPLT3;AP2/EREBP#086;OsPLT7;AP2/EREBP#033;OsPLT4;SHAT1;OsPLT1;RSI
AP2	DEG	2	RSR1;AP2/EREBP#052
AP2	Down	2	RSR1;AP2/EREBP#052
ARF	ALL	26	OsARF16;OsARF2;OsARF3;OsARF4;OsARF5;OsARF6a;OsARF7;OsARF8;OsARF9;OsAI
ARF	DEG	2	OsARF2;Os08g0520550

1) TF_Family：transcription factor gene family
2) type：the group to which it belongs (ALL: all transcription factors in that transcription factor family, DEG: differential transcription factors in that transcription factor family, Up: up-regulated differential transcription factors in that transcription factor family, Down: down-regulated differential transcription factors in that transcription factor family)
3) num：the number of transcription factors in the group.
4) gene_name：names of transcription factor genes within this group

2.7.3 Bar chart illustrating the distribution of transcription factor gene families

The statistics of transcription factors in each transcription factor family and the number of up-and down-regulated differential transcription factors in that family are shown below:



Bar chart illustrating the distribution of transcription factor gene families

The horizontal coordinates indicate each transcription factor family; the vertical coordinates represent the number of genes; all transcription factors contained in the transcription factor family are represented in red; up-regulated transcription factors within the transcription factor family are represented in pink; and down-regulated differential transcription factors within the transcription factor family are represented in blue.

2.7.4 Table of predicted transcription factor target genes

In transcriptome analysis, differentially expressed genes in differential combinations usually incorporate transcription factor genes. Investigation of these differential transcription factors and their target genes may help us identify the causes of differentially expressed genes. Transcription factors typically bind to specific regions in the DNA sequence of a gene, which are referred to as transcription factor binding sites (TFBS). It takes two steps to predict the target genes regulated by a transcription factor: 1. First, we need to know the characteristics of the binding sequences of the transcription factor; 2. Based on these characteristics of the binding sequences, we search in the upstream promoter region of the gene. The transcription factor likely regulates the target gene if a sequence matching the binding characteristics is found.

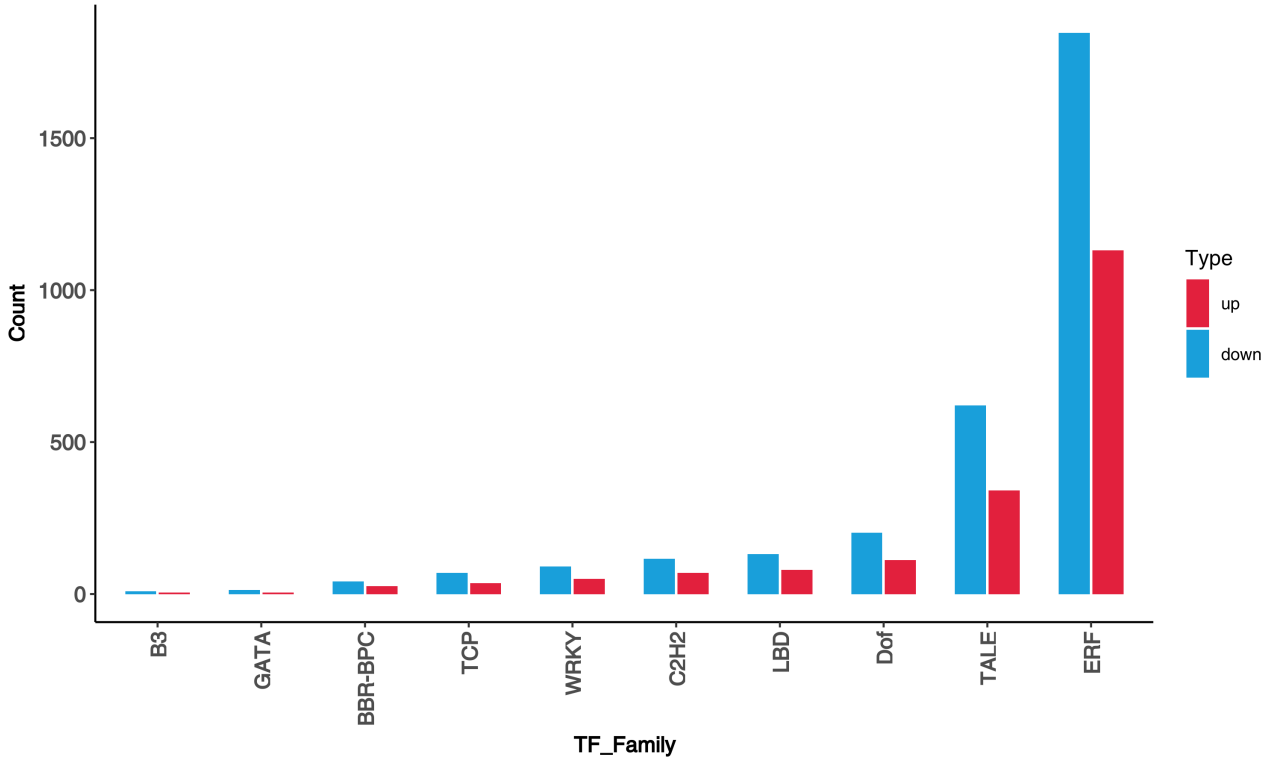
We obtained the binding sequence signatures for our transcription factors from the plantTFDB database and used FIMO to scan the 2-kb region upstream of the TSS of each gene to derive the predictions of the target genes of the transcription factors. We then screened the differential genes and their upstream transcription factors for the corresponding differential combinations from these predictions, as shown in the table below:

Table 12 Table of predicted transcription factor target genes

TF_Family	Motif_id	TF_gene_name	TF_gene_regulated
bZIP	MP00470	Os02g0132500	down
bZIP	MP00470	Os02g0132500	down
bZIP	MP00470	Os02g0132500	down
bZIP	MP00470	Os02g0132500	down
bZIP	MP00470	Os02g0132500	down

1) TF_Family：transcription factor gene family
2) Motif_id：Motif name
3) TF_gene_name：name of the transcription factor
4) TF_gene_regulated：Up- or down-regulation of transcription factor expression ### Bar chart of differential target genes in transcription factor families

The number of differential target genes corresponding to transcription factors in each gene family was counted, and a bar chart was plotted with the top 10 gene families in terms of the number of differential target genes (all gene families were presented if the number of gene families was less than 10), as shown in the figure below:



Bar chart showing the number of differential target genes corresponding to transcription factors in gene families

The horizontal coordinate indicates the gene family of the transcription factor, while the vertical coordinate indicates the number of target genes. The red color in the plot represents that these target genes show an up-regulation trend in the differential group. In contrast, the blue indicates that these target genes show a down-regulation trend in the differential group.

2.7.5 Enrichment analysis table of transcription factors corresponding to differential target genes

Significant enrichment analysis applies a hypergeometric test to detect which transcription factors are significantly enriched for differential target genes. The formula for the hypergeometric distribution is shown below:

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

Where N represents the number of transcription factor target genes in all genes, n represents the number

of differential genes in N, M represents the number of a transcription factor target gene in N, and m represents the number of a transcription factor differential target gene in M.

Table 13 Enrichment analysis table of transcription factors corresponding to differential target genes

TF_name	GeneRatio	BgRatio	pvalue
Os10g0419300(HSF)	9/4947(0.18%)	34/35618(0.1%)	0.0384417
Os04g0597300(WRKY)	36/4947(0.73%)	197/35618(0.55%)	0.0502036
Os07g0640900(HSF)	9/4947(0.18%)	36/35618(0.1%)	0.0535663
Os02g0265200(WRKY)	8/4947(0.16%)	38/35618(0.11%)	0.1483901
Os01g0242200(C2H2)	3/4947(0.06%)	10/35618(0.03%)	0.1517896

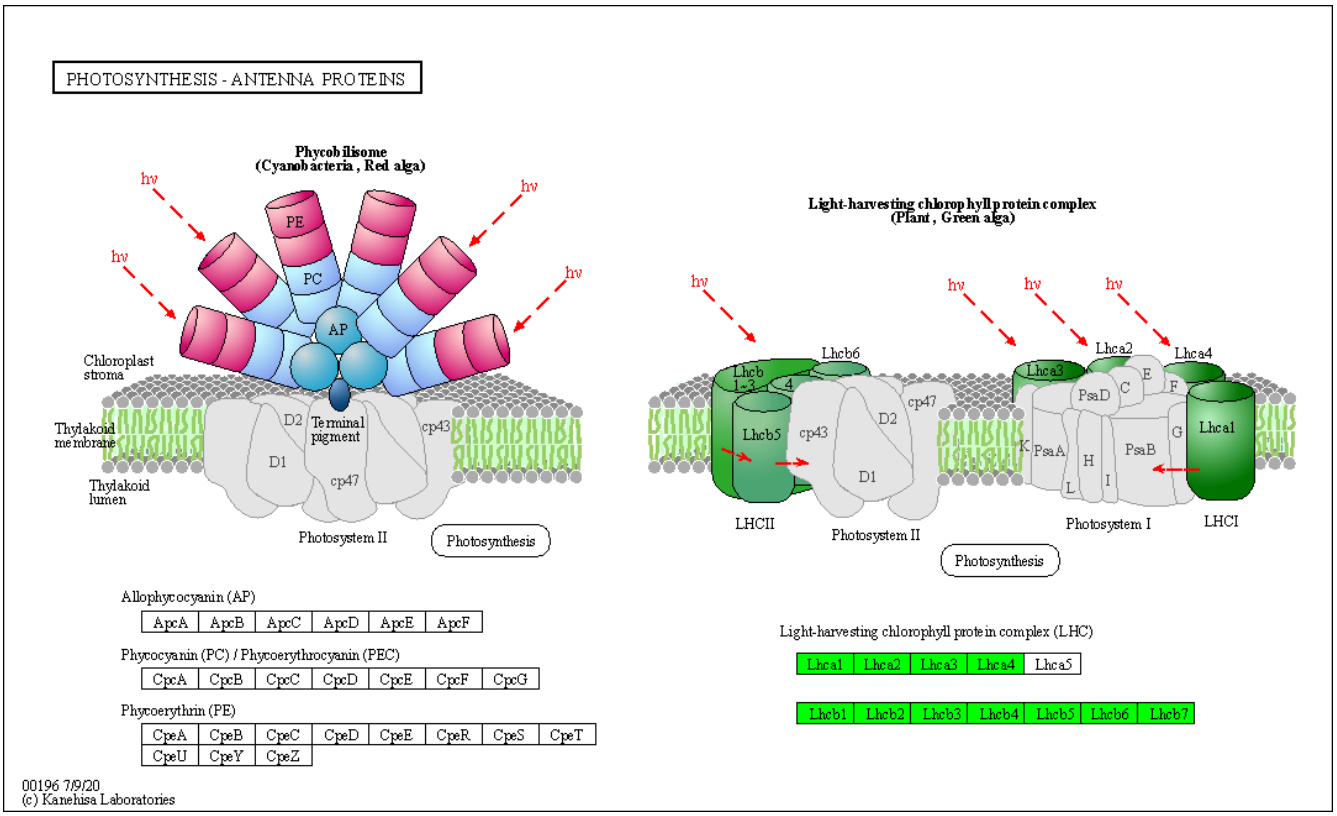
1) TF_name：Transcription factor names and corresponding gene families
2) GeneRatio：The ratio of the number of differential target genes to the total number of differential genes for this transcription factor
3) BgRatio：Ratio of the number of all target genes of this transcription factor to the total number of background genes
4) pvalue：p-value of the significance test

2.8 Differential Gene Function Annotation and Enrichment Analysis

Different gene products within an organism interact to perform biological functions, and pathway annotation analysis of differentially expressed genes helps in further understanding gene functions. The Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg) is a comprehensive database that integrates information on genomics, biological pathways, diseases, drugs, chemicals, and more[14]. KEGG seamlessly combines genomic information with high-level functional information, providing systematic analysis for the vast amounts of data generated by genome sequencing and other high-throughput experimental techniques.

2.8.1 Differentially Expressed Gene KEGG Enrichment Pathway Map

The results of the annotation of the KEGG pathway of differentially expressed genes are shown below：



Differentially Expressed Gene KEGG Enrichment Pathway Map

For the treatment group, enzymes marked in red boxes are associated with up-regulated genes and enzymes marked in green boxes are associated with down-regulated genes. The enzymes marked in blue boxes are related to both up-regulated and down-regulated genes, and the number in the box represents the enzyme number (EC number). While the whole pathway consists of complex biochemical reactions catalyzed by multiple enzymes, the enzymes related to differentially expressed genes in this pathway map are marked with different colors. Based on the differences between the study subjects, we focused on the differential expression of genes related to certain metabolic pathways, to explain the root cause of phenotypic differences through the pathway.

2.8.2 KEGG Enrichment of Differentially Expressed Genes

Pathway enrichment analysis takes the pathways in the KEGG database as units and finds pathways that are significantly enriched in differentially expressed genes compared to the whole genomic background by applying the hypergeometric test. The hypergeometric distribution is calculated using the formula shown

below:

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

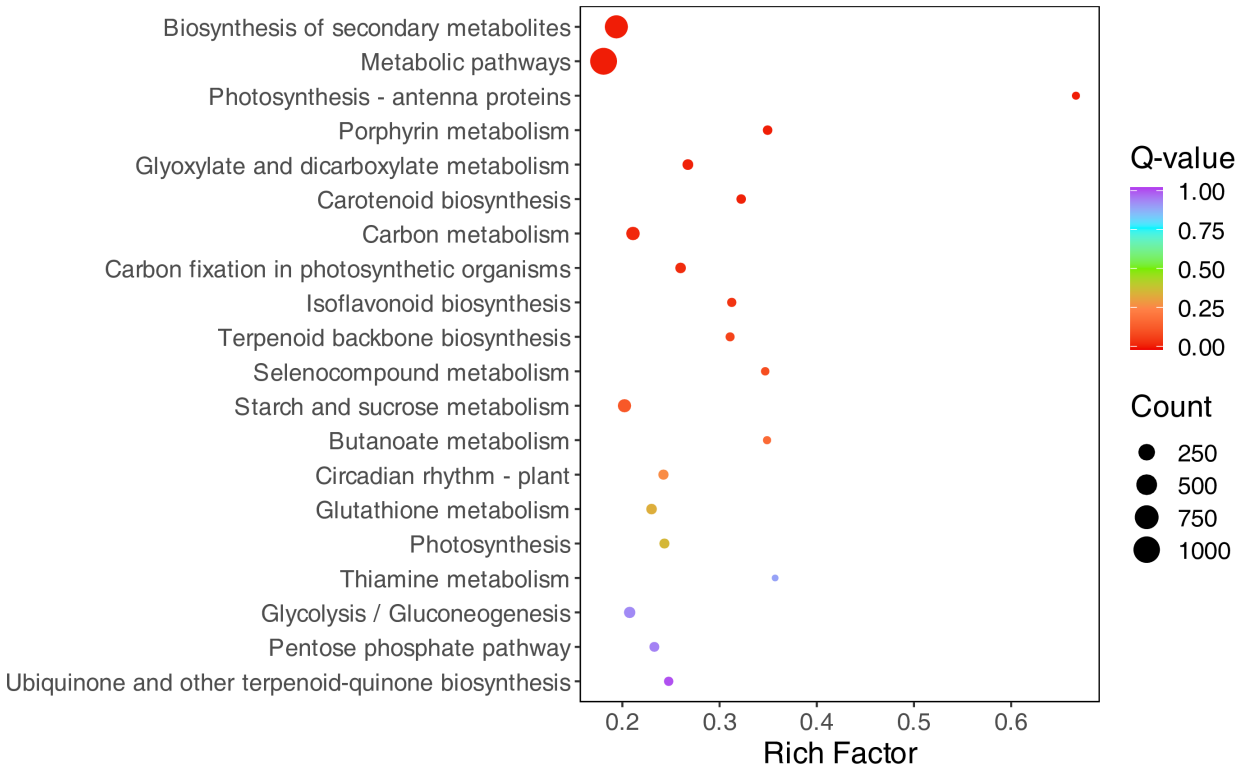
where N represents the number of genes with KEGG annotation in all genes, n represents the number of differentially expressed genes in N, M represents the number of genes in a KEGG pathway in N, and m represents the number of differentially expressed genes in a KEGG pathway in M. The results of KEGG enrichment are listed in the table below:

Table 14 KEGG Enrichment Result

ID	KEGG_level_2	Description
ko01110	Global and overview maps	Biosynthesis of secondary metabolites
ko01100	Global and overview maps	Metabolic pathways
ko00196	Energy metabolism	Photosynthesis - antenna proteins
ko00860	Metabolism of cofactors and vitamins	Porphyrin metabolism
ko00630	Carbohydrate metabolism	Glyoxylate and dicarboxylate metabolism
ko00906	Metabolism of terpenoids and polyketides	Carotenoid biosynthesis
ko01200	Global and overview maps	Carbon metabolism
ko00710	Energy metabolism	Carbon fixation in photosynthetic organisms
ko00943	Biosynthesis of other secondary metabolites	Isoflavonoid biosynthesis
ko00900	Metabolism of terpenoids and polyketides	Terpenoid backbone biosynthesis

- ID：KEGG pathway name
- Description：KEGG pathway name
- GeneRatio：ratio of the number of differentially expressed genes annotated to this pathway to the number of differentially expressed gene with annotations

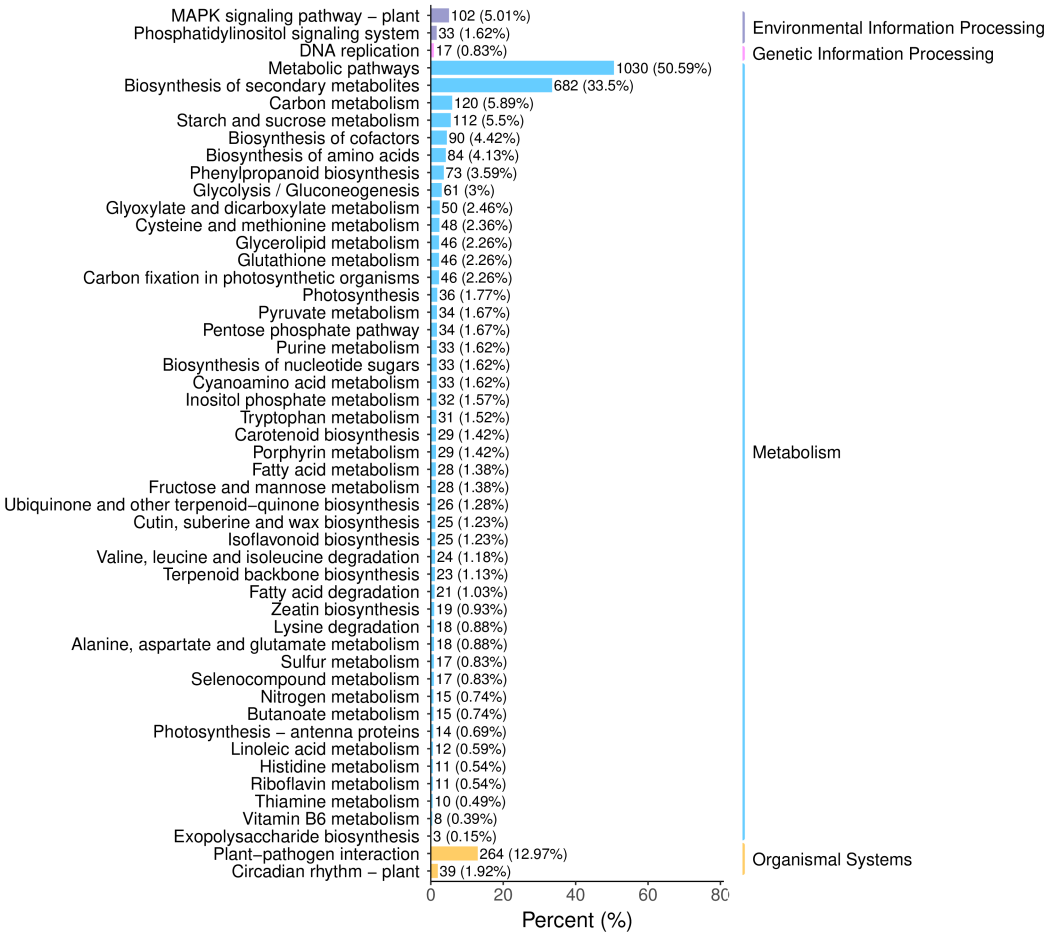
A scatter plot is a graphical presentation of the results of KEGG enrichment analysis. In this plot, the degree of KEGG enrichment is measured by rich factor, Q-value and the number of differentially expressed genes enriched in this pathway. The rich factor is the ratio of the number of differentially expressed genes enriched in the pathway to the number of all genes annotated to the pathway. The larger the rich factor, the stronger the enrichment. A smaller Q-value indicates a more significant enrichment. We selected the 20 most significantly enriched pathway entries for display in this plot, or all of them if there are less than 20 enriched pathway entries.



KEGG Enrichment Results Scatter Plot

The vertical coordinate indicates the KEGG pathway. The horizontal coordinate indicates the Rich factor. The larger the rich factor, the stronger the enrichment. The larger the dot, the greater the number of differentially expressed genes enriched in the pathway. The redder the color of the dot, the more significant the enrichment.

The 50 KEGG pathways with the lowest q-values in the enrichment analysis results were selected and the enrichment entries were plotted in a bar chart as follows:



Differentially Expressed Gene KEGG Enrichment Bar Chart

The horizontal coordinate indicates the number of differentially expressed genes annotated to the pathway, and the vertical coordinate indicates the name of the KEGG pathway. The number in the graph indicates the number of differentially expressed genes annotated to the pathway. The number in parentheses is the ratio of the number of differentially expressed genes annotated to that entry to the number of differentially expressed genes with annotations. The rightmost label represents the classification to which the KEGG pathway belongs.

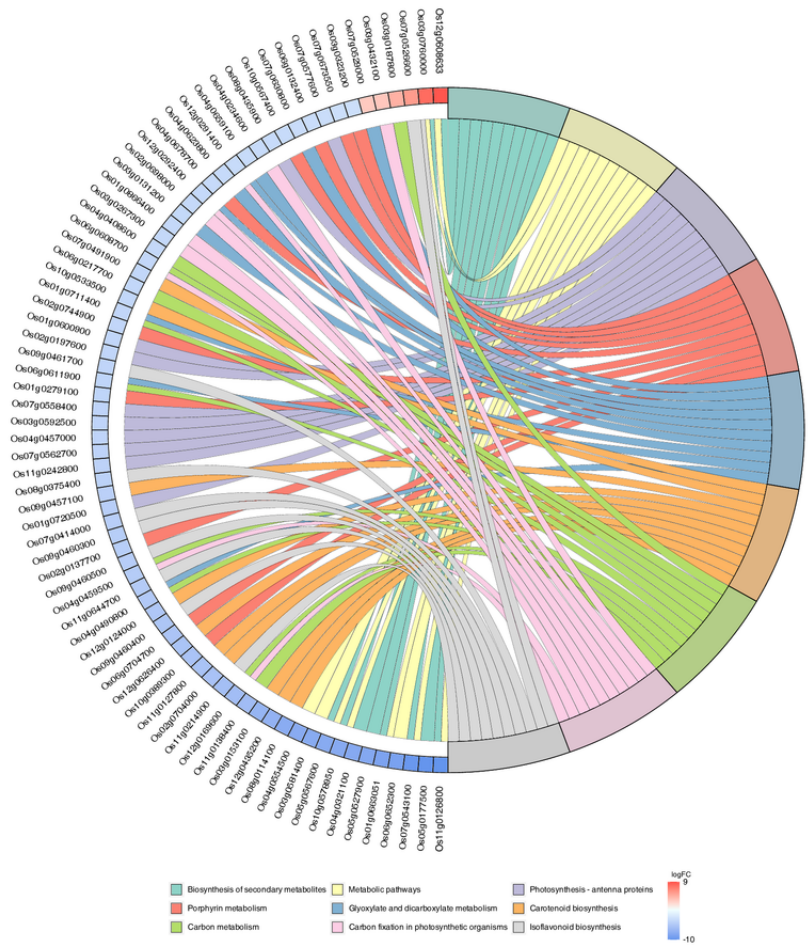
If there are multiple comparison combinations, we will provide multi-combination KEGG enrichment scatter plots (if the number of differential groupings is greater than 10, only 10 differential groupings will be displayed; if the number of differential groupings is less than 10, all of them will be displayed). The KEGG enrichment results of each comparison combination will be sorted by q-value, and the merged set of

15 pathways with the smallest q-value of each comparison combination will be displayed.

Multi-Compare KEGG Enrichment Scatter Plot

The horizontal coordinate indicates the comparison combination, and the vertical coordinate indicates the pathway with gene enrichment. The size of the dot represents the number of differentially expressed genes enriched to the pathway (the larger the dot, the more differentially expressed genes are enriched to the pathway). The color of the dot represents the significance of the enrichment to the pathway (the darker the red color of the dot, the more significant the enrichment).

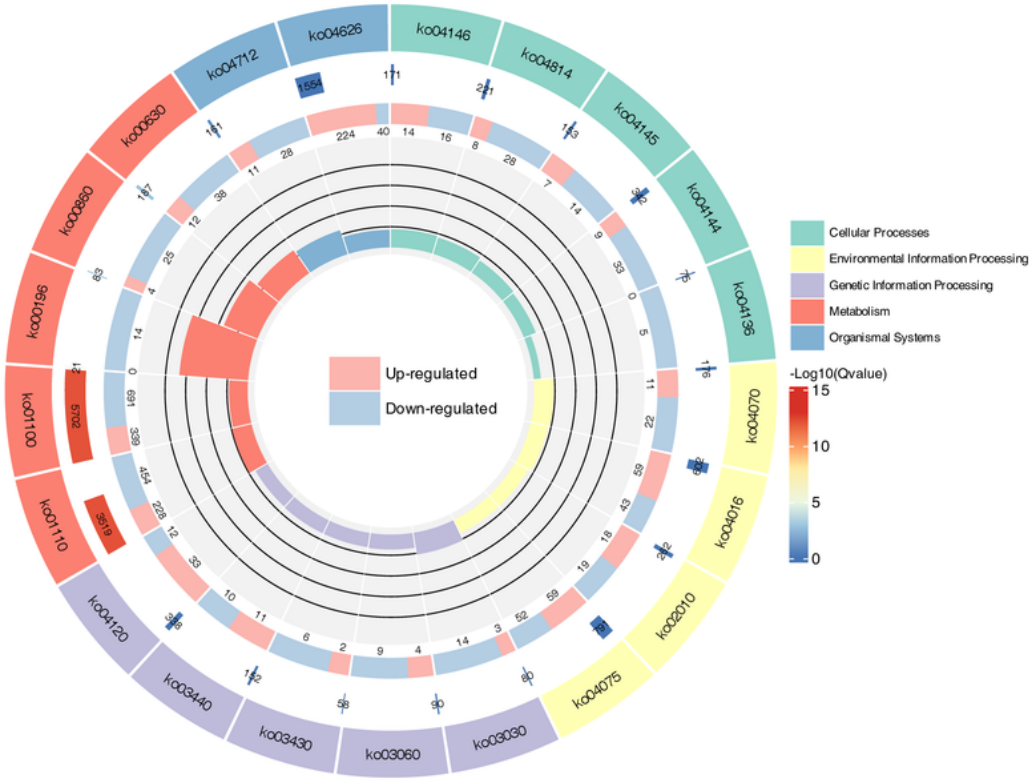
The nine pathways with the smallest q-value were selected to plot the enrichment analysis chord diagram. The diagram is divided into left and right sides: the left side shows the 10 genes with the largest |logFC| in each classification; the right side of the diagram shows the 9 pathways with the most significant enrichment; the middle line represents the correspondence between pathways and genes; the legend of the heatmap at the bottom right indicates the logFC values of genes, with red being up-regulated genes and blue being down-regulated genes; the shade of color indicates the logFC size, with darker colors indicating a larger fold of difference.



KEGG Enrichment Chord Diagram

The 9 most significantly enriched pathways are shown on the right side of the figure, the 10 genes with the largest fold change |logFC| in each pathway are shown on the left side, and the middle line represents the correspondence between pathways and genes.

We selected the 5 KEGG pathways with the smallest q-value in each KEGG_level_1 (if there are less than 5 enriched pathways in each KEGG_level_1, all of them will be shown) to plot the KEGG enrichment circular plot.

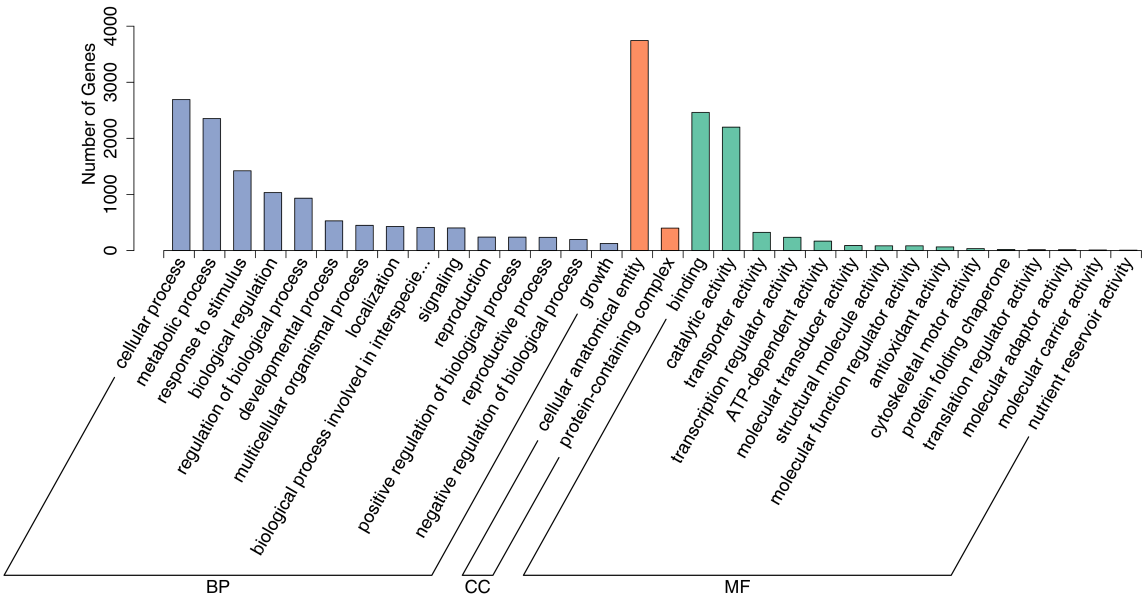


Differentially Expressed Gene KEGG Enrichment Circular Plot

From outside to inside, the first circle shows the KEGG pathways, with different colors representing different KEGG classifications; the second circle shows the number of background genes belong to that classification and the qvalue, where the more genes the longer the bar, the more significant the enrichment the darker the red color; the third circle is a bar chart of the proportion of up- and down-regulated genes, with light red representing the proportion of up-regulated genes and light blue representing the proportion of down-regulated genes, and specific values shown below; the fourth circle shows the RichFactor value for each classification (the number of foreground genes in that classification divided by the number of background genes), with each grid of the background auxiliary line indicating 0.2.

2.8.3 GO Analysis of Differentially Expressed Genes

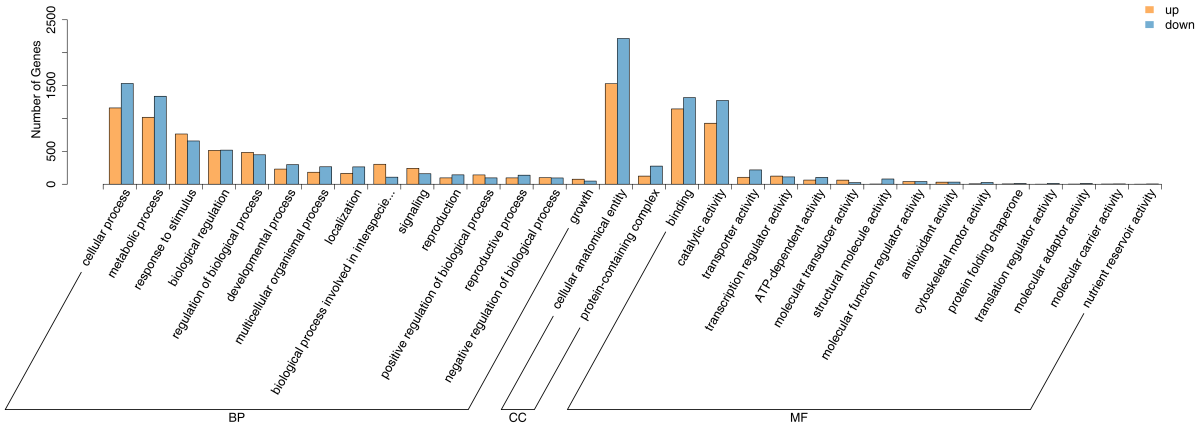
Gene Ontology[15]（GO）is an international standard classification system for gene function. As a database established by the Gene Ontology Consortium (GOC), it aims to establish a linguistic vocabulary standard that is applicable to various species, qualifies and describes the functions of genes and proteins, and can be updated as research progresses. GO is divided into three components: molecular function, biological process, and cellular component. The differentially expressed genes were sorted by the number of genes annotated with level 2 GO terms from largest to smallest. The top 15 GO terms (all GO terms if less than 15) were taken from the three categories of biological process , cellular component and molecular function, respectively, to draw a GO classification bar chart.



Classification Plot of Differentially Expressed Genes Annotated with Level 2 GO Terms

The horizontal coordinate indicates level 2 GO terms, and the vertical coordinate is the number of differentially expressed gene annotated with that GO term.

The classification statistics results of the up- and down-regulation of differentially expressed genes annotated with GO terms are shown below:



Bar Chart of Up- and Down-Regulation of Expressed Genes Annotated with Level 2 GO Terms

The horizontal coordinate indicates level 2 GO terms, and the vertical coordinate is the number of up/down-regulated differentially expressed genes for that GO term, with yellow indicating up-regulation and blue indicating down-regulation.

2.8.4 GO Enrichment Analysis of Differentially Expression Expressed Genes

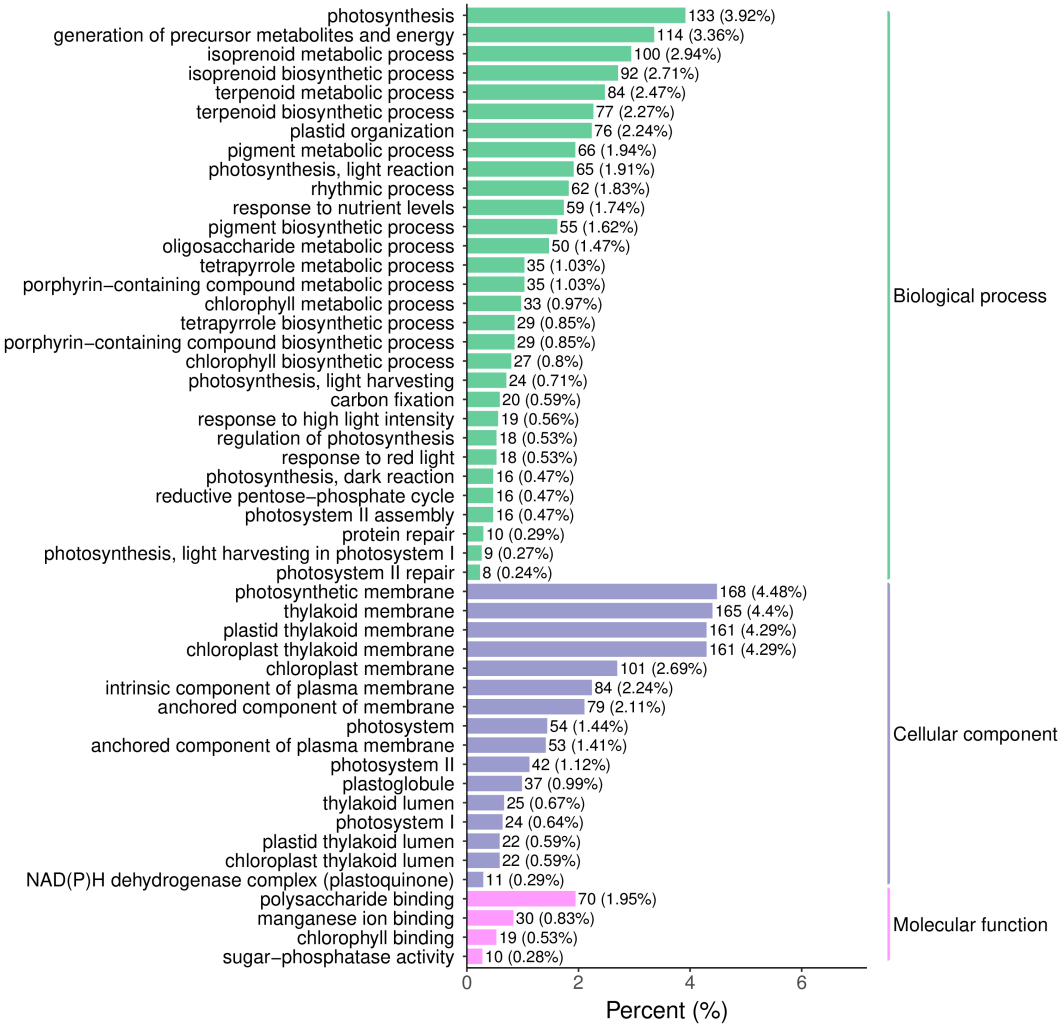
After screening the differentially expressed genes based on the experimental objectives, enrichment analysis was performed to study the distribution of differentially expressed genes in Gene Ontology in order to elucidate the functional representations of the differences in the experimental samples. The principle of ordinary GO enrichment analysis is hypergeometric distribution. GO-Term enrichment analysis takes GO terms in the GO database as units and applies hypergeometric tests to identify GO terms that are significantly enriched in differentially expressed genes compared to the whole genomic background. The results of the enrichment analysis are shown in the table below:

Table 15 GO Enrichment Analysis of Differentially Expressed Genes

GO_level_1	GO	Description	DiffRatio
Biological process	GO:0015979	photosynthesis	133/3396 3.92%
Biological process	GO:0019684	photosynthesis, light reaction	65/3396 1.91%
Biological process	GO:0009765	photosynthesis, light harvesting	24/3396 0.71%
Biological process	GO:0010207	photosystem II assembly	16/3396 0.47%
Biological process	GO:0042440	pigment metabolic process	66/3396 1.94%
Biological process	GO:0006779	porphyrin-containing compound biosynthetic process	29/3396 0.85%
Biological process	GO:0015995	chlorophyll biosynthetic process	27/3396 0.8%
Biological process	GO:0033014	tetrapyrrole biosynthetic process	29/3396 0.85%
Biological process	GO:0006091	generation of precursor metabolites and energy	114/3396 3.36%
Biological process	GO:0015994	chlorophyll metabolic process	33/3396 0.97%

- GO_level_1：GO ontology type
- GO：GO term ID
- Description：function description for the GO term
- DiffRatio：ratio of the number of differentially expressed genes annotated with the GO term to the total number of differentially expressed genes

The 50 GO terms with the lowest q-value from the enrichment analysis were selected to plot bar charts of the enrichment entries, as shown in the following figures:

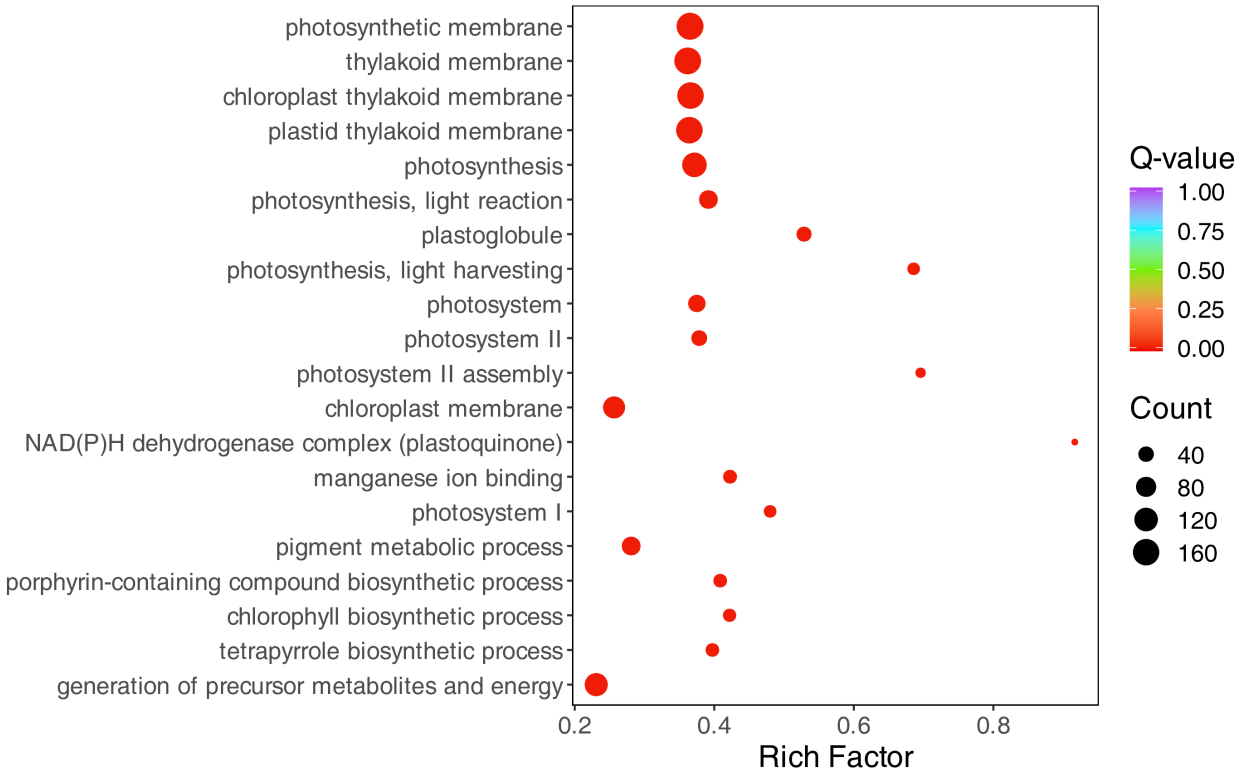


Differentially Expressed Gene GO Enrichment Bar Chart

The horizontal coordinate indicates the number of differentially expressed genes annotated with the term, and the vertical coordinate indicates the name of the GO term. The number in the graph indicates the number of differentially expressed genes annotated with the term. The number in parentheses is the ratio Goals the number of differentially expressed genes annotated with that term to the number of differentially expressed genes with annotations. The rightmost label represents the classification to which the GO term belongs.

A scatter plot is a graphical presentation of the results of GO enrichment analysis. In this plot, the degree of GO enrichment is measured by rich factor, Q-value and the number of differentially expressed genes enriched in this entry. The rich factor is the ratio of the number of differentially expressed genes

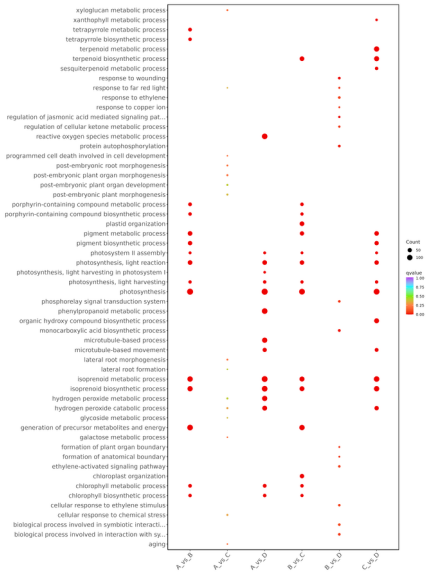
enriched in the pathway to the number of all genes annotated to the pathway. The larger the rich factor, the stronger the enrichment. A smaller Q-value indicates a more significant enrichment. We selected the 20 most significantly enriched GO terms for display in this plot, or all of them if there are less than 20 enriched GO terms.



Differentially Expressed Gene GO Enrichment Scatter Plot

The vertical coordinate indicates the GO term and the horizontal coordinate indicates the Rich factor. The larger the rich factor, the stronger the enrichment. The larger the dot, the greater the number of differentially expressed genes enriched in the pathway. The darker the red color of the dot, the more significant the enrichment.

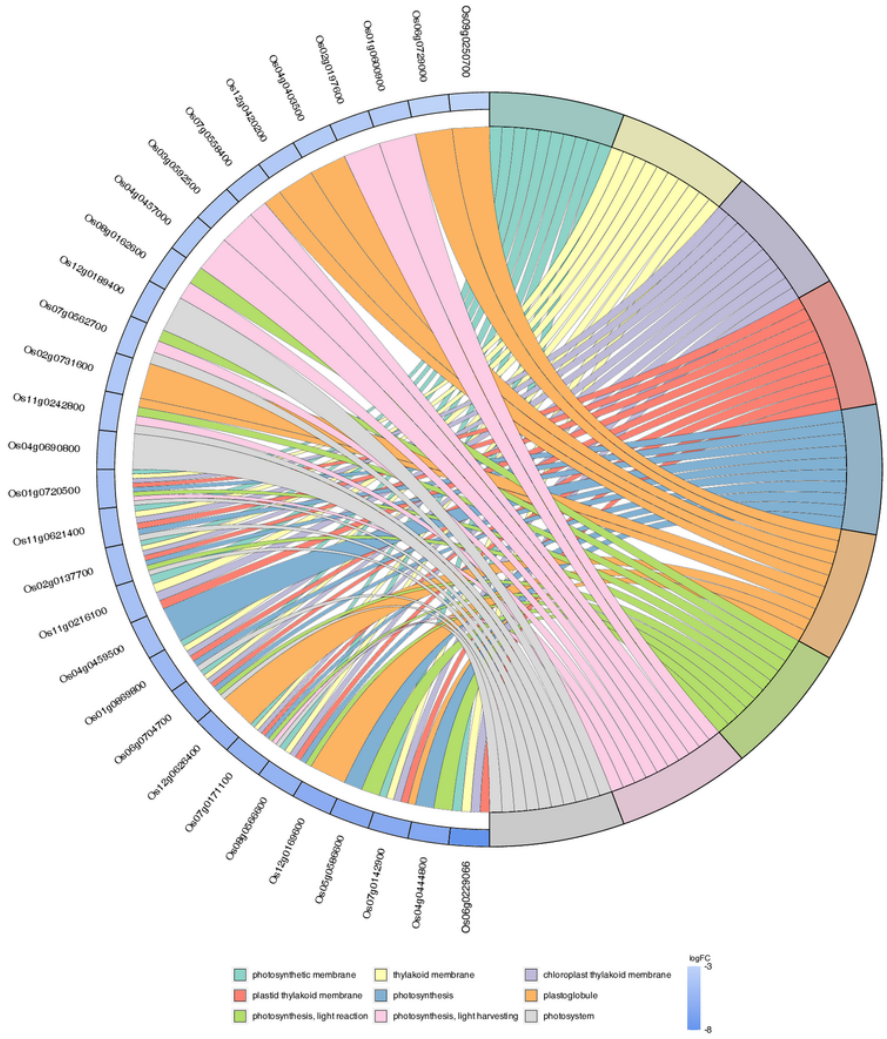
If there are multiple comparison combinations, we will provide multi-combination GO enrichment scatter plots (if the number of differential groupings is greater than 10, only 10 differential groupings will be displayed; if the number of differential groupings is less than 10, all of them will be displayed). The GO enrichment results of each comparison combination will be sorted by q-value, and the merged set of 15 entries with the smallest q-value of each comparison combination will be displayed.



Multi-Compare GO Enrichment Scatter Plot

The horizontal coordinate indicates the comparison combination, and the vertical coordinate indicates the GO term with gene enrichment. The size of the dot represents the number of differentially expressed genes enriched to the GO term (the larger the dot, the more differentially expressed genes are enriched to the pathway). The color of the dot represents the significance of the enrichment to the GO term (the darker the red color of the dot, the more significant the enrichment).

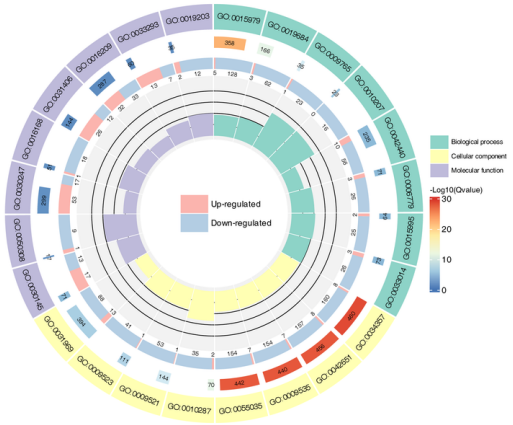
The nine GO terms with the smallest q-value were selected to plot the enrichment analysis chord diagram. The diagram is divided into left and right sides: the left side shows the 10 genes with the largest |logFC| in each classification; the right side of the diagram shows the 9 GO terms with the most significant enrichment; the middle line represents the correspondence between GO terms and genes; the legend of the heatmap at the bottom right indicates the logFC values of genes, with red being up-regulated genes and blue being down-regulated genes; the shade of color indicates the logFC size, with darker colors indicating a larger fold of difference.



Differentially Expressed Gene GO Enrichment Chord Diagram

The 9 most significantly enriched GO terms are shown on the right side of the figure, the 90 genes with the largest fold change |logFC| in these 9 GO terms are shown on the left side, and the middle line represents the correspondence between pathways and genes.

We selected the 8 GO terms with the smallest q-value in the three major GO categories (biological process, cellular component and molecular function) (all GO terms are shown in case of less than 8 enriched GO terms from the three major GO categories) to plot the GO enrichment circular plot.

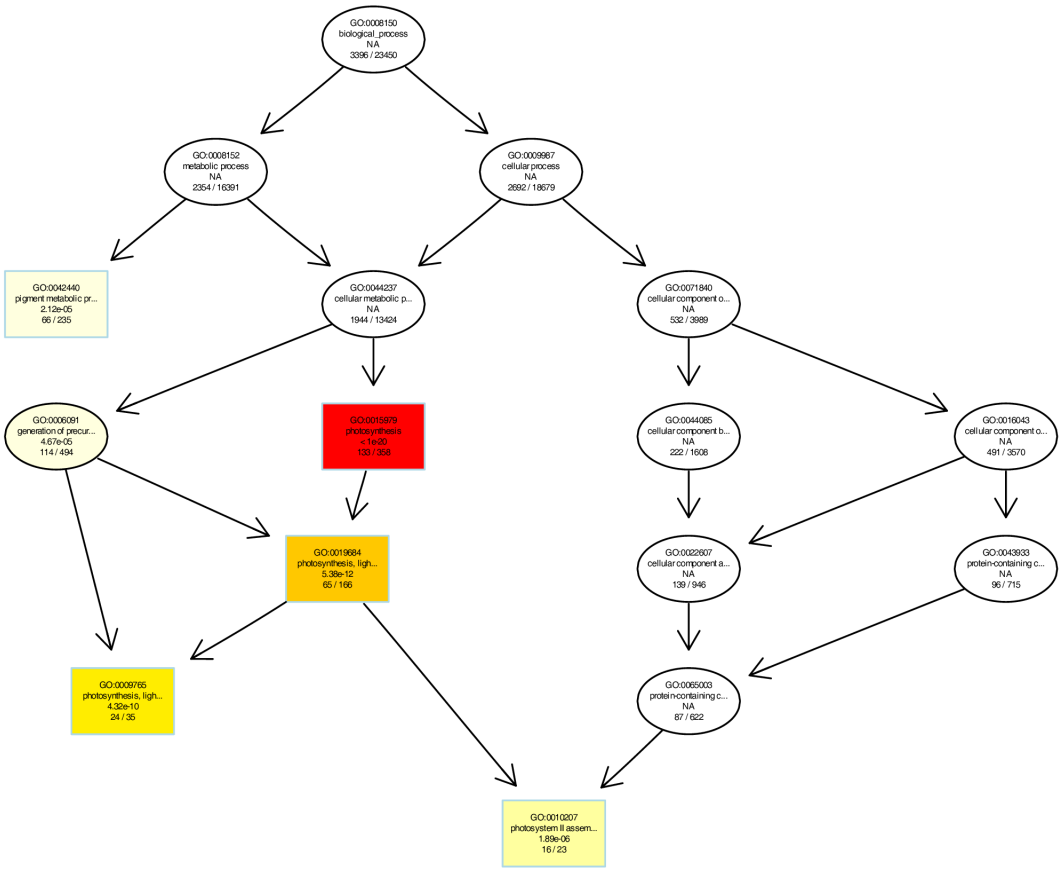


Differentially Expressed Gene GO Enrichment Circular Plot

From outside to inside, the first circle shows the GO terms, with different colors representing different GO classifications; the second circle shows the number of background genes belong to that classification and the qvalue, where the more genes the longer the bar, the more significant the enrichment the darker the red color; the third circle is a bar chart of the proportion of up- and down-regulated genes, with light red representing the proportion of up-regulated genes and light blue representing the proportion of down-regulated genes, and specific values shown below; the fourth circle shows the RichFactor value for each classification (the number of foreground genes in that classification divided by the number of background genes), with each grid of the background auxiliary line indicating 0.2.

2.8.5 GO Enrichment Level Analysis of Differentially Expression Expressed Genes

The enrichment analysis was performed on the differentially expressed genes between samples, and the enriched terms were taken to plot topGO directed acyclic graphs. The topGO directed acyclic graph visualizes enriched GO nodes (terms) and their hierarchical relationships for differentially expressed genes, and is a graphical representation of the results of differentially expressed gene GO enrichment analysis, with branches representing inclusion relationships and increasingly specific functional descriptions defined from top to bottom. The topGO molecular function directed acyclic graphs of differentially expressed genes between samples are shown below:



Directed Acyclic Graph of Enriched GO Terms

Each node represents a GO term, with the rectangle representing the top 5 selected GO terms with the highest enrichment, and the ellipse representing the contained nodes. The colors of the rectangles and ellipses represent the relative enrichment. From bright yellow to dark red indicates a decreasing p-value, i.e., increasing significance, while white represents non-significance. Each node corresponds to 4 lines of data, indicating the ID of the GO term, the function, the corrected P-value, the number of differentially expressed genes annotated with the GO term and the total number of genes.

2.8.6 KOG Analysis of Differentially Expressed Genes

The Clusters of Orthologous Genes (COG) database[16]（https://www.ncbi.nlm.nih.gov/COG/ is a protein database created and maintained by the NCBI. It is constructed based on the evolutionary relationships

of encoded proteins in complete genomes of bacteria, algae, and eukaryotes. Using the similarity of protein sequences, the database classifies proteins into different groups, assigning each group a unique identifier representing a homologous protein. The COG database consists of two parts: COG and KOG. The former clusters homologous proteins in prokaryotes, making it suitable for annotating prokaryotic genes, while the latter clusters homologous proteins in eukaryotes, making it suitable for annotating eukaryotic genes. Protein sequences or cDNA sequences are aligned to the KOG database using the Diamond software, and annotations from the KOG database are then extracted.

2.8.6.1 KOG Annotation of Differentially Expressed Genes

After aligning differentially expressed genes to the KOG database, relevant annotation information is extracted based on the database protein IDs, as shown in the table below:

Table 16 Differentially Expressed Genes KOG Annotation

query	subject	evalue	KOG
Os04g0444800	At5g49740	0	KOG0039
Os04g0444800	At5g49740	0	KOG0039
Os08g0189900	At5g39110	0	NA
Os08g0190100	At5g39110	0	NA
Os11g0707000	At2g39730	0	KOG0651
Os04g0127200	At4g10540	0	NA
Os09g0553900	At1g42550	0	NA
Os02g0744900	At1g74470	0	NA
Os10g0409400	At1g70370	0	NA
Os03g0265900	At4g37300	0	NA

- query：Differentially expressed genes ID
- subject：KOG ID
- evalue：The expected value of the reliability of diamond comparison results, the lower the value, the more reliable the comparison
- KOG：KOG ID

2.8.6.2 Classification Statistics of KOG Annotations

Based on the KOG annotation results, the number of differentially expressed genes contained in each KOG functional classification is counted, as shown in the table below:

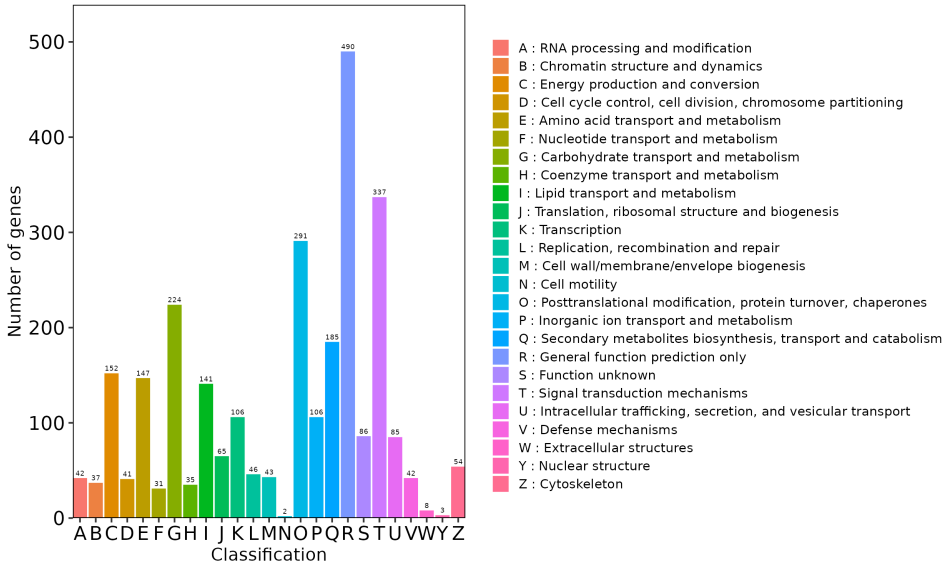
Table 17 Differentially Expressed Gene KOG Classification

Classification	Code	CodeFunction
CELLULAR PROCESSES AND SIGNALING	D	Cell cycle control, cell division, chromosome partitioning
CELLULAR PROCESSES AND SIGNALING	M	Cell wall/membrane/envelope biogenesis
CELLULAR PROCESSES AND SIGNALING	N	Cell motility
CELLULAR PROCESSES AND SIGNALING	O	Posttranslational modification, protein turnover, chaperones
CELLULAR PROCESSES AND SIGNALING	T	Signal transduction mechanisms

- Classification：The first-level classification of KOG
- Code：Functional classification of KOG IDs, encoded with single letters
- CodeFunction：Description of the functional classification of KOG

2.8.6.3 Bar Chart of Classification Statistics for KOG Annotations

A bar chart is generated using the classification statistics information from KOG annotations, as shown in the figure below:



Bar Chart of Classification Statistics for KOG Annotations

The abscissa represents the functional classification (Code) of KOG ID, the ordinate represents the number of differential genes included, and different classifications are indicated by different colors. The legend is Code plus a description of its function.

2.9 Gene Set Enrichment Analysis (GSEA)

Conventional enrichment analysis based on hypergeometric distribution relies on significantly upregulated or downregulated genes, which can easily miss genes with biologically significant differences in expression that are not statistically significant. Gene Set Enrichment Analysis (GSEA)[17] does not require specifying a specific threshold for differentially expressed genes. Instead, it ranks genes based on their differential expression between two groups of samples and uses statistical methods to test whether a predefined gene set is enriched at the top or bottom of the ranked list. The principles are as follows:

(1)Based on the expression data of all genes, calculate the differential expression (signal-to-noise) of each gene in two groups, ClassA and ClassB. Then, sort genes in descending order of differential expression between the two phenotypes to create a sorted gene list.

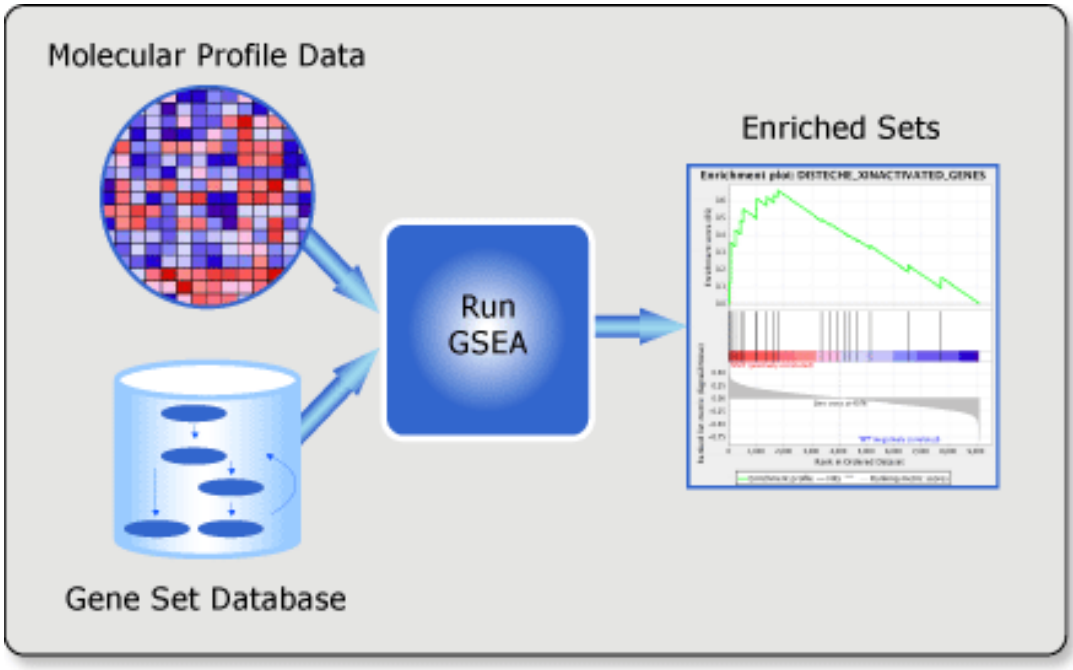
(2)Determine whether the genes in gene set S are enriched at the top or bottom of the sorted list.

(3) Calculate the Enrichment Score (ES) for gene set S. The calculation proceeds by starting with the first gene in the target gene list L and computing a cumulative statistic. When encountering a gene within gene set S, the statistic is increased, and when encountering a gene not in gene set S, the statistic is

decreased, with the magnitude of the increment depending on the gene's correlation with the phenotype. The highest peak in the cumulative statistic becomes the Enrichment Score (ES).

(4) Calculate the significance level (nominal p-value) of ES. An empirical phenotype-based permutation test is used to calculate the nominal p-value of ES, preserving the complex correlations in the original expression data.

(5) Multiple hypothesis testing. Taking into account the size of the gene set, the ES for each gene set is normalized to obtain the Normalized Enrichment Score (NES). False discovery rate (FDR) is calculated to control the false positive rate.



GSEA

GSEA analysis is applied on the KEGG pathway

Table 18 KEGG GSEA

ko_ID	SIZE	ES	NES
ko00750	21	0.6716270	2.188919
ko00261	20	0.6299651	2.105924
ko00860	73	0.5806948	2.012102
ko00900	68	0.5652509	1.956910
ko00195	115	0.6896130	1.954996
ko00130	92	0.5235105	1.929256
ko00260	133	0.3882534	1.923728
ko00710	148	0.4964949	1.862915
ko00670	21	0.7050098	1.859192
ko01240	414	0.3693058	1.828304

- ko_ID：KEGG Pathway ID corresponding to the gene set
- SIZE：Number of genes included in the gene set (greater than 15 and less than 5000)
- ES：Enrichment Score, a measure of enrichment
- NES：Normalized Enrichment Score, a standardized enrichment score that takes into account the size of the gene set

GO GSEA Results（Currently the default is to use the GO secondary classification term as the gene set）

Table 19 GO GSEA

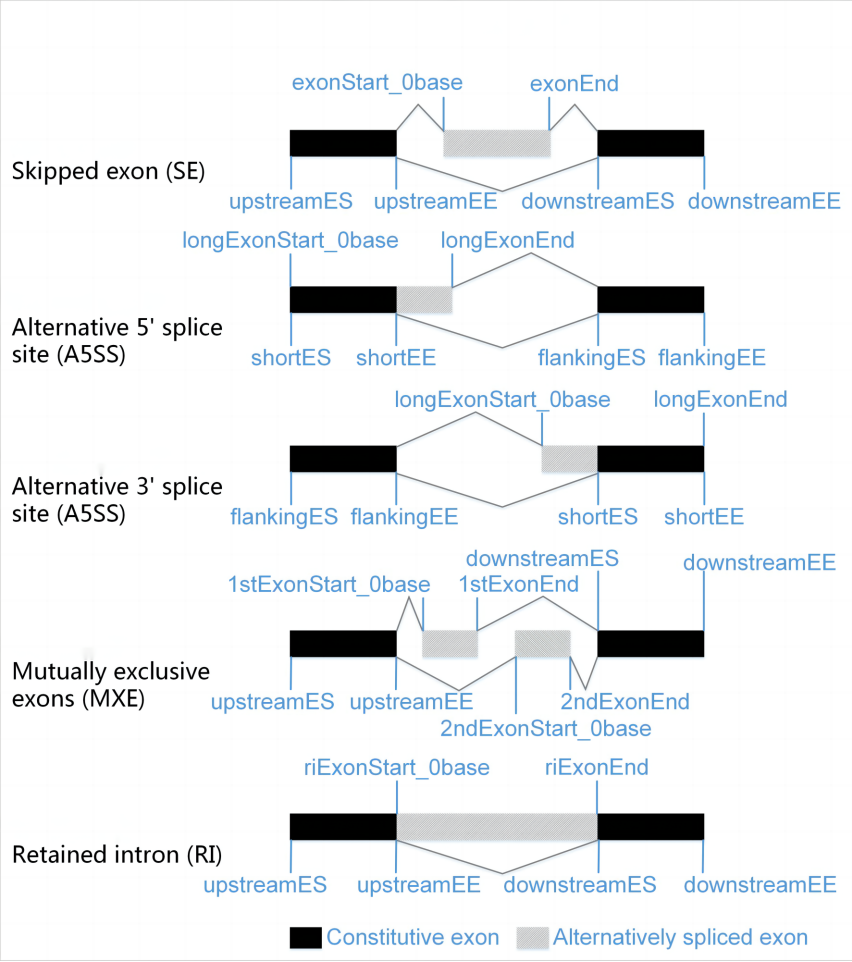
GO_ID	SIZE	ES	NES
GO:0005198	551	0.4743186	1.642832
GO:0003774	129	0.5071854	1.633752
GO:0098754	139	0.4023021	1.601840
GO:0005215	1665	0.3317157	1.559122
GO:0048511	227	0.2990366	1.540665
GO:0044183	92	0.3330815	1.363433
GO:0051179	2834	0.2358377	1.360565
GO:0045182	155	0.2445140	1.257090
GO:0032991	3232	0.2258511	1.244754
GO:0098772	563	0.2433788	1.172992

- GO_ID：GO Term ID corresponding to the gene set
- SIZE：Number of genes included in the gene set (greater than 15 and less than 5000)
- ES：Enrichment Score, a measure of enrichment
- NES：Normalized Enrichment Score, a standardized enrichment score that takes into account the size of the gene set

2.10 Alternative Splicing Analysis

Alternative splicing (AS) is a common form of gene expression in most eukaryotic cells. The gene sequence of eukaryotic cells contains introns and exons, where introns are removed by RNA spliceosomes after the gene is transcribed into mRNA precursors, while exons are retained in mature mRNAs. An unspliced RNA may be spliced with multiple exon splicing forms, therefore allowing a gene to be translated into different protein isoforms at different times and in different environments, and thereby increasing the complexity or adaptability of the system in its physiological status. Alternative splicing analysis of transcriptome data in this project was implemented using rMATS[18].rMATS quantifies the expression of alternative splicing events in different biological replicates, and calculates the P value using likelihood-ratio test to represent the difference of alternative splicing events between two groups of samples. To control the false discovery rate (FDR), the Benjamini Hochberg method is then used to adjust the P-values for multiple hypothesis testing. Alternative splicing events with FDR less than 0.05 are considered as differential alternative splicing. There

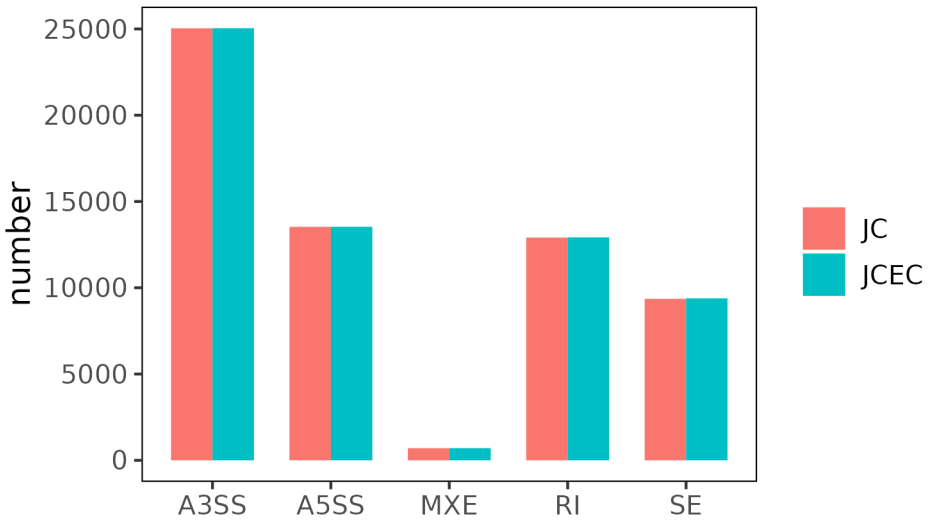
are five types of alternative splicing events identified by rMATS in this study, as shown below: （1）Skipped exon (SE)（2）Alternative 5’ splice site (A5SS)（3）Alternative 3’ splice site (A3SS)（4）Mutually exclusive exons (MXE) （5）Retained intron (RI)



Plot of Alternative Splicing Events Identified by rMATS

2.10.1 Classification and Statistics of Alternative Splicing Events

For each differential grouping, we analyzed the types of alternative splicing events and calculated their numbers using rMATS. We then computed the expressions of each category of alternative splicing events separately, and finally performed a differential analysis of those alternative splicing events. rMATS employs two quantification methods, JC and JCEC. JC only uses reads across splicing junctions, while JCEC uses reads across splicing junctions as well as reads that fully mapped to optional exons. Statistics on types and numbers of alternative splicing events for each differential grouping are shown in the following figures:



Statistical Chart of Types and Numbers of Alternative Splicing Events

The horizontal coordinate indicates the different alternative splicing event types and the vertical coordinate indicates the number of alternative splicing events. JC only uses reads across splicing junctions, while JCEC uses reads across splicing junctions as well as reads that fully mapped to optional exons for quantification.

Table 20 Statistics on Types and Numbers of Alternative Splicing Events

AS_type	EventNum.JC	SigEventNum.JC	EventNum.JCEC	SigEventNum.JCEC
A3SS	25035	1167 (611;556)	25040	1169 (619;550)
A5SS	13530	626 (370;256)	13534	625 (374;251)
MXE	705	97 (39;58)	705	98 (40;58)
RI	12909	1102 (606;496)	12921	1161 (642;519)
SE	9360	900 (620;280)	9386	939 (652;287)

- AS_type：type of the alternative splicing event
- EventNum.JC：the total number of alternative splicing events quantified using the JC method
- SigEventNum.JC：the number of differential alternative splicing events quantified using the JC method. The number before the semicolon in parentheses is the number of up-regulated differential alternative splicing events, and the number after the semicolon is the number of down-regulated differential alternative splicing events

- EventNum.JCEC：the total number of alternative splicing events quantified using the JCEC method
- SigEventNum.JCEC：the number of differential alternative splicing events quantified using the JCEC method. The number before the semicolon in parentheses is the number of up-regulated differential alternative splicing events, and the number after the semicolon is the number of down-regulated differential alternative splicing events

2.10.2 Alternative Splicing

The screening criteria for alternative splicing events was FDR < 0.05. The results of alternative splicing analysis of the skipped exon (SE) type by JC quantification are shown in the following table:

Table 21 Skipped Exon (SE) Alternative Splicing Analysis Results by JC Quantification

GeneID	exonStart_0base	exonEnd	upstreamES	upstreamEE
Os12g0516300	20062924	20062966	20061913	20062267
Os12g0459300	16099758	16100032	16099511	16099662
Os12g0404400	12124557	12124704	12124075	12124139
Os12g0404400	12131036	12131153	12124075	12124139
Os12g0255200	8729314	8729425	8729062	8729241
Os12g0226400	6895550	6895717	6890657	6890835
Os11g0303200	11397790	11398118	11397509	11397648
Os11g0303200	11398550	11398641	11396899	11397648
Os11g0256900	8467311	8467728	8466831	8466997
Os11g0251400	8168856	8169200	8165632	8166577

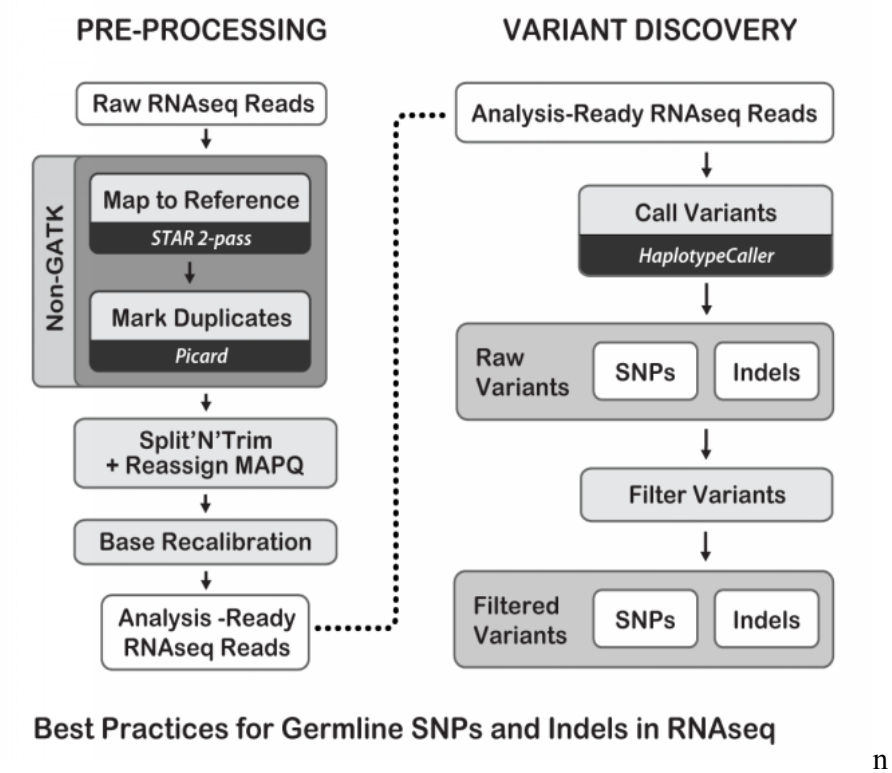
- GeneID：ID of the gene where the alternative splicing occurred
- exonStart_0base：the start position of the skipped exon event (the base prior to the starting position of the skipped exon, the starting point of the diagonally striped part in the schematic of the alternative splicing event identified by rMATS)
- exonEnd：the termination position of the skipped exon event (the ending site of the skipped exon, the termination point of the diagonally striped part in the schematic of the alternative splicing event identified by rMATS)
- upstreamES：the start position of the exon upstream of the skipped exon event (the start position of the black exon on the left in the schematic of the alternative splicing event identified by rMATS)

- upstreamEE：the termination position of the exon upstream of the skipped exon event (the end point of the black exon on the left in the schematic of the alternative splicing event identified by rMATS)

2.11 SNP and InDel Analysis

Single Nucleotide Polymorphism (SNP) refers to a genetic marker formed by a single nucleotide variation on the genome. SNP markers are characterized by rich polymorphism, wide distribution and high genetic stability. They allow high-throughput, automated assays with the advantages of low cost and high efficiency, thus serving as an important link between biological phenotypes and genotypes. InsertionDeletion (InDel) refers to the insertion or deletion of a nucleotide fragment of different sizes at the same locus of the genome between individuals of closely related species or the same species, i.e., the insertion or deletion of one or more bases at a locus of a sequence compared to another homologous sequence. InDel markers have been widely used in genetic analysis of plant and animal populations and molecular assisted breeding because of their excellent stability, high polymorphism and simple genotyping system. In general, SNPs are single nucleotide variants with a variation frequency greater than 1 %, and most of the InDels are within 50 bp in length.

Since some of the mRNAs are subject to RNA editing, i.e., insertion, deletion or substitution of bases in the coding region of the transcribed RNAs, resulting in polymorphic gene expression products. Based on the alignment results, the RNA editing results for SNP and single nucleotide substitution are the same. Therefore, the SNPs and InDels detected by transcriptome sequencing data inevitably contain the products of RNA editing. SNP and InDel detection in this study was implemented using GATk[19]and annotation was implemented using ANNOVAR[20].The flow of SNP and InDel detection is shown in the following figure:



n

图 1 1:

2.11.1 Detection of Variable Sites

After detecting SNPs and InDels using GATK, the variable sites were annotated using ANNOVAR to obtain the analysis results and annotation information of SNPs and InDels. Since the analysis results for InDels are consistent with SNPs, only some of the results for SNPs are presented in the report, as shown in the table below:

Table 22 SNP Loci and Annotations

Chr	Pos	Ref	Alt	location	Gene
1	17710	C	T	exonic	Os01g0100500
1	18823	T	C	intronic	Os01g0100500
1	23025	A	T	5' UTR	Os01g0100600
1	34057	C	T	exonic	Os01g0100800
1	35928	G	T	exonic	Os01g0100900
1	63971	C	T	exonic	Os01g0101300
1	151499	A	G	downstream	Os01g0102700
1	172975	C	T	ncRNA_exonic	Os01g0103050
1	172993	G	A	ncRNA_exonic	Os01g0103050
1	194698	C	T	3' UTR	Os01g0103700

- Chr：the chromosome where the SNP locus is located
- Pos：the coordinates of the SNP locus
- Ref：the reference base corresponding to the SNP locus
- Alt：the variant base corresponding to the SNP locus
- location：the type of gene element where the SNP locus is located, where exonic refers to the exon in the coding region
- Gene：the gene where the SNP is located or multiple genes in close proximity

2.11.2 Statistics of the Regions of the Variable Sites

The number of SNPs of different types was counted according to the types of genetic elements in which the SNPs were located. In this case, exonic refers to exons in the coding region. Since the analysis for the InDel markers is similar to that for the SNP markers, the report only shows the statistical results for the SNP markers, as shown in the table below:

Table 23 Statistics of the Regions of the Variable Sites

Sample	exonic	5' UTR	3' UTR	intronic	splicing	upstream	downstream	intergenic
A1	32350	3253	11365	4324	64	1308	2019	3358
A2	29033	2995	10877	4313	65	1196	1833	2968
A3	28666	2902	10834	4116	62	1150	1814	2930
B1	14625	1268	5255	974	11	459	443	862
B2	27764	2782	10610	3808	58	1081	1693	2751
B3	29064	3010	10803	4598	75	1169	1774	3184
C1	28645	2886	10787	3908	57	1155	1841	2890
C2	29659	3091	10997	4794	75	1248	1837	3236
C3	22092	1869	9367	1755	24	716	1288	1593
D1	25077	2362	9973	2586	36	916	1454	2272

- Sample：sample name
- exonic：coding region exon
- 5' UTR：5' end untranslated region (UTR)
- 3' UTR：3' end untranslated region (UTR)
- intronic：intron
- splicing：indicates that the SNP is located within 2 bp of the splice site
- upstream：upstream of the gene, the variant site within 1 kb upstream of the transcription start site
- downstream：downstream of the gene, the variant site within 1 kb downstream of the transcription termination site
- intergenic：intergenic region

2.11.3 Statistics on the Functions of Variable Sites

The number of synonymous mutations, missense mutations, nonsense mutations, terminator codon mutations and unknown types of SNPs were counted based on their effects on the gene. The report only shows the statistical results for the SNP markers, as shown in the table below:

Table 24 Statistics on the Functions of Variable Sites

Sample	synonymous SNV	nonsynonymous SNV	stopgain	stoploss	unknown
A1	15842	16227	223	74	27444
A2	13970	14793	214	67	25906
A3	13845	14561	208	64	25419
B1	7406	7108	87	31	9869
B2	13403	14111	198	63	24307
B3	13998	14796	217	66	26239
C1	13818	14566	201	71	25135
C2	14228	15155	223	66	26991
C3	10882	11034	133	52	17817
D1	12331	12536	162	56	20929

- Sample：sample name
- synonymous SNV：synonymous mutation
- nonsynonymous SNV：missense mutation
- stopgain：nonsense mutation
- stoploss：terminator codon mutation
- unknown：unknown type, indicating that the mutation is located in the non-coding region and the effect is unknown

2.12 Weighted Gene Co-expression Network Analysis (WGCNA)

The WGCNA algorithm is a system biology algorithm for constructing gene co-expression networks based on high-throughput messenger RNA (mRNA) expression data, which is widely used in biomedical fields worldwide. The WGCNA algorithm first assumes that the gene network obeys a scale-free distribution, and defines the gene co-expression correlation matrix, the adjacency function of the gene network, and then calculates the dissimilarity coefficients of different nodes, and constructs the hierarchical clustering tree accordingly. Different clads (branches) of this clustering tree represent different gene modules, with high degree of gene co-expression within a same module and low degree of gene co-expression in different modules. Finally, the association between modules and specific phenotypes or diseases is explored for the purpose of identifying target genes and gene networks.

2.12.1 Sequencing Data Filtering

Before starting WGCNA, the input FPKM expression file must be filtered. We used the varFilter function of the R language genefilter package to remove genes with low expression and genes with constant expression in all samples to improve the accuracy of network construction. The list of filtered genes is as follows:

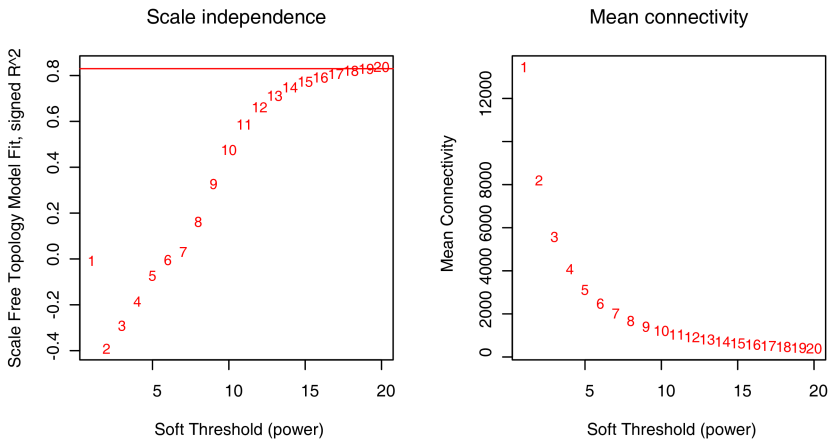
Table 25 Filtered FPKM File

geneID	A1	A2	A3
ENSRNA049466596	0.0000	0.7128	0.0000
ENSRNA049466739	0.6555	0.0000	0.0000
ENSRNA049467888	9.8329	18.0867	10.5535
ENSRNA050013657	1.3111	0.0000	0.0000
ENSRNA050013665	0.7492	0.0000	0.7538
Os01g0100100	8.6677	11.1077	11.1540
Os01g0100400	3.4436	4.4192	4.6736
Os01g0100500	15.8625	18.8216	19.6964
Os01g0100600	7.0342	6.8605	7.1963
Os01g0100700	144.4706	146.4781	135.3085

2.12.2 Soft Threshold Selection

WGCNA first calculates the correlation coefficient (Pearson Correlation Coefficient) between any two genes. In order to measure whether two genes have similar expression patterns, it is generally required to set a threshold value for screening, with those above the threshold considered similar. However, if the threshold value is set to 0.8, it is difficult to demonstrate a significant difference between 0.8 and 0.79. Therefore, a weighted value of correlation coefficient is applied when performing WGCNA, i.e., the Nth power is taken for the gene correlation coefficient. This approach reinforces the strong correlation and attenuates the weak or negative correlation, making the connections between genes in the network obey scale-free network distribution, which is more biologically significant. All horizontal axes in the graphs below represent the weighting parameter β, which is the soft threshold. The vertical axis in the left panel represents the square of the correlation coefficient in the corresponding network. Sometimes it appears negative because it is multiplied by the negative direction of the slope column value, so it is adequate to focus only on the positive values. The higher the square of the correlation coefficient, the more the network approximates a scale-free network. We have

set a threshold value of 0.85 for the square of the correlation coefficient. The vertical axis of the right panel represents the mean value of all gene adjacency functions in the corresponding gene module. The optimal β value is the soft threshold used for the subsequent analysis.

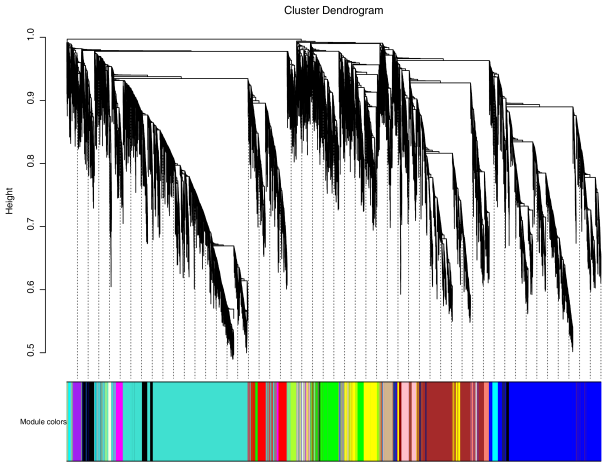


Schematic Diagram of Soft Threshold Selection

All horizontal axes in the graphs represent the weighting parameter β, which is the soft threshold. The vertical axis in the left panel represents the square of the correlation coefficient in the corresponding network. Sometimes it appears negative because it is multiplied by the negative direction of the slope column value, so it is adequate to focus only on the positive values. The higher the square of the correlation coefficient, the more the network approximates a scale-free network. The vertical axis of the right panel represents the mean value of all gene adjacency functions in the corresponding gene module.s

2.12.3 Module Hierarchical Clustering

WGCNA constructs a dendrogram based on the correlation of expression between genes and divides the genes into different modules. The threshold for merging modules together was set at a value of 0.25. The minimum number of genes allowed in a module was set to 50. Each color in the diagram indicates that the genes corresponding to this color belong to the same module in the clustering tree. If some genes always have similar expression changes in a physiological process or in different tissues, these genes may be functionally related and can be defined as a module. For the upper half of the dendogram, the vertical distance represents the distance between two nodes (genes) and the horizontal distance is meaningless.

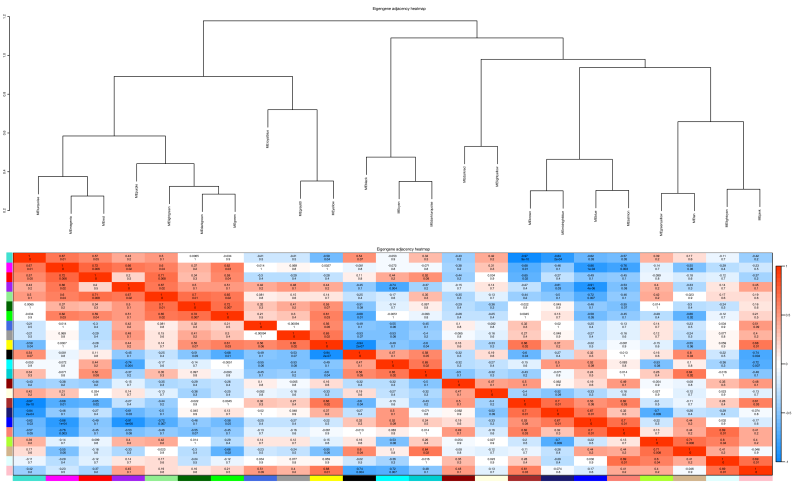


Module Hierarchical Cluster Dendrogram

The figure is divided into two parts: the upper part shows the clustering dendrogram of genes, and the lower part shows the module clustering results, with each color representing a module.

2.12.4 Inter-Module Correlation Heatmap

The inter-module correlation heatmap can be divided into two parts, with the upper part clustering the modules according to their characteristic values called eigengenes. The vertical coordinates represent the degree of dissimilarity of the nodes. Each row and column in the lower half of the graph represents a module. The darker the color of the square (the redder), the stronger the correlation; the lighter the color of the square, the weaker the correlation.

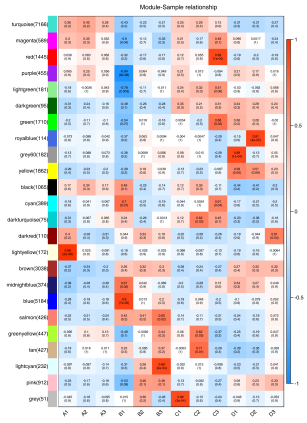


Inter-Module Correlation Heatmap

The inter-module correlation heatmap can be divided into two parts, with the upper part clustering the modules according to their eigengenes. The vertical coordinates represent the degree of dissimilarity of the nodes. Each row and column in the lower half of the graph represents a module. The darker the color of the square (the redder), the stronger the correlation; the lighter the color of the square, the weaker the correlation.

2.12.5 Sample-Module Correlation Heatmap

In general, if one module shows significantly higher correlation with the sample than the other modules, it means that this one module probably correlates most strongly with that sample, as shown in the figure below:

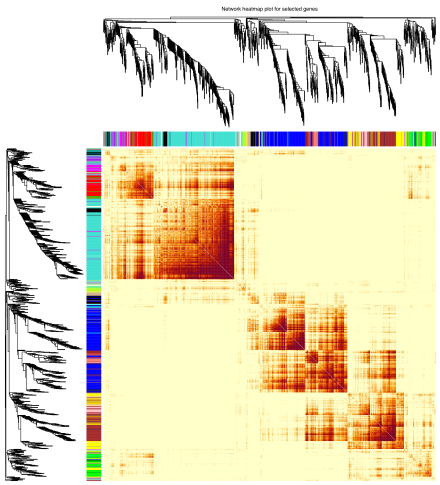


Sample-Module Correlation Heatmap

Sample-Module Correlation Heatmap: The horizontal coordinate represents the sample, while the vertical coordinate represents the module. The value in parentheses is the number of module genes. The color shades in the figure indicate the correlation level, with red being a positive correlation and blue being a negative correlation.

2.12.6 Module Gene Clustering Heatmap

Each clad in the dendrogram represents a gene, and the darker the color of each node (white → yellow → red) represents the stronger correlation between the two genes the two genes in the corresponding row and column. The results are shown in the figure below.

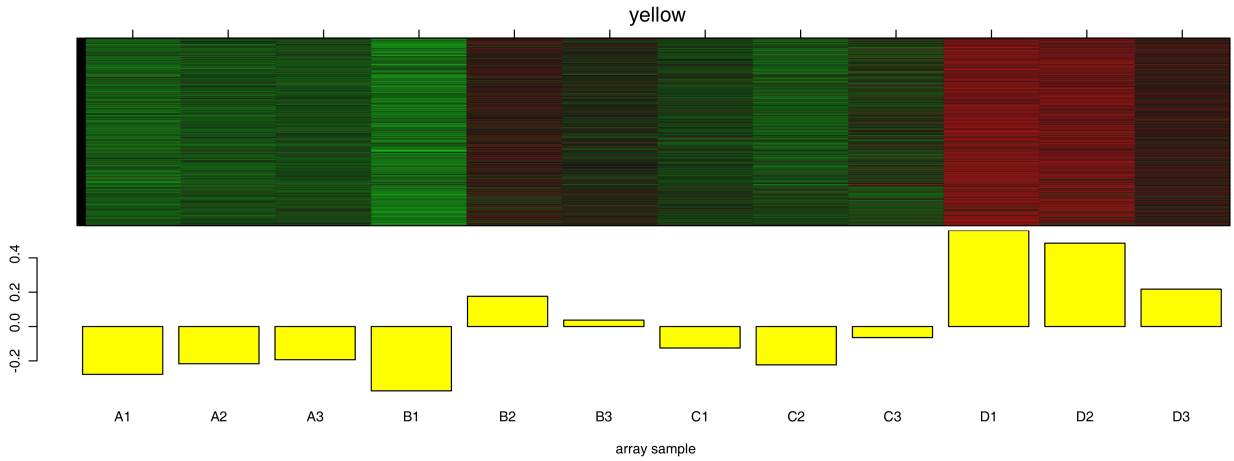


Module Gene Clustering Heatmap

Module gene clustering heatmap, each tree-like plot representing a module and each clad representing a gene, and the darker the color of each node (white → yellow → red) represents the stronger to communicate between the two genes the two genes in the corresponding row and column.

2.12.7 Module Gene Expression Patterns

WGCNA can generate a module gene expression pattern map, with the upper half being the clustered heatmap of genes within that module (red for high expression and green for low expression) and the lower half being the expression pattern of module eigengenes in different samples. This graph shows the expression trends of the module genes in different samples (only the top 10 modules are shown if the number of modules is greater than 10, and all modules are shown if the number of modules is less than 10):



Module Gene Expression Pattern Map

The upper half is a clustering heatmap of genes within the module, with high expression in red and low expression in green. The lower half shows the expression patterns of the module eigengenes in different samples.

2.12.8 Gene List by Module

Connectivity values, expression information and 7 database annotations were added to the gene list of each module obtained from WGCNA. Connectivity values indicate the strength of correlation or association between genes (usually only calculated within a module), often referred to as connectivity or degree, or expressed as k value. In general, the genes with the highest connectivity (k-value) in a module are regarded as hub genes. Only the top 10 modules are displayed if the number of modules is greater than 10, while all the modules are displayed if the number of modules is less than 10:

Table 26 Network Node Gene List by Module

geneID	moduleColors	kTotal	kWithin	kOut	kDiff
Os01g0102900	yellow	1716.3715	360.61871	1355.7528	-995.1341
Os01g0104200	yellow	1350.4186	303.97122	1046.4474	-742.4761
Os01g0104800	yellow	418.0534	81.18880	336.8646	-255.6758
Os01g0104900	yellow	644.1623	133.84908	510.3132	-376.4641
Os01g0106750	yellow	1137.5592	338.43081	799.1284	-460.6976
Os01g0106800	yellow	823.4737	292.81325	530.6605	-237.8472
Os01g0107000	yellow	686.3223	210.38773	475.9346	-265.5469
Os01g0107400	yellow	722.1262	138.62358	583.5026	-444.8790
Os01g0111700	yellow	314.7762	47.02754	267.7487	-220.7211
Os01g0121100	yellow	1074.6552	244.89439	829.7608	-584.8664

- geneID：gene number
- moduleColors：module to which the gene belongs
- kTotal：total gene connectivity
- kWithin：gene connectivity within the module
- kOut：gene connectivity outside the module, which is calculated by kTotal minus kWithin
- kDiff：difference between kWithin and kOut

2.12.9 Relationship between Network Nodes of Each Module

The interaction relationships of genes within each module in WGCNA can be exported and subsequently imported to Cytoscape software to generate a network map. If the number of modules is greater than 10, only the first 10 modules are displayed, and if the number of modules is less than 10, all of them are displayed:

Table 27 List of Network Node Relationships for Each Module

fromNode	toNode	weight
Os01g0102900	Os01g0104200	0.1965394
Os01g0102900	Os01g0106750	0.1456428
Os01g0102900	Os01g0106800	0.1800702
Os01g0102900	Os01g0107000	0.1139693
Os01g0102900	Os01g0121100	0.2373194
Os01g0102900	Os01g0128000	0.2149932
Os01g0102900	Os01g0136800	0.1363127
Os01g0102900	Os01g0137950	0.2246248
Os01g0102900	Os01g0138000	0.1503685
Os01g0102900	Os01g0140500	0.2037365

- fromNode：network node gene 1
- toNode：network node gene 2
- weight：the edge weights of the adjacency matrix, which represents the strength of the connection between two nodes (genes)

2.13 Protein Interaction Network

We applied the interactions in STRING protein-protein interaction database (http://stringdb.org) to analyze the protein-protein interaction network constructed using differentially expressed genes. We first aligned the sequences in the target gene set to the protein sequences of the reference species (or proximate species) contained in the string database by applying diamond blastx, to obtain protein-protein interactions based on the alignment results.

Table 28 Protein Interaction Network

gene1	gene2	combined_score
Os01g0101200	Os10g0419500	920
Os01g0101200	Os10g0419400	916
Os01g0101200	Os11g0484000	945
Os01g0101700	Os05g0170950	727
Os01g0101700	Os06g0650900	861
Os01g0101700	Os11g0703900	777
Os01g0101700	Os03g0821100	777
Os01g0101700	Os01g0840100	777
Os01g0101700	Os03g0276500	776
Os01g0101700	Os05g0460000	777

- gene1：No.1 differentially expressed gene
- gene2：No.2 differentially expressed gene
- combined_score：database interaction score

3 Appendix

3.1 Article Citations and Acknowledgements

Article Citations and Acknowledgements If your research project uses the sequencing and analysis services of MetwareBio, we would appreciate it if you would cite or mention Metware Biotechnology Inc. in the Method section or Acknowledgements section when publishing your paper. For reference, statements such as

Methods:

The cDNA libraries were sequenced on the Illumina sequencing platform by Metware Biotechnology Inc. (Boston, USA).

Acknowledgements：

We are grateful to/thank USA Metware Biotechnology Inc for assisting in sequencing and/or bioinformatics analysis.

3.2 Experiments and Methods

Methods in Englsih：src/appendix/RNAseq_methodsEnglish.pdf

3.3 Format Description for Results Files

Table 29

File Type	File Description	Opening method
*fa/*fasta	Sequence file	Use commands such as less/more under Linux; use text editors such as notepad under
*fq/*fastq	reads file	Use commands such as less/more under Linux; use text editors such as notepad under
xls/txt	Form Documents	Use commands such as less/more under Linux; use text editors such as notepad under
png/pdf	Image files	Windows using image viewer, Adoble Read, etc.

3.4 Analysis Software List and Version Information

Table 30

Analysis	Software	Versions	Parameters
Data QC	fastp	0.23.2	–n_base_limit 15,–qualified_quality_phred
Data Comparison	hisat2	2.2.1	Default Parameters
Genetic quantification	featureCounts	2.0.3	Default Parameters
Alternative Splicing Analysis	rMATS	4.1.2	Default Parameters
Screening for Differentially Expressed Genes	DESeq2	1.38.3	|log2foldchang| >=1 && FDR <0.05
Screening for Differentially Expressed Genes	edgeR	3.40.2	|log2foldchang| >=1 && FDR <0.05
Venn plot	VennDiagram	1.6.20	Default Parameters
Enrichment Analysis	clusterProfiler	v4.6.0	Default Parameters
GSEA Analysis	clusterProfiler	v4.6.0	Default Parameters
WGCNA Analysis	WGCNA	1.71	CutHeight = 0.25

Reference

1. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90. doi:10.1093/bioinformatics/bty560.

2. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nature methods. 2015;12. doi:10.1038/nmeth.3317.

3. Robinson JT, ThorvaldsdÃ³ttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011;29. doi:10.1038/nbt.1754.

4. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from rna-seq reads. Nature biotechnology. 2015;33. doi:10.1038/nbt.3122.

5. Huson DH, Buchfink B. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2015. https://doi.org/10.1038/nmeth.3176.

6. Yi Z, Chen J, Sun H, Rosli HG, Pombo MA, Zhang P, et al. iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Molecular Plant. 2016;9:1667–70. https://doi.org/10.1016/j.molp.2016.09.014.

7. Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LGG, Rensing SA, Kersten B, Mueller-Roeber B. PlnTFDB: Updated content and new features of the plant transcription factor database. Nucleic Acids Res. 2010;38:D822–7. doi:10.1093/nar/gkp805.

8. Jin J, Zhang H, Kong L, Gao G, Luo J. PlantTFDB 3.0: A portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res. 2014;42:D1182–7. doi:10.1093/nar/gkt1016.

9. Hu H, Miao Y-R, Jia L-H, Yu Q-Y, Zhang Q, Guo A-Y. AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. doi:10.1093/nar/gky822.

10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology. 2014;15. doi:10.1186/s13059-014-0550-8.

11. Varet H, Brillet-Guéguen L, Coppée J-Y, Dillies M-A. SARTools: A DESeq2- and EdgeR-based r pipeline for comprehensive differential analysis of RNA-seq data. PLOS ONE. 2016;11:e0157022. doi:10.1371/journal.pone.0157022.

12. Smyth GK. EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139. https://doi.org/10.1093/bioinformatics/btp616.

13. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England). 2014;30. doi:10.1093/bioinformatics/btt656.

14. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic acids research. 2008;36. doi:10.1093/nar/gkm882.

15. Ashburner M, Ball CA, Blake JA, Botstein D, Cherry JM. Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nature Genetics. 2000;25:25–9. https://doi.org/10.1038/75556.

16. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28:33–6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102395/.

17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102. doi:10.1073/pnas.0506580102.

18. Shen S, Park JW, Lu Z-x, Lin L, Henry MD, Wu YN, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-seq data. PNAS. 111:E5593–601. doi:10.1073/pnas.1419161111.

19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. Genome research. 2010;20. doi:10.1101/gr.107524.110.

20. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164. doi:10.1093/nar/gkq603.