F17FTSAPJT0170_FUNxovR(HiSeq RNA-Seq)

2017/5/24

@2017 BGI All Rights Reserved

Table of Contents

Results

1 QC and Reads Mapping

In the project numbered F17FTSAPJT0170_FUNxovR, we sequenced 6 samples of Aspergillus_Nidulans species using RNA-Seq technology[1][2], averagely generating 7,417,146 raw sequencing reads and then 7,416,375 clean reads after filtering low quality (see Data Filtering in method page). Table1 briefly summarizes information of sequencing data for each sample. Distribution charts of base composition and base quality on clean reads are also presented, respectively in Figure1.

After filtering, clean reads are mapped to reference using HISAT [3]/ Bowtie2 [4] tool (see Reads Mapping in method page for details). The average mapping ratio with reference gene is 66.75% and Table2 lists seperate mapping rate for each sample. The average genome mapping ratio is 92.92% corresponding to Table3. We conducted strict quality control for each sample from several aspects as illustrated in Table4, to evalute whether the sequencing data are qualified.

Table 1 Summary of sequencing data for each sample (Download)

Sample	Sequencing Strategy	Raw Data Size (bp)	Raw Reads Number	Clean Data Size (bp)	Clean Reads Number	Clean Data Rate (%)
Cont_24h_1_C1_24h	SE50	393,684,963	8,034,387	393,638,168	8,033,432	99.98
Cont_48h_1_C1_48h	SE50	357,848,960	7,303,040	357,809,809	7,302,241	99.98
Diploid_24h_1_D1_24h	SE50	340,837,728	6,955,872	340,795,735	6,955,015	99.98
Diploid_48h_1_D1_48h	SE50	330,045,723	6,735,627	330,021,762	6,735,138	99.99
Humic_24h_1_H1_24h	SE50	350,334,173	7,149,677	350,291,739	7,148,811	99.98
Humic_48h_1_H1_48h	SE50	407,889,524	8,324,276	407,857,135	8,323,615	99.99

Clean Data Rate (%)=Clean Reads Number/Raw Reads Number

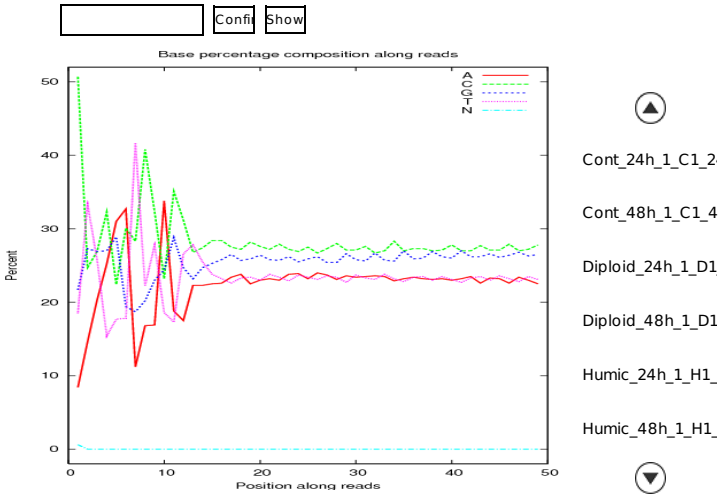

Figure 1 Distribution of base composition on clean reads.

X axis is base positions along reads. Y axis is base quality value. Each dot in the image represents the number of total bases with certain quality value of the corresponding base along reads. Darker dot color means greater bases number. If the percentage of the bases with low quality (< 20) is very high, then the sequencing quality of this lane is bad. If one sample name appears several times in the right box, that means it is consisted of more than one sequencing lane.

Table 2 Alignment statistics of reads align to reference gene (Download)

Sample	Total Reads	Total Mapped Reads (%)	Unique Match(%)	Multi-position Match (%)	Total Unmapped Reads (%)
Cont_24h_1_C1_24h	8,033,432	63.00	62.73	0.27	37.00
Cont_48h_1_C1_48h	7,302,241	62.54	62.29	0.25	37.46
Diploid_24h_1_D1_24h	6,955,015	69.25	68.89	0.36	30.75
Diploid_48h_1_D1_48h	6,735,138	68.72	68.50	0.23	31.28
Humic_24h_1_H1_24h	7,148,811	69.71	69.38	0.33	30.29
Humic_48h_1_H1_48h	8,323,615	67.26	66.97	0.29	32.74

Total Mapped Reads (%) = Unique Match (%) + Multi-position Match (%).

Table 3 Alignment statistics of reads align to reference genome (Download)

Sample	Total Reads	Total Mapped Reads (%)	Unique Match(%)	Multi-position Match (%)	Total Unmapped Reads (%)
Cont_24h_1_C1_24h	8,033,432	86.68	83.99	2.69	13.33
Cont_48h_1_C1_48h	7,302,241	88.81	86.09	2.72	11.19
Diploid_24h_1_D1_24h	6,955,015	94.23	91.38	2.85	5.77
Diploid_48h_1_D1_48h	6,735,138	96.43	93.79	2.64	3.57
Humic_24h_1_H1_24h	7,148,811	95.17	92.36	2.81	4.83
Humic_48h_1_H1_48h	8,323,615	96.20	93.42	2.78	3.80

Total Mapped Reads (%) = Unique Match (%) + Multi-position Match (%).

Table 4 QC items for each sample (Download)

Sample	Clean Read1 Q20(%) >= 95	Clean Read1 Q30(%) >= 90	Clean Reads >= 5 (M)	Gene Unique Mapping Ratio(%) >= 80	Genome Mapping Ratio(%) >= 50
Cont_24h_1_C1_24h	98.9 (Y)	96.6 (Y)	8.03 (Y)	99.57 (Y)	86.68 (Y)
Cont_48h_1_C1_48h	98.9 (Y)	96.7 (Y)	7.30 (Y)	99.60 (Y)	88.81 (Y)
Diploid_24h_1_D1_24h	98.9 (Y)	96.6 (Y)	6.96 (Y)	99.48 (Y)	94.23 (Y)
Diploid_48h_1_D1_48h	98.9 (Y)	96.8 (Y)	6.74 (Y)	99.68 (Y)	96.43 (Y)
Humic_24h_1_H1_24h	98.9 (Y)	96.8 (Y)	7.15 (Y)	99.53 (Y)	95.17 (Y)
Humic_48h_1_H1_48h	98.9 (Y)	96.7 (Y)	8.32 (Y)	99.57 (Y)	96.2 (Y)

'Y' means sample passed this QC item and 'N' means failed.

2 Sequencing Saturation and Reads Randomness

Sequencing data saturation analysis is used to measure whether the depth of sequencing data is sufficient for informatic analysis. With the number of sequenced reads increasing, the number of identified genes is also increased. However, when the number of sequenced reads reaches a certain amount, the growth curve of identified genes flattens, indicating that the number of identified genes tends to reach the saturation. Figure2 displays saturation analysis for each sample.

The distribution of the reads on reference gene reflects whether each part of the gene body are evenly sequenced. If the randomness is good, the reads in every position(from 5' terminal to 3' terminal) would be evenly distributed. If the randomness is poor, reads preference to specific gene region will directly affect subsequent bioinformatics analysis. We use the distribution of reads on the reference genes to evaluate the fragmentation randomness, showing as Figure3.

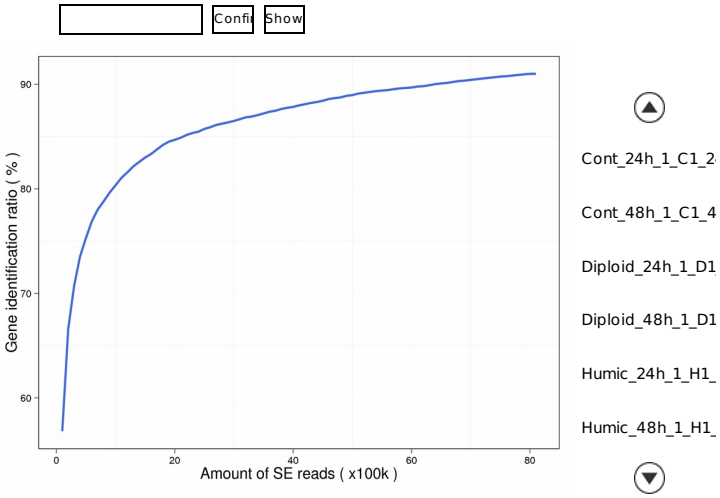

Figure 2 urve of sequencing saturation.

X-axis shows the number of clean reads, units is 100 k -- extreme value is currently the volume of sequencing. Y-axis shows the ratio of identified gene number to number of total gene reported in database.

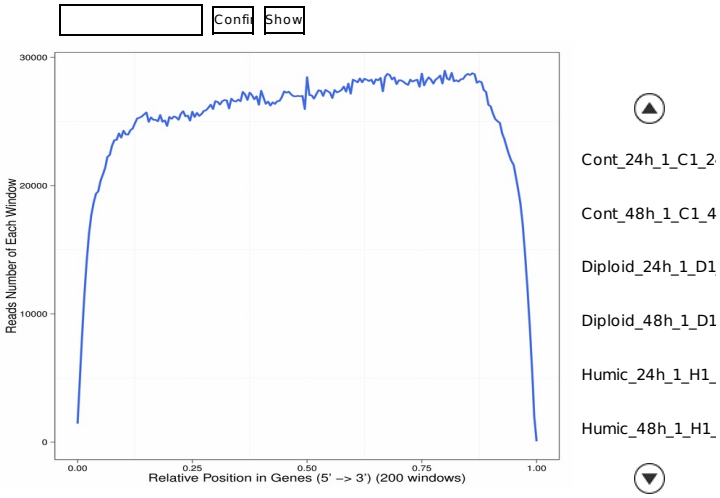

Figure 3 Reads distribution on reference gene.

Because of varialbe lengths of reference genes, the average length of genes is divided into N equal parts. Each equal part is called a window. X-axis shows the relative position of genes, and Y-axis shows the number of reads in each window.

3 Gene Expression

Genes expression level is quantified by a software package called RSEM [5](see Gene Quantification in method page). We counted the number of identified expressed genes and calculated its proportion to total gene number in database for each sample as Figure4. Meanwhile, the distribution of gene number on different expression level for each sample are shown as Figure5, from which would help us can get a general idea about how genes express at high and low level. The following listed file suffixed with *gene.* FPKM *.xls* are results of gene expression for each sample (see File Format of Gene Expression Resultin help page). The file *all.gene.* FPKM *.xls* is a expression table for all samples with brief gene

description and annotation.

Table 5 Cont_24h_1_C1_24h.gene.FPKM.xls (Download)
Table 6 Cont_48h_1_C1_48h.gene.FPKM.xls (Download)
Table 7 Diploid_24h_1_D1_24h.gene.FPKM.xls (Download)
Table 8 Diploid_48h_1_D1_48h.gene.FPKM.xls (Download)
Table 9 Humic_24h_1_H1_24h.gene.FPKM.xls (Download)
Table 10 Humic_48h_1_H1_48h.gene.FPKM.xls (Download)
Table 11 all.gene.FPKM.xls (Download)

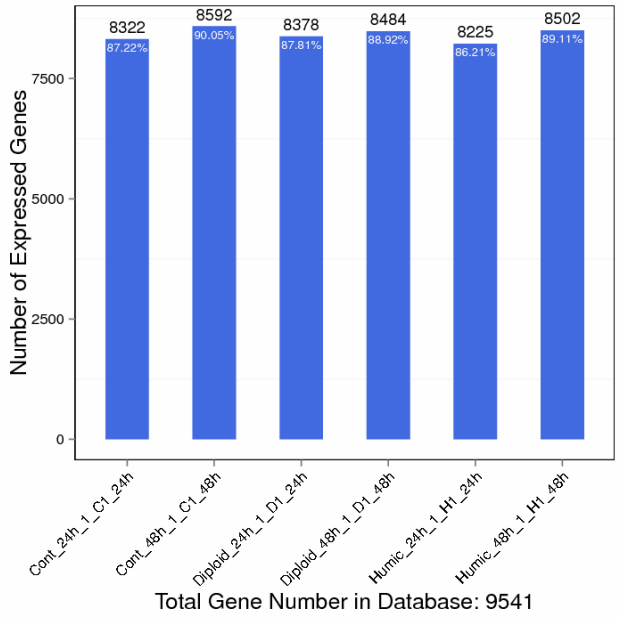

Figure 4 Number of identied genes.

X-axis is sample name. Y-axis is number of identified expressed genes. The proportion at the top of each bar equals expressed genes number devided by total gene number reported in database.

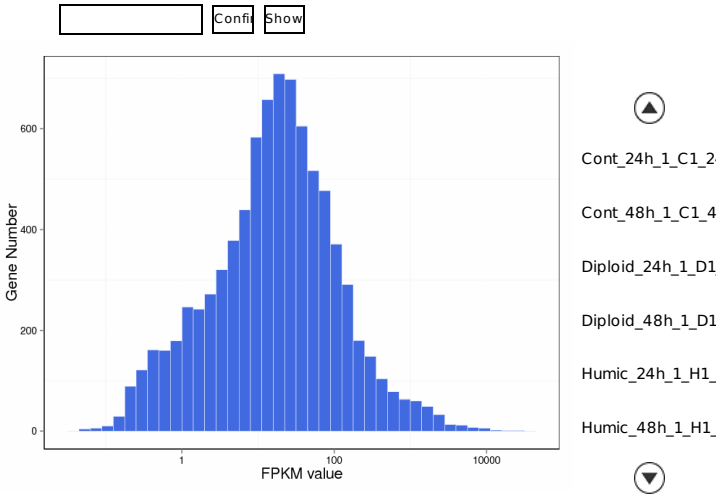

Figure 5 Histogram distribution of genes on expression level of each sample.

X-axis is FPKM value (the coordinate has been changed by logarithm for better view). Y-axis is gene number of corresponding FPKM.

4 Deep Analysis on Sample Level

For multiple samples, we can do more deep analysis based on gene expression to do a comprehensive assess on the whole project.

Correlation bewteen Samples

Biological replicates are required for almost every biological experiment, and high-throughput sequencing technology is no exception [6]. The correlation of gene expression level among samples is a key criterion to test whether the experiments are reliable and whether the samples chosen are reasonable. We calculate correlation value bewteen each two samples based on normalized expression result and draw correlation heatmap as Figure6 (click All.correlation.stat.xls to see concrete values).Cluster tree presenting the distance among samples is also built as Figure7.

Clustering of Gene Expression

Genes with similar expression patterns usually have same functional correlation. So we perform clustering analysis of gene expression patterns with cluster [7] [8] and javaTreeview software [9] according to the provided cluster plans. Please read detailed report named *expCluster_en.html* unber project result folder *BGI_result/3.QuantitativeAnalysis/GeneCluster* (see the section How to Read Report of Clustering Analysis in method page). We also provide complete intersection and union gene expression heatmap for each cluster plan as Figure8 and Figure9 respectively.

Venn Charts of Genes

What's more, we can draw venn chart to display common genes between(among) samples as Figure10 according to the provided plans.

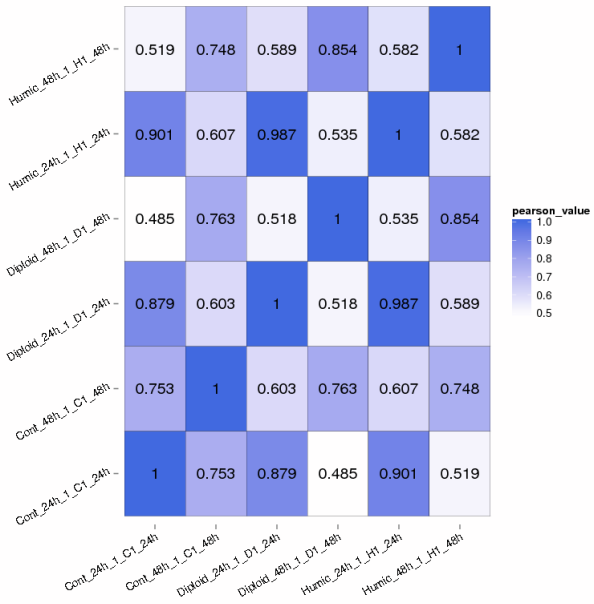

Figure 6 Heatmap of correlation coefficient values acrossing samples.

Gradient color barcode at the right top indicates the minimum value in white and the maximum in blue. If one sample is highly similar with another one, the correlation value between them is very close to 1.

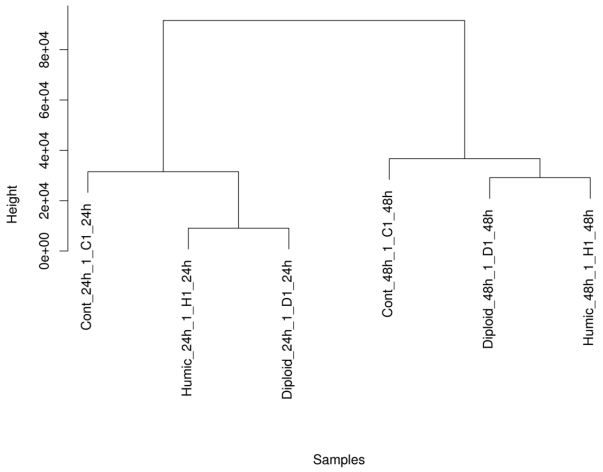

Figure 7 Cluster tree involving all samples.

The distances of expressed gene are calculated by euclidean method. Meanwhile, the algorithm of Sumof Squares of Deviations is used to calculate the distance between samples so that cluster tree can be build. Y axis means height in the cluster tree. When samples have similar height values, they are easily to be gathered.

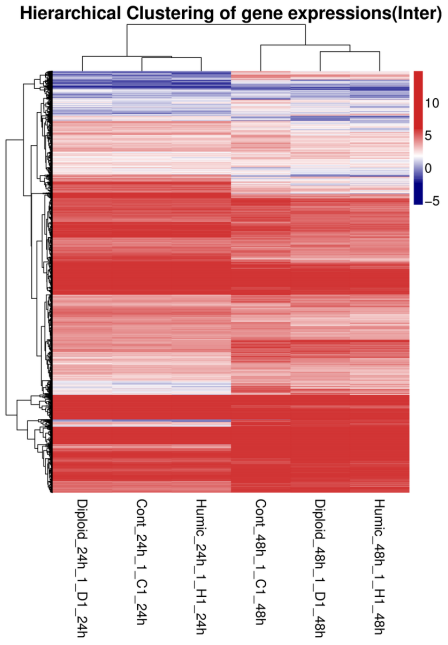

Figure 8 Intersection heatmap of gene expression for each cluster plan.

Only genes that expressed in all samples of cluster plan are used to build this heatmap. Gradient color barcode at the right top indicates log2(FPKM) value. Each row represents a gene and each column represents a sample(for some reason in R method, if there is only one sample, the sample name doesn't appear in bottom). Genes with similar expression value are clustered both at row and column level.

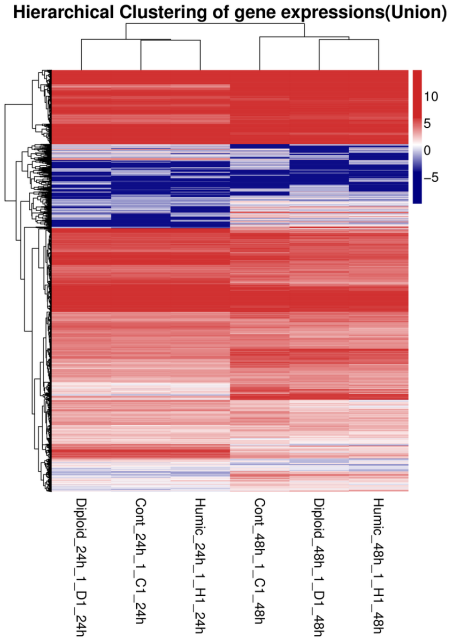

Figure 9 Union heatmap of gene expression for each cluster plan.

Genes that expressed in at least one sample of cluster plan are used to build this heatmap. Non-expressed genes value will be replaced with a very small value 0.001. Gradient color barcode at the right top indicates log2(FPKM) value. Each row represents a gene and each column represents a sample. Genes with similar expression value are clustered both at row and column level.

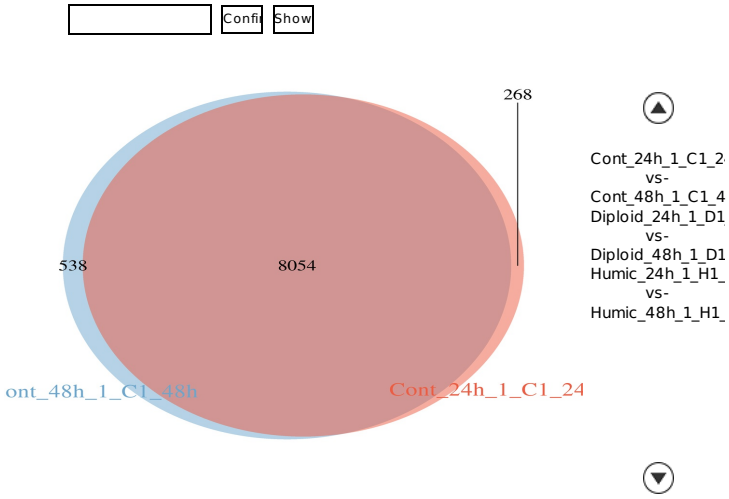

Figure 10 Venn chart of co-expressed genes between (among) samples.

For two samples, the figure is proportional, but can't be proportional when sample number is more than 2. The pipeline supports at most 5 samples to draw one venn chart.

5 Screening Differentially Expressed Genes (using possionDis)

DEGs screening is aimed to find differential expressed genes between samples and perform further function analysis on them. We use possion distribution method to do this analysis(see Screening DEGs using Possion Distribution Method in method page).

All expressed genes of each pairwise are stored in *.*GeneDiffExp.xls* and screened DEGs

are stored *in* *.*GeneDiffExpFilter.xls*. They have the same file format and are listed as following (The name before "-VS-" is control and after it is treatment case. Please see DEGs screening Format (using possionDis) in help page):

Table 12 Cont_24h_1_C1_24h-VS-Cont_48h_1_C1_48h.GeneDiffExp.xls (Download)
Table 13 Cont_24h_1_C1_24h-VS-Cont_48h_1_C1_48h.GeneDiffExpFilter.xls (Download)
Table 14 Cont_24h_1_C1_24h-VS-Diploid_24h_1_D1_24h.GeneDiffExp.xls (Download)
Table 15 Cont_24h_1_C1_24h-VS-Diploid_24h_1_D1_24h.GeneDiffExpFilter.xls (Download)
Table 16 Cont_24h_1_C1_24h-VS-Diploid_48h_1_D1_48h.GeneDiffExp.xls (Download)
Table 17 Cont_24h_1_C1_24h-VS-Diploid_48h_1_D1_48h.GeneDiffExpFilter.xls (Download)
Table 18 Cont_24h_1_C1_24h-VS-Humic_24h_1_H1_24h.GeneDiffExp.xls (Download)
Table 19 Cont_24h_1_C1_24h-VS-Humic_24h_1_H1_24h.GeneDiffExpFilter.xls (Download)
Table 20 Cont_24h_1_C1_24h-VS-Humic_48h_1_H1_48h.GeneDiffExp.xls (Download)
Table 21 Cont_24h_1_C1_24h-VS-Humic_48h_1_H1_48h.GeneDiffExpFilter.xls (Download)
Table 22 Cont_48h_1_C1_48h-VS-Diploid_24h_1_D1_24h.GeneDiffExp.xls (Download)
Table 23 Cont_48h_1_C1_48h-VS-Diploid_24h_1_D1_24h.GeneDiffExpFilter.xls (Download)
Table 24 Cont_48h_1_C1_48h-VS-Diploid_48h_1_D1_48h.GeneDiffExp.xls (Download)
Table 25 Cont_48h_1_C1_48h-VS-Diploid_48h_1_D1_48h.GeneDiffExpFilter.xls (Download)
Table 26 Cont_48h_1_C1_48h-VS-Humic_24h_1_H1_24h.GeneDiffExp.xls (Download)
Table 27 Cont_48h_1_C1_48h-VS-Humic_24h_1_H1_24h.GeneDiffExpFilter.xls (Download)
Table 28 Cont_48h_1_C1_48h-VS-Humic_48h_1_H1_48h.GeneDiffExp.xls (Download)
Table 29 Cont_48h_1_C1_48h-VS-Humic_48h_1_H1_48h.GeneDiffExpFilter.xls (Download)
Table 30 Diploid_24h_1_D1_24h-VS-Diploid_48h_1_D1_48h.GeneDiffExp.xls (Download)
Table 31 Diploid_24h_1_D1_24h-VS-Diploid_48h_1_D1_48h.GeneDiffExpFilter.xls (Download)
Table 32 Diploid_24h_1_D1_24h-VS-Humic_24h_1_H1_24h.GeneDiffExp.xls (Download)
Table 33 Diploid_24h_1_D1_24h-VS-Humic_24h_1_H1_24h.GeneDiffExpFilter.xls (Download)
Table 34 Diploid_24h_1_D1_24h-VS-Humic_48h_1_H1_48h.GeneDiffExp.xls (Download)
Table 35 Diploid_24h_1_D1_24h-VS-Humic_48h_1_H1_48h.GeneDiffExpFilter.xls (Download)
Table 36 Diploid_48h_1_D1_48h-VS-Humic_24h_1_H1_24h.GeneDiffExp.xls (Download)
Table 37 Diploid_48h_1_D1_48h-VS-Humic_24h_1_H1_24h.GeneDiffExpFilter.xls (Download)
Table 38 Diploid_48h_1_D1_48h-VS-Humic_48h_1_H1_48h.GeneDiffExp.xls (Download)
Table 39 Diploid_48h_1_D1_48h-VS-Humic_48h_1_H1_48h.GeneDiffExpFilter.xls (Download)
Table 40 Humic_24h_1_H1_24h-VS-Humic_48h_1_H1_48h.GeneDiffExp.xls (Download)
Table 41 Humic_24h_1_H1_24h-VS-Humic_48h_1_H1_48h.GeneDiffExpFilter.xls (Download)

For result list of each control-treatment pair above(*.GeneDiffExp.xls), we draw scatter plots of all expressed genes as Figure11 and volcano graph as Figure12 to present the distribution of DEGs in screening threshold dimensions. At last, an histogram is drawed to show significant up-down regulation gene numbers in each pairwise as Figure13.

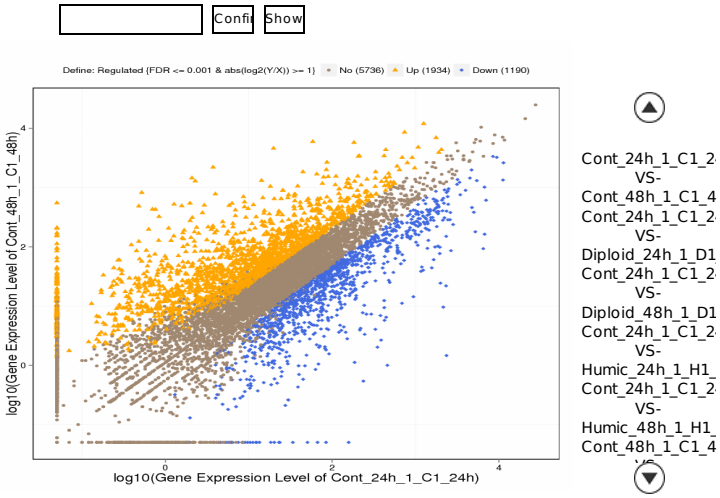

Figure 11 Scatter plots of all expressed genes in each pairwise.

X-axis and Y-axis present log2 value of gene expression. Blue means down-regulation gene, orange means up-regulation gene and brown means non-regulation gene. If a gene expressed just in one sample, its expression value in another sample will be replaced by the minimum value of all expressed genes in control and case samples. Screening threshold is on top legend.

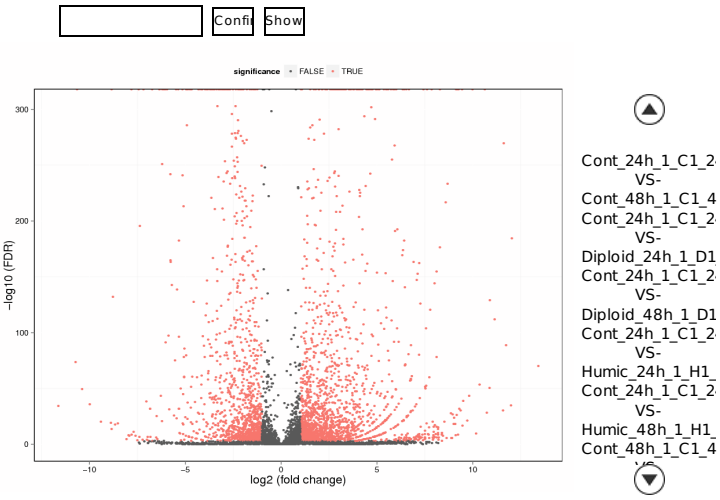

Figure 12 Volcano graph of all expressed genes in each pairwise.

X-axis and Y-axis present threshold value in log transform. Each dot is a differential expressed genes. Dots in red mean significant DEGs which passed screening threshold and black dots are non-significant DEGs. Threshold can be known from Figure11 above.

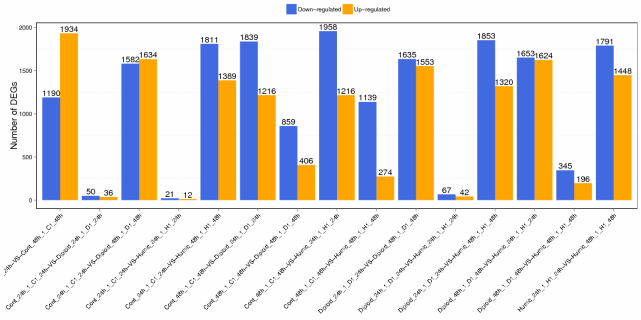

Figure 13 Statistic of differentially expressed genes.

X axis represents pairwise and Y axis means number of screened DEGs. Blue bar denotes down-regulated genes and orange bar for the up-regulated.

6 Screening Differentially Expressed Genes (using Noiseq)

DEGs screening is aimed to find differential expressed genes between samples and perform further function analysis on them. We use NOISeq method to do this analysis(see \title{Screening DEGs using NOISeq method} in method page). For NOISeq method, samples should be firstly grouped so that comparison between every two groups as a control-treatment pairwise can be done later. The provided group information is as following:

```
24h:Cont_24h_1_C1_24h,Diploid_24h_1_D1_24h,Humic_24h_1_H1_24h
48h:Cont_48h_1_C1_48h,Diploid_48h_1_D1_48h,Humic_48h_1_H1_48h
Cont:Cont_24h_1_C1_24h,Cont_48h_1_C1_48h
Diploid:Diploid_24h_1_D1_24h,Diploid_48h_1_D1_48h
Humic:Humic_24h_1_H1_24h,Humic_48h_1_H1_48h
```

All expressed genes of each pairwise are stored in *.*GeneDiffExp.xls* and screened DEGs are stored *in* *.*GeneDiffExpFilter.xls*. They have the same file format and are listed as following (The name before "-VS-" is control and after it is treatment case. Please see DEGs screening Format (using Noiseq) in help page):

Table 42 24h-VS-48h.GeneDiffExp.xls (Download)
Table 43 24h-VS-48h.GeneDiffExpFilter.xls (Download)
Table 44 Cont-VS-Diploid.GeneDiffExp.xls (Download)
Table 45 Cont-VS-Diploid.GeneDiffExpFilter.xls (Download)
Table 46 Cont-VS-Humic.GeneDiffExp.xls (Download)
Table 47 Cont-VS-Humic.GeneDiffExpFilter.xls (Download)
Table 48 Diploid-VS-Humic.GeneDiffExp.xls (Download)
Table 49 Diploid-VS-Humic.GeneDiffExpFilter.xls (Download)

For result list of each control-treatment pair above(*.GeneDiffExp.xls), we draw scatter plots of all expressed genes as Figure14 and volcano graph as Figure15 to present the distribution of DEGs in screening threshold dimensions. At last, an histogram is drawed to show significant up-down regulation gene numbers in each pairwise as Figure16.

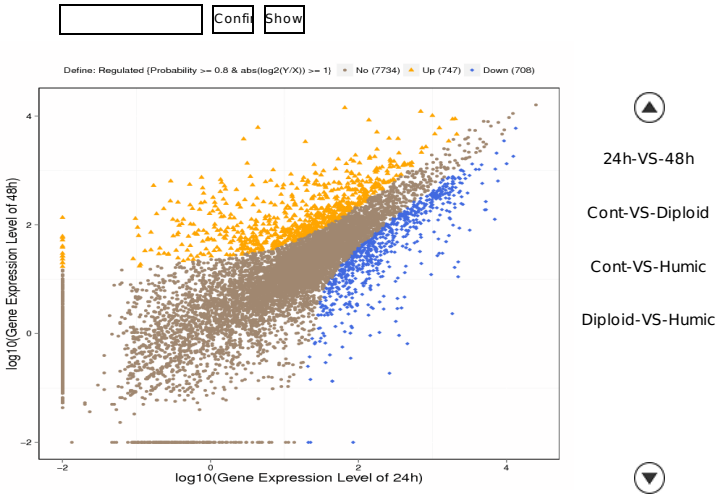

Figure 14 Scatter plots of all expressed genes in each pairwise.

X-axis and Y-axis present log2 value of gene expression. Blue means down-regulation gene, orange means up-regulation gene and brown means non-regulation gene. If a gene expressed just in one sample, its expression value in another sample will be replaced by the minimum value of all expressed genes in control and case samples. Screening threshold is on top legend.

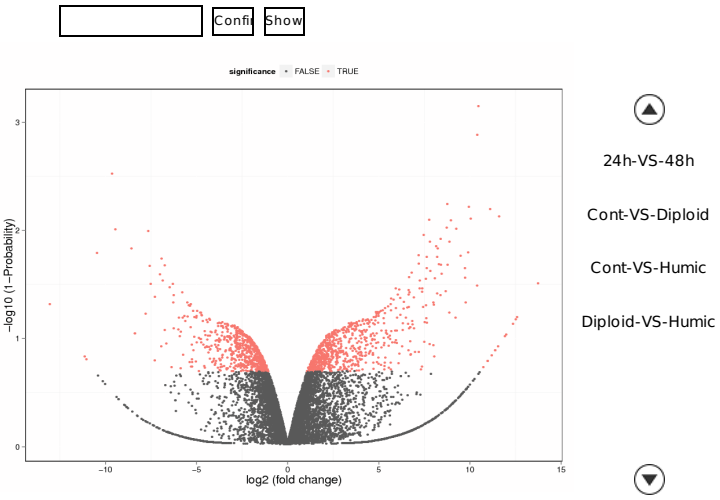

Figure 15 Volcano graph of all expressed genes in each pairwise.

X-axis and Y-axis present threshold value in log transform. Each dot is a differential expressed genes. Dots in red mean significant DEGs which passed screening threshold and black dots are non-significant DEGs. Threshold can be known from Figure14 above.

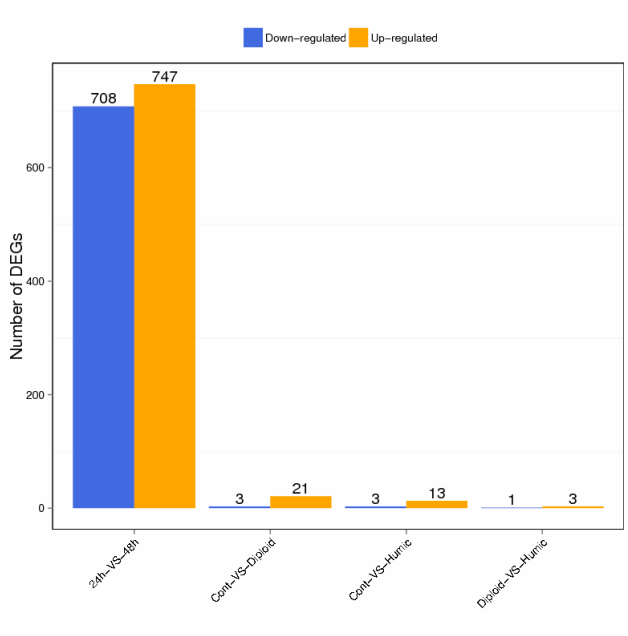

Figure 16 Statistic of differentially expressed genes.

X axis represents pairwise and Y axis means number of screened DEGs. Blue bar denotes down-regulated genes and orange bar for the up-regulated.

7 Clustering Analysis of DEGs

Genes with similar expression patterns usually have same functional correlation. So we perform clustering analysis of differentially expressed genes with cluster [7] [8] and javaTreeview software [9] according to the provided cluster plans for DEGs. Please read report named *cluster_en.html* under project result folder *BGI_result/3.QuantitativeAnalysis/GeneDiff_Function/Cluster/* (see How to Read Report of Clustering Analysis in help page). We also provide complete intersection and union DEGs heatmap for each cluster plan as Figure17 respectively(intersection DEGs heatmap is empty).

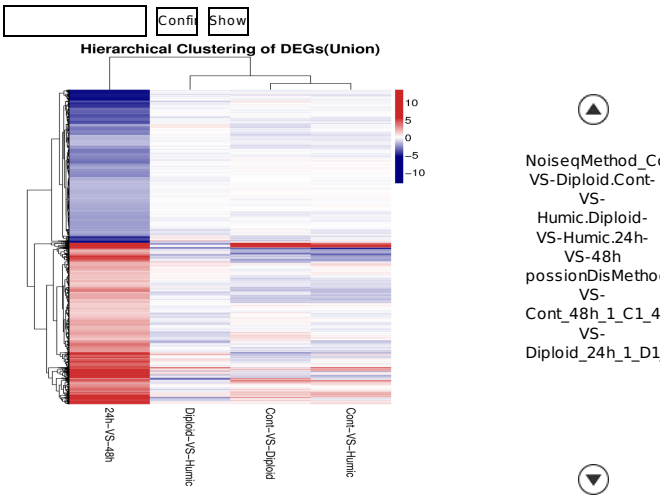

Figure 17 Union heatmap of DEGs for each cluster plan.

Genes that expressed in all pairwises of cluster plan and differentially expressed in at least one pairwise are used to build this heatmap. Gradient color barcode at the right top indicates log2(FC) value (FC, Foldchange of expression in treatment case to expression in control case). Each row represents a DEG and each column represents one condition pairwise(for some reason in R method, if there is only one pairwise, the pairwise name doesn't appear in bottom). DEGs with similar foldchange value are clustered both at row and column level.

8 Gene Ontology Analysis of DEGs

Annotation analysis of Gene Ontology (GO) are performed for screened DEGs and then generate a report named *GOView.html* under the project folder *BGI_result/3.QuantitativeAnalysis/GeneDiff_Function/GO/* (see the section Gene Ontology Annotation in method page and the section How to Read Report of GO Annotation in help page). After getting GO annotation for DEGs, we use WEGO software [10] to do GO functional classification as Figure18 to help understand the distribution of gene functions of the specie from the macro level.

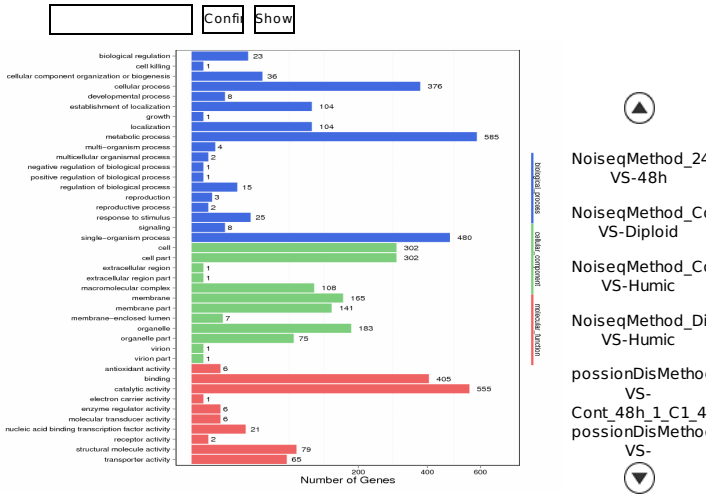

Figure 18 GO functional classification on DEGs for each pairwise.

X axis means number of DEGs (the number is presented by its square root value). Y axis represents GO terms. All GO terms are grouped in to three ontologies: blue is for biological process, brown is for cellular component and orange is for molecular function.

9 Pathway Enrichment Analysis of DEGs

Genes usually interact with each other to play roles in certain biological functions. We perform pathway enrichment analysis of DEGs based on KEGG database [11] and generate a report for DEGs in each pairwise respectively, stored in project result folder *BGI_result/3.QuantitativeAnalysis/GeneDiff_Function/Pathway/* (see the section KEGG Pathway Enrichment in method page and the section How to Read Report of Pathway Enrichment in help page). In addition, we generate a scatter plot for the top 20 of KEGG enrichment results as Figure19 and a bar plot for the statistics of KEGG terms types as Figure20.

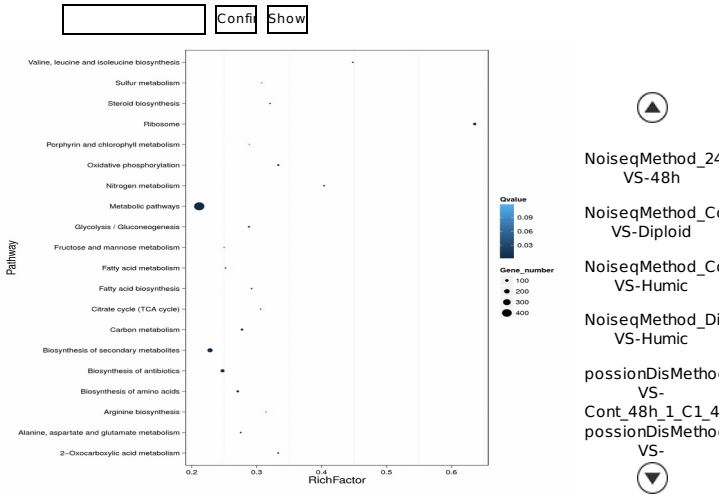

Figure 19 Statistics of pathway enrichment of DEGs in each pairwise.

RichFactor is the ratio of differentially expressed gene numbers annoted in this pathway term to all gene numbers annoted in this pathway term. Greater richFator means greater intensiveness. Qvalue is corrected pvalue ranging from 0~1, and less Qvalue means greater intensiveness. We just display the top 20 of enriched pathway terms.

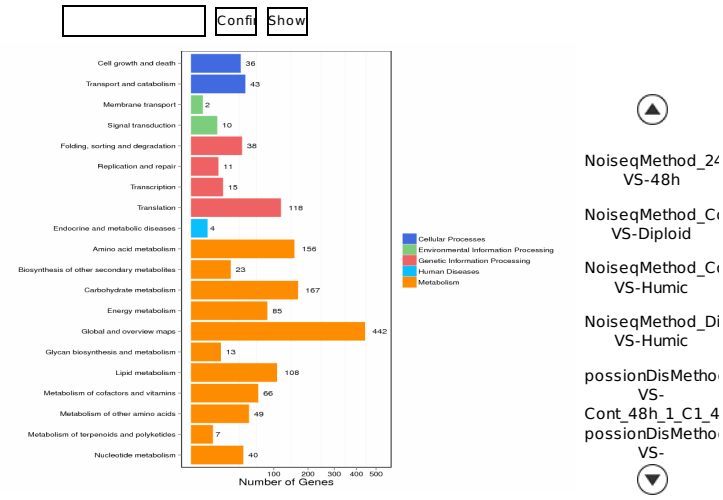

Figure 20 KEGG classification on DEGs for each pairwise.

X axis means number of DEGs. Y axis represents second KEGG pathway terms. All second pathway terms are grouped in top pathway terms indicated in different color.

Methods

1 Experiment Pipeline

Each step in experiment process (like sample test, library construction and sequecing) influences data quality and quantity, and then directly affect bioinformatics analysis results. To get high realiable sequencing data, we carry out strict quality control in each experiment step. The experiment pipeline is described as Figure1.

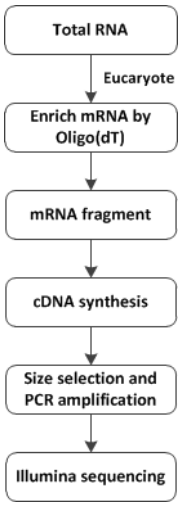

Figure 1 RNA-Seq experimental process.

The total RNA samples are first treated with DNase I to degrade any possible DNA contamination. Then the mRNA is enriched by using the oligo (dT) magnetic beads. Mixed with the fragmentation buffer, the mRNA is fragmented into short fragments. Then the first strand of cDNA is synthesized by using random hexamer-primer. Buffer, dNTPs, RNase H and DNA polymerase I are added to synthesize the second strand. The double strand cDNA is purified with magnetic beads. End reparation and 3'-end single nucleotide A (adenine) addition is then performed. Finally, sequencing adaptors are ligated to the fragments. The fragments are enriched by PCR amplification. During the QC step, Agilent 2100 Bioanaylzer and ABI StepOnePlus Real-Time PCR System are used to qualify and quantify of the sample library. The library products are ready for sequencing via Illumina HiSeq^{TM}2000 or other sequencer when necessary.

2 Bioinformatics Pipeline

After getting raw data, we will do each bioinformitics analysis as the client appoints on contract. Figure2 demonstrates a complete pipeline for RNA-Seq (Quantification) project.

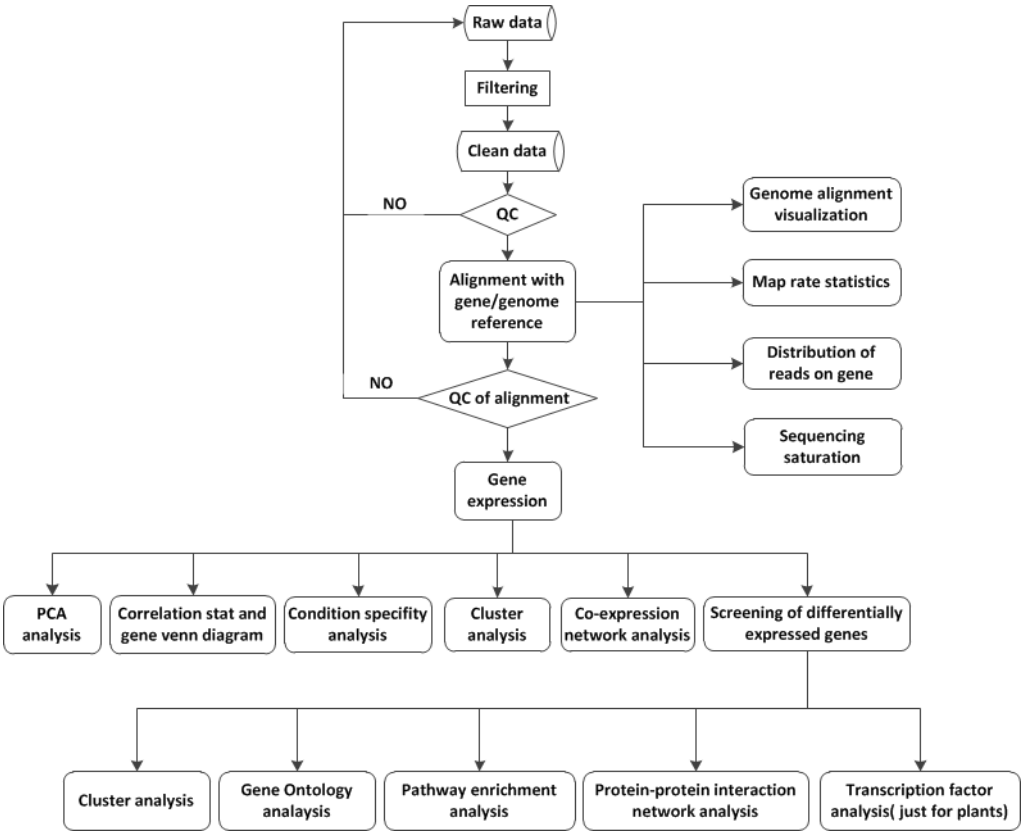

Figure 2 Bioinformatics analysis pipeline.

Primary sequencing data that produced by Illumina HiSeq^{TM}2000, called as raw reads, is subjected to quality control (QC) to determine if a resequencing step is needed. After QC, raw reads are filtered into clean reads which will be aligned to the reference sequences. QC of alignment is performed to determine if resequencing is needed. The alignment data is utilized to calculate distribution of reads on reference genes and mapping ratio. If alignment result passes QC, we will proceed with downstream analysis including gene expression and deep analysis based on gene expression (PCA/correlation/screening differentially expressed genes and so on). Further, we also can perform deep analysis based on DEGs, including Gene Ontology (GO) enrichment analysis, KEGG pathway enrichment analysis, cluster analysis, protein-protein interaction network analysis and finding transcription factor.

3 Data Filtering

We define "dirty" raw reads as reads which contain the sequence of adaptor, high content of unknown bases and low quality reads. They need to be removed before downstream analysis to decrease data noise. Filtering steps are as follows:

1) Remove reads with adaptors;

2) Remove reads in which unknown bases are more than 10%;

3) Remove low quality reads (the percentage of low quality bases is over 50% in a read, we define the low quality base to be the base whose sequencing quality is no more than 5).

After filtering, the remaining reads are called "clean reads" and stored as FASTQ format [12](see FASTQ Format in help page).

4 Reads Mapping

In general, the higher ratio of alignment, indicating that the closer the genetic relationship between sample and reference species. The lower rate may be due to low similarity with reference species or there are other pollutions.

We use Bowtie2 [4] to map clean reads to reference gene and use HISAT [3] to reference genome. Their alignment parameters change a little according different sequencing strategy (PE or SE):

```
Bowtie2 parameters for PE reads: -q --phred64 --sensitive --dpad 0 --gbar 99999999 --mp 1,1 --np 1 --score-min L,0,-0.1 -I 1 -X 1000 --no-mixed --no-discordant -p 16 -k 200
Bowtie2 parameters for SE reads: -q --phred64 --sensitive --dpad 0 --gbar 99999999 --mp 1,1 --np 1 --score-min L,0,-0.1 -p 16 -k 200
HISAT parameters for PE reads: -p 8 --phred64 --sensitive --no-discordant --no-mixed -I 1 -X 1000
HISTA parameters for SE reads: -p 8 --phred64 --sensitive -I 1 -X 1000
```

To learn about concrete meaning of Bowtie2 parameters, please refer to "Options" section in website http://computing.bio.cam.ac.uk/local/doc/bowtie2.html#. And refer to "Options" section in website http://ccb.jhu.edu/software/hisat2/manual.shtml to know about HISAT .

5 Gene Quantification

RSEM [5] is a quantification tool that computed Maximum likelihood abundance estimates using the Expectation Maximization (EM) algorithm for its statistical model, including the modeling of paired-end (PE) and variable-length reads, fragment length distributions, and quality scores, to determine which transcripts are isoforms of the same gene.

FPKM method is used in calculated expression level, the formula is shown as following formula:

$$FPKM = \frac{10^6 C}{NL/10^3}$$

Given to be the expression of gene A , C to be number of fragments that are aligned to gene A , N to be total number of fragments that are aligned to all genes, and L to be number of bases on gene A. The FPKM method is able to eliminate the influence of different gene length and sequencing discrepancy on the calculation of gene expression. Therefore, the calculated gene expression can be directly used for comparing the difference of gene expression among samples.

6 Screening DEGs using Possion Distribution Method

Referring to "The significance of digital gene expression profiles"[13], we have developed a strict algorithm to identify differentially expressed genes between two samples. Denote the number of unambiguous clean tags (which means reads in RNA_Seq) from gene A as x, given every gene's expression occupies only a small part of the library, x yields to the Poisson distribution:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$ (λ is the real transcripts of the gene)

The total clean tag number of the sample 1 is N_1, and total clean tag number of sample 2 is N_2; gene A holds x tags in sample 1 and y tags in sample 2. The probability of gene A expressed equally between two samples can be calculated with:

$$2\sum_{i=0}^{i=y} p(i \mid x)$$

$$or\ 2\times(1-\sum_{i=0}^{i=y} p(i \mid x))\ (if\ \sum_{i=0}^{i=y} p(i \mid x) > 0.5)$$

$$p(y \mid x) = (\frac{N_2}{N_1})^y \frac{(x+y)!}{x!\,y!(1+\frac{N_2}{N_1})^{(x+y+1)}}$$

We do correction on P-value corresponds to differential gene expression test using bonferroni method [14]. Since DEG analysis generate a large multiplicity problems in which thousands of hypothesis (is gene x differentially expressed between the two groups) are tested simultaneously, correction for false positive (type I errors) and false negative (type II) errors are performed using FDR method [15]. Assume that we have picked out R differentially expressed genes in which S genes really show differential expression and the other V genes are false positive. If we decide that the error ratio "Q = V / R" must stay below a cutoff (e.g. 5%), we should preset the FDR to a number no larger than 0.05. We use ' FDR ≤ 0.001 and the absolute value of Log2Ratio ≥ 1' as the default threshold to judge the significance of gene expression difference. More stringent criteria with smaller FDR and bigger fold-change value can be used to identify DEGs.

7 Screening DEGs using NOISeq

NOISeq method [16] can screen differentially expressed genes between two groups, showing a good performance when comparing it to other differential expression methods, like Fisher's Exact,Test(FET),edgeR,DESeq and baySeq. NOISeq maintains good True Positive and False Positive rates when increasing sequencing depth, while most other methods show poor performance. What's more, NOISeq models the noise distribution from the actual data, so it can better adapt to the size of the data set, and is more effective in controlling the rate of false discoveries.

First, NOISeq uses sample's gene expression in each group to calculate log2(foldchange) M and absolute different value D of all pair conditons to build noise distribution model.

$$M^i = log_2(\frac{x_1^i}{x_1^i})\ and\ D^i = |x_1^i - x_2^i|$$

Second, for gene A, NOISeq computes its avearge expression "Control_avg" in control group and average expression "Treat_avg" in treatment group. Then the foldchange (M_A=log2((Treat_avg)/(Control_avg))) and absolute different value D (D_A=|Congrol_avg-Treat_avg|) will be got. If M_A and D_A diverge from noise distribution model markedly, gene A will be defined as a DEG. There is a probability value to assess how M_A and D_A both diverge from noise distribution model:

$$P_A = P\left(M_A \geq \{M\} \;\&\&\; D_A \geq \{D\} \right)$$

Finally, we screen differentially expressed genes according to the following default criteria: Foldchange ≥2 and diverge probability ≥0.8.

8 Gene Ontology Annotation

Gene Ontology (GO), which is an international standard gene functional classification system, offers a dynamic-updated controlled vocabulary, as well as a strictly defined concept to comprehensively describe properties of genes and their products in any organism. GO has three ontologies: molecular function, cellular component and biological process. The basic unit of GO is GO-term. Every GO-term belongs to a type of ontology.

GO enrichment analysis provides all GO terms that significantly enriched in a list of DEGs, comparing to a genome background, and filter the DEGs that correspond to specific biological functions. This method firstly maps all DEGs to GO terms in the database (http://www.geneontology.org/), calculating gene numbers for every term, then uses hypergeometric test to find significantly enriched GO terms in the input list of DEGs, based on 'GO::TermFinder' (http://www.yeastgenome.org/help/analyze/go-term-finder), we have developed a strict algorithm to do the analysis, and the method used is described as follow:

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

Where N is the number of all genes with GO annotation; n is the number of DEGs in N; M is the number of all genes that are annotated to certain GO terms; m is the number of DEGs in M. The calculated p-value goes through Bonferroni Correction [14], taking corrected p-value ≤ 0.05 as a threshold. GO terms fulfilling this condition are defined as significantly enriched GO terms in DEGs. This analysis is able to recognize the main biological functions that DEGs exercise.

9 KEGG Pathway Enrichment

Pathway-based analysis helps to further understand genes biological functions. KEGG [11] (the major public pathway-related database) is used to perform pathway enrichment analysis of DEGs. This analysis identifies significantly enriched metabolic pathways or signal transduction pathways in DEGs comparing with the whole genome background. The calculating formula is the same as that in GO analysis. Here N is the number of all genes that with KEGG annotation, n is the number of DEGs in N, M is the number of all genes annotated to specific pathways, and m is the number of DEGs in M.

Help

1 FASTQ Format

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

```
@A80GVTABXX:4:1:2587:1979#ACAGTGAT/1
NTTTGATATGTGTGAGGACGTCTGCAGCGTCACCTTTATCGGCCATGGT
+
BMMTKZXUUUddddddddddddddddddddddddddddddaddddddd^WYYU
```

The first and third lines are sequences names generated by the sequence analyzer; the second line is sequence; the fourth line is sequencing quality value, in which each letter corresponds to the base in line 2; the base quality is equal to ASCII value of the character in line 4 minus 64, e.g. the ASCII value of c is 99, then its base quality value is 35. Table1 demonstrates the relationship between sequencing error rate and the sequencing quality value. Specifically, if the sequencing error rate is denoted as E and base quality value is denoted as Q, the relationship is as following formula:

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

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

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

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

Sequencing Error Rate(%)	Sequencing Quality Value	Character
1.00	20	T
0.10	30	^
0.01	40	h

More detaild information about FASTQ format can be got in website http://en.wikipedia.org/wiki/FASTQ_format.

2 BAM Format

Mapping results are stored in BAM file, which is binary equivalent of SAM file. SAM, short for Sequencing Alignment/Map, is human-readable text file with the format illustrated in Table2. And each bit in the FLAG field is defined as Table3. Samtools can do format conversion between BAM and SAM, and support more complex tasks like variant calling and alignment viewing as well as sorting, indexing, data extraction. Please refer to http://samtools.sourceforge.net/samtools.shtml#5 go get more details about samtools.

Table 2 Column description of SAM format (Download)

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Table 3 Flag description in SAM format (Download)

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

If necessary, BAM files of genome mapping result can be provided for clients in pipeline. We also recommend using IGV (Ietegrative Genomics Viewer) tool to visualize BAM file in different scales. IGV supports loading of multiple samples to do comparison in the same scale, and can view distribution of reads on the Exon, Intron, UTR, and Intergenic regions, which makes it very convenient and intuitional. Figure2 is an example and please read the brief user of IGV that we provided in project result. More information about IGV tool is available in website http://www.broadinstitute.org/software/igv/UserGuide.

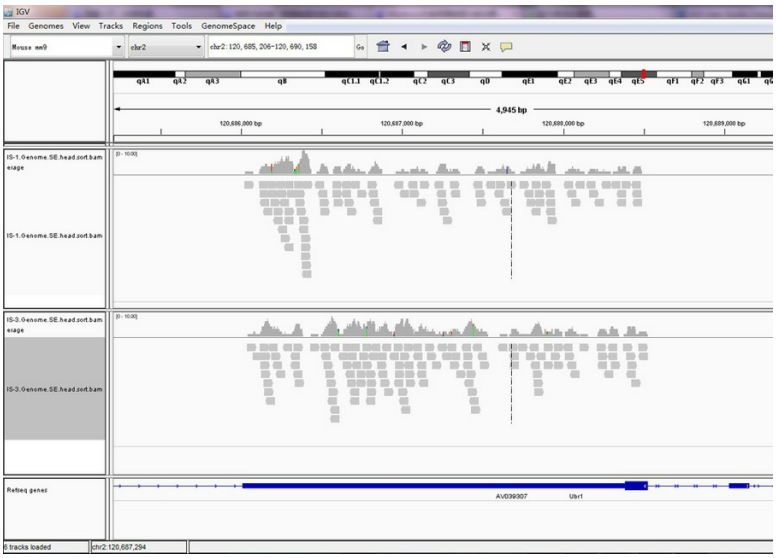

Figure 2 A screenshot of IGV interface.

This example loads two samples into IGV tool in window operation system.

3 File Format of Gene Expression Result

Gene expression result of each sample is stored in tab-seperated text file named *.*gene.* FPKM *.xls*(* presents sample name) with the format description in Table4.

Table 4 Format description of gene expression result file (Download)

Field	Description
gene_id	gene ID number
transcript_id(s)	trascript list of gene, seperated by comma
length	length of gene after model regulation
expected_count	support reads number to this gene after model regulation
FPKM	FPKM value of this gene

4 DEGs screening Format (using possionDis)

The result of differentially expressed genes which are screened by possion distribution method in each control-treatment pairwise is stored in tab-seperated text file named *.*GeneDiffExpFilter.xls*(* presents pairwise name) with the format description in Table5.

Table 5 Format description of DEGs screening result file (Download)

Field	Description
geneID	Identity of gene
geneLength	Gene length
sample1-Expression	Reads number that uniquely mapped to gene(sample sample1)
sample2-Expression	Reads number that uniquely mapped to gene(sample sample2)
sample1-FPKM	Gene expression in sample sample1
sample2-FPKM	Gene expression in sample sample2
log2 Ratio(sample2/sample1)	Log2 (folds of differentially expressed)
Up-Down-Regulation(sample2/sample1)	Gene up or down regulation (compare to sample1) in sample sample2
P-value	P-value from difference test
FDR	FDR value
Symbol	Gene symbol
Description	Brief gene description
KEGG Orthology	KEGG annotation
GO Component	GO Compoment annotation
GO Function	GO Function annotation
GO Process	GO Process annotation
Blast nr	NR annotation

5 DEGs screening Format (using Noiseq)

The result of differentially expressed genes which are screened by NOISeq method in each control-treatment pairwise is stored in tab-seperated text file named *.*GeneDiffExp.xls* (* presents pairwise name) with the format description in Table6.

Table 6 Format description of DEGs screening result file (Download)

Field	Description
GeneID	Identity of gene
geneLength	Gene length
Means-groupA	mean expression (FPKM) of groupA
Means-groupB	mean expression (FPKM) of groupB
log2Ratio(s2/s1)	Log2(folds of mean expression in two groups)
Up-Down-Regulation(groupB/groupA)	Gene up or down regulation (compared to groupA) in sample groupB
Probability	probability of difference
Symbol	Gene symbol
Description	Brief gene description
KEGG Orthology	KEGG annotation
GO Component	GO Compoment annotation
GO Function	GO Function annotation
GO Process	GO Process annotation
Blast nr	NR annotation

6 How to Read Report of Clustering Analysis

Since we use the same tool (cluster [7] [8] and javaTreeView [9]) to do gene expression clustering and DEGs foldchange clustering, they share nearly the same report format. So we just do illustration of this report based on DEGs clustering result. Make sure that your computer has installed Java and use IE brower to open *cluster_en.html*. The interface is as the Figure3.

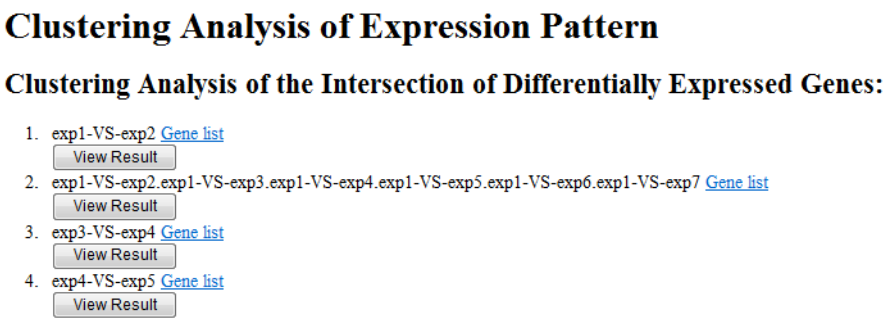

Figure 3 The webpage interface of clustering analysis report.

Each cluster plan which is consisted of more than one pairwise, has two types clustering results: intersection and union. Click "Gene list" to see what genes are used to do clustering and their foldchange values in each pairwise. Click the button "View Result", the JavaTreeView will work and interactive clustering tree interface will appear as Figure4.

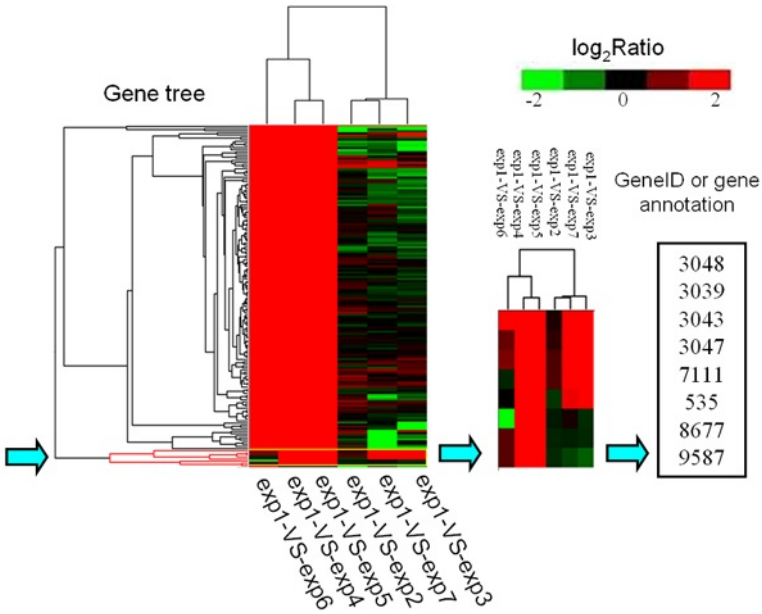

Figure 4 Clustering image of gene foldchange levels.

Each column represents an experimental condition (e.g. exp1-VS-exp2), each row represents a gene. Log2(Foldchange) values are shown in different colors. Red means up regulation and green means down regulation. When the line pointed by the left arrow is clicked, the color of the branches derived from the clicked line changes to red. And the corresponding genes or gene annotations are shown on the right. The middle is just an amplification of the hierarchical clustering of the chosen genes.

Please refer to website: http://jtreeview.sourceforge.net/manual.html to get more operating instructions.

7 How to Read Report of GO Annotation

Make sure that the computer has installed java and use IE brower to open GOView.html.

The left navigation includes three types of GO terms for each control-treatment pairwise (C: cellular component, P: biological process, F: molecular function). Click one of them, the enriched GO terms result will be listed as Figure5.

Gene Ontology term	Cluster frequency	Genome frequency of use	Corrected P-value	Expressi Profile
BLOC complex (view genes)	2 out of 82 genes, 2.4%	8 out of 16090 genes, 0.0%	0.03943	View Res
cytosol (view genes)	2 out of 82 genes, 2.4%	15 out of 16090 genes, 0.1%	0.14450	View Res
cytosolic part (view genes)	2 out of 82 genes, 2.4%	15 out of 16090 genes, 0.1%	0.14450	View Res
intracellular part (view genes)	67 out of 82 genes, 81.7%	11513 out of 16090 genes, 71.6%	1	View Res

Figure 5 Significantly enriched GO terms in DEGs.

Column 1 is GO term name. Column 2 is the ratio of DEGs enriched to this GO term. Column 3 is the ratio of genes enriched to this GO term in background database. Column 4 is Corrected P-value which indicates the degree of enrichment and the smaller Corrected P-value, the more significantly DEGs enriched to this GO term. The result list has been sorted by Corrected P-value. Column 5 is clustering of foldchange value for these enriched DEGs using the tools cluster [7] [8] and javaTreeView [9] (see the section How to Read Report of Clustering Analysis in help page).

Click the term name 'BLOC complex' in Figure5, you can go to http://amigo.geneontology.org/amigo for more information when the computer is Internet-connected. Click 'view genes' in Figure5, you can get gene IDs that enriched to this GO term as Figure6.

BLOC complex	63915, 100526837
cytosol	63915, 100526837

Figure 6 Gene ID list related to GO terms.

There are two DEGs enriched to the term 'BLOC complex': 63915, 100526837.

8 How to Read Report of Pathway Enrichment

Open html report for pathway enrichment result and the enriched KEGG pathways will be listed as Figure7.

1. sample3-VS-sample4

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

Figure 7 Pathway enrichment analysis of DEGs.

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

Click pathway name 'Leukocyte transendothelial migration' in Figure7, you can get gene IDs that enriched to it as Figure8.

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

Figure 8 Gene ID list related to pathway.

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

Furtherly, detecting the most significant pathways, the enrichment analysis of DEG pathway significance, allows us to see detailed pathway information in KEGG database. For example, clicking the hyperlink on 'Leukocyte transendothelial migration' in Figure8 will get detailed information as shown in Figure9.

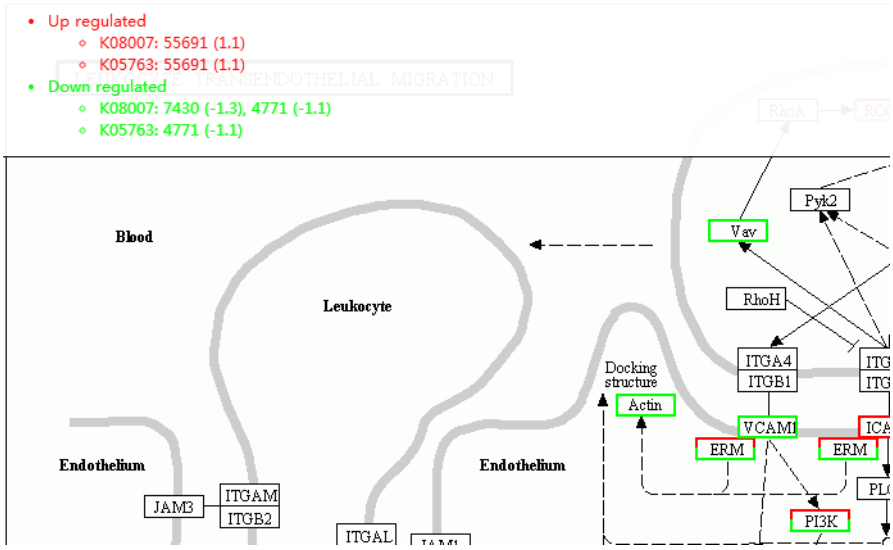

Figure 9 An example of KEGG pathway of 'Leukocyte transendothelial migration'.

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

FAQs

Does it need biology repeat? If so, how many times are needed?

Yes, it needs at least two biological repeats, more than 3 times is much better. Article of Hansen in July 2011 showed that biological difference is gene expression characteristic, no related to the detection technology, data dispose either. High-effected magazines may refuse the draft if we don't set biological repeats.

Must genome reference be provided using RNA-Seq method?

No, but reference sequence is needed. Unigene, mRNA and CDS can be treat as reference sequence.

What is the relationship between genome size and the recommended amount of sequencing data using RNA-Seq method?

The recommended sequencing data is mainly related to gene number. Though different species diverse in genome size, there is little difference in coding gene number of general species (about 30,000). So we generally recommend 10M clean reads for HiSeq and Ion Proton platform, 20M clean reads for CG platform.

When preparing library, why you use RNA fragmentation instead of cDNA fragmentation?

Please refer to the reference 'RNA-Seq: a revolutionary tool for Transcriptomics'.

What information can we get from the *.md5sum files?

In the Linux or Unix environment, md5sum is a program used to calculate and check the result files. *.md5sum files are generated by the computer program md5sum which is commonly used to verify the integrity of files.

How to understand the figure in randomness analysis? What's the criterion for randomness?

Randomness is one of criterions for sequencing quality. At present, there is no criterion to evaluate the randomness. Generally speaking, if the randomness is good, the reads would be evenly distributed on reference sequence.

In the Gene Expression Difference Analysis, for example 1vs2, how to understand the up-regulated and down-regulated?

1vs2 means sample 1 is control and sample 2 is case. In the corresponding files 1-VS-2.GeneDiffExp.xls and 1-VS-2.GeneDiffExpFilter.xls, if a gene is up, it means the expression of this gene is up-regulated in sample 2 compared to sample 1.

In the figure of pathway enrichment analysis, why the number of gene is not equal to colored borders?

Because each border in figure represents one kind of enzyme, and the function of an enzyme is participation of several genes, one border maybe related to many genes.

Why is gene mapping rate always lower than genome mapping rate?

The reason may be following: 1)Gene database that used in pipeline is not completed; 2)There are new transcripts in sequencing data; 3)Sequencing reads comes from noncoding regions, resulting in the situation that they can't map to gene as well as genome.

References

[1] Wang Z., et al. (2009). RNA-Seq: a revolutionary tool for Transcriptomics. Nature Reviews Genetics, 10(1): 57-63.

[2] Mortazavi, A., et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, 5(7):621-8.

[3] Kim D, Langmead B and Salzberg SL.(2015). HISAT: a fast spliced aligner with low memory requirements. Nature Methods,12(4):357-60. doi: 10.1038/nmeth.3317

[4] Langmead B, et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3): 25-34.

[5] Li B and Dewey C N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics, 12(1): 323.

[6] Hansen et al. (2011). Sequencing technology does not eliminate biological variability. Nat Biotechnol, 29(7):572-3.

[7] Eisen, M. B., et al. (2001). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, (1998)95(25): 14863-8. 2001.29: 1165-1188.

[8] M. J. L. de Hoon, et al. (2004). Open Source Clustering Software. Bioinformatics, 20(9): 1453-1454.

[9] Saldanha, A. J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics, 20(17): 3246-8.

[10] Ye, J., et al. (2006). WEGO: a web tool for plotting GO annotations. Nucleic Acids Res, 34(Web Server issue): W293-7.

[11] Kanehisa, M., et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res, 36 (Database issue): D480-4.

[12] Cock P., et al.(2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38(6): 1767-1771.

[13] Audic, S. and J. M. Claverie. (1997). The significance of digital gene expression profiles.

Genome Res, 10: 986-95.

[14] Abdi, H. (2007). " The bonferroni and Sidak corrections for multiple comparisons. " In N.J. Salkind (ed.). Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage.

[15] Benjamini, Y. and D. Yekutieli. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29: 1165-1188.

[16] Sonia Tarazona, et al. (2011). Differential expression in RNA-seq: A matter of depth. Genome Research, 21:2213–2223.

2015 Copyright BGI All Rights Reserved 粤ICP备 12059600
Technical Support E-mail:info@bgitechsolutions.com
Website: www.bgitechsolutions.com

华大基因
BGI

基因科技造福人类

联系我们

服务热线：400-706-6615
网址：www.bgitechsolutions.com
邮　箱：info@bgitechsolutions.com
地址：广东省深圳市盐田区北山工业区11栋（邮编：518083）

本结题报告仅供客户学习、交流和研究使用，请勿用于商业用途，违者必究。
版权声明：本结题报告版权属于深圳华大基因股份有限公司所有，未经本公司书面许可，任何其他个人或组织均不得以任何形式将本结题报告中的各项内容进行复制、拷贝、编辑或翻译为其他语言。本结题报告中的所有商标或标志均属于深圳华大基因股份有限公司及其提供者所有。

2017年01月