RNA-seq Analysis without Reference Genome

Demo Report

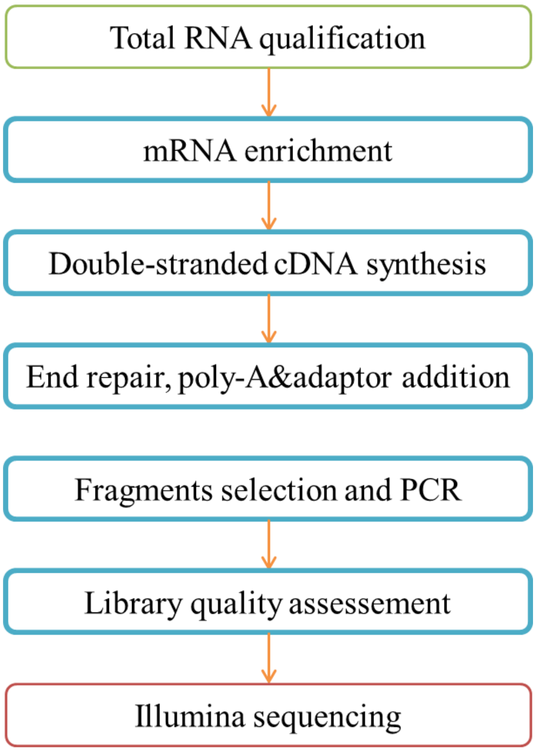
Overseas Department

Augest 2, 2017

Contents

1 Library Preparation and Sequencing

From the RNA sample to the final data, each steps, including sample test, library preparation, and sequencing, influences the quality of the data, and data quality directly impacts the analysis results. To guarantee the reliability of the data, quality control (QC) is performed at each step of the procedure. The workflow is as follows:


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

1.1 Total RNA Sample QC

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

(1) Nanodrop: preliminary quantitation

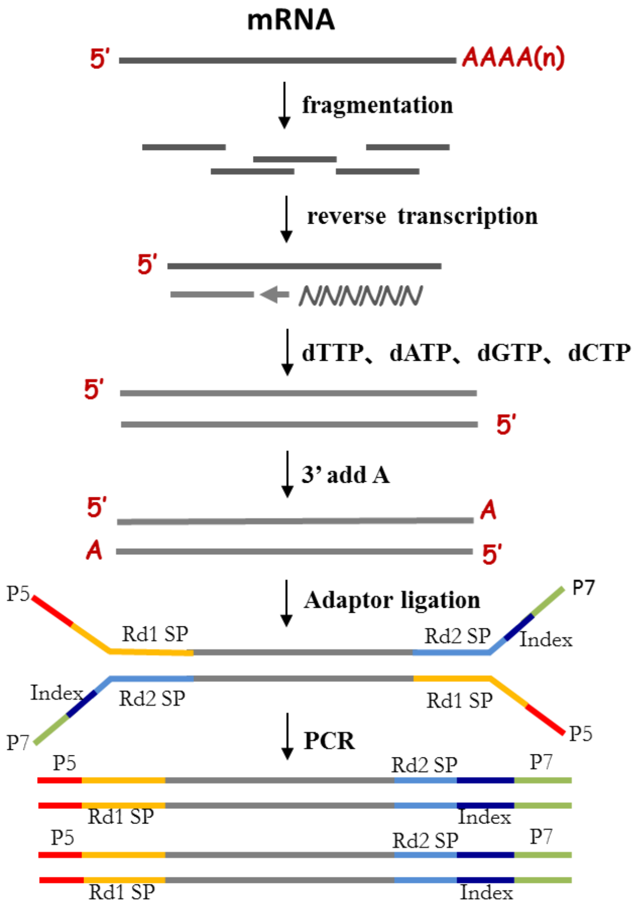
(2) Agarose Gel Electrophoresis: tests RNA degradation and potential contamination

(3) Agilent 2100: checks RNA integrity and quantitation

1.2 Library Construction and Quality Assessement

After the QC procedures, mRNA from eukaryotic organisms is enriched from total

RNA using oligo(dT) beads. For prokaryotic samples, rRNA is removed using a specialized kit that leaves the mRNA. The mRNA from either eukaryotic or prokaryotic sources then fragmented randomly in fragmentation buffer, followed by cDNA synthesis using random hexamers and reverse transcriptase. After first-strand synthesis, a custom second-strand synthesis buffer (Illumina) is added, with dNTPs, RNase H and Escherichia coli polymerase I to generate the second strand by nick-translation and AMPure XP beads is used to purify the cDNA. The final cDNA library is ready after a round of purification, terminal repair, Atailing, ligation of sequencing adapters, size selection and PCR enrichment. Library concentration was first quantified using a Qubit 2.0 fluorometer (Life Technologies), and then diluted to 1 ng/μl before checking insert size on an Agilent 2100 and quantifying to greater accuracy by quantitative PCR (Q-PCR) (library activity >2 nM). The workflow chart is as follows:

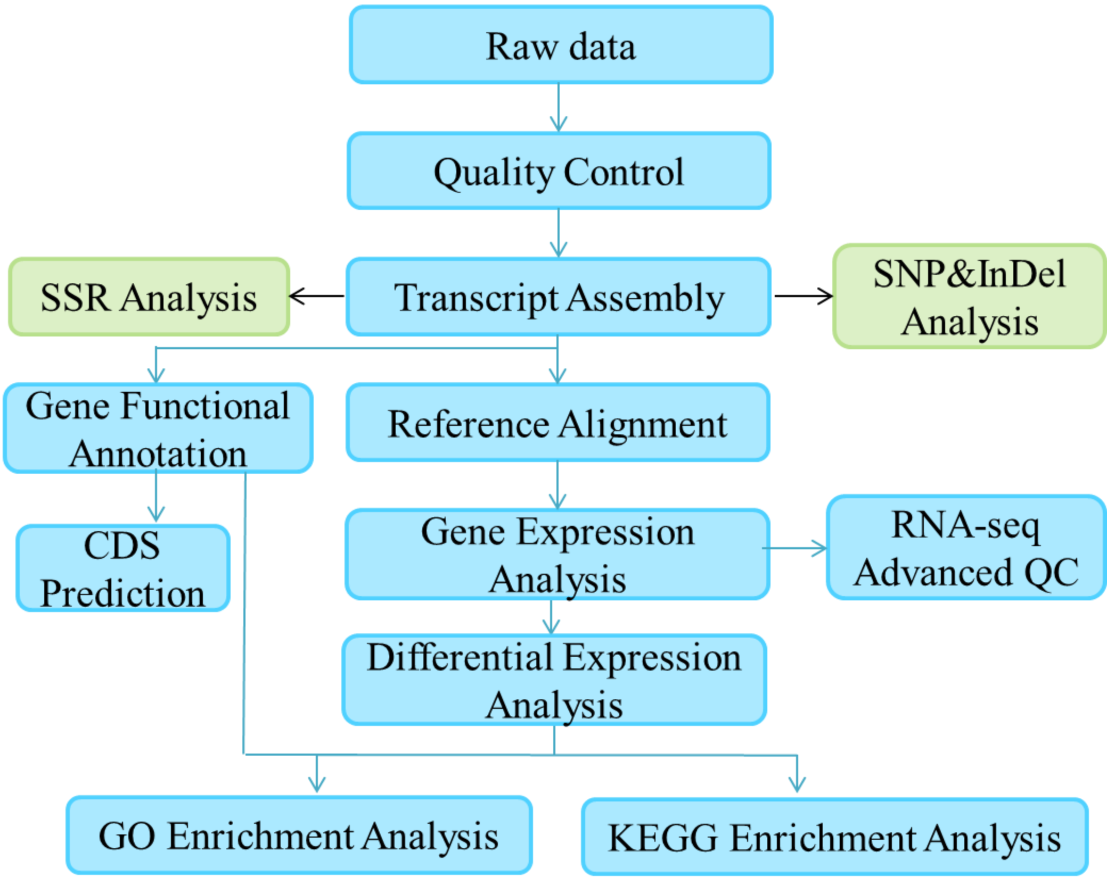


1.3 Sequencing

Libraries are fed into HiSeq machines according to activity and expected data volume.

2. Analysis Workflow

The analysis workflow for data without a reference genome is as follows:



3. Project Results

1 Raw Data

The original raw data from Illumina HiSeq^{TM} are transformed to Sequenced Reads by base calling. Raw data are recorded in a FASTQ file, which contains sequence information (reads) and corresponding sequencing quality information.

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHHFAE0ECFFD/AEHH
```

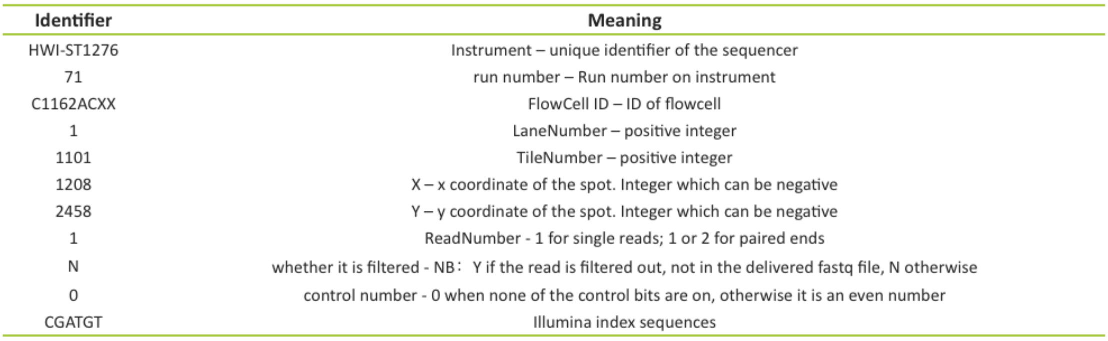
Line 1 begins with a '@' character and is followed by the Illumina Sequence Identifiers and an optional description.

Line 2 is the raw sequence read.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier and description.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of characters as there are bases in the sequence (Cock et al.).

Illumina Sequence Identifier details:



Identifier	Meaning
HWI-ST1276	Instrument – unique identifier of the sequencer
71	run number – Run number on instrument
C1162ACXX	FlowCell ID – ID of flowcell
1	LaneNumber – positive integer
1101	TileNumber – positive integer
1208	X – x coordinate of the spot. Integer which can be negative
2458	Y – y coordinate of the spot. Integer which can be negative
1	ReadNumber - 1 for single reads; 1 or 2 for paired ends
N	whether it is filtered - NB： Y if the read is filtered out, not in the delivered fastq file, N otherwise
0	control number - 0 when none of the control bits are on, otherwise it is an even number
CGATGT	Illumina index sequences

2 Data Quality Control

2.1 Error Rate

The error rate for each base can be transformed by the Phred score as in equation 1 (equation 1: Qphred = -10log10(e)). Base Quality and Phred score relationship with the Illumina CASAVA v1.8 software:

Phred score	Base Calling error rate	Base Calling correct rate	Q-sorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

Sequencing error rate and base quality depend on the sequencing machine, reagent availability, and the samples. Error rate increases as the sequencing reads are extended and sequencing reagents become more and more scarce. Additionally, the first six bases have a relatively high error rate due to the random hexamers used in priming cDNA synthesis (Jiang et al.).

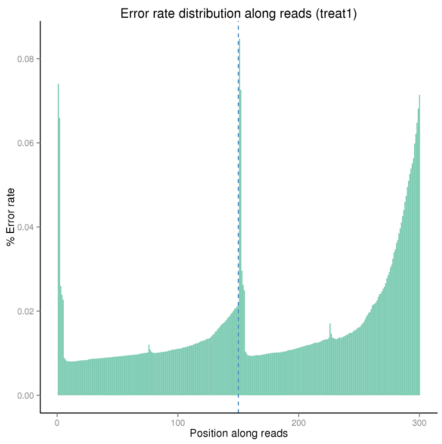


Figure 1 Error Rate Distribution

The x-axis shows the base position along each sequencing read and the y-axis shows the base error rate.

2.2 GC Content Distribution

GC content distribution is evaluated to detect potential AT/GC separation, which affects subsequent gene expression quantification. Theoretically, G should equal C, and A should equal T throughout the whole sequencing process for non-stranded libraries, whereas AT/GC separation is normally observed in stranded libraries. For DGE (Digital Gene Expression) libraries, a large variation of sequencing error in the first 6-7 bases is allowed due to the use of random primers in library construction.

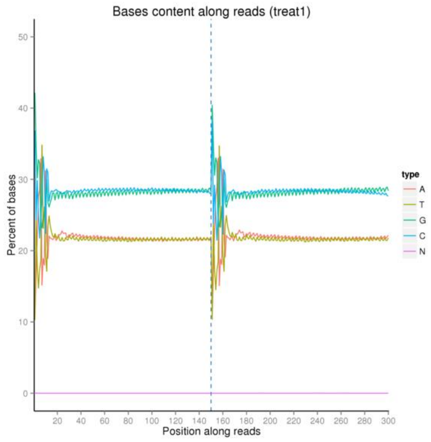


Figure 2 GC content distribution

The x-axis shows each base position within a read, and the y-axis shows the percentage of each base, with each base represented by a different color.

2.3 Data Filtering

Raw reads are filtered to remove reads containing adapters or reads of low quality, so that downstream analyses are based on clean reads. The filtering process is as follows:

(1) Discard reads with adaptor contamination.

(2) Discard reads when uncertain nucleotides constitute more than 10 percent of either read (N > 10%).

(3) Discard reads when low quality nucleotides (base quality less than 20) constitute more than 50 percent of the read.

RNA-seq Adapter sequences (Oligonucleotide sequences of adapters from TruSeq^{TM} RNA and DNA Sample Prep Kits):

NEBNext® Ultra^{TM} RNA Library Prep Kit

RNA 5' Adapter (RA5):

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

RNA 3’ Adapter (RA3):

5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(6-nucleotide index)ATCTCGTATGCCGTCTTCTGCTTG-3'

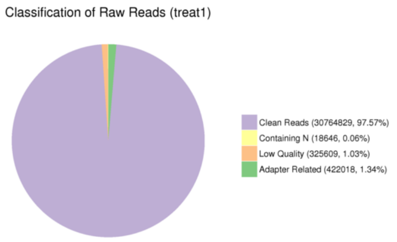


Figure 3 Raw Data

Results are shown as percentage of total raw reads.

(1) Adapter related, reads that had adapter contamination.

(2) Containing N, reads in which uncertain nucleotides constituted more than 10 percent of the read.

(3) Low quality, reads in which low quality nucleotides constituted more than 50 percent of the read.

(4) Clean reads, reads that passed quality control

2.4 Data Quality Control Summary

Table 1 Data Production

Sample	Raw Reads	Clean reads	Clean bases	Error(%)	Q20(%)	Q30(%)	GC(%)
CK1	44126070	43155820	6.47G	0.02	96.45	91.46	56.79
CK2	61749942	60582254	9.09G	0.02	96.51	91.57	56.40
CK3	59696672	58348958	8.75G	0.02	96.51	91.57	56.66
treat1	63062204	61529658	9.23G	0.02	96.45	91.46	56.61
treat2	53673536	52433140	7.86G	0.02	96.64	91.88	56.30
treat3	62389420	60731546	9.11G	0.02	95.38	88.54	57.45

(1) Sample name: the names of samples

(2) Raw Reads: the original sequencing reads counts

(3) Clean Reads: number of reads after filtering

(4) Clean Bases: clean reads number multiply read length, saved in G unit

(5) Error Rate: average sequencing error rate, which is calculated by Qphred=-10log10(e)

(6) Q20: percentages of bases whose correct base recognition rates are greater than 99% in total bases

(7) Q30: percentages of bases whose correct base recognition rates are greater than 99.9% in total bases

(8) GC content: percentages of G and C in total bases

3 Transcriptome Reconstruction

3.1 Transcriptome Reconstruction

For samples in the absence of a reference genome, clean reads need to be assembled to get a reference sequence for the following analysis. Trinity (Grabherr et al., 2011) is the software chosen to complete the transcriptome reconstruction process.

Trinity is developed by Broad Institute and Hebrew University of Jerusalem. It is a professional transcriptome assembler software (comprising modules entitled Inchworm, Chrysalis and Butterfly). The workflow of Trinity is as follows:

Inchworm: Constructs a k-mer dictionary from all sequenced reads (in practice, k = 25), selects the most frequent seeding k-mer in the dictionary and extends the seed in each direction to form a contig assembly.

Chrysalis: Chrysalis clusters minimally overlapping Inchworm contigs into sets of connected components, and constructs complete de Bruijn graphs for each component. Each component defines a collection of Inchworm contigs that are likely to be derived from alternative splice forms or closely related paralogs.

Butterfly: Butterfly reconstructs plausible, full-length, linear transcripts by reconciling the individual de Bruijn graphs generated by Chrysalis with the original reads and paired ends. It reconstructs distinct transcripts for splice isoforms and paralogous genes, and resolves ambiguities stemming from errors or from sequences >k bases long that are shared between transcripts. The final assembled result file: TRINITY.fasta.

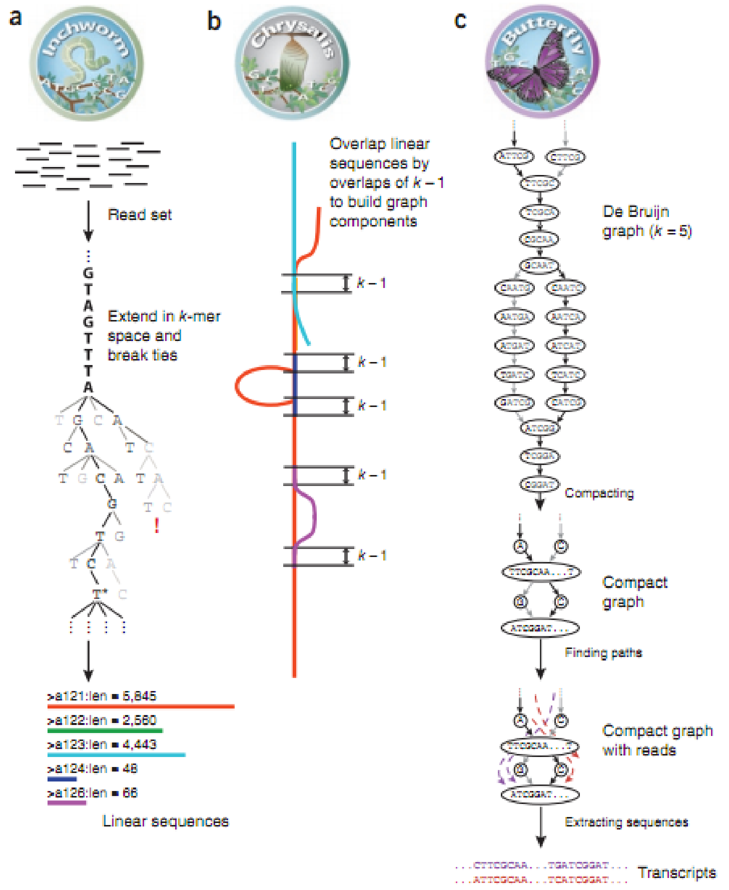


Figure 1 Overview of Trinity. (a) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a *k*-mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each *k*-mer present only once in the contigs. (b) Chrysalis pools contigs (colored lines) if they share at least one *k* – 1-mer and if reads span the junction between contigs, and then it builds individual de Bruijn graphs from each pool. (c) Butterfly takes each de Bruijn graph from Chrysalis (top), and trims spurious edges and compacts linear paths (middle). It then reconciles the graph with reads (dashed colored arrows, bottom) and pairs (not shown), and outputs one linear sequence for each splice form and/or paralogous transcript represented in the graph (bottom, colored sequences).

The sequencing data of assembled transcriptome is recorded under FASTA format as follows:

```
>c13_g1_i1 len=263 path=[369:0-108 477:109-148 65:149-262]
TGAAGAGGGAGGAGGCGAATTGGGTTTGGCGTGGCTGCTGTTAAGGGGCTGCAAGAGG
TG
GAAAGGAGGACAGAGAAGATGGAAAGATGGAGACAAGGACTGATCTGGGTGGTAGCAA
CA
GTACCTGGAAGTGGGTGTTTGGAGAAAGGGCGAAAGATGTGGTCTCTGGGAATGGCGA
TG
GAATGGGCAGCAGCAGCAGCAGGAGTCCAGGACAGGTAGTAGCAGTGGCGGAAATTAT
AC
CTGGGATAAGGCCCAGATCTCTG
>c14_g1_i1 len=249 path=[131:0-128 260:129-248]
```

```
ATTTAGTTCTCCTATGATCAGATTTTTTAGCTCATCTTTTCTTAAAGATTTGGTCTGTTG
AGAGGAAACATGAGAAATCAGGGAAACCCCCAAAATGACATCTAGTAATTGGAACTATTA
TTCAGACCCTGATTGAGGGTATTGGCAAACAATTCCTATGACAAAAGTGCAATAGCTATT
TCATTATGAGTTATGAGGAGATAGTGTTGTCAACTTTCTATTCCCCATGGCATTTAATCA
TTTTATGAG
```

Line 1 starts with '>' character and followed by the id number of the transcript;'len=' shows the length of the transcript, which is the base number of the transcript; 'path' includes the pathway information from de Bruijn Graph Subcomponet. From line 2 to the end encodes the sequence information of the transcript. More detailed explanation could be found from Trinity's website http://trinityrnaseq.github.io.

3.2 Hierarchical Clustering

Corset (Davidson et al.,2014) works by clustering contigs based on shared reads, and separates contigs when different expression patterns between samples are observed. Corset also uses the read information to filter out contigs with a low number of mapped read (less than 10 reads by default).

The sequencing data of hierarchical clustering transcriptome is recorded on FASTA format as follows:

```
>Cluster-19196.1_c113952_g2_i2
TCATTATGAGTTATGAGGAGATAGTGTTGTCAACTTTCTATTCCCCATGGCATTTAATCA
ATTTAGTTCTCCTATGATCAGATTTTTTAGCTCATCTTTTCTTAAAGATTTGGTCTGTTG
CAAGAAGAAGACTGGTGAAGGAGGCCACCAAGACACCTACGAGCACTCTGACGGAGTT
AA
AAGCATCAGTGGCTCAGATGGGAGAGACTGTACATACAACAACTGTTGCCCGGGTGCTT
C
TCCAGTCGAAGCTGTATAGGAGGGTGAAGGCAAAGAGAAAGCCACTGTTGAAAAAGCTC
A
TATGAAATCTCGCCTGCATTTCGCCCAAAGGCTGCGCTAGACTCCAAGGTCAATTGGAA
G
AAGGTTCTTTGGTCTGATGAGACTAAAATTTATTTATTTG

>Cluster-22704.0_c60010_g1_i1
CACACACACACACACACACGAGGCTAAGGATTGGTGAGAGGCTGAGTCACAGGTGCTG
CCC
TCTAGTGGTGCATGCTGCTCTTCACCCTGTTTGCTCACGCTGGGCTCAGGTCTGGGTTA
TC
CGCTGATGACATGGGATGGTGTTCCACACAGCAGACTGACCACAGTGACCCCCCAACA
```

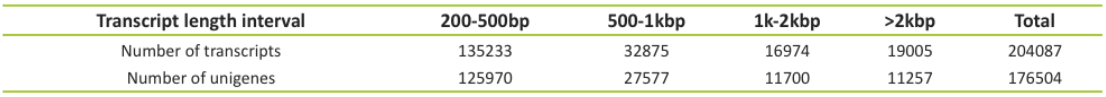
CAG
CGGATGGACCACAAGTGACAGACACTCCAACAGAGCC

Header line starts with '>' character and followed by the sequence id, "Clusters-X.Y_c_g_i". "X" means the super cluster ID, Each super cluster contains all contigs that share one or more reads with another contig in the same super cluster. "Y" means the cluster number of the super cluster ID. "c_g_i" is the transcript id assembled by Trinity. More detailed explanation could be found from Corset's website https://github.com/Oshlack/Corset/wiki.

3.3 Transcript Length Distribution

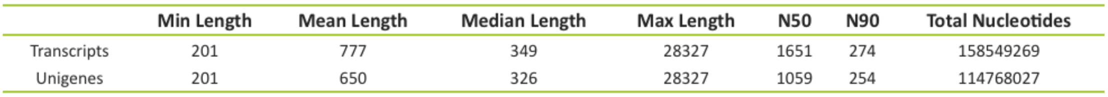
Clean reads are de novo assembled by Trinity to get assembly transcriptome. Then Corset will perform Hierarchical Clustering to remove redundence. Afterwards the longest transcripts of each cluster will be selected as unigenes. Length distribution information of transcripts and unigenes are listed in the following tables:

Table 2 Overview of the number of transcripts and unigenes in different length intervals



Transcript length interval	200-500bp	500-1kbp	1k-2kbp	>2kbp	Total
Number of transcripts	135233	32875	16974	19005	204087
Number of unigenes	125970	27577	11700	11257	176504

Table 3 Overview of the length distribution of transcripts and unigenes



	Min Length	Mean Length	Median Length	Max Length	N50	N90	Total Nucleotides
Transcripts	201	777	349	28327	1651	274	158549269
Unigenes	201	650	326	28327	1059	254	114768027

The N50 length is defined as the length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs, and for which the collection of all contigs of that length or shorter contains at least half of the total of the lengths of the contigs. (When more than one value of length meets both these criteria then the N50 is the average of the longest and shortest lengths that meet these criteria.) The N90 statistic is smaller than or equal to the N50 statistic; it is the length for which the collection of all contigs of that length or longer contains at least 90% of the total of the lengths of the contigs, and for which the collection of all contigs of that length or shorter contains at least 10% of the total of the lengths of the contigs.

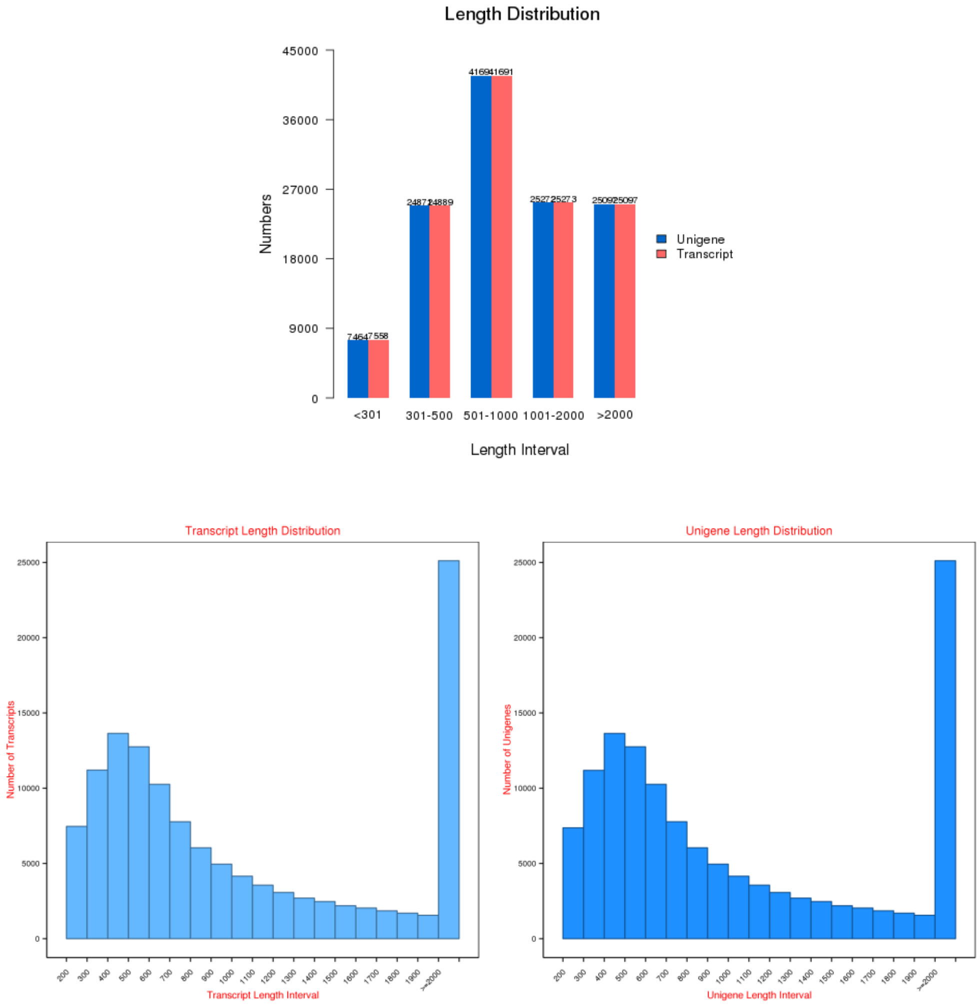


Figure 4 Length distribution of transcripts and unigene

X-axis indicates length interval of transcript and unigene; Y-axis indicates the frequency of transcript and unigene in each length interval.

3.4 Gene Functional Annotation

3.4.1 Gene Functional Annotation

To achieve comprehensive gene functional annotation, seven databases are applied by Novogene. The function and characteristics of the seven databases are as follows:

Nr (NCBI non-redundant protein sequences): it is the formal protein sequence databases of NCBI, which includes protein sequence information from GenBank, PDB (Protein Data Bank), Swiss-Prot, PIR (Protein Information Resource), PRF (Protein Research Foundation) etc.

Nt (NCBI nucleotide sequences): it is the formal nucleotide sequence database of NCBI. It includes nucleotide sequence from GenBank, EMBL and DDBJ (but does not contain EST, STS, GSS, WGS, TSA, PAT, HTG).

Pfam (Protein family): it is the most comprehensive collection of protein domains and families, represented as multiple sequence alignments and as profile hidden Markov models. Many proteins are composed of structural domains while the protein sequence of a specific structural domain possess a certain degree of conservative property. In Pfam database, proteins are classified into different protein families according to their structural domains, and the HMM statistical model of each family's amino acid sequence is established by alignment of the protein sequences. According to the reliability of annotations, PFAM families are classified into two categories, Pfam-A and Pfam-B. Pfam-A family consists of a curated seed alignment containing a small set of representative members of the family, profile hidden Markov models (profile HMMs) built from the seed alignment and an automatically generated full alignment which contains all detectable protein sequences belonging to the family, as defined by profile HMM searches of primary sequence databases. Pfam-B entries are automatically generated from the ProDom database, and are represented by a single alignment. Through HMMER3 program, the established HMM model can be searched to annotate unigenes. More details: http://pfam.sanger.ac.uk/.

KOG/COG: Both COG (Cluster of Orthologous Groups of proteins) and KOG (euKaryotic Orthologous Groups) are based on NCBI's gene orthologous relationships. COG is specific to prokaryotes while KOG is specific to eukaryotes. According to their evolutionary relationships, COG/KOG divides the homologous genes from different species into different ortholog clusters. The COG collection currently consists of 138,458 proteins, which form 4873 COGs and the current KOG set consists of 4852 clusters of orthologs, which include 59,838 proteins. As genes from the same ortholog own the same function, the functional annotation can be shared to the other members from the same COG/KOG clusters. More details could be found from the following website: http://www.ncbi.nlm.nih.gov/COG/.

Swiss-Prot: A manually annotated and reviewed protein sequence database. It's a high quality protein sequence database, which brings together experimental results, computed features and scientific conclusions. More details could be found from the following website: http://www.ebi.ac.uk/uniprot/.

KEGG (Kyoto Encyclopedia of Genes and Genome): KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. It contains KEGG PATHWAY, KEGG DRUG, KEGG DISEASE, KEGG MODULE, KEGG GENES, KEGG GENOME etc.

And KO system (KEGG ORTHOLOG) combines each KEGG annotation system. KEGG has established a complete KO annotation system which can accomplish the function annotation of the genome/transcriptome of a newly sequenced species. More details could be found from the following website: http://www.genome.jp/kegg/ .

GO (Gene Ontology): GO is the established standard for the functional annotation of gene products. GO vocabulary is a controlled vocabulary used to classify the following functional attributes of gene products: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). GO term is the basic unit of GO system. Each term has a unique identifier. The relationship between the GO term of each ontology can form a Directed Acyclic Topology. More details could be found from the website: http://www.geneontology.org/.

The software and parameters used in each database are as follows:

NR, NT, SwissProt, KOG: NCBI blast 2.2.28+. For NR, NT and SwissProt databases, the evalue threshold is 1e-5 (Each unigene shows top10 alignment results), and 1e-3 for KOG. We will show the top 10 for each unigene;

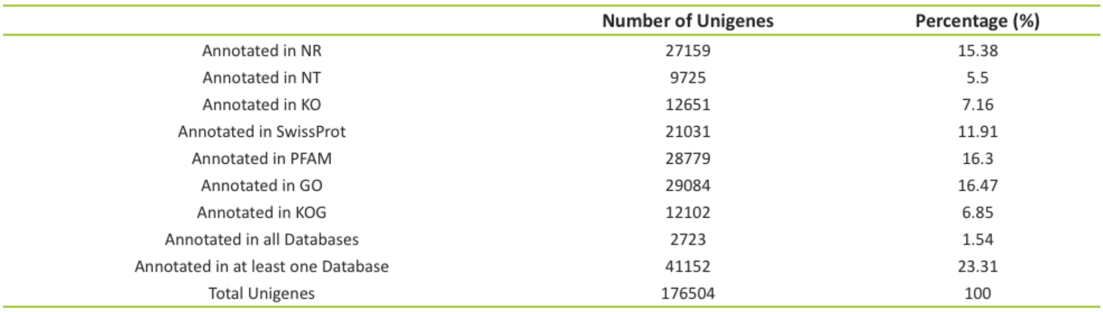
PFAM, the prediction of protein structure domain: HMMER 3.0 package, hmmscan, the evalue threshold is 0.01;

GO: based on the protein annotation results of NR and Pfam: Blast2GO v2.5 (Götz et al., 2008) and novogene script, the evalue threshold is 1e-6;

KEGG: KAAS, KEGG Automatic Annotation Server, the evalue threshold is 1e-10;

The statistics of successfully annotated genes by each database are shown in Table4 The Ratio of Successfully Annotated Genes.

Table4 The Ratio of Successfully Annotated Genes



	Number of Unigenes	Percentage (%)
Annotated in NR	27159	15.38
Annotated in NT	9725	5.5
Annotated in KO	12651	7.16
Annotated in SwissProt	21031	11.91
Annotated in PFAM	28779	16.3
Annotated in GO	29084	16.47
Annotated in KOG	12102	6.85
Annotated in all Databases	2723	1.54
Annotated in at least one Database	41152	23.31
Total Unigenes	176504	100

(1) The number of genes successfully annotated in NR and its percentage in total Unigene number.

(2) The number of genes successfully annotated in NT and its percentage in total Unigene number.

(3) The number of genes successfully annotated in KO and its percentage in total Unigene number.

(4) The number of genes successfully annotated in Swissprot and its percentage in total Unigene number.

(5) The number of genes successfully annotated in Pfam and its percentage in total Unigene number.

(6) The number of genes successfully annotated in GO and its percentage in total Unigene number.

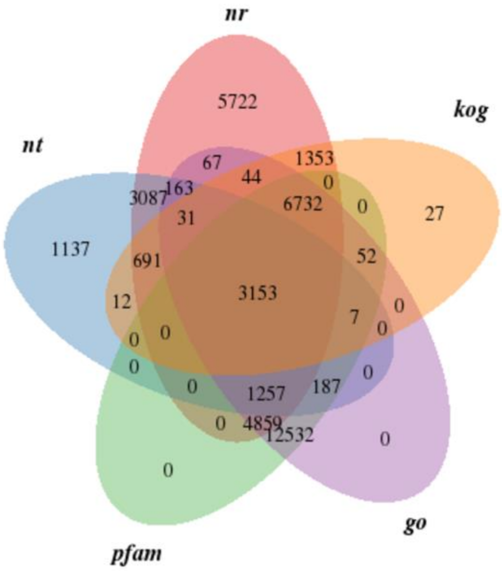
(7) The number of genes successfully annotated in KOG and its percentage in total Unigene number.

(8) The number of genes successfully annotated in all the seven databases and its percentage in total Unigene number.

(9) The number of genes successfully annotated in at least one database and its percentage in total Unigene number.

(10) Total Unigene number and total unigene percentage.

The venn diagram is mapped with 5 selected database annotation result from 7 database results:



4.2 Gene Annotation Results through Nr Database

Table 5 Part of Gene Annotation Results through Nr Database

Gene ID	Gene Length	NR GI	NR ID	NR Score	NR Evalue
Cluster-82483.0	414	658834540	XP_008420017.1	242	2.69382e-21
Cluster-74250.27752	729	657545700	XP_008278143.1	570	9.70023e-69
Cluster-19196.1	280	12666718	CAC28060.1	156	8.03317e-13
Cluster-74250.17473	3446	554862959	XP_005941562.1	1439	7.5488e-167
Cluster-74250.17477	583	768949162	XP_011613755.1	464	6.67989e-50

(1) The gene ID from Corset result.

(2) The longest length generated by the gene transcript.

(3) The GenBank ID of the annotated nucleotide.

(4) The NR ID of the annotated nucleotide.

(5) Alignment score based on a specific score matrix.

(6) Expected value calculated according to the score, query sequence's length and library size. Evalue essentially represents the false positive rate, the smaller is the better.

The species distribution, E-value distribution and similarity distribution plots are as follows:

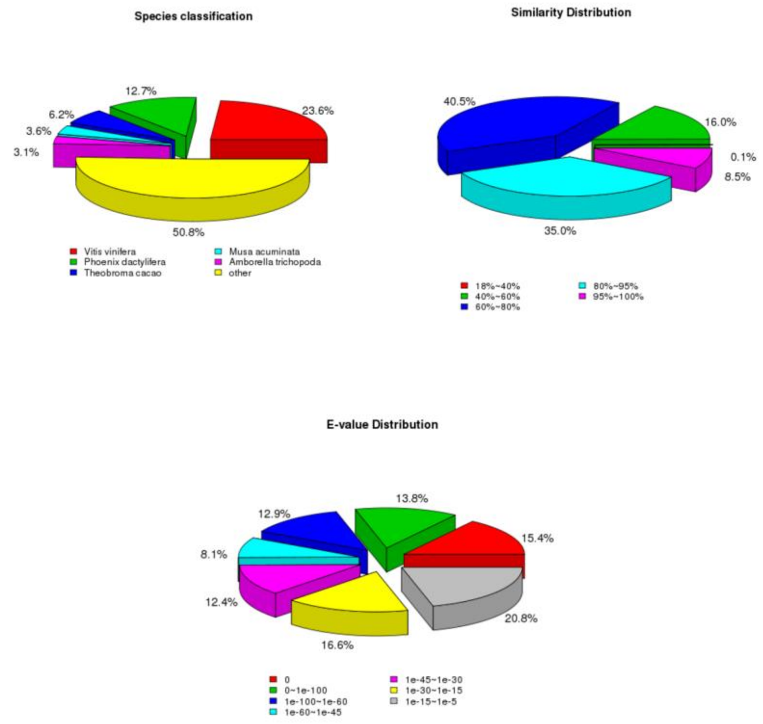


Figure 5 Nr Distribution

Figure 1 The Species Distribution. Figure 2 The E-value Distribution. Figure 3 The Similarity Distribution.

4.3 GO Classification

After GO annotation, the successfully annotated genes will be grouped into three main GO domains: Biological Process (BP), Cellular Component (CC), Molecular Function (MF).

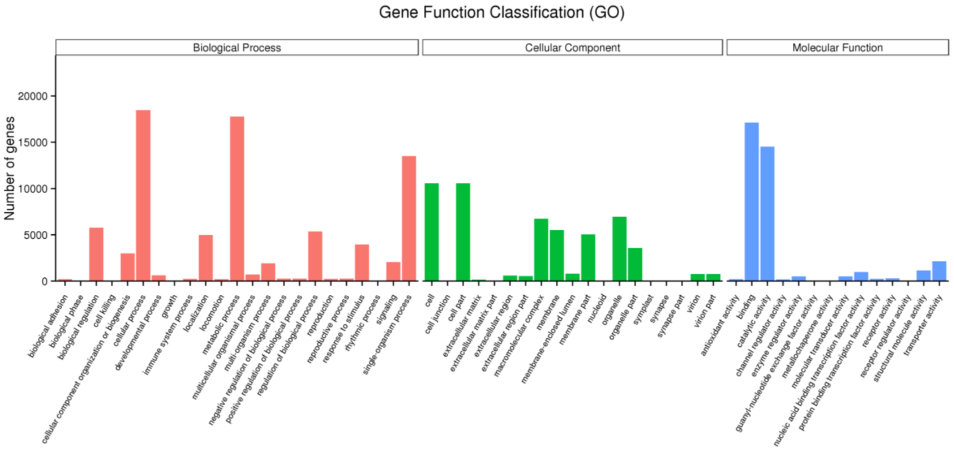


Figure 6 GO Classification

X-axis is the GO term under the three main GO domains; Y-axis is the number and percentage of the genes annotated in the term (include its sub-term).

4.4 KOG Classification

KOG is divided into 26 groups. Figure 4.3 shows the classifications of the genes successfully annotated in KOG.

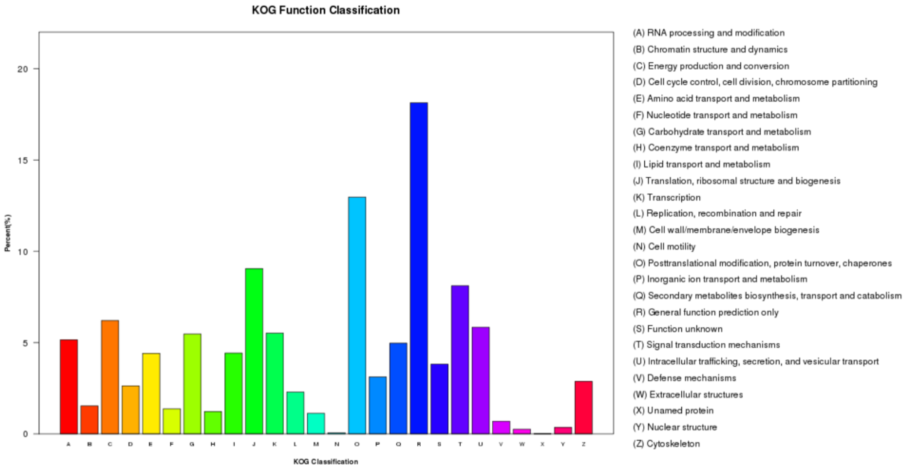


Figure 7 KOG Classification

X-axis is the names of the 26 KOG group; Y-axis is the percentage of genes annotated under this group in the total annotated genes.

4.5 KEGG Classification

The genes successfully annotated in KEGG can be classified according to the KEGG pathway they joined in.

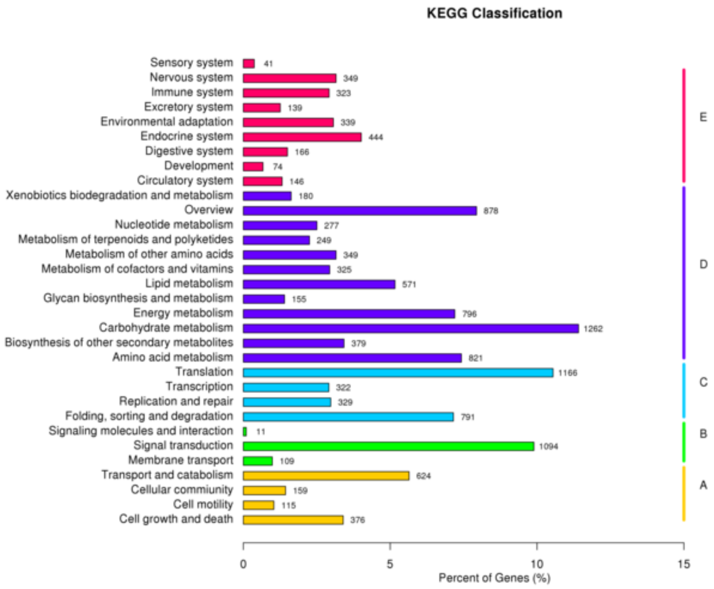


Figure8 KEGG Classification

Y-axis is the names of KEGG pathways; X-axis is the number of the genes annotated in the pathway and the ratio between the number in this pathway and the total number of annotated genes. The KEGG metabolic pathways gene involved in are divided into 5 branches: A: Cellular Processes, B: Environmental Information Processing, C: Genetic Information Processing, D: Metabolism, E: Organismal Systems.

5 CDS Prediction

5.1 CDS Prediction

CDS prediction can be divided into two steps: 1. BLAST unigene according to the priority of NR and Swissprot databases. If the information matched, CDS is extracted from unigene sequences and translated into peptide sequences based on the standard codon table (from 5' to 3'). 2. Unigenes with no hits in BLAST are analyzed with ESTScan (3.0.3) to predict their coding regions and determine their sequence direction. Part of results are shown as follows:

CDS extracted from BLAST:

```
>Cluster-82483.0;orf1    len=369    frame:-3    start:370    end:2    gi|617492671|ref|XP_007578527.1|
PREDICTED: protein PAT1 homolog 1 isoform X1 [Poecilia formosa]
GTGCTGAGAGAAAACGGCTTCTGTTTTTATTATTGTGGTGCAATGACTTCTCTATTTGCT
GTTTTTGGTCAGAACTCTCCTCTGTGTCGCGGCCCGTTTCCTCCCGGAGTTGGTCCGGT
C
CTGTCTCAGATCCAACGTGCCCAGCTGCTCAACTCTCAGGTGGCTGGTTTCCCCCACGG
T
GGGCCTCCTTTGTTACCAGGTGGTGGCTTCAGGCCGTTCTTCGGGGGCCCTCCTCCTC
CA
CACGGTCACCGAATGGGTCCGCCGCCCCCTCACGGCCCCCCCAACCACACGCCACCCA
TT
CGGCACAACACCACCCACCTCCACCCTCAGCATCGCCGCATGCTCACGCAGCGCATGC
AG
AACCGAGGA
>Cluster-74250.27752;orf1   len=477   frame:1   start:100   end:576   gi|657545700|ref|XP_008278143.1|
PREDICTED: protein N-terminal asparagine amidohydrolase [Stegastes partitus]
CCAGGTGATTTGAGTTGCACCTTGCTTAAGGAAATGCCTTTGTTTATTCAAAATAGAGGA
CTTGGCCGCATAAGCTCGACGGGGGAACTATTCGAAAAATATCCACATTTACAGGAAAAT
GCAAGAACATTTCGCTCCAAGCCGCTTGTGGATGTCGACCGAAAGTGCCTCTTGTATGT
C
CAACAGAGAGAGTTTGCTGCAACAACACCAGCAGACAACAGTGTTTCAGTAATTGGATCT
GATGATGCCACCACCTGCCATTTGGTTGTGCTGCGACACACTGGAAGTGGAGCTGTTTG
C
CTTGCTCACTGTGATGGTTCCAGTACCAGGTCTGAAGTCCCGCTCCTTGTGAGAGCTGT
C
ATGTCACTGAGTAACGTCAGTAAGGAGGGCAGGTATGAAACAGGCAAACATAAAAGTGC
T
CCTCTTATTCAGATGTTCTTTATTTCCTTTCCTTTGCTGTCTGTTACGCAGGCTTGA
```

header: >(sequence ID, the only identifier for the gene): (orf1, pridicted ORF id, one gene may have two or more pridicted ORFs) len: (the base length of this ORF) frame: (the reading frame of this ORF in the original gene, '-' represents the minus strand) start: (the start position of this ORF in the original gene) end: (the end position of this ORF in the original gene) (descriptions of the gene blasted protein)

CDS predicted by ESTScan:

```
>Cluster-23194.0; len=120 start:820 end:700; minus strand
ATGGACCCGACGCAGTCTTCGGGTCGGAGCAGCTGTAAACCTCCTGCTGTGCTCTCGTC
G
AAGGAGTCGGTCAGCTTGACTCTTGGACAGAGACGAGGTTTAGTCACTGCTCAGGCGTA
```

A
>Cluster-76416.0; len=194 start:1 end194
ATGATCAATGGAAGGGGGATGAGGATGCAAATGGATGGTTTGATTATGCCTCTGCAAAA
A
AATGATGGTGATTTCAATTCAGCAATGGCTGTAATTTTCCAGGCTAAAATGGGACTTGCC
AATTCAAGGGAGAATGGATTTAAAGGGAAGCATGCATGGAAAGTGAGCCCCATAGGTGT
T
TTCATCATTTAA

header: >(sequence ID, the only identifier for the gene); (a represents another ORF of this gene) len: (the base length of this ORF) start: (the start position of this ORF in the original gene) end: (the end position of this ORF in the original gene)(minus strand represents that this ORF is the minus strand in the original gene, Otherwise, it's the positive strand)

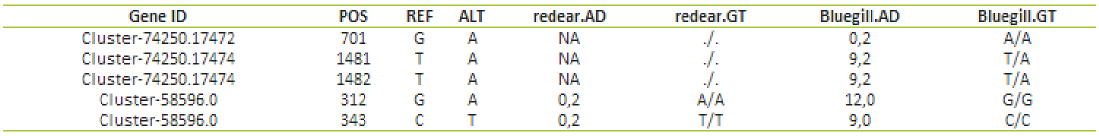
6 SNP and INDEL

6.1 SNP and INDEL

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

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

Table 6 SNP Results



Gene ID	POS	REF	ALT	redear.AD	redear.GT	Bluegill.AD	Bluegill.GT
Cluster-74250.17472	701	G	A	NA	./.	0,2	A/A
Cluster-74250.17474	1481	T	A	NA	./.	9,2	T/A
Cluster-74250.17474	1482	T	A	NA	./.	9,2	T/A
Cluster-58596.0	312	G	A	0,2	A/A	12,0	G/G
Cluster-58596.0	343	C	T	0,2	T/T	9,0	C/C

(1) Gene ID of SNP.

(2) Position of SNPs.

(3) Reference genotype.

(4) SNP genotype (Alternative genotype).

(5) Lettered columns show the genotype of each sample in the locus. The number before "," represents the number of reads supporting REF. The number after "," represents the number of reads supporting ALT. Number “0” means that there is no read supporting the locus.

3.7. SSR Analysis

3.7.1 SSR Analysis

Simple sequence repeats (SSR) or microsatellites are the repetitive nucleotide sequences of motifs of length 1-6 base pairs. They are scattered throughout the genomes of all the known organisms ranging from viruses to eukaryotes. MISA(v1.0, default parameters; Minimum number of repeats of each unit size is: 1-10; 2-6; 3-5; 4-5; 5-5; 6-5) is used for the SSR detection of unigenes. More details can be found from the following website:http://pgrc.ipk-gatersleben.de/misa/misa.html.

Table 7 SSR analysis results

Gene ID	SSR nr.	SSR type	SSR	size	start	end	ssr_position
Cluster-23194.0	1	p2	(AC)6	12	340	351	utr5
Cluster-74250.17473	1	p1	(G)10	10	2390	2399	utr5
Cluster-37938.0	1	p2	(AC)6	12	27	38	cds
Cluster-74250.17476	1	p2	(CT)7	14	10	23	utr3
Cluster-74250.17476	2	p1	(T)10	10	666	675	cds

(1) Gene ID of SSR analysis.

(2) SSR ID of each unigene.

(3) SSR type: c, Complex repetitive type; p1, Mono-base repeat; p2, Di-bases repeat; p3, the three Tri-base repeat.

(4) The repeat sequence.

(5) The size of repeated sequence.

(6) The start position of repeated sequence.

(7) The end position of repeated sequence.

(8) The position of repeat regions.

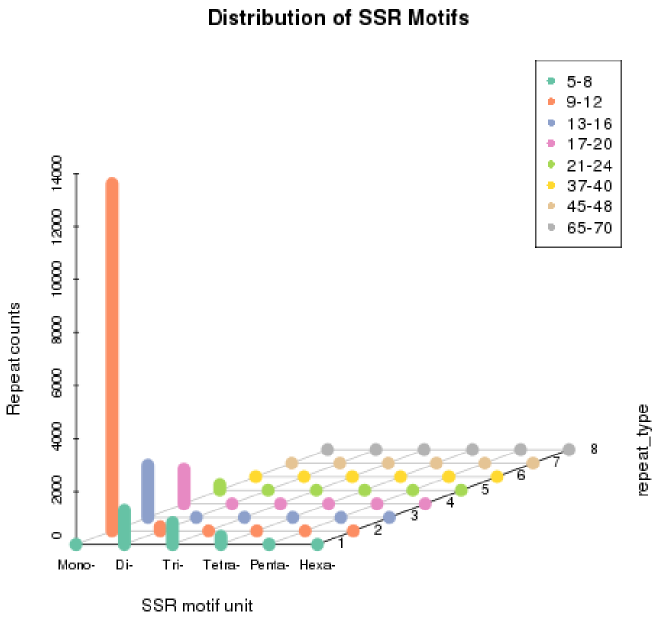


Figure 9 Distribution of SSR Motifs

The y-axis number is correspond with the number of repeat counts with different color,the z-axis is the number of SSR.

3.7.2 SSR Primer Design

Primer3 (2.3.5 version, the default parameters) is used to design SSR primer.

Table 8 Primer Design of SSR

ID	FORWARD PRIMER1 (5'-3')	Tm	size	REVERSE PRIMER1 (5'-3')	Tm	size	PRODUCT 1 size (bp)	start (bp)	end (bp)
Cluster-23194.0	AGTCGGACTGATACCTCGCT	60.107	20	ACTGTACACGGCAGGCTAAC	60.038	20	174	215	388
Cluster-74250.17473	AATCCAGAACGGGCCGTATC	59.894	20	GGGAAACTACCTTGGGGAGC	60.034	20	271	2319	2589
Cluster-37938.0	GGACATGCAACCGCCACC	61.748	18	ACATGAACATTTGGGTGCGC	60.038	20	222	0	221
Cluster-74250.17476	AGACAAACGTGCACCTGAGT	59.822	20	TTGTCTATGTGCTGGCCAGG	60.035	20	244	639	882
Cluster-74250.17475	CAGCATGAGCCGAGCTGATA	59.967	20	TGTCAAAGGTCTTCCGAGGC	59.965	20	254	1065	1318

(1) Gene id.

(2) The sequence of forword primer1.

(3) The annealing Temperature of primer1.

(4) The size of primer1.

(5) The sequence of reverse primer1.

(6) The annealing Temperature of primer2.

(7) The size of primer2.

(8) The size of the product.

(9) The start position of primer.

(10) The end position of primer.

3.8 Gene Expression Analysis

3.8.1 Reference Alignment

De novo transcriptome filtered by Corset is used as a reference(ref). RSEM(Li et al., 2011) will map reads back to transcriptome and quantify the expression level. The summary of mapping results are shown as follows:

Table 9 Overview of the Alignment Situation

Sample name	Total reads	Total mapped
CK1	55328578	45123170(81.55%)
CK2	52874640	42972508(81.27%)
CK3	51646040	42121134(81.56%)
treat1	62753296	51296906(81.74%)
treat2	53963142	44079300(81.68%)
treat3	56704144	46621518(82.22%)

(1) Sample name

(2) Clean reads number.

(3) Total number of reads that can be mapped to the reference genome.

3.8.2 Summary of Gene Expression Levels

To calculate the gene expression level, RSEM analysed the mapping results of Bowtie, and then got the read count for each gene of each sample. Furthermore, converted them into FPKM value. In RNA-seq, FPKM, short for the expected number of Fragments Per Kilobase of transcript sequence per Millions base pairssequenced, is the most commonest method of estimating gene expression levels, which takes into account the effects of both sequencing depth and gene length oncounting of fragments. The results (part of all results) are shown in Table10.

Table 10 Gene Expression Summary

Gene Id	Sample Name	Read Count	FPKM
Cluster-0.0	redear	12.00	1.51
Cluster-1.0	redear	9.00	0.77
Cluster-10.0	redear	86.42	1.74
Cluster-100.0	redear	6.00	1.26
Cluster-1000.0	redear	9.00	0.84

(1) Gene ID

(2) Sample Name

(3) The read count value of each sample.

(4) The FPKM value of each sample.

3.8.3 FPKM Density Distribution

Density distribution of FPKM can display the overall gene express levels.

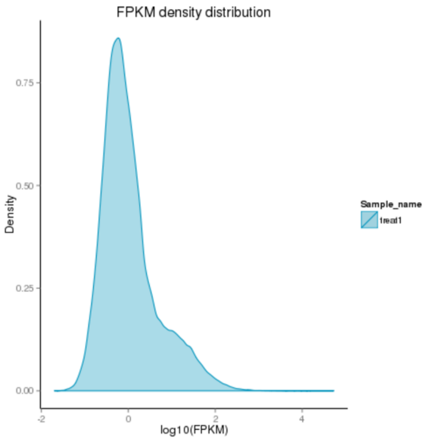


Figure 10 FPKM Density Distribution

The x-axis shows log10(FPKM) and the y-axis shows the density of log10(FPKM).

3.9 RNA-seq Advanced QC

3.9.1 Sample Correlation

Biological replicates are necessary for any biological experiment, including those involving RNA-seq technology (Hansen et al.). In RNA-seq, replicates have a two-fold purpose. First, they demonstrate whether the experiment is repeatable, and secondly, they can reveal differences in gene expression between samples. The correlation between samples is an important indicator for testing the reliability of the experiment. The closer the correlation coefficient is to 1, the greater the similarity of the samples. ENCODE suggests that the square of the Pearson correlation coefficient should be larger than 0.92, under ideal experimental conditions.

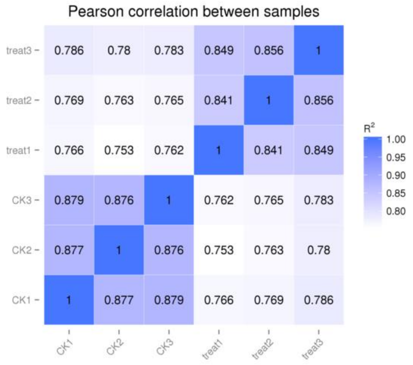


Figure 11 Sample Correlation

If the samples are more than 4 groups, then only present the scatter diagrams between biological replicates. The scatter diagrams demonstrate the correlation coefficient between samples; R2, the square of the Pearson coefficient. Heat maps of the correlation coefficient between samples are also shown.

3.10 Gene Expression Difference Analysis

3.10.1 Comparision between Gene Expression Levels

To compare gene expression levels under different conditions, an FPKM distribution diagram and box plot are used. For biological replicates, the final FPKM would be the mean value.

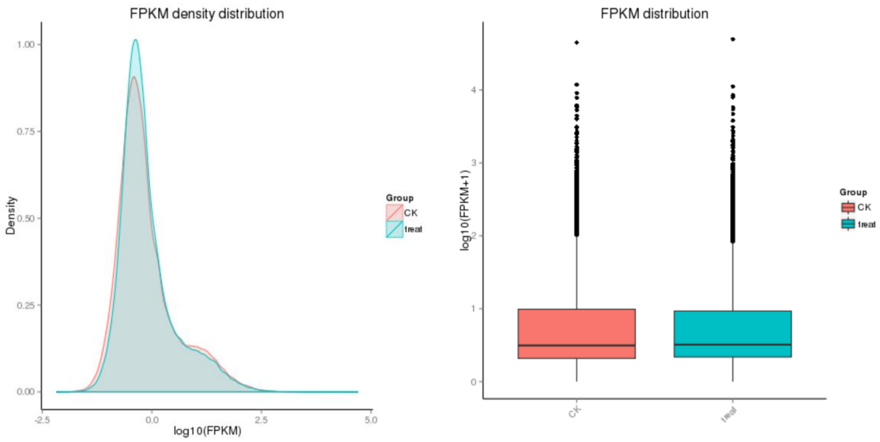


Figure 12 Different Gene Expression Levels under Different Experimental Conditions

Figure 1: The x-axis is gene's log10(FPKM) value. The y-axis is the density of log10(FPKM)value.

Fugure 2: Box plot of FPKM, the x-axis is the sample names and the y-axis is log10(FPKM+1).

3.10.2 List of Differentially Expressed Genes

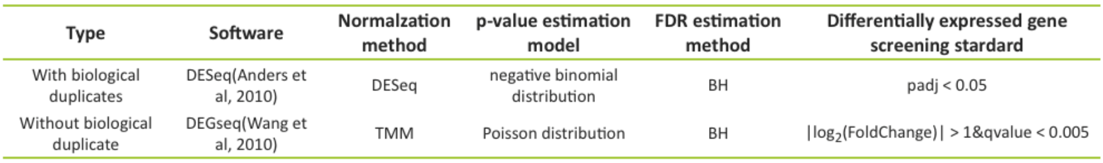
The input data for differential gene expression analysis are readcounts from gene expression level analysis. The differential gene expression analysis contains three steps:

1)Readcounts Normalization;

2)Model dependent p-value estimation;

3)FDR value estimation based on multiple hypothesis testing.

Different softwares and parameter sets are applied in different situations. The analysis methods are listed below:



Type	Software	Normalzation method	p-value estimation model	FDR estimation method	Differentially expressed gene screening stardard
With biological duplicates	DESeq(Anders et al, 2010)	DESeq	negative binomial distribution	BH	padj < 0.05
Without biological duplicate	DEGseq(Wang et al, 2010)	TMM	Poisson distribution	BH	$\|log_2(FoldChange)\| > 1$&qvalue < 0.005

The readcount value of the ith gene in the jth sample is Kij, then

Negative binomial distribution: $Kij \sim NB(\mu ij,\sigma ij2)$

Poisson distribution: $Kij \sim P(\mu ij)$

Table 11 Differentially Expressed Genes

Gene Id	redear	Bluegill	log2FoldChange	p.value	q.value. Storey.et.al..2003.
Cluster-1262.0	5.5193079374755	0	3.4645	4.6461e-05	0.00025763
Cluster-28980.3	5.059622078436	0	3.339	9.7512e-05	0.00049344
Cluster-44242.2	10.9329972433092	0	4.4506	1.3886e-08	1.4716e-07
Cluster-44973.2	4.30210566738947	0.804665442368278	2.4186	0.00044383	0.0019627
Cluster-46055.4	5.43033648088721	0.262857377840304	4.3687	6.1625e-05	0.00032961

(1) Gene ID

(2) treat: The read count values of sample1after normalization.

(3) CK: The read count values of sample2 after normalization.

(4) log2(Group1/Group2)

(5) The p-value.

(6) The p-value after normalization. The smaller the p-adjusted value is, the more significant is the difference.

3.10.3 Filtering the Differential Gene Expression

Volcano plots can be used to infer the overall distribution of differentially expressed genes.

For experiments with biological replicates, as the DESeq already eliminates the biological variation, our threshold is normally set as: padj < 0.05.

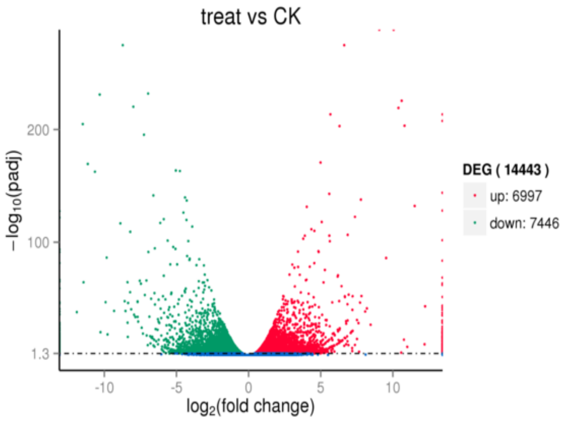


Figure 13 Volcano Plot

The x-axis shows the fold change in gene expression between different samples, and the y-axis shows the statistical significance of the differences. Statistically significant differences are represented by red dots.

3.10.4 Venn Diagram of Expression Gene

When there are only 2 samples or groups, venn diagram of expression genes will be plotted.

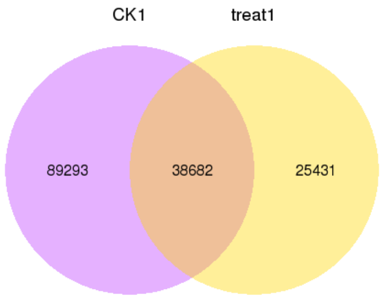


Figure 14 Venn Diagram of Expression Gene

The sum of the numbers in each circle is the total number of genes expressed within a group, and the overlap represents the genes expressed in common between groups. Use Fpkm > 0.3 as the criterion.

3.10.5 The Venn Diagram of Differentially Expressed Genes

The Venn diagram presents the number of genes that are uniquely expressed differentially within each group, with the overlapping regions showing the number of genes that are expressed in two or more groups. (The diagram depicts only the results for groups 2, 3, 4 and 5).

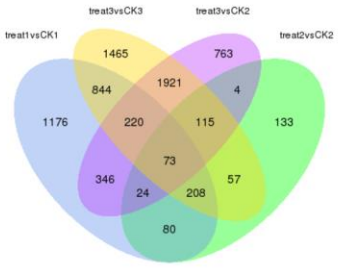


Figure 15 Venn diagram of differentially expressed genes

The sum of the numbers in each circle is the total number of genes expressed within a group, and the overlap represents the genes expressed in common between groups.

3.10.6 Cluster Analysis of Gene Expression Differences

Cluster Analysis is used to find genes with similar expression patterns under various experimental conditions. By clustering genes with similar expression patterns, it may be possible to discern unknown functions of previously characterized genes or the function of unknown genes. In hierarchical clustering, areas of different colors denote different groups (clusters) of genes, and genes within each cluster may have similar functions or take part in the same biological process. In addition to the FPKM cluster, the H-cluster, K-means and SOM are also used to cluster the log2(ratios). Genes within the same cluster exhibit the same trends in expression levels under different conditions.

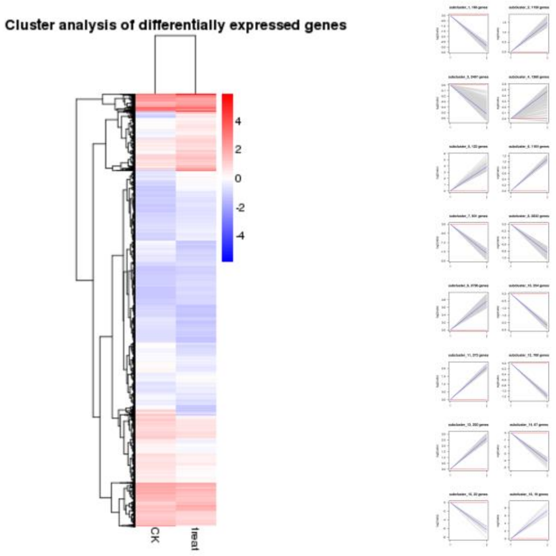


Figure 16 Cluster Analysis

Figure 1: the overall results of FPKM cluster analysis, clustered using the log10(FPKM+1) value. Red denotes genes with high expression levels, and blue denotes genes with low expression levels. The color range from red to blue represents the log10(FPKM+1) value from large to small. Figure 2: log2(ratios) line chart. Each grey line in a subline chart represents the relative expression value of a gene cluster under different experimental conditions, and the blue line represents the mean value. The x-axis shows the experimental condition and the y-axis shows the relative expression value.

3.11 GO Enrichment Analysis of DEGs

3.11.1 GO Enrichment Analysis of DEGs

Gene Ontology (http://www.geneontology.org/) is a major bioinformatics initiative to unify the presentation of gene and gene product attributes across all species. DEGs refer to differentially expressed genes.

Table 12 Significantly Enriched GO terms in DEGs

GO accession	Description	Term type	Over represented p-Value	Corrected p-Value	DEG item	DEG list
GO:0003824	catalytic activity	molecular_function	7.2525e-14	4.1237e-10	3572	6898
GO:0006629	lipid metabolic process	biological_process	1.0523e-12	2.9916e-09	492	6898
GO:0072330	monocarboxylic acid biosynthetic process	biological_process	3.5491e-12	6.7268e-09	107	6898
GO:0006631	fatty acid metabolic process	biological_process	1.5176e-11	2.1573e-08	115	6898

(1) Gene Ontology entry.

(2) Detailed description of Gene Ontology.

(3) GO types, including cellular_component, biological_process and molecular_function.

(4) p-value in hypergenometric test.

(5) Corrected P-value; GO with corrected p-value < 0.05 are significantly enriched in DEGs.

(6) Number of DEGs with GO annotation.

(7) Number of all reference genes with GO annotation.

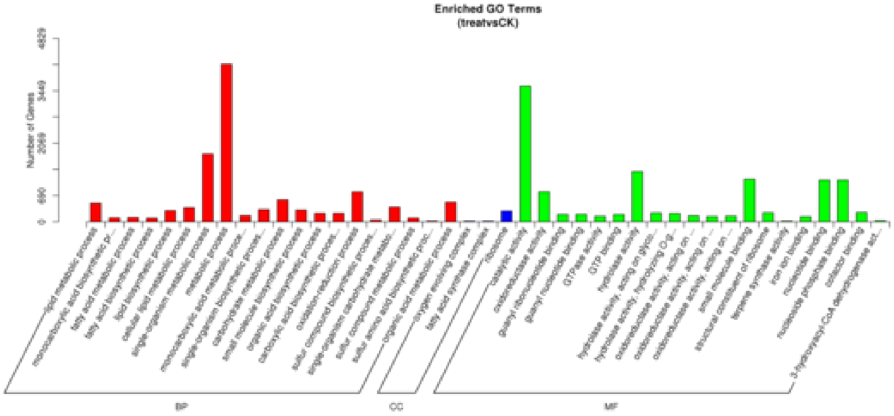


Figure 17 GO enrichment Bar Chart of DEGs

The x-axis shows GO term in the sub-level of the GO three main domains, and the y-axis shows the number of the differential expression genes annotated in this term and the ratio between this number and the total number of annotated differential expression genes. From left to right is the three main GO domains: biological_process, cell_composition and molecular_function.

3.11.2 GO Enrichment DAG Figure

DAG (Directed Acyclic Graph, DAG) can visually display the enriched GO term of differential expression genes and its hierarchy. (Figure17) illustrates the topGO DAGs. Branch means hierarchical relationship and the function ranges become more and more specific from top to bottom. DAG of biological process, molecular function and cellular component are shown respectively in the report.

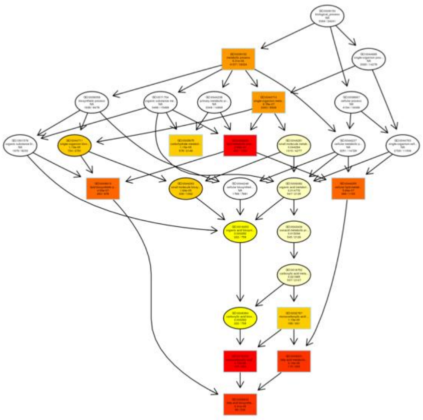


Figure 18 Illustration of topGO DAG

Each node represents a GO term, and TOP10 GO terms are boxed. The darker the color is, the higher is the enrichment level of the term. The name and p-value of each term are present on the node.

3.12 KEGG Pathway Enrichment Analysis

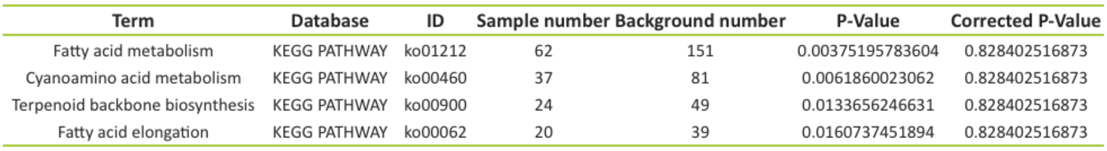
3.12.1 KEGG Pathway Enrichment Analysis

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

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

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

Table 13 KEGG Enrichment List



Term	Database	ID	Sample number	Background number	P-Value	Corrected P-Value
Fatty acid metabolism	KEGG PATHWAY	ko01212	62	151	0.00375195783604	0.828402516873
Cyanoamino acid metabolism	KEGG PATHWAY	ko00460	37	81	0.0061860023062	0.828402516873
Terpenoid backbone biosynthesis	KEGG PATHWAY	ko00900	24	49	0.0133656246631	0.828402516873
Fatty acid elongation	KEGG PATHWAY	ko00062	20	39	0.0160737451894	0.828402516873

(1) Term: Description of KEGG pathways

(2) ID: KEGG ID.

(3) Sample Number: Number of DGEs with pathway anntation.

(4) Background Number: Number of all reference genes with pathway annotation.

(5) P-value: P-value in hypergeometric test.

(6) Corrected P-value: Pathways with corrected p-value < 0.05 are significantly enriched in DEGs.

3.12.2 KEGG Enrichment Scattered Plot

KEGG enrichment scattered plot shows the DEGs enrichment analysis results in KEGG pathway. The degree of KEGG enrichment is measured by Rich factor, q-value and the number of genes enriched in this pathway. Rich factor refers to the ratio of the DEGs number in the pathway and the number of all genes annotated in the pathway. Q-value is the pvalue after normalization and its range is [0,1]. The smaller q-value is, the more significant the enrichment is. The top20 significantly DEGs enriched pathways are displayed in the report. If the enriched pathways are less than 20, all enriched pathways are displayed.

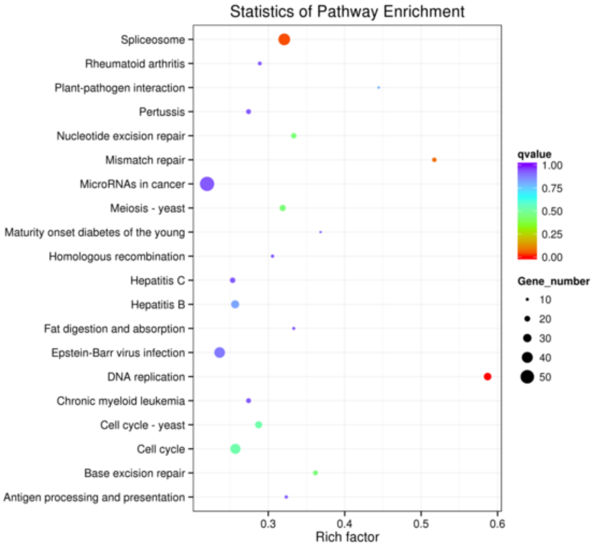


Figure 19 KEGG Enrichment Scatter Plot of DEGs

The y-axis represents the name of the pathway and the x-axis represents the Rich factor. Dot size represents the number of different genes and the color indicates the q-value.

3.12.3 KEGG Enrichment Pathway

KEGG enrichment pathway shows the DEGs significantly enriched pathways. In the diagram, if this node contains up-regulated genes, the KO node is labeled in red. If the node contains up-regulated genes, the KO node is labeled in green. If the node contains both up and down-regulated genes, the labeled color is yellow.

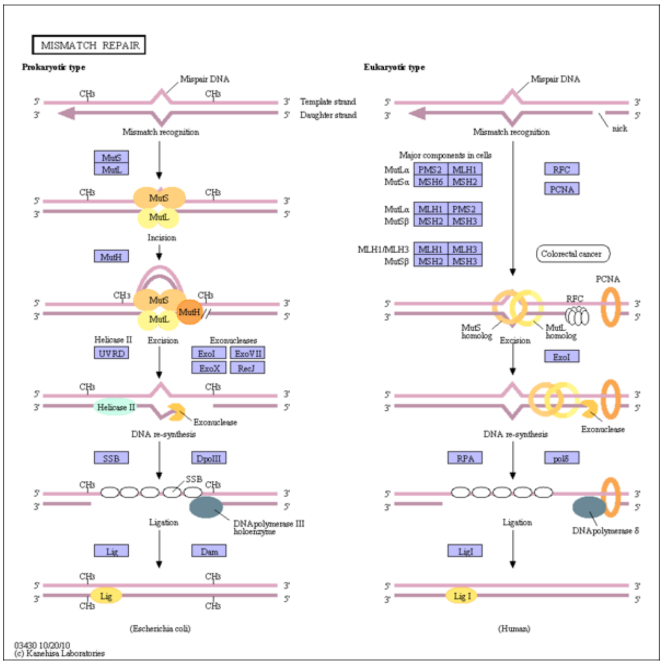


Figure 20 Diagram Showing Significantly Enriched KEGG Pathway

4 Appendix

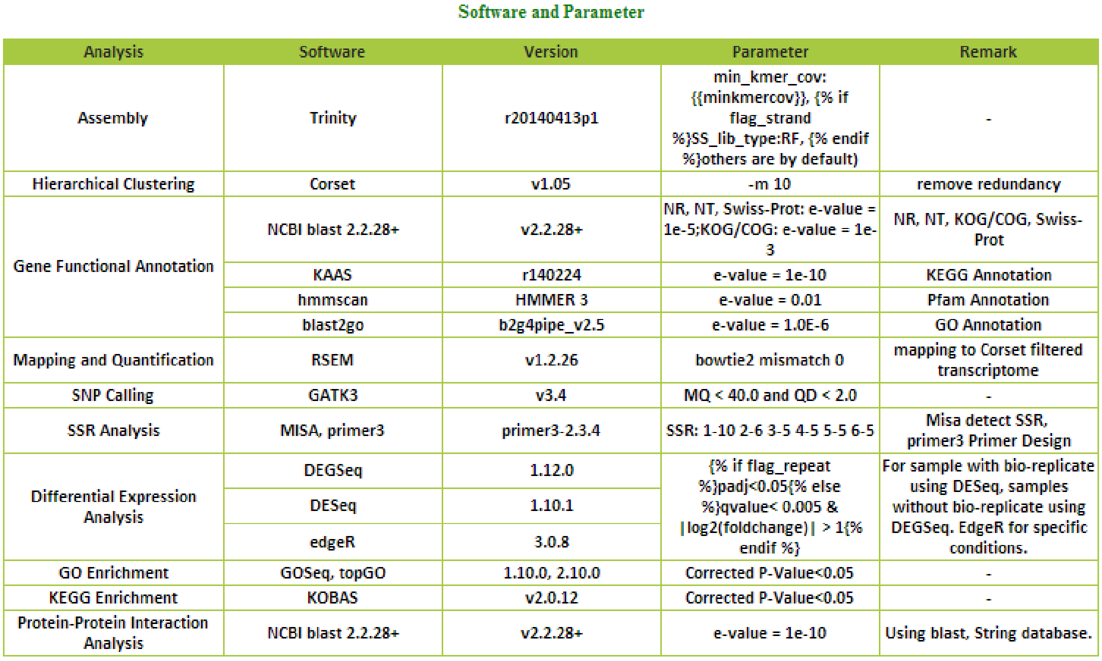
4.1 Result Directory Lists

Click to open the result directory. (Note: Please make sure the report directory and the result directory is under the same directory). Result Directory Lists: html

```
../../NHHWXXXXXX_species_results
├── 1. RawData: the rawdata of sequencing
├── 2. QC:  the results of quality control
│     ├── 2.1. ErrorRate: the results of ErrorRate distribution
│     ├── 2.2. GC: the results of GC distribution
│     ├── 2.3. ReadsClassification: Reads Composition display
│     └── 2.4. CleanData_QCsummary: the results of cleandata
├── 3. TranscriptomeAssembly:  the results of TRINITY
│     ├── 3.1. AssembledTranscriptome: the sequences of TRINITY  and Corset
│     └── 3.2. AssemblyINFO: the results of trinity and unigene
├── 4. GeneFunctionalAnnotation: the results of gene functional annotation
│     ├── 4.1. GeneFunctionalAnnotation: the results of gene functional annotation
│     ├── 4.2. GOclassification: GO annotation results
│     ├── 4.3. KOGclassification: KOG annotation results
│     └── 4.4. KEGGclassification: KEGG annotation results
├── 5. CDSprediction: CDS prediction results
├── 6. SNPcalling:  SNP/InDel analysis results
├── 7. SSRdetection: SSR analysis results
├── 8. GeneExprQuantification:  gene expression analysis
│     ├── 8.1. GeneExprQuantification: gene expression analysis
│     └── 8.2. GeneExpContrast: the comparison chart of Gene expression level
├── 9. RNA-seqQC: RNA-seq quality control evaluation
│     └── 9.1. Correlation: the results of correlation betwenn samples
├── 10. DiffExprAnalysis: the results of differential expression analysis
│     ├── 10.1.  DiffExprAnalysis: the results of differential expression analysis
│     ├── 10.2.  DEGsFilter:  differential genes selection（Volcanic diagram）
│     ├── 10.3.  VennDiagram: the diagram of venn
│     ├── 10.4.  DEGcluster: the cluster results of differential genes
│     └── 10.5.  DEGannotation:  the annotation results of differential genes
├── 11. DEG_GOenrichment: GO enrichment analysis results of differential genes
└── 12. DEG_KEGGenrichment:KEGG enrichment analysis results of differential
genes(all-all differential genes, up-up-regulation differential gene,
down-down-regulation differential gene)
```

4.2 Software List

Software and Parameter



Software and Parameter

Analysis	Software	Version	Parameter	Remark
Assembly	Trinity	r20140413p1	min_kmer_cov: {{minkmercov}}, {% if flag_strand %}SS_lib_type:RF, {% endif %}others are by default)	-
Hierarchical Clustering	Corset	v1.05	-m 10	remove redundancy
Gene Functional Annotation	NCBI blast 2.2.28+	v2.2.28+	NR, NT, Swiss-Prot: e-value = 1e-5;KOG/COG: e-value = 1e-3	NR, NT, KOG/COG, Swiss-Prot
	KAAS	r140224	e-value = 1e-10	KEGG Annotation
	hmmscan	HMMER 3	e-value = 0.01	Pfam Annotation
	blast2go	b2g4pipe_v2.5	e-value = 1.0E-6	GO Annotation
Mapping and Quantification	RSEM	v1.2.26	bowtie2 mismatch 0	mapping to Corset filtered transcriptome
SNP Calling	GATK3	v3.4	MQ < 40.0 and QD < 2.0	-
SSR Analysis	MISA, primer3	primer3-2.3.4	SSR: 1-10 2-6 3-5 4-5 5-5 6-5	Misa detect SSR, primer3 Primer Design
Differential Expression Analysis	DEGSeq	1.12.0	{% if flag_repeat %}padj<0.05{% else %}qvalue< 0.005 & |log2(foldchange)| > 1{% endif %}	For sample with bio-replicate using DESeq, samples without bio-replicate using DEGSeq. EdgeR for specific conditions.
	DESeq	1.10.1		
	edgeR	3.0.8		
GO Enrichment	GOSeq, topGO	1.10.0, 2.10.0	Corrected P-Value<0.05	-
KEGG Enrichment	KOBAS	v2.0.12	Corrected P-Value<0.05	-
Protein-Protein Interaction Analysis	NCBI blast 2.2.28+	v2.2.28+	e-value = 1e-10	Using blast, String database.

4.3 Novofinder Manual

We developed a powerful software Novofinder to help customer browse and integrate bioinforamtic results. Results could be accessed through Gene id, key word of gene function, expression level and other customer defined way. Through Novofinder, customer could easily get well-organized gene sequence, functional annotation, SNP/InDel, SSR, quantification, differentially expressed gene analysis and enrichement results. Hope novofinder could accelerate your research process!

Novofinder Manual: PDF

5 References

Cock P J A, Fields C J, Goto N, et al. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research 38, 1767-1771. (FASTQ)

Hansen K D, Brenner S E, Dudoit S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic acids research 38, e131-e131. (Biases)

Jiang L, Schlesinger F, Davis C A, et al. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome research 21, 1543-1551. (spike-in)

Grabherr M G, Haas B J, Yassour M, et al. (2011).Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29, 644-652. (Trinity)

Nadia M D, Alicia O. (2014). Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biology 15, 1-14. (Corset)

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

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

Moriya Y, Itoh M, Okuda S, et al. (2007). KAAS: an automatic genome annotation and pathway reconstruction server[J]. Nucleic acids research, 2007, 35(suppl 2): W182-W185. (KAAS)

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

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

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

Li B, Dewey C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, doi:10.1186/1471-2105-12-323. (RSEM)

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

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

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

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

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

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

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