生工® Sangon Biotech

生工生物

Denovo 转录组

项目报告

合同编号	
客户单位	
报告时间	

适用范围

本项目分析报告适用于无参考基因组转录组项目，不同样本数分析内容会略有差别。

目录

1. 名词解释

Bp:base-pair，碱基对，读长的单位，每一个 bp 指一对互补的碱基。

Read：序列，测序数据中每一条序列就是一个 read。

Raw_reads: 原始数据。

Clean_reads：QC 之后的数据。

Fastq: 序列数据存储的标准格式之一，每 4 行为一条 read 的信息。包含测序 read 名，序列，正反链标示，序列质量值。

Pair-end 测序：双端测序，两端均测序，随后合并成一条 read。

Single-end 测序：单端测序，只测一端，即为一条 read。

质量评分：指的是一个碱基的错误概率的对数值，即质量评分越高，错误概率越小。

QC：Quality control，即质量控制。

滑窗法：检测一个窗口内的碱基质量值，如果满足条件则向前移动一个单位继续检测，如果不满足条件即做删除处理，随后继续移动到下一个单位进行检测，直到检测完所有的数据。

测序接头：序列在上机测序的时候需要在两端各加上一段人工序列，当序列片段比实际测序读长短时，3’端会测到接头序列，该段序列在分析之前需要去除掉。

N：表示未知碱基，在测序的时候，当某个碱基无法确定为某个碱基时，该位判定为 N，某条序列中 N 越多说明该序列质量越低，一般该种序列需要剔除掉。

Isoform：单条转录本，同 transcript，每条 isoform 可以编码一种蛋白。

Unigene：同基因，对拼接的 isoform 进行聚类，序列类似的 isoform 聚类一类，该类称为 Unigene 基因，一条 Unigene 可编码几条 Isoform。

N50：将 transcript 从长到短排序，依次累加 transcript 碱基数，当累计碱基数达到 transcript 总碱基数的 50%时的 transcript 的长度。

N90：将 transcript 从长到短排序，依次累加 transcript 碱基数，当累计碱基数达到 transcript 总碱基数的 90%时的 transcript 的长度。

可变剪切：可变剪切（或选择性剪切）是一个过程，即主要基因或者 mRNA 前体转录所产生的 RNA 的外显子以多种方式通过 RNA 剪切进行重连，由此产生的不同的 mRNA 可能被翻译成不同的蛋白质构体，因此，一个基因可能编码多种蛋白质。

Novel 转录本：新的转录本，相较于与已知转录本而言。

SSR：短片段重复序列，该类序列在物种的种群中有很高的多样性，该类序列可用作分子标记。

NR 数据库：NR (NCBI non-redundant protein sequences) 是 NCBI 官方的蛋白序列数据库，它包括了 GenBank 基因的蛋白编码序列，PDB(Protein DataBank)蛋白数据库、SwissProt 蛋白序列及来自 PIR（Protein Information Resource）和 PRF（Protein Research Foundation）等数据库的蛋白序列。

NT 数据库：NT(NCBI nucleotide sequences) 是 NCBI 官方的核酸序列数据库，包括了 GenBank，EMBL 和 DDBJ（但不包括 EST,STS,GSS,WGS,TSA,PAT,HTG 序列）的核酸序列。

PFAM 数据库：Pfam (Protein family)是最全面的蛋白结构域注释的分类系统。蛋白质是由一个个结构域组成的，而每个特定结构域的蛋白序列具有一定保守性。

KOG/COG: COG 是 Clusters of Orthologous Groups of proteins 的简称，KOG 为 euKaryotic Ortholog Groups。这两个注释系统都是 NCBI 的基于基因直系同源关系，其中 COG 针对原核生物，KOG 针对真核生物。

Swiss-Prot：（A manually annotated and reviewed protein sequence database）搜集了经过有经验的生物学家整理及研究的蛋白序列。详见 http://www.ebi.ac.uk/uniprot/。

KEGG：KEGG 是 Kyoto Encyclopedia of Genes and Genomes 的简称，是系统分析基因产物和化合物在细胞中的代谢途径以及这些基因产物的功能的数据库。它整合了基因组、化学分子和生化系统等方面的数据,包括代谢通路（KEGG PATHWAY）、药物（KEGG DRUG）、疾病（KEGG DISEASE）、功能模型（KEGG MODULE）、基因序列（KEGG GENES）及基因组（KEGG GENOME）等等。详见 http://www.genome.jp/kegg/。

GO：(Gene Ontology)是一套国际标准化的基因功能描述的分类系统。GO 分为三大类 ontology：生物过程（Biological Process）、分子功能（Molecular Function）和细胞组分(Cellular Component)，分别用来描述基因编码的产物所参与的生物过程、所具有的分子功能及所处的细胞环境。GO 的基本单元是 term，每个 term 有一个唯一的标示符（由“GO:”加上 7 个数字组成，例如 GO:0072669）；每类 ontology 的 term 通过它们之间的联系（is_a, part_of, regulate）构成一个有向无环的拓扑结构。详见 http://www.geneontology.org/。

CDS：编码区，指的是转录本中真正编码蛋白质的区域，一般首为起始密码子，终为终止密码子。

Mapping：序列比对，将测序的短序列与参考序列比较，找出短序列在参考序列中的准确位置。

均一化分析：均一化分析是用于评估转录组测序建库时对 mRNA 的打断是否随机，若不随机则可能对后续的分析会产生较大偏好性。

测序饱和度曲线：测序饱和度曲线用于反映基因表达水平定量对数据量的要求。表达量越高的基因，就越容易被准确定量；反之，表达量低的基因，需要较大的测序数据量才能被准确定量。当曲线达到饱和，说明测序数据量已满足定量要求。

FPKM：FPKM（Fragment Per Kilo bases per Million mapped Reads）是每百万 reads 中来自某一基因每千碱基长度的 reads 数目，FPKM 同时考虑了测序深度和基因长度对 reads 计数的影响，FPKM 用于评估基因的表达量。

样品间相关性分析：衡量样本间相关性，相关系数越接近 1，表明样品之间表达模式的相似度越高。若样品中有生物学重复，通常生物重复间相关系数要求较高。

热图：通过颜色深浅来可视化数据大小，每一个颜色块表示一个数值，一般颜色越深说明数值越大。

密度曲线：用来衡量数据的分布，数据在某个区域越集中，则该区域的面积越大。

PCA 分析：PCA 分析（Principal Component Analysis）是一种研究数据相似性和差异性的可视化方法。进过一系列的计算之后，选择主要的，排在前几位的特征值，对样本之间的关系进行描述。

韦恩图：又叫文氏图，用于反映不同数据集合的共性及特异性。

SNP/Indel：SNP 为单碱基核酸突变，Indel 表示插入和缺失。

Pvalue：统计学检验的 P 值，P 值越小说明样本间差异越大

FDR：多重假设检验校正后的 P 值，在做多次检验的时候为控制假阳性率需对 P 值再做校正，一般 P 值越小，FDR 值也越小。

Foldchange：表达量差异倍数，一般差异倍数越大，说明表达差异越大。

火山图：火山图（Volcano Plot）在一张图中显示了两个重要的指标（Fold change/p-Value），可以非常直观且合理地筛选出在两样本间发生差异表达的基因。

MA 图：横坐标 X 轴表示 log 均值，即（log2(A)+log2(B)）/2, 纵坐标为代表 log（Foldchange），即 log2(B/ A), 据此图可看出差异基因分布在高表达基因或者低表达基因。

表达模式聚类：对所有的差异基因进行聚类分析，该分析可以将表达模式相近的基因聚到一起，筛选出特定表达模式的基因类。

功能富集分析：对差异基因做检验，看差异基因在不同功能类下的分布，通过此分析可推断差异基因主要的功能及生物学意义。

共表达网络：基因共表达网络分析（Gene Co-expression Network Analysis）是根据基因表达信号值的动态变化，计算基因间的共表达关系，来建立基因转录调控模型，得到基因间的表达调控关系及调控方向，从而寻找一个或多个物种在不同发育阶段，或者不同组织在不同条件或处理下的全部基因表达调控网络模型以及关键基因，从而系统的研究生物体复杂的生命现象。

蛋白互作网络：蛋白间存在相互作用，对差异基因构建蛋白互作网络，可筛选出候选的关键差异基因。

2. 相关软件及数据库

2.1 软件

FastQC：http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/，版本 0.11.5。

cutadapt：https://pypi.python.org/pypi/cutadapt/1.2.1，版本 1.2.1。

Prinseq：http://prinseq.sourceforge.net/，版本 0.19.5。

blast+ ：

http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download，

版本 2.28。

Trinity：http://trinityrnaseq.github.io/，版本 r20140717。

bowtie2：http://bowtie-bio.sourceforge.net/bowtie2/，版本 2.2.3。

samtools：http://samtools.sourceforge.net/，版本 0.1.18。

bwa：http://bio-bwa.sourceforge.net/，版本 0.7.5a。

jellyfish：http://www.cbcb.umd.edu/software/jellyfish/，版本 2.0。

MISA：http://pgrc.ipk-gatersleben.de/misa/，版本 1.0。

primer3：http://primer3.sourceforge.net/，版本 2.3.6。

Tophat:（http://ccb.jhu.edu/software/tophat/），版本 2.0.11。

KAAS：http://www.genome.jp/tools/kaas/，版本 1.0。

OrrPredictor：http://www.proteomics.ysu.edu/tools/OrfPredictor.html/，版本 1.0。

RSeQC：（http://rseqc.sourceforge.net/），版本 2.6.1。

R：https://www.r-project.org 版本 3.1.2。

RSEM：http://deweylab.biostat.wisc.edu/rsem/，版本 1.0。

cufflinks：http://cole-trapnell-lab.github.io/cufflinks/，版本 2.2.1。

MATS：http://rnaseq-mats.sourceforge.net/ 版本 1.0

Bioconductor：http://www.bioconductor.org/。

R 包：qvalue，pheatmap，scatterplot3d，gplots，topGO，RColorBrewer，VennDiagram，DESeq，edgeR，Rgraphviz ，Statistics.R perl，pathview，WGCNA。

Cytoscape：http://www.cytoscape.org/ 版本 3.3.0

2.2 数据库

NR：ftp://ftp.ncbi.nlm.nih.gov/blast/db/nr.*tar.gz，NR(NCBI non-redundant protein sequences) 是 NCBI 官方的蛋白序列数据库

NT：ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt.*tar.gz，NT(NCBI nucleotide sequences) 是 NCBI 官方的核酸序列数据库

KOG/COG：ftp://ftp.ncbi.nlm.nih.gov/pub/COG/ COG 是 Clusters of Orthologous Groups of proteins 的简称，KOG 为 euKaryotic Ortholog Groups。这两个注释系统都是 NCBI 的基于基因直系同源关系，其中 COG 针对原核生物，KOG 针对真核生物。

Swiss-Prot：http://www.uniprot.org/downloads （A manually annotated and reviewed protein sequence database）搜集了经过有经验的生物学家整理及研究的蛋白序列。

TrEMBL：http://www.uniprot.org/downloads Uniprtot 中的另外一个蛋白数据库，该数据库收集了大量的蛋白序列。

CDD、Pfam 数据库：ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/cdd.tar.gz。

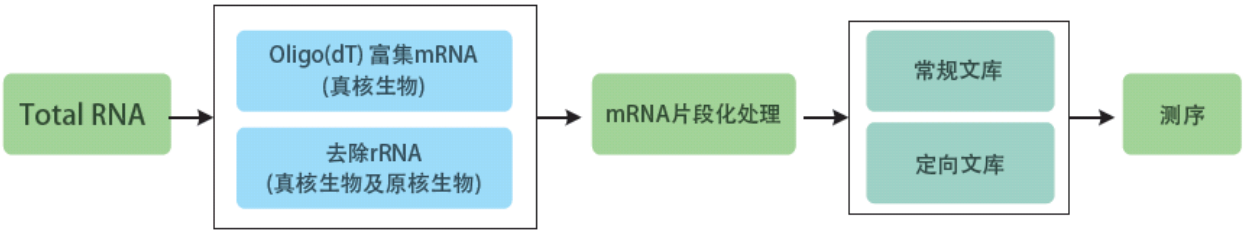
KEGG：http://www.kegg.jp/ KEGG 是 Kyoto Encyclopedia of Genes and Genomes 的简称，是系统分析基因产物和化合物在细胞中的代谢途径以及这些基因产物的功能的数据库。

Ensembl：欧洲基因组数据库，与 NCBI，UCSC 并为三大基因组数据库，其中可下载大部分物种的基因组序列及相关注释文件；动物基因组:http://asia.ensembl.org/index.html，其它物种基因组：http://ensemblgenomes.org/。

Biomartview：http://www.ensembl.org/biomart/martview/f0c1eaeff9cf930ca8180723e05ede99，可用于导出物种相关数据库信息。

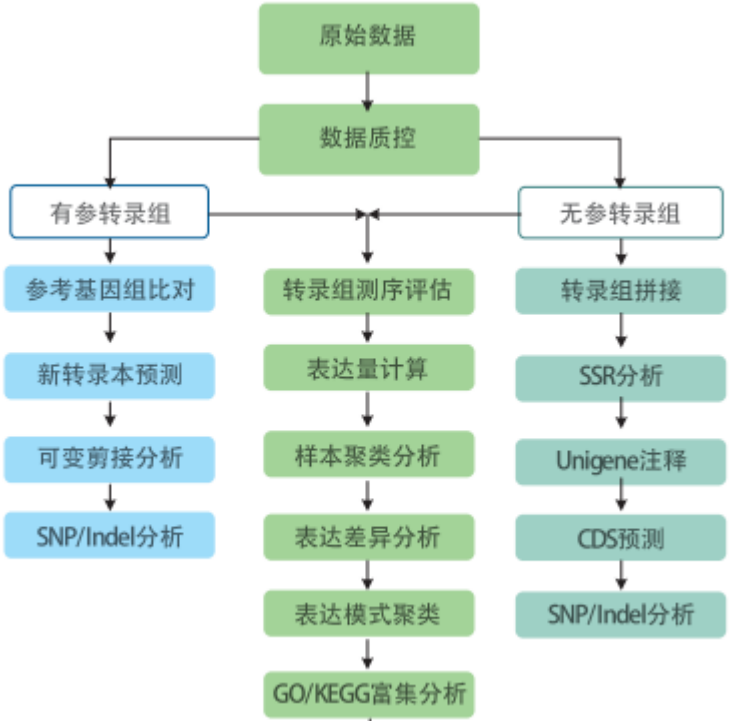
STRING：http://string-db.org/，蛋白互作数据库。

3. 实验流程



4. 分析流程

4.1 分析流程图



4.2 详细分析内容列举

分析项目	详细分析内容	说明
1. 数据质控	1）原始数据 Q 值分布图 2）原始数据 GC 含量分布图 3）去除测序接头及低质量碱基 4）原始数据及 QC 后数据统计 5）污染检测统计	
2. 转录组拼接	1）转录本及 Unigene 长度统计，N50/N90 2）GC 含量分布图 3）转录本及 Unigene 长度累积曲线 4）Unigene 可变剪切统计	仅无参转录组做
3. SSR 分析	1）SSR 查找结果统计 2）SSR 类型统计 3）SSR 引物设计 4）SSR 密度分布图	仅无参转录组做

4. Unigene 注释	1）NT/NR/PFam/Swissprot/TrEMBL/COG/KOG 数据库注释 2）GO/KEGG 数据库注释 3）注释物种分布饼图 4）GO/KEGG/COG/KOG 功能分布条形图	
5. CDS 预测	1）CDS 长度统计 2）CDS 编码蛋白序列 3）CDS 长度占比统计	
6. SNP/Indel 分析	1）各样本 SNP 汇总列表 2）SNP 密度分布图图 3）SNP 突变图谱	两个以上样本才做
7. 转录组测序评估	1）Mapping 结果统计 2）均一化分析 3）测序饱和度分析 4）基因 coverage 分析 5）建库长度评估	
8. 表达量计算	1）各样本表达量盒状图 2）各样本表达量密度曲线 3）表达量区间分布图 4）样本间共同表达韦恩图	共同表达韦恩图需要样本数大于 1 小于 6
9. 样本聚类分析	1）样本距离热图 2）样本聚类树图 3）PCA 分析 3D/2D 图 4）样本间距离矩阵，样本间相关性分析	样本数需大于 2 个才做 PCA 分析
10. 表达差异分析	1）差异基因数目分布图 2）样本间表达量盒状图 3）样本间表达量密度曲线 4）样本间表达量散点图 5）样本表达量 MA 图 6）火山图 7）差异基因韦恩图	差异基因韦恩图需比较对数目大于 1 小于 6
11. 表达模式聚类	1）差异基因表达量热图 2）样本间距离热图 3）各表达模式基因集表达量折线图 4）共有差异基因统计 5）foldchange 热图	共有差异基因统计及 foldchange 热图 需要比较对数目大于 1
12. GO/KEGG 富集	1）富集分析统计表	Pvalue 热图需比较对数目大

分析	2）各 Term 差异基因数目条形图 3）GO 有向无环图 4）富集到的 term 数目统计表 5）pathway 上色图 6）富集的 term Pvalue 热图 7）富集散点图	于 1

4.3 分析步骤及方法简介

1. 测序质量评估及质控

1）采用 FastQC 对测序原始序列做质量评估

2）去除 3’端测序接头

3）去除融合后的 reads 尾部质量值在 20 以下的碱基

4）切除 reads 中含 N 部分序列：长度阈值 35bp

5）对序列进行污染评估，看其是否有污染

2. 转录本拼接

1）将各样本过滤之后序列进行合并，之后进行 de novo 拼接，使用软件 Trinity，使用 paired-end 的拼接方法。对拼接序列去重复，取长度大于 200bp 的序列，每个 Loci（c*_g*_）下最长的转录本作为 Unigene

2）统计转录本及 Unigene 长度，GC 含量等

3. SSR 分析

1）采用 MISA 对 Unigene 及 Transcript 进行 SSR 检测，对不同 SSR 类型在基因与转录本的密度分布进行统计

2）采用 primer3 对 SSR 进行引物设计

4. 基因功能注释

1）将 Unigene 基因序列分别与 NR、NT、KOG、CDD、PFAM、Swissprot、TrEMBL、GO、KEGG 库进行比对，取相似度>30%，且 e<1e-5 的注释结果

2）采用 KAAS 做 KEGG 数据库注释

3）统计各数据库注释结果

5. RNA-seq 测序评估

1）将 QC 后序列比对到拼接后或者参考转录本中，比对采用 bowtie2

2）采用 RSeQC 做 RNASeq 测序评估，评估内容包括如下：

i. Mapping 统计

ii. 均一化分析

iii. 建库长度评估

iv. 测序饱和度评估

v. 基因 coverage 评估

6. 表达量统计及样本间聚类分析

1）计算各样本各基因的表达量，无参采用 RSEM，有参采用 cufflinks

2）对样本做样本聚类分析，聚类采用皮尔森相关系数值，并做 PCA 分析，上述内容均采用 R 来分析

7. SNP/Indel 分析

1）基于比对结果对各样本做 SNP/INDEL calling，采用软件为 samtools，

2）对得到 SNP 及 INDEL 结果进行过滤，过滤条件为：

a) QUAL 值大于 20

b) 覆盖度大于 2

8. 差异表达分析

1）无生物学重复样本分析方法如下：

参照 Audic S.等人发表在 Genome Research 上的基于测序的差异基因检测方法(Audic, 1997)对差异检验的 p value 作多重假设检验校正，采用的方法为 FDR，在分析中，差异表达基因定义为 p ≤0.01 且倍数差异在 2 倍以上的基因

2）有生物学重复样本筛选方法如下：

采用 DEseq 进行差异分析,筛选阈值为 qvalue<0.001 且|FoldChange|>2

差异分析软件：DESeq、edgeR

9. 差异基因表达模式聚类分析

1）合并所有比较对间的差异表达基因

2）对该基因集做聚类分析，获得表达模式相近的基因集

3）筛选出表达模式相近的基因

10. GO 富集分析

1） GO 富集分析方法为 GOseq（Young et al, 2010），此方法基于 Wallenius non-central hyper-geometric distribution

2） 筛除出显著富集的 GO

3） 绘制 topGO 有向无环图

11. KEGG 富集分析

1）Pathway 显著性富集分析以 KEGG Pathway 为单位，应用超几何检验，找出与整个基因组背景相比， 在差异表达基因中显著性富集的 Pathway。该分析的计算公式如下：

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

2）筛选显著富集的 pathway

3）对富集的 pathway 进行上色

5. 分析结果展示

5.1 测序质量评估及质控

5.1.1 测序质量评估

本次测序采用 Hiseq PE150 模式（双端测序 PE：paired-end），每一个样本分别有 R1.fastq 和 R2.fastq 两个文件，分别代表 5’ -> 3’和 3’->5’的测序结果。R1.fastq 与 R2.fastq 中的文件行数是一致的，且根据 reads name 一一对应。

FASTQ: Fastq 是 Solexa 测序技术中一种反映测序序列的碱基质量的文件格式。每条 read 包含 4 行信息。第一行以“@”开头，随后是序列标示和相关的描述信息，第三行以“+” 开头，随后是序列描述信息或者什么都不加；第二行为碱基序列，第四行是质量信息，与第二行中的碱基序列一一对应，根据评分体系不同每个字符的含义所表示的数字有所差别。例如：

@SEQ_ID

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

质量评分：质量评分指的是一个碱基的错误概率的对数值。其最初在 Phred 拼接软件中定义与使用，其后在许多软件中得到使用。其质量得分与错误概率的对应关系见下表：

Phred quality scores are logarithmically linked to error probabilities		
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %

对于每个碱基的质量编码标示，不同的软件采用不同的方案，本项目中使用的方案是，Phred quality score，值的范围从 0 到 62 对应的 ASCII 码从 64 到 126，得分在 0 到 40 之间。

软件：FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/（用于统计序列原始信息及绘图）

结果目录：1_QC/

All_sample_raw_data_infor.xls：所有样本原始数据统计，结果如下：

Table 5.1 原始数据统计

	T1	T2	T3	C1
Total Reads Count(#):	18378142	16325862	15371472	20300126
Total Bases Count(bp):	2297267750	2040732750	1921434000	2537515750
Average Read Length(bp):	125	125	125	125
Q30 Bases Count(bp):	2031505919	1829033708	1707271182	2196988467
Q30 Bases Ratio(%):	88.52638275	89.67873518	88.89941823	86.65272547
Q20 Bases Count(bp):	2167326744	1938468838	1817158772	2368715685
Q20 Bases Ratio(%):	94.36476832	94.9988858	94.58239784	93.35877918
Q10 Bases Count(bp):	2296581658	2040100471	1920851953	2536763846
Q10 Bases Ratio(%):	99.97000127	99.97000131	99.9700013	99.97000131
N Bases Count(bp):	686092	632279	582047	751904
N Bases Ratio(%):	0.000299	0.00031	0.000303	0.000296
GC Bases Count(bp):	1100978831	860463579	912167966	1309330083
GC Bases Ratio(%):	47.92557729	42.16444211	47.47329161	51.59889482

注：若样本数目较多，此处只会截取部分样本数据，完整数据请见结果文件夹中的对应文件。

Total Reads Count：样本所有 reads 数目，为 reads1 与 reads2 数目之和

Total Base Count：所有碱基数目，即数据量

Average Read Length：平均序列长度

Q30 Base Count：碱基质量在 30 以上的数目

Q30 Base Ratio：Q30 碱基比例

N Base Count：N 碱基的数目

N Base Ratio：N 碱基比例

GC Base Count：GC 碱基数目

GC Base Ratio：GC 含量

各样本碱基质量图如下：

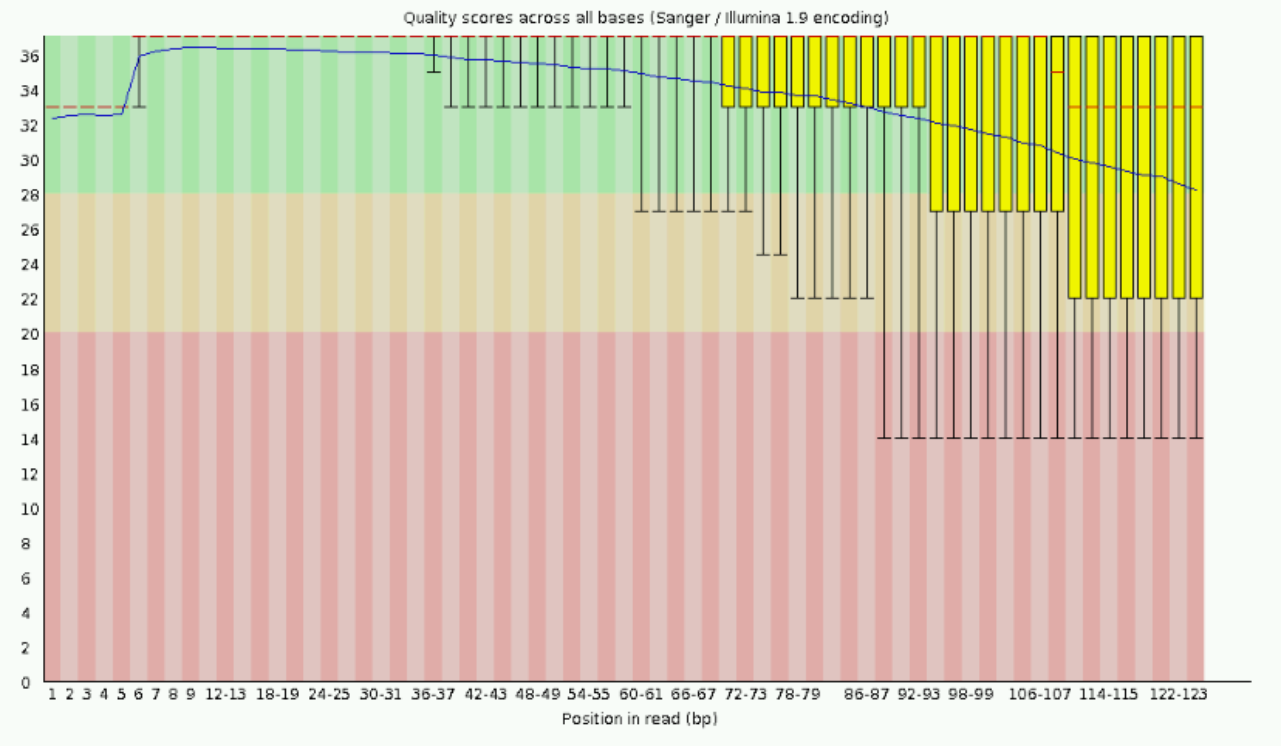


图 5.1 各位置碱基质量分布图

注：若样本数据较多，此处只展示某个样本的 Read1 质量分布，其它样本数据见 1_QC/Sample/*fastqc.zip 文件。

说明：横坐标表示测序位置，纵坐标为测序质量值。图中横轴代表位置，纵轴 quality。红色表示中位数，黄色是 25%-75%区间，触须是 10%-90%区间，蓝线是平均数。Hiseq 测序是双端测序，每条 read 长度 150bp。随着测序的进行，酶的活性会逐步下降，因此到达一定测序长度后，碱基质量值也会随之下降。从图 5.1 可知，中位值均在 Q20 以上，因此该文库碱基质量良好，可用于后续分析。本分析会对所有数据进行质控，后续只取 Q20 以上的数据进行分析。

各样本碱基 GC 含量分布图如下：

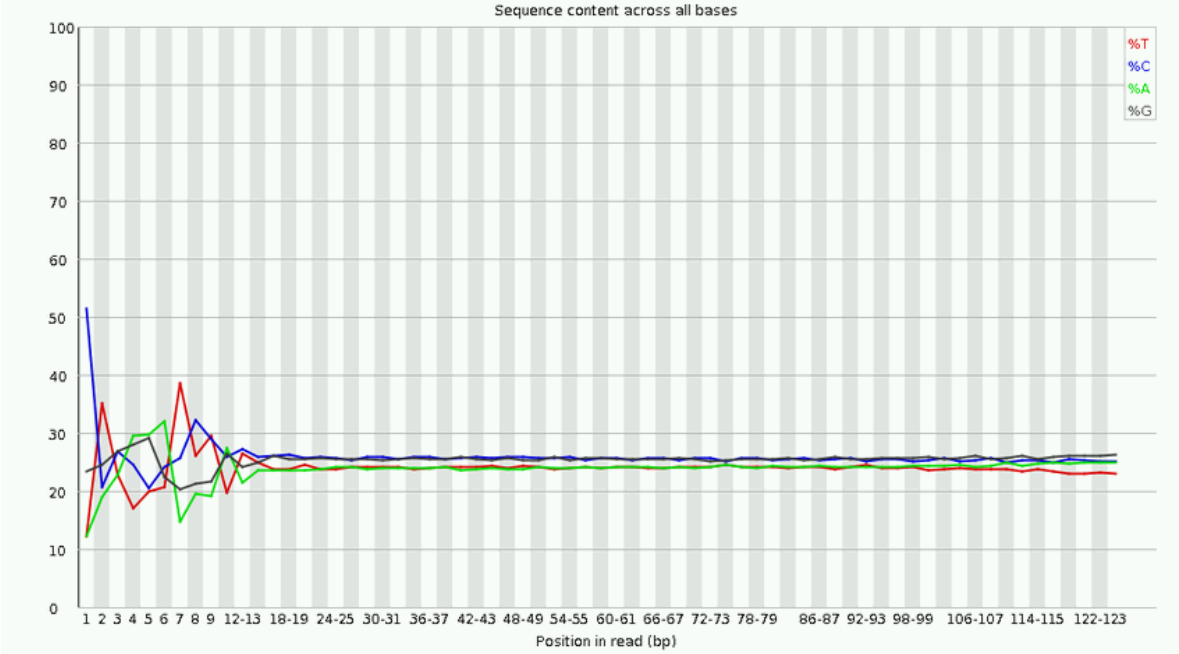


图 5.2 各位置碱基 GC 含量分布图

注：若样本数据较多，此处只展示某个样本的 Read1 质量分布，其它样本数据见 1_QC/Sample/*fastqc.zip 文件。

说明：横坐标是 reads 碱基坐标，纵坐标是所有 reads 的 A、C、G、T 碱基分别占的百分比。在文库较均匀随机的情况下，四种颜色的分界线应该波动极小，呈一条直线，但一般测序前几个碱基由于测序尚不大稳定，前几个碱基 ACGC 含量会有波动。

5.1.2 数据质控

对于 Hiseq 双端测序原始序列 3’端可能带有 adaptor 接头序列，以及一些少量低质量序列和杂质序列，为了提高后续分析质量和可靠性，对原始序列进行去接头、质量剪切、污染评估等处理。

数据质控步骤：

1）去除 3’端测序接头，采用的软件为 cutadapt，Read1 3’端测序接头为 AGATCGGAAGAGCACACGTCTGAAC，Read2 3’端测序接头为 AGATCGGAAGAGCGTCGTGTAGGGA。

2）去除融合后的 reads 尾部质量值在 20 以下的碱基。设置 10bp 的端口，如果窗口内的平均质量值低于 20，从窗口开始去除后端的碱基

3）切除 reads 中含 N 部分序列：长度阈值 35bp

4）对序列进行污染评估，看其是否有污染，方法为：随机从 QC 之后序列中抽取 100000 条序列进行 blast 比对，比对数据库为 NCBI NT 数据库，取 evalue <=1e-10 并且相似度>90%,coverage >80%的比对结果，计算其物种分布。

去除测序接头软件：cutadapt（ https://pypi.python.org/pypi/cutadapt/1.2.1）

主要参数设置：-O 10 -min_len 35 -a AGATCGGAAGAGCACACGTCTGAAC

质量控制使用软件：Prinseq（http://prinseq.sourceforge.net/）

主要参数设置：-trim_qual_left 20 -trim_qual_right 20 -trim_qual_window 10 -trim_qual_step 1 -min_len 35

污染评估软件：blast+

（http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download）

主要参数设置：-evalue 1e-10 -num_threads 40

结果目录：1_data_for_analysis/

All_sample_QC_infor.xls：所有样本 QC 之后结果统计，详细结果如下：

表 5.2 QC 之后结果统计

	T1	T2	T3	C1
Raw_sequences	18378142	16325862	15371472	20300126
Raw_bases	2297267750	2040732750	1921434000	2537515750
Raw_mean_length	125	125	125	125
Good_sequences	18308982	16274660	15317772	20215208
Good_ratio	99.62	99.69	99.65	99.58
Good_bases	2163702760	1941275976	1820805470	2373628392
Good_mean_length	118.18	119.28	118.87	117.42

注：若样本数目较多，此处只会截取部分样本数据，完整数据请见结果文件夹中的对应文件。

Raw_sequences：原始序列数目，为 Read1 与 Read2 数目之和

Raw_bases：原始序列碱基数目

Raw_mean_length：原始数据序列平均长度

Good_sequences：QC 之后剩余的序列数目

Good_ratio：QC 之后剩余序列数目比例

Good_bases：剩余序列总碱基数目

Good_mean_length：QC 之后序列平均长度

***blast_out.best_species_count.xls：**污染评估数据结果，详细如下：

表 5.3 污染评估结果

Species	Reads_number
Ovis aries	4397
Bos taurus	532
Capra hircus	419
Tursiops truncatus	58
Sus scrofa	50
Pantholops hodgsonii	39
Ceratotherium simum simum	29
Orcinus orca	28
Odobenus rosmarus divergens	27
Homo sapiens	13
Bubalus bubalis	13
Escherichia coli	12
Gorilla gorilla gorilla	8
Ailuropoda melanoleuca	6
uncultured bacterium	6
uncultured organism	6
Nomascus leucogenys	5
Escherichia fergusonii ATCC 35469	4
Dasypus novemcinctus	4
Capra caucasica	3

注：此表展示的是 NT 数据库比对后前 20 的物种，若样本数据较多，此处只展示某个样本的污染评估结果，其它样本数据见 1_QC/Sample/*_blast_out.best_species_count.xls 文件。

说明：此表可以反应出原始样品测序序列有无明显污染其它物种序列，一般看前几个物种，若前几个物种与样品物种符合或者非常近源（有些样品所测物种在数据库中并无该物种的相关序列信息，此时考虑其近缘物种），则样品无明显污染。

5.2 denovo 转录本拼接

5.2.1 方法说明

将各样本过滤之后序列进行合并，之后进行 de novo 拼接，使用软件 Trinity，版本 trinityrnaseq_rr20140717，使用 paired-end 的拼接方法。对拼接序列去重复，取长度大于 200bp 的序列，每个 Loci（c*_g*_）下最长的转录本作为 Unigene（软件的 Chrysalis clusters 模块）。Trinity 拼接过程基本如下：

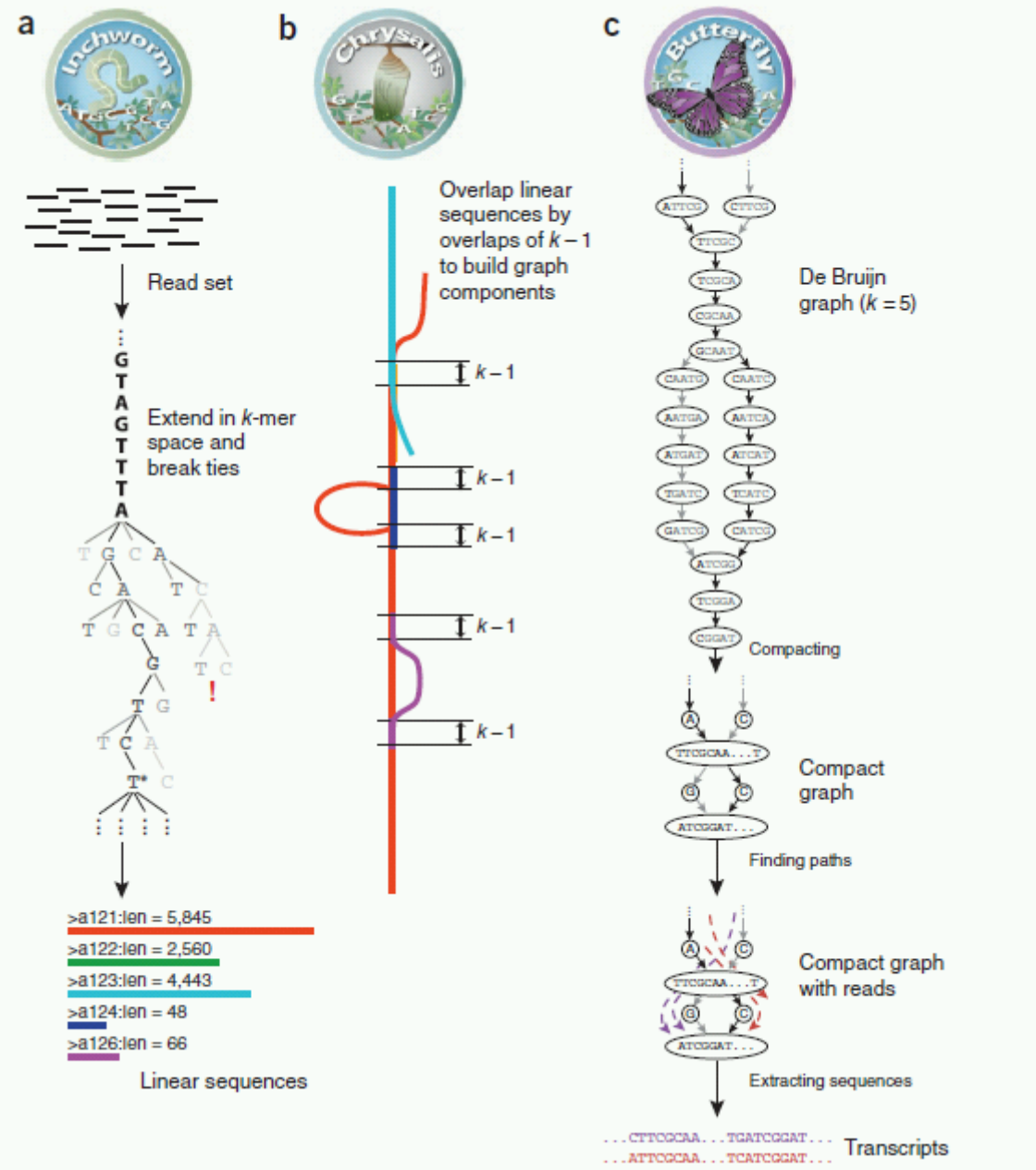


Figure 1 Overview of Trinity. (**a**) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a *k*-mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each *k*-mer present only once in the contigs. (**b**) Chrysalis pools 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).

使用软件: **Trinity** （http://trinityrnaseq.github.io/）。

软件参数设置：--min_contig_length 200 --seqType fq

5.2.2 结果展示

结果目录：2_assembly/

assembly_result.xls: 拼接结果统计，结果见表 5.4。

表 5.4 拼接结果统计

	All_num	>=500bp	>=1000bp	N50	N90	Max_len	Min_len	All_len	Mean_len
Transcript	66307	26238	14710	1568	293	20348	201	53888535	812.71
Unigene	59774	20839	10654	1231	271	20348	201	42202107	706.03

说明：N50：将 transcript 从长到短排序，依次累加 transcript 碱基数，当累计碱基数达到 transcript 总碱基数的 50%时的 transcript 的长度，Unigene 同；N90 以相似的方法统计。N50 参数在 RNA-Seq 项目中仅具有参考价值，不是评估结果好坏的客观标准。

Transcript.fa：所有拼接转录本序列，fasta 格式，文本文件，可用 excel 或者写字板打开，格式如下：

>c0_g1_i1 len=202 path=[133:0-100 389:101-201]

GCAGCAATCATGGAACTAGCCATATTAAATTTTTATCCGATTTACTTCGCAATAAGCACT

AGTAAATGATGCAACACTGAAGCAAACGTTGATGCAAGCTATTCTCAACAAAAGACGATG

AAGTGAAAGAGGAATTGAGATAAGTTAGTTGAATTAGAAGATAACAGAACTTACCTAATG

TTAAAACAATAATAATAGATCG

>c5_g1_i1 len=207 path=[284:0-155 439:156-206]

ATTTGGCACTGGAAACTCAATTAGAATAGCTAAAAACGTTTCGTCCTTTAATCAAGATAG

TCTACTGGATACGATGAAAGCTCATTTTCCTAATGCTGTTGCCACCAATGAAACTAATGA

AGAAATTGTTTACCAACTAAGCTCAGGAATTAAATCCACAGCAGATGTATCAACAGCAAC

TTTGGCTCAACTCTGTGATTCCCTGGA

>c6_g1_i1 len=245 path=[127:0-38 165:39-76 63:77-244]

GAGTTGCTGATGATATTCCTACGAGGAAAATACATGATCTTTGAAATTTAAACTGTTCGA

CCACTCGACCTCCGATGATAGGCCCAAGAGCAAATCCTAATGAAAGGACAGATACAAATA

GACTGCTTATCAATCCAGAGGCATCAGGATCATCCGCAAAACTGGTTAAATCTTTTTCCC

TCATTGCTTGAGTGAATGTAGCTACATAACCAATGGCAAAACCAATGCCAATAAACATTT

GACAA

其中大于号>后紧跟转录本的 id 号，len=后面为转录本的长度，即该转录本的碱基数，path 为从 de Bruijn Graph subComponent 中经历的路径。其后为该转录的碱基序列。每个转录本的 id 号构成都为 c*_g*_i*， 其中 c 为拼接过程形成的 de Bruijn Graph Component, g 为 subcomponet，可以看作为广

泛意义上的基因，i 代表转录本。详细见 Trinity 官方网站：http://trinityrnaseq.github.io/。

Unigene.fa：Unigene 序列，格式如上。

***_GC_content.pdf**：Trancsript /Unigene GC 含量分布图，作图源文件为*_GC_content.xls，如下图：

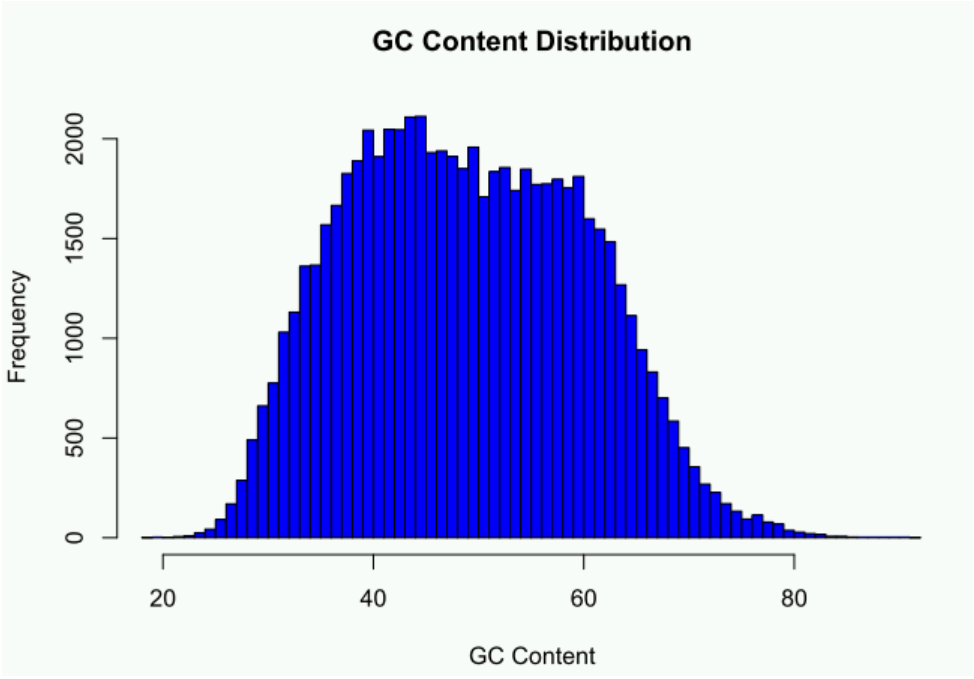


图 5.3 Transcript GC 含量分布图

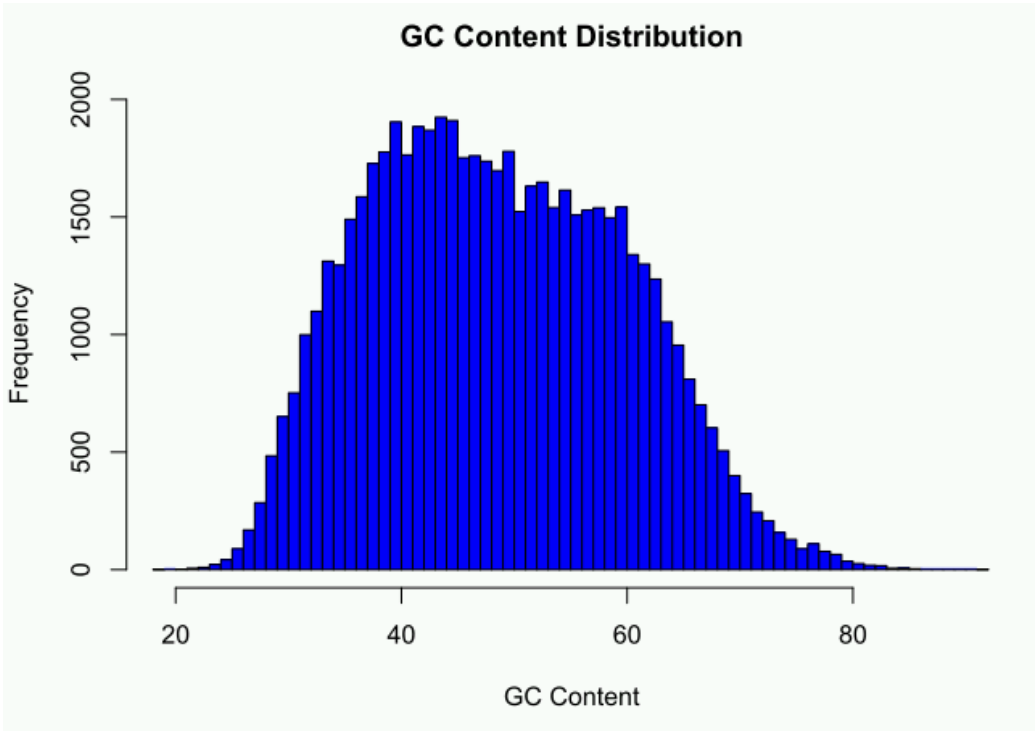


图 5.4 Unigene GC 含量分布图

说明：横坐标为 GC 含量，纵坐标为序列数目，该图可以看出样本转录本序列有无 GC 偏好性。

***_Len_Dis.pdf：**Trancsript /Unigene 长度分布图，作图源文件为*_len_distribution.xls，展示如下：

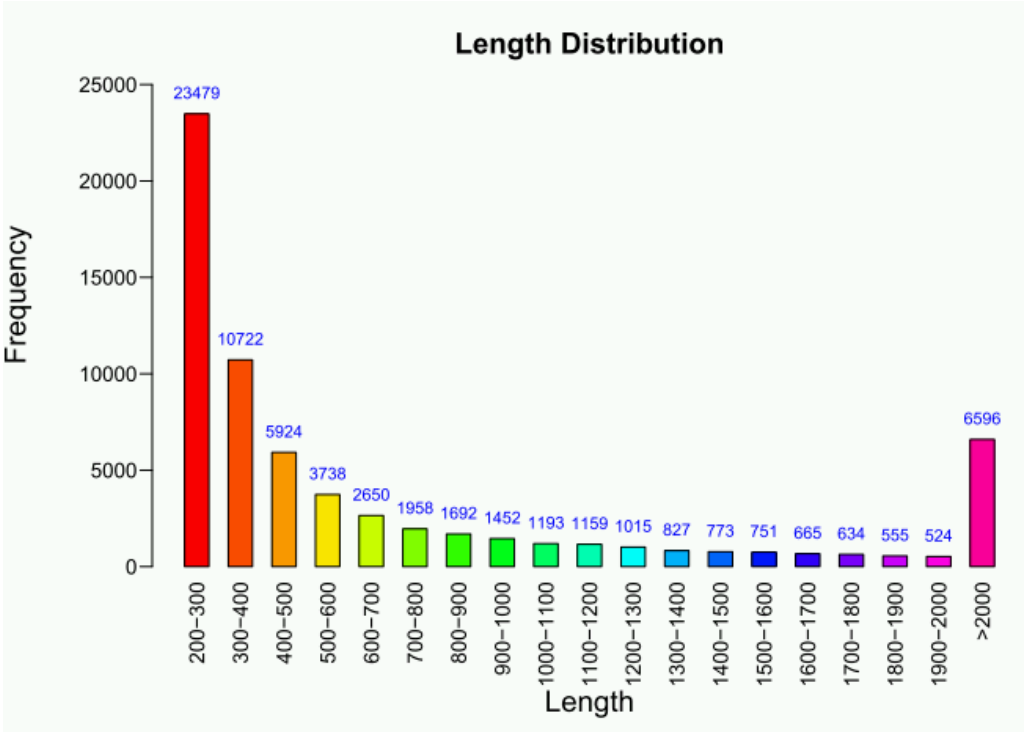


图 5.5 Transcript 长度分布图

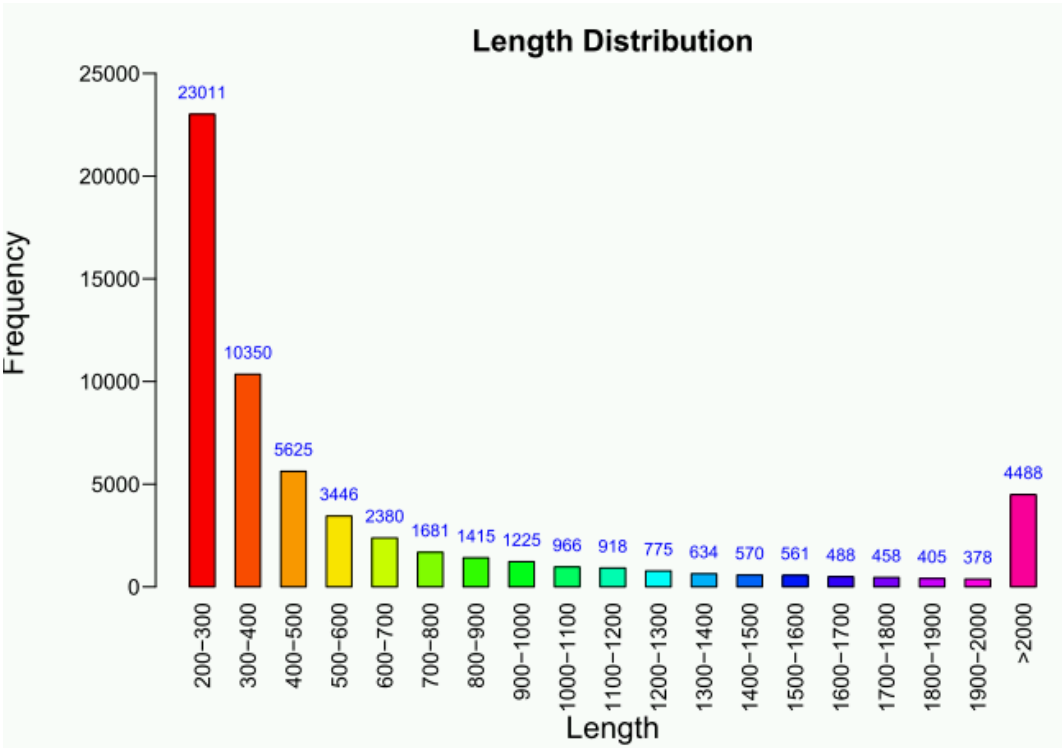


图 5.6 Unigene 长度分布图

说明：横轴表示长度区间，纵轴表示在某个区间内的 Trancript/Unigene 数目。

***_Len_accumulate.pdf**：Trancsript/Unigene 长度累积分布图，作图源文件为*_len_accumulate.xls，详细展示如下图：

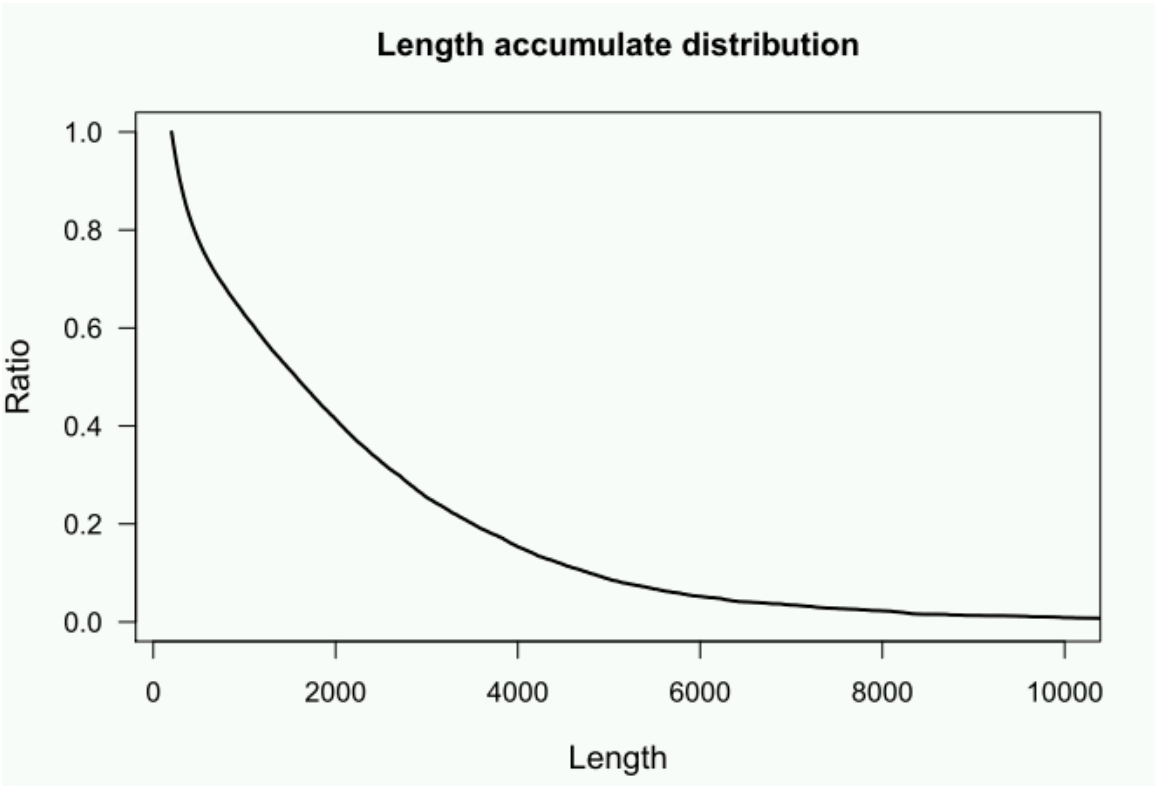


图 5.7 Transcript 长度累积分布图

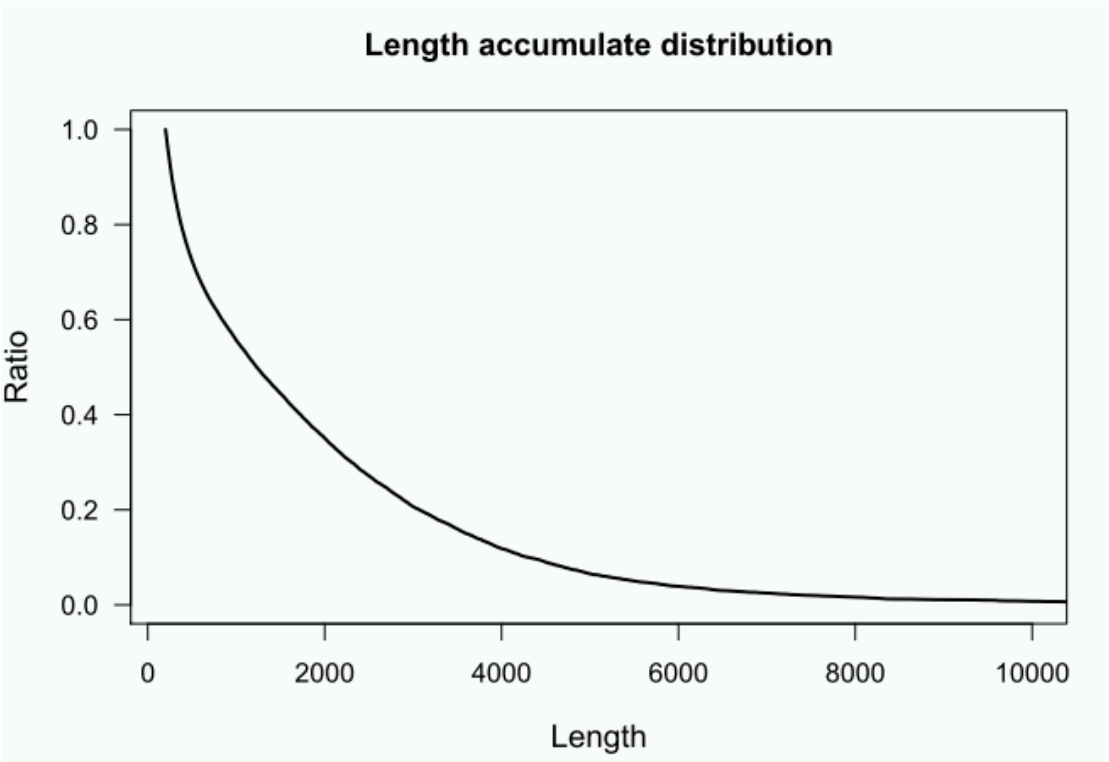


图 5.8 Unigene 长度累积分布图

说明：长度累积分布图，横坐标表示 Transcript/Unigene 序列长度，纵坐标表示大于某长度的序列总长度占 Transcript/Unigene 总长度的比例。该图可反应出如 N50，N90 等值。

Unigene_isoforms_num_count.pdf：Unigene 下面 isoform 数目分布图，作图源文件为 Unigene_isoforms_num.xls，展示如下：

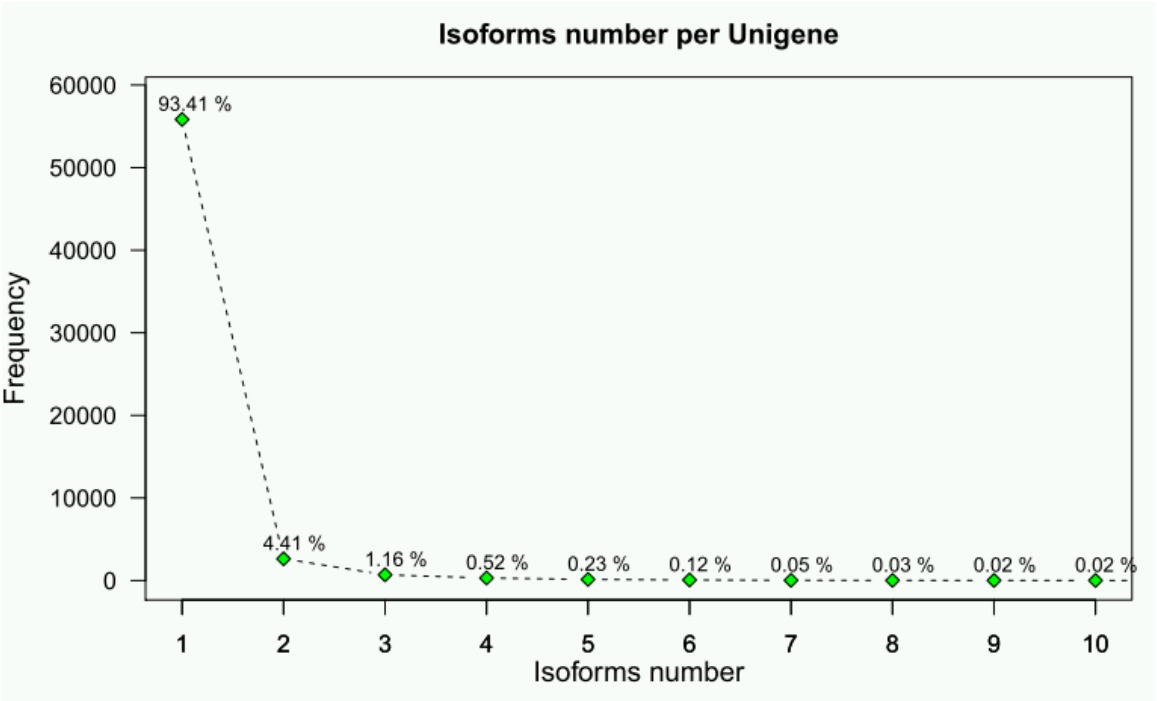


图 5.9 Unigene 下 Isoform 数目分布图

说明：上图横坐标表示每个 Unigene 下面 isoform 的数目，纵坐标表示 Unigene 数目。该图可反映出 Unigene 下可变剪切发生频率。

5.3 SSR 分析

5.3.1 SSR 分析

采用 MISA（1.0 版，默认参数）对 Unigene 及 Transcript 进行 SSR 检测，对不同 SSR 类型在基因与转录本的密度分布进行统计。

所用软件：MISA （http://pgrc.ipk-gatersleben.de/misa/）

软件参数设置：

参数	说明	参数设置
1	单碱基连续重复次数	10
2	双碱基连续重复次数	6
3	3 碱基连续重复次数	5

4	4 碱基连续重复次数	5
5	5 碱基连续重复次数	5
6	6 碱基连续重复次数	5
Max_difference_between_2_SSRs	两个 SSR 间最大间隔长度(bp)	100

结果目录：3_SSR/

Unigene.fa.ssr.xls：Unigene SSR 寻找结果，详细见下表；

表 5.5 Unigene SSR 结果

ID	SSR nr.	SSR type	SSR	size	start	end
c4228_g1	1	p1	(A)11	11	917	927
c1907_g1	1	p1	(T)10	10	1144	1153
c11962_g1	1	p3	(CAG)5	15	2637	2651
c11962_g1	2	p3	(TGC)7	21	4420	4440
c55732_g1	1	p1	(T)10	10	51	60
c12786_g4	1	p2	(TG)7	14	9	22
c8581_g1	1	p1	(G)11	11	1032	1042
c11692_g2	1	p3	(ACC)8	24	1071	1094
c46697_g1	1	p3	(TTG)5	15	275	289
c44301_g1	1	p2	(TG)9	18	1	18
c11580_g1	1	p2	(GA)6	12	402	413
c30643_g1	1	p1	(A)11	11	19	29
c11388_g1	1	p1	(T)11	11	2253	2263
c11388_g1	2	p1	(A)11	11	2530	2540
c45016_g1	1	p1	(A)12	12	70	81
c4441_g2	1	p1	(T)13	13	1	13
c54743_g1	1	c	(AG)6acag	36	102	137
c10658_g1	1	p1	(A)14	14	373	386
c6823_g1	1	p1	(A)10	10	207	216
c2861_g1	1	p1	(A)13	13	3	15

注：上表只展示了前 20 个 SSR 的结果，其他请参阅相关文件。

ID：做 SSR 分析的基因的 id

SSR nr.：SSR 给每个相同 id 的转录本的编号（不需要关注）

SSR type：SSR 类型：c，复杂重复类型；p1，单碱基重复；p2，两个碱基重复；p3 三个碱基重复……

SSR：重复序列

Size：重复序列的大小

Start：重复序列的开始碱基位置

End：重复序列的结尾碱基位置

Unigene.fa_density.pdf：SSR 密度分布图，作图源文件为 Unigene.fa_density.xls，详细展示如下图：

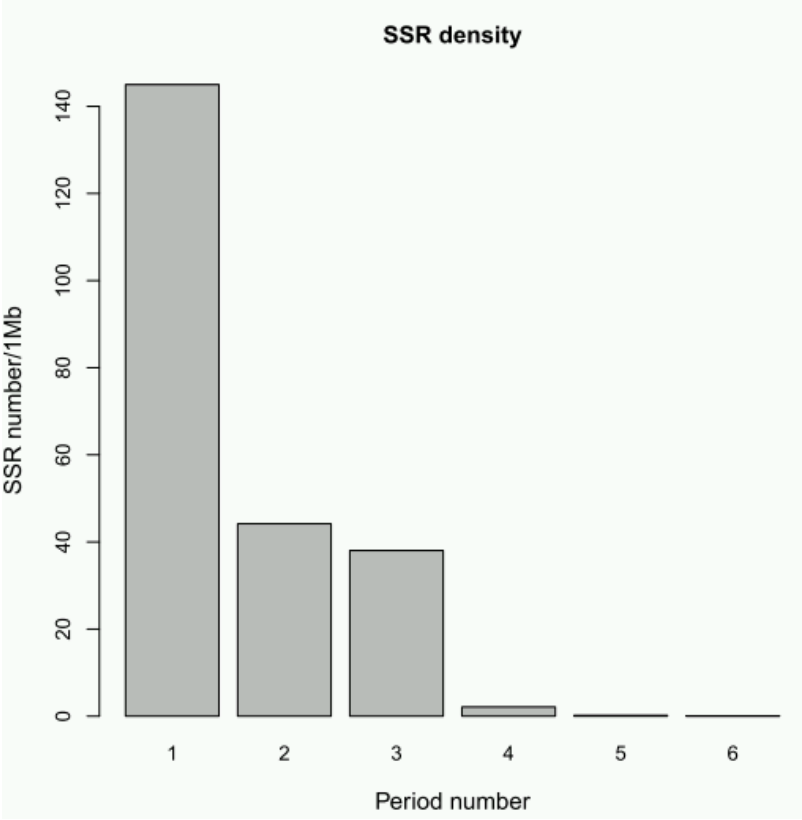


图 5.10 Unigene 下 SSR 密度分布图

说明：横坐标为不同的 SSR 类型，纵坐标为每百万碱基中 SSR 的个数。

5.3.2 SSR 引物设计

找出 SSR 标记之后，采用 Primer3（默认参数）进行 SSR 引物设计。

所用软件：primer3 http://primer3.sourceforge.net/，采用默认参数。

结果目录：3_SSR/

Unigene.fa.ssr.primer.xls：Unigene SSR 引物设计结果，结果展示如下表：

表 5.6 Unigene SSR 引物设计结果

ID	FORWARD PRIMER1 (5'-3')	Tm(°C)	REVERSE PRIMER1 (5'-3')	Tm(°C)
c4228_g1	GTAGCCATCAACTGCAGCAA	60.019	TAGCCTCTGTCTGCCAACCT	60.012
c1907_g1	CAGACTTGGGTAGCACAGCA	60.049	GTCATGGAGCATGGACACAG	60.121
c11962_g1	GCAGTTCCTTTTGCTTTTGC	60.003	GCTGTCATCTGGCTCAACAA	59.992
c11962_g1	GCAGTTCCTTTTGCTTTTGC	60.003	GCTGTCATCTGGCTCAACAA	59.992
c55732_g1	TGGTGGGGAGCTGATAGGTA	60.474	GAGACCCCCAGTCTCTGGAT	60.469
c12786_g4	CTTCCCAGGGCAAAGATGTA	60.066	TCGTCTCACAGCAGATTTGG	59.984
c8581_g1	AGACCTACACGGAGGTGGTG	60.026	CACTGTGGTGAAGGGGAAGT	60.002
c11692_g2	CTCTGCTGCCTTTACCTTGG	60.008	ACCCCAAAGAGAGAAGGGAA	60.045
c46697_g1	TTTTTAAAATGGGGCATCCA	60.123	GAAAGCTTGGCTGTCCTTTG	59.993
c44301_g1	TTCACCCTGGCAATTAAAGG	59.931	CGCTGGCCTTGTTAACATTT	60.131
c11580_g1	GTGAGGAGTCAGCGAGTTCC	59.993	CCCTCACCTTTTCCTCTTCC	60.045
c30643_g1	TGTCACCAAGCAAAGGACAA	60.278	AATCTATGCTTCCTGGGCCT	60.06
c11388_g1	GTTCAACGAGGAGAAGCAGG	59.989	GGTCTGGGCAGAGTGAAGAG	59.986
c11388_g1	GTTCAACGAGGAGAAGCAGG	59.989	GGTCTGGGCAGAGTGAAGAG	59.986
c45016_g1	GGAAGGTCTGTGCAGTGTCA	59.872	TGGATGACAGTCTCTGTCGC	59.988
c4441_g2	GCTGTGCTTTCATTTGCTACA	59.126	TCATACATGGTTTGATGGACA	57.232
c54743_g1	GACCAAAAACCACTCTCCCA	59.943	AGGGTACCTGGGTGTCACAG	59.876
c10658_g1	AACCAATTACACTCTGGCCG	59.993	GCCAAAACAAAAGGATTCCA	59.916
c6823_g1	AGACAACCAGCCCACAATTC	59.973	TGTTCCCTGTGGACGTTGTA	60.001
c2861_g1	ATTCCCACCACCTGTCTCAG	59.962	ATATGCCCTGCGTAAACTGG	59.982

注：上图展示的仅为部分结果，详细请参阅对应文件夹中的文件，另外上述所有展示的结果均为 Unigene 的分析结果，Transcript 的分析结果见 3_SSR/Transcript*命令的文件。

5.4 Unigene 注释

5.4.1 各数据库比对

将 Unigene 序列与公共数据 gene 进行比较，通过 gene 的相似性进行功能注释。基因相似性比对主要基于 BLAST 算法。BLAST，全称 Basic Local Alignment Search Tool，即"基于局部比对算法的搜索工具"，由 Altschul 等人于 1990 年发布。Blast 能够实现比较两段核酸或者蛋白序列之间的相似性的功能，它能够快速的找到两段序列之间的相似序列并对比对区域进行打分以确定相似性的高低。将 Unigene 基因序列分别与 NR、NT、KOG、CDD、PFAM、Swissprot、TrEMBL、GO、KEGG 库进行比对，取相似度>30%，且 e<1e-5 的注释，合并基因得到的所有注释详细信息。

各数据库说明如下：

NR：NR (NCBI non-redundant protein sequences) 是 NCBI 官方的蛋白序列数据库，它包括了 GenBank 基因的蛋白编码序列，PDB(Protein DataBank)蛋白数据库、SwissProt 蛋白序列及来自 PIR（Protein Information Resource）和 PRF（Protein Research Foundation）等数据库的蛋白序列。

NT：NT(NCBI nucleotide sequences) 是 NCBI 官方的核酸序列数据库，包括了 GenBank，EMBL 和

DDBJ（但不包括 EST,STS,GSS,WGS,TSA,PAT,HTG 序列）的核酸序列。

PFAM：Pfam (Protein family)是最全面的蛋白结构域注释的分类系统。蛋白质是由一个个结构域组成的，而每个特定结构域的蛋白序列具有一定保守性。PFAM 将蛋白质的结构域分为不同的蛋白家族，通过蛋白序列的比对建立了每个家族的氨基酸序列的 HMM 统计模型。PFAM 家族按注释结果可靠性分为两大类：手工注释的可靠性高的 Pfam-A 家族和程序自动产生 Pfam-B 家族。我们通过 HMMER3 程序，搜索已建好的蛋白结构域的 HMM 模型，对 unigene 进行了蛋白家族的注释。详见 http://pfam.sanger.ac.uk/。

KOG/COG: COG 是 Clusters of Orthologous Groups of proteins 的简称，KOG 为 euKaryotic Ortholog Groups。这两个注释系统都是 NCBI 的基于基因直系同源关系，其中 COG 针对原核生物，KOG 针对真核生物。COG/KOG 结合进化关系将来自不同物种的同源基因分为不同的 Ortholog 簇，目前 COG 有 4873 个分类，KOG 有 4852 个分类。来自同一 ortholog 的基因具有相同的功能，这样就可以将功能注释直接继承给同一 COG/KOG 簇的其他成员。详见 http://www.ncbi.nlm.nih.gov/COG/。

Swiss-Prot（A manually annotated and reviewed protein sequence database）搜集了经过有经验的生物学家整理及研究的蛋白序列。详见 http://www.ebi.ac.uk/uniprot/。

KEGG 是 Kyoto Encyclopedia of Genes and Genomes 的简称，是系统分析基因产物和化合物在细胞中的代谢途径以及这些基因产物功能的数据库。它整合了基因组、化学分子和生化系统等方面的数据，包括代谢通路（KEGG PATHWAY）、药物（KEGG DRUG）、疾病（KEGG DISEASE）、功能模型（KEGG MODULE）、基因序列（KEGG GENES）及基因组（KEGG GENOME）等等。KO（KEGG ORTHOLOG）系统将各个 KEGG 注释系统联系在一起，KEGG 已建立了一套完整 KO 注释的系统，可完成新测序物种的基因组或转录组的功能注释。详见 http://www.genome.jp/kegg/。

GO(Gene Ontology)是一套国际标准化的基因功能描述的分类系统。GO 分为三大类 ontology：生物过程（Biological Process）、分子功能（Molecular Function）和细胞组分(Cellular Component)，分别用来描述基因编码的产物所参与的生物过程、所具有的分子功能及所处的细胞环境。GO 的基本单元是 term，每个 term 有一个唯一的标示符（由“GO:”加上 7 个数字组成，例如 GO:0072669）；每类 ontology 的 term 通过它们之间的联系（is_a, part_of, regulate）构成一个有向无环的拓扑结构。详见 http://www.geneontology.org/。

各数据库及功能注释所用到的软件及方法：

NT：NCBI blast 2.2.28+，blastn

NR、SwissProt、TrEMBLe 序列数据库的比对：NCBI blast 2.2.28+，blastx;

CDD、COG/KOG、PFAM：NCBI blast2.2.28+，rpsblast;

GO 功能注释：基于 Swissprot 和 TrEMBL 两部分的蛋白注释结果及 GO 数据库通过自写脚本获取 GO 注释信息;

KEGG 相关注释：KAAS，KEGG Automatic Annotation Server。

结果目录：4_Annotation/

Annotation_statistics.xls：各数据库注释比例统计，结果如下：

表 5.7 各数据注释比例统计

Database	Number of Unigenes	Percentage(%)
Annotated in CDD	14031	23.47
Annotated in KOG	13503	22.59
Annotated in NR	23653	39.57
Annotated in NT	38629	64.63
Annotated in PFAM	12781	21.38
Annotated in Swissprot	21476	35.93
Annotated in TrEMBL	22833	38.2
Annotated in GO	20597	34.46
Annotated in KEGG	7793	13.04
Annotated in at least one database	39103	65.42
Annotated in all database	5555	9.29
Total Unigenes	59774	100

Annotated in CDD: CDD 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in KOG: KOG 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in NR: NR 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in PFAM: Pfam 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in Swissprot: Swissprot 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in TrEMBLe: TrEMBLe 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in GO: GO 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in KEGG: KO 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in at least one Database: 在以上 8 个数据库中至少 1 个数据库注释成功的蛋白数目及其占总蛋白数的比例

Annotated in all Databases: 在以上 8 个数据库中都注释成功的蛋白数目及其占总蛋白数的比例

Total Unigenes: 总的蛋白条数,占总蛋白比例为 100%

Annotation_ratio.pdf：各数据注释比例折线图，作图源文件为 Annotation_statistics.xls，展示如下

图：

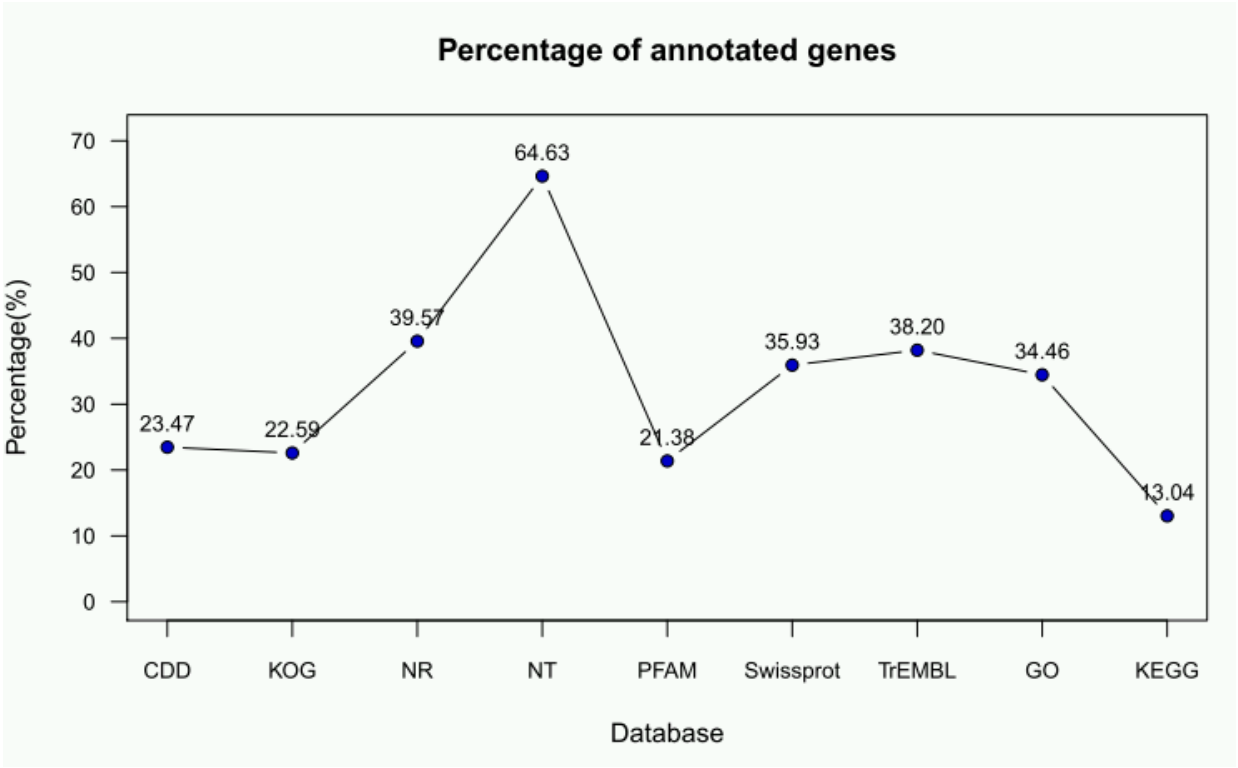


图 5.11 数据库注释比例折线图

nr_species_count.pdf：NR 数据库注释物种统计饼图，作图源文件为 nr_species_count.xls，展示如下图：

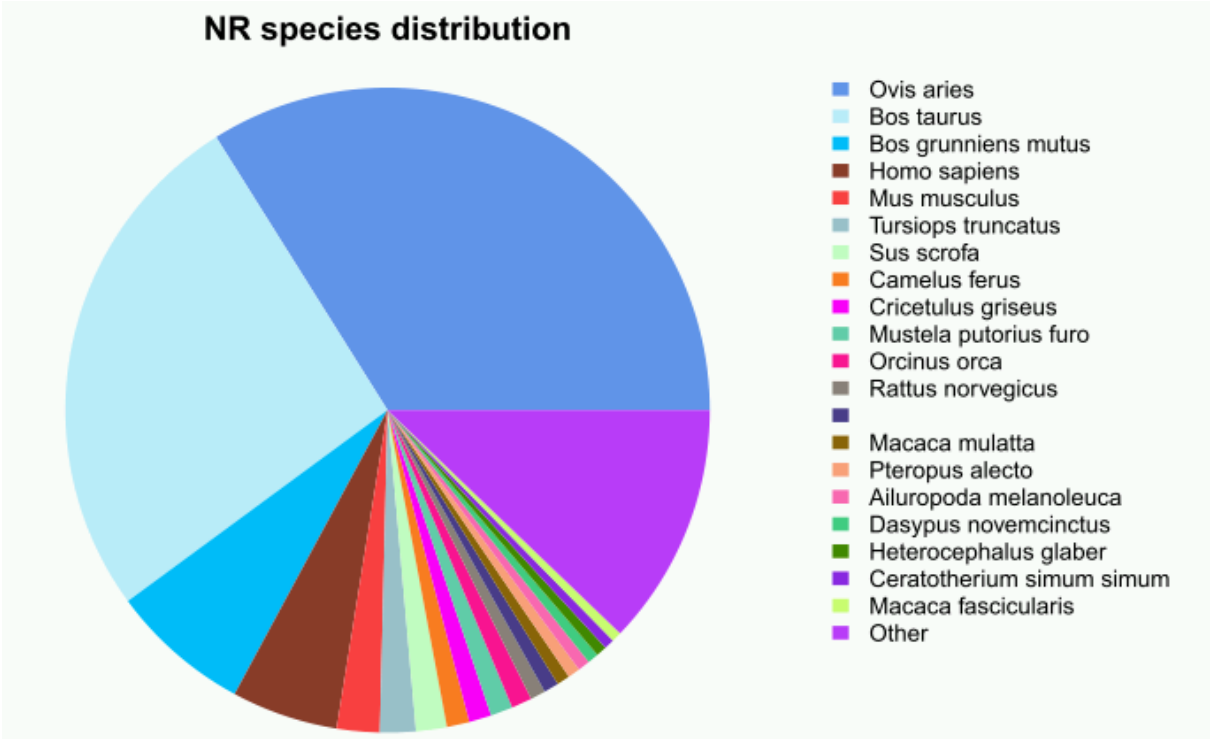


图 5.12 NR 数据库物种分布饼图

注：默认只展示丰度前 20 的物种

Venn_diagram_for_annotation.pdf：各数据库注释上的基因韦恩图，默认为绘制 NR、KEGG、Swissprot、KOG/COG 之间的韦恩图，展示如下图：

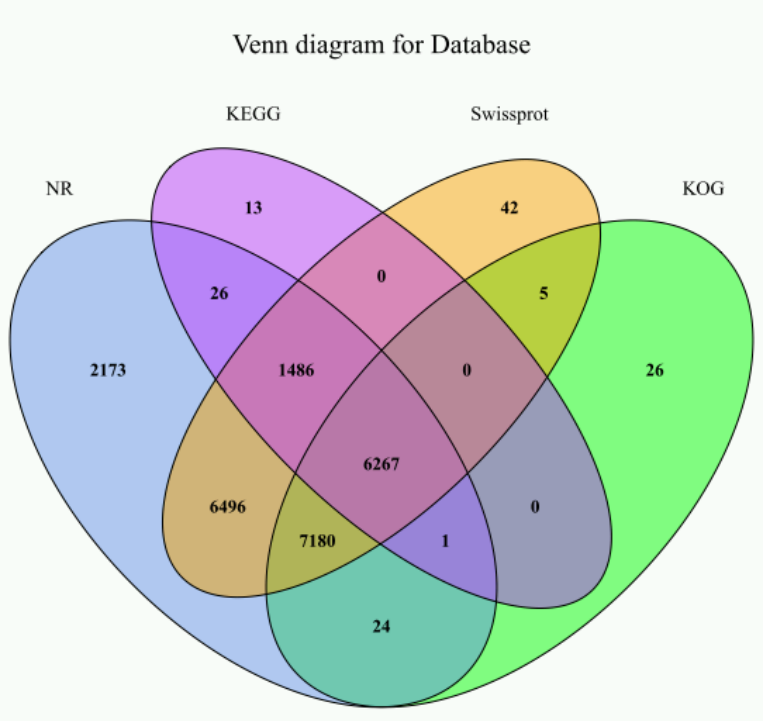


图 5.13 各数据库注释韦恩图

5.4.2 COG、KOG 注释

将 Unigene 序列比对到 COG/KOG 数据库中，原核物种比对 COG 数据库，真核物种比对到 KOG 数据库，基于比对结果计算各功能类下面注释上的基因数目。

结果目录：4_Annotation/KOG/

KOG_code_count.xls/COG_code_count.xls：各功能类基因数目统计列表，结果如下表：

表 5.8 COG/KOG 各功能类基因数统计表

Code	Name	Gene_num	Gene_ratio
T	Signal transduction mechanisms	2210	16.37
R	General function prediction only	1955	14.48
K	Transcription	1397	10.35
S	Function unknown	1175	8.7
O	Posttranslational modification, protein turnover, chaperones	1031	7.64
Z	Cytoskeleton	761	5.64
U	Intracellular trafficking, secretion, and vesicular transport	759	5.62
J	Translation, ribosomal structure and biogenesis	495	3.67
A	RNA processing and modification	477	3.53
L	Replication, recombination and repair	356	2.64

注：上表展示的仅为前 10 的功能类，完整列表请看相关文件

KOG_Categories.pdf/COG_Categories.pdf：COG /KOG 注释功能类条形图，绘图源文件为 KOG_code_count.xls/COG_code_count.xls，结果展示如下：

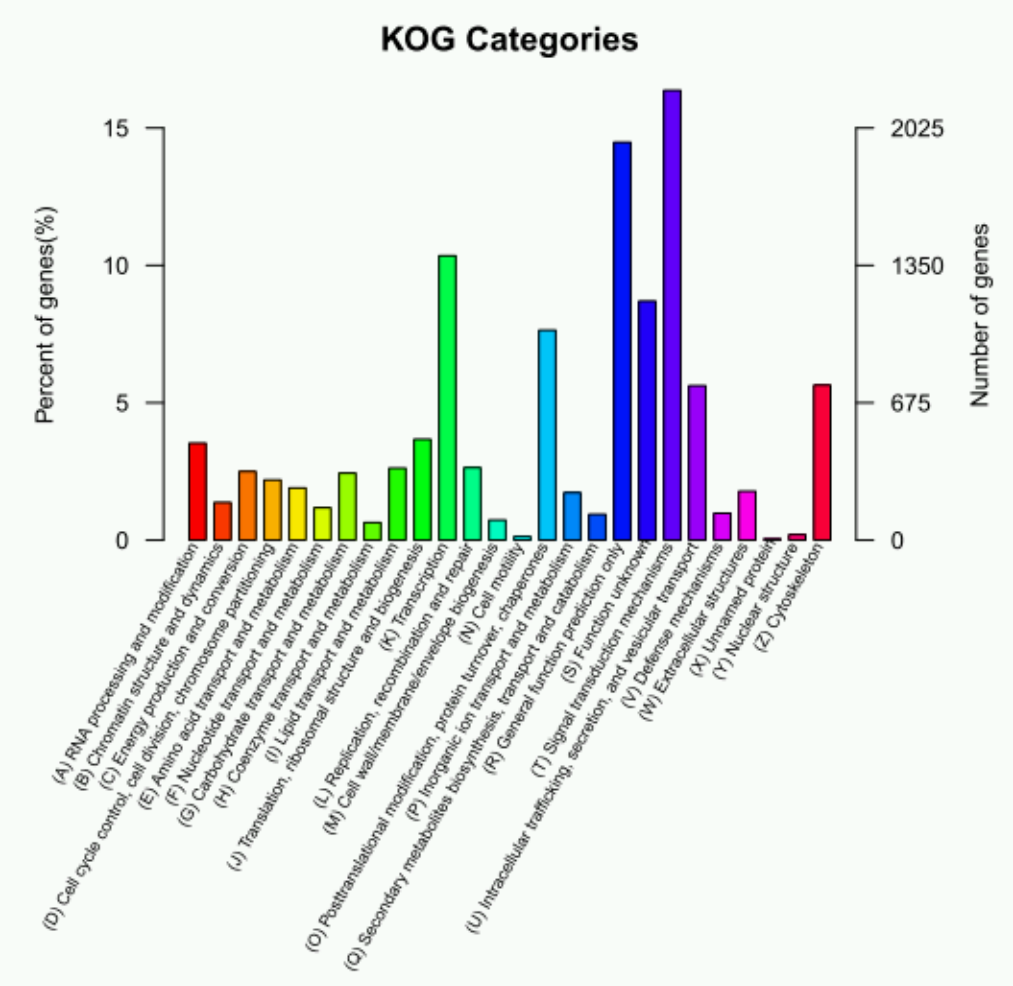


图 5.14 KOG 注释条形图

注：各字母意义：

[S] Function unknown

[Z] Cytoskeleton

[Y] Nuclear structure

[W] Extracellular structures

[V] Defense mechanisms

[U] Intracellular trafficking, secretion, and vesicular transport

[T] Signal transduction mechanisms

[R] General function prediction only

[Q] Secondary metabolites biosynthesis, transport and catabolism

[P] Inorganic ion transport and metabolism

[O] Posttranslational modification, protein turnover, chaperones

[N] Cell motility

[M] Cell wall/membrane/envelope biogenesis

[L] Replication, recombination and repair

[K] Transcription

[J] Translation, ribosomal structure and biogenesis

[I] Lipid transport and metabolism

[H] Coenzyme transport and metabolism

[G] Carbohydrate transport and metabolism

[F] Nucleotide transport and metabolism

[E] Amino acid transport and metabolism

[D] Cell cycle control, cell division, chromosome partitioning

[C] Energy production and conversion

[B] Chromatin structure and dynamics

[A] RNA processing and modification

5.4.3 GO 注释

对得到的基因进行 GO 分类，统计基因在 Biological Process, Cellular Component, Molecular Function 三个类别的各 GO term。此分析是基于 blast uniprot 的结果（即合并与 swissprot 和 trembl 的结果），利用得到的 uniprot 号比对 GO term。

所用软件：自写程序

结果目录：4_Annotation/GO/

GO_classification_level2.pdf：GO 分类在 level2 水平上基因分别条形图，结果如下图：

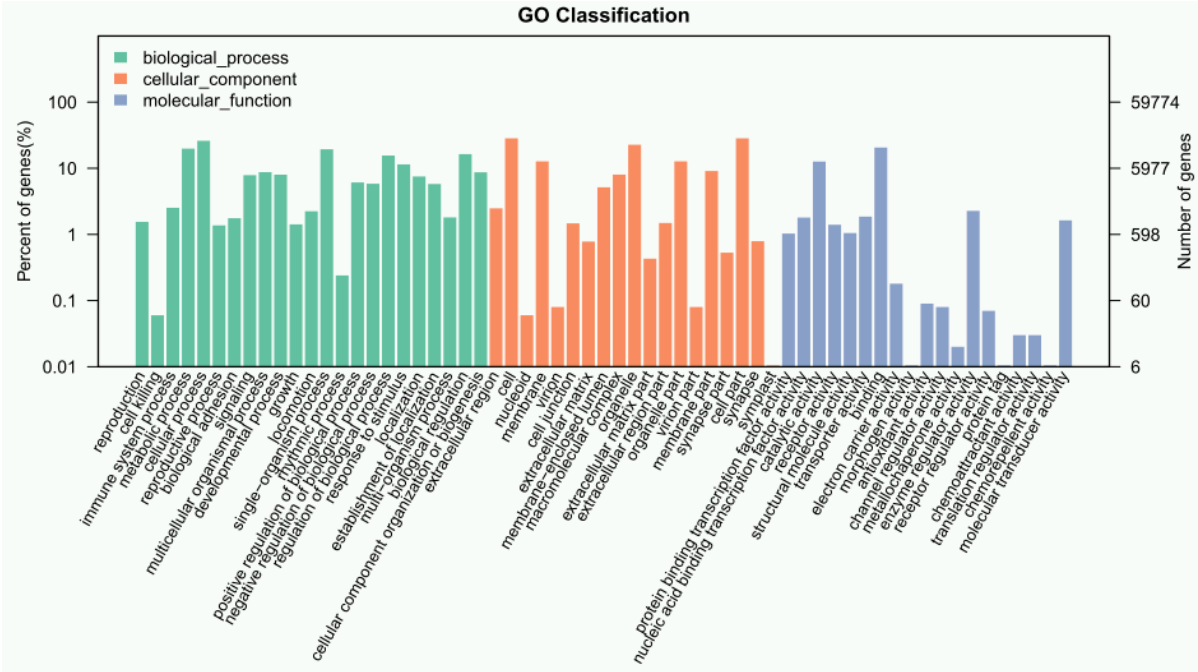


图 5.15 level2 水平 GO 注释上的基因分布

说明：横坐标表示 Level2 水平上 GO term，不同颜色代表不同 GO 类，分为三大类，纵坐标表示注释到该 term 上面的 Unigene 数目及比率，纵坐标采用对数坐标。

5.4.4 KEGG 注释

对得到的基因进行 KEGG Pathway 分析，利用 KAAS 预测得到对应的 KO 号，然后利用 KO 号对应到 KEGG pathway 上，分析基因与 KEGG 中酶注释的关系文件以及映射到 pathway 的信息。

结果目录：4_Annotation/KEGG/

kegg_annot2.xls：各 pathway 注释上的 Unigene 数目，结果如下：

表 5.9 各 pathway 注释上的 Unigene 数目

Pathway_ID	Pathway_name	Gene_num
ko04151	PI3K-Akt signaling pathway	217
ko05200	Pathways in cancer	205
ko05166	HTLV-I infection	181
ko04010	MAPK signaling pathway	165
ko05169	Epstein-Barr virus infection	157
ko05016	Huntington's disease	156
ko04141	Protein processing in endoplasmic reticulum	155
ko03013	RNA transport	146
ko05205	Proteoglycans in cancer	145
ko05010	Alzheimer's disease	144

注：上表展示的仅为前 10 的 pathway，完整列表请看相关文件

KEGG_Categories.pdf：pathway 注释分类结果，绘图源文件为 kegg_annot_catar.xls，展示如下图：

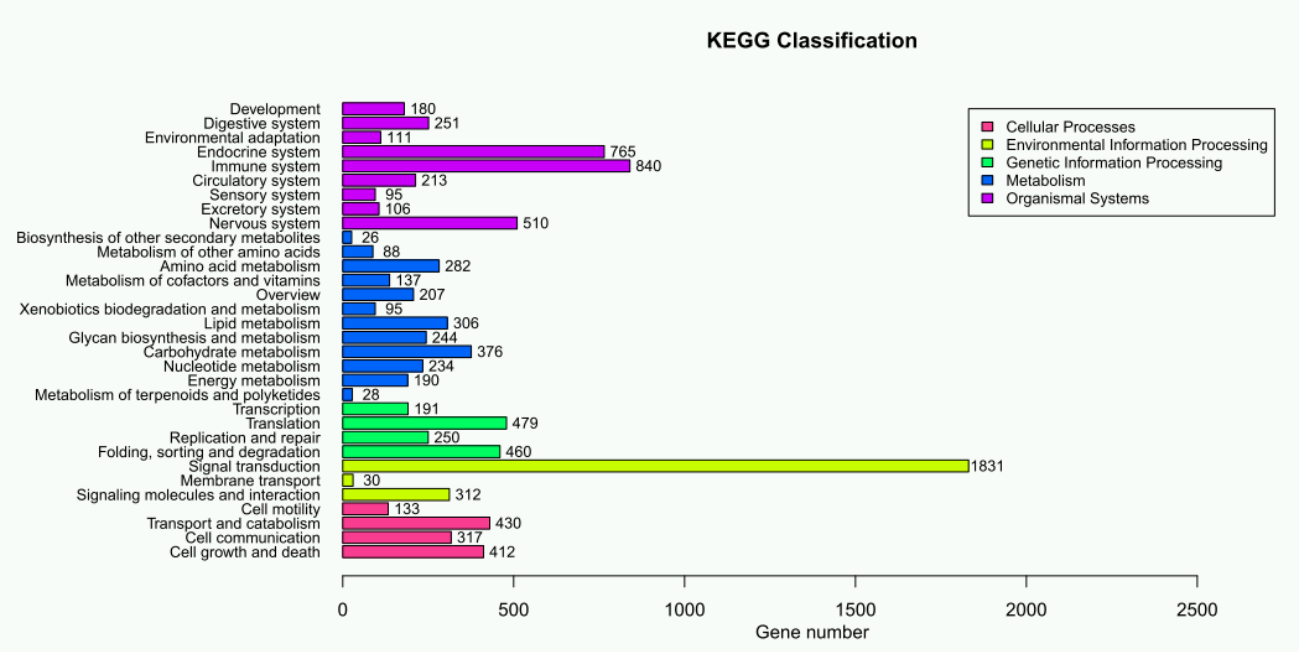


图 5.16 pathway 注释分类条形图

说明：纵坐标为 KEGG 代谢通路的名称，横坐标为注释到该通路下的基因个数，将基因根据参与的 KEGG 代谢通路分为 5 个分支：细胞过程（A，Cellular Processes），环境信息处理（B，Environmental Information Processing），遗传信息处理（C，GeneticInformation Processing），代谢（D，Metabolism），有机系统（E，Organismal Systems）。

5.4.5 CDS 预测

获取 NR 数据库最佳比对结果，通过该结果确定 Unigene 的 ORF 读码框，然后根据标准码子表确定其 CDS 及编码的氨基酸序列，未比对上的 Unigene 通过 OrfPredict 软件预测其 CDS 序列。

所用软件：OrrPredictor，(http://www.proteomics.ysu.edu/tools/OrfPredictor.html/)

参数设置：默认参数。

结果目录：4_Annotation/CDS_predict/

CDS_Len_Dis.pdf：CDS 长度分布图，绘图源文件为 cds_length_distribution.xls，结果如下：

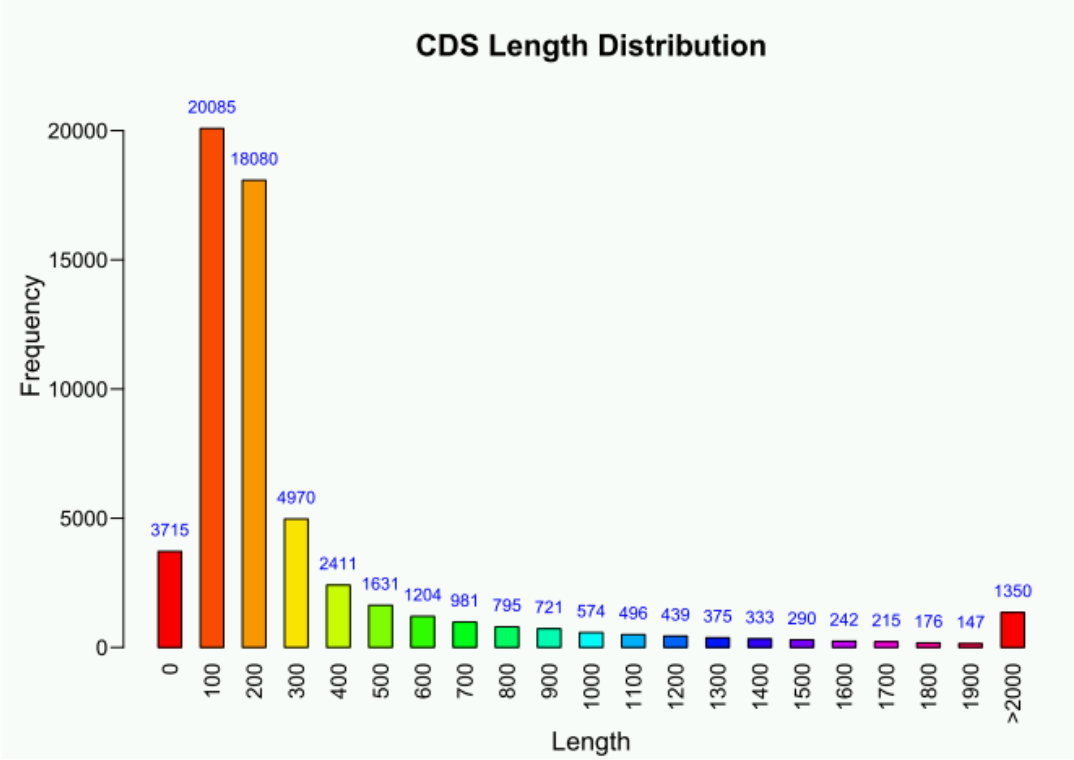


图 5.17 CDS 长度分布图

CDS_length_ratio.pdf：CDS 区域占 Unigene 基因长度比例分布图，绘图源文件为 cds_length_ratio.xls，结果如下：

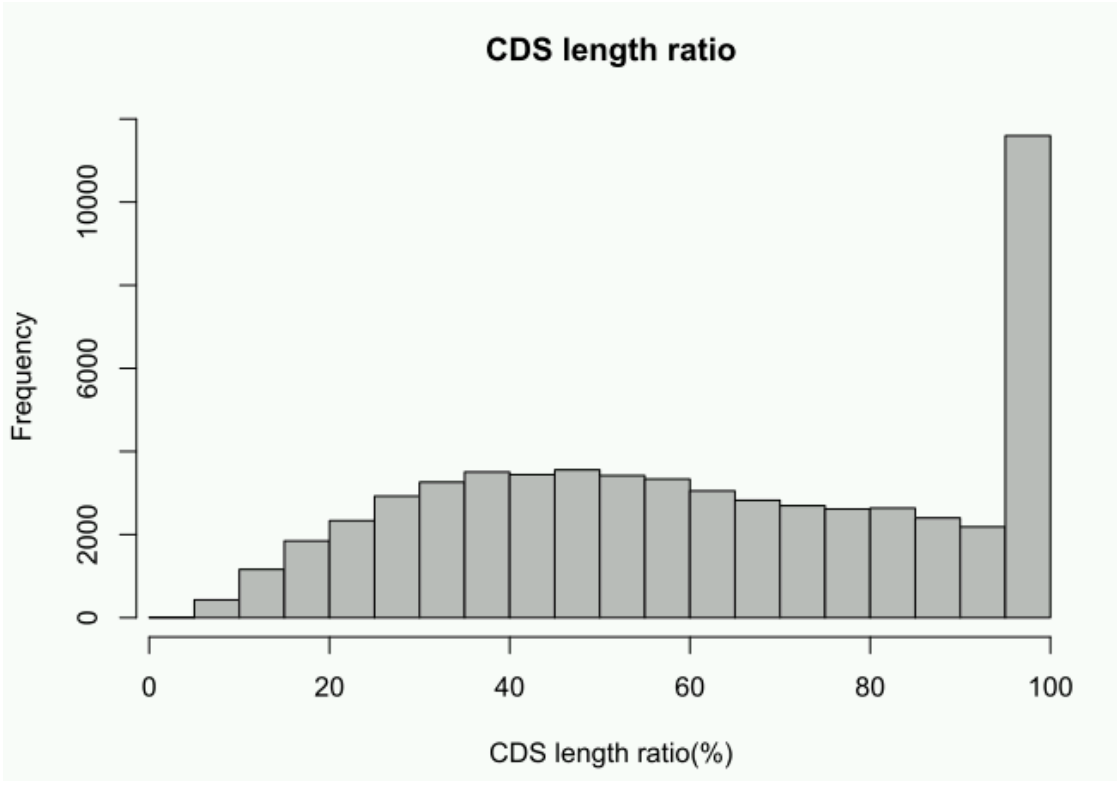


图 5.18 CDS 区域占 Unigene 基因长度比例分布图

5.5 RNASeq 测序评估

5.5.1 Mapping 结果统计

以 Trinity 拼接得到的转录组作为参考序列，将每个样品的 clean reads（pair-end 序列）对参考序列做 mapping。该过程采用了 RSEM 软件（Li et al.,2011），RSEM 中使用到的 bowtie 参数：mismatch 2。比对之后计算各样本的 Mapping 比率，Mapping 比率计算采用软件 RSeQC。

Mapping 软件：bowtie（http://bowtie-bio.sourceforge.net/index.shtml）

关键参数设置：-v 2 -S

Mapping 统计软件：RSeQC（http://rseqc.sourceforge.net/）bam_stat.py 模块，RSeQC 为一款专门用于做 RNAseq 数据 QC 的软件。

关键参数设置：默认参数

结果目录：5_RNASeq_evaluation/

All_sample_mapping_stastics.xls：所有样本的 Mapping 比例统计，结果如下表：

表 5.10 各样本 Mapping 统计结果

Sample_name	T1	T2	T3	C1
Total reads	18188380	16165306	15195530	20008028
Total mapped	15925180(87.56%)	14679042(90.81%)	13350436(87.86%)	17066092(85.30%)
Mutiple mapped	1841204(10.12%)	260146(1.61%)	1063204(7.00%)	2686146(13.43%)
Unique mapped	14083976(77.43%)	14418896(89.20%)	12287232(80.86%)	14379946(71.87%)
Read1 mapped	7041988(38.72%)	7209448(44.60%)	6143616(40.43%)	7189973(35.94%)
Read2 mapped	7041988(38.72%)	7209448(44.60%)	6143616(40.43%)	7189973(35.94%)
Mapped to '+'	7041988(38.72%)	7209448(44.60%)	6143616(40.43%)	7189973(35.94%)
Mapped tp '-'	7041988(38.72%)	7209448(44.60%)	6143616(40.43%)	7189973(35.94%)
Non-splice reads	14083976(77.43%)	14418896(89.20%)	12287232(80.86%)	14379946(71.87%)
Splice reads	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)
Reads mapped in proper pairs	14083976(77.43%)	14418896(89.20%)	12287232(80.86%)	14379946(71.87%)

注：若样本数目较多，此处只会截取部分样本数据，完整数据请见结果文件夹中的对应文件。

Total Reads：所有的序列数目，为 QC 之后的 pair-end 序列，single_end 序列未加入分析；

Total Mapped：比对上的序列数目及比例

Mutiple mapped：比对到多个地方的序列数目及比例

Unique Mapped：唯一比对的序列数及比例，后面各列统计的均为 Unique mapped 结果

Read1 Mapped：Read1 Mapped 上的序列数及占比，此处只算 Unique Mapped 序列

Read2 Mapped：Read2 Mapped 上的序列数及占比，此处只算 Unique Mapped 序列

Mapped to '+'：比对到正向的序列数及比例

Mapped to '-'：比对到反向的序列数及比例

Non-splice reads：非 splice Mapped 序列比对，此处比对的是转录本，基本都为 Non-splice reads

Splice reads：splice Mapped 序列比对，此处比对的是转录本，无 splice reads

Reads mapped in proper pairs：PE reads 一起 Mapped 上的数目及比例。

5.5.2 均一化分析

均一化分析是用于评估转录组测序建库时对 mRNA 的打断是否随机，若不随机则可能对后续的分析会产生较大偏好性，计算方法为将每条转录本均均一化为 100 长度，计算每个位置的覆盖度，之后计算所有转录本在这 100 长度位置上的覆盖度均值，看其是否均一。根据转录组建库实验的特点，转录本产生的测序序列距离转录本的 5'端和 3'端越近，测序深度越低，但总体的均一化程度比较高。符合上述特点的转录本测序序列可判定为一次合格的转录组测序。

均一化分析软件：RSeQC inner_distance.py 模块及 geneBody_coverage.py 模块，参数默认。

结果目录：5_RNASeq_evaluation/

All.geneBodyCoverage.curves.pdf：所有样本均一化分布曲线图，绘图源文件为 All.geneBodyCoverage.txt，结果展示如下：

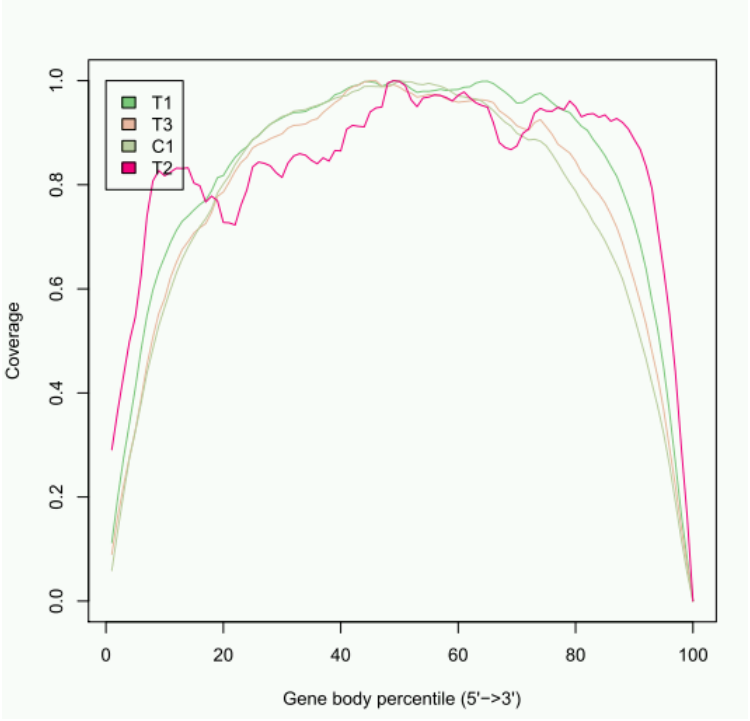


图 5.19 所有样本均一化分布曲线

说明：横坐标为距离转录本 5'端的相对位置（以百分比表示），纵坐标为覆盖深度的平均值，从

图中可知该次测序符合正常 RNASeq 测序特点，为合格测序。

All.geneBodyCoverage.heatMap.pdf：均一化分布热图，与上图展示的结果类似，采用热图模式。

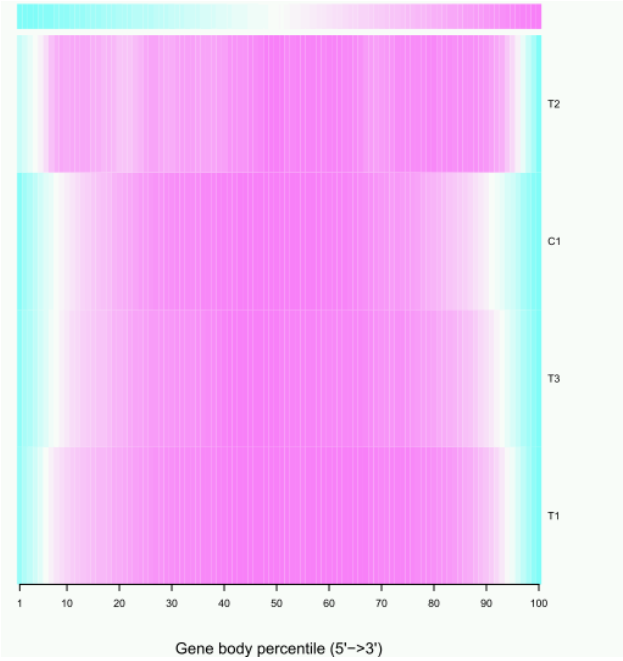


图 5.20 均一化分布热图

说明：横坐标为距离转录本 5’端的相对位置，纵轴为样本，每一个颜色块表示平均覆盖度，颜色越红覆盖度越高，一般正常测序中间颜色块较红，越往两端颜色越绿。

***/*.geneBodyCoverage.curves.pdf**：某样本均一化分布曲线，每个样本对应的文件夹中均会有该文件。

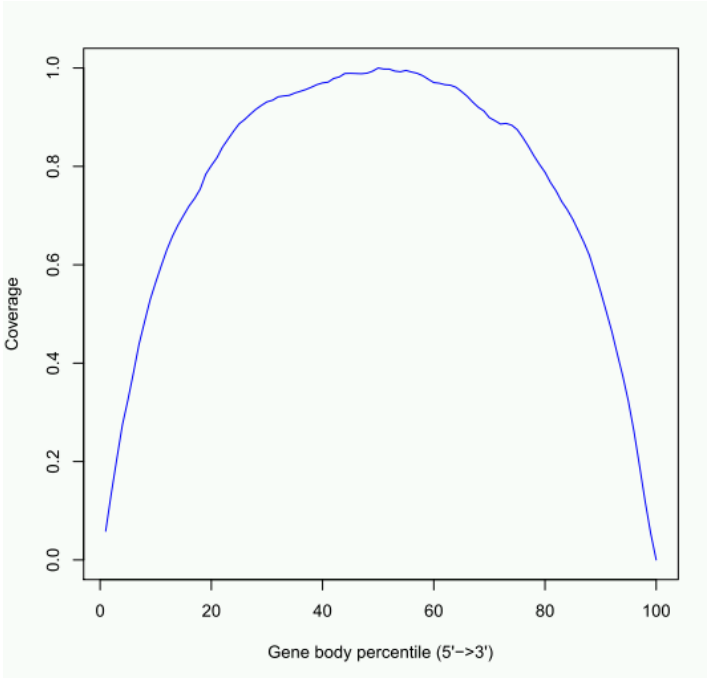


图 5.21 单样本均一化分布曲线

注：每个样本均会有一个该文件，在对应的样本文件夹中，上面展示的只是其中一个样本的结果，其他样本的见对应的样本名文件夹。

***/*.inner_distance_plot.pdf：**某样本建库片段 inner 距离直方图，inner_distance 即 Read1 与 Read2 之间的距离。结果如下：

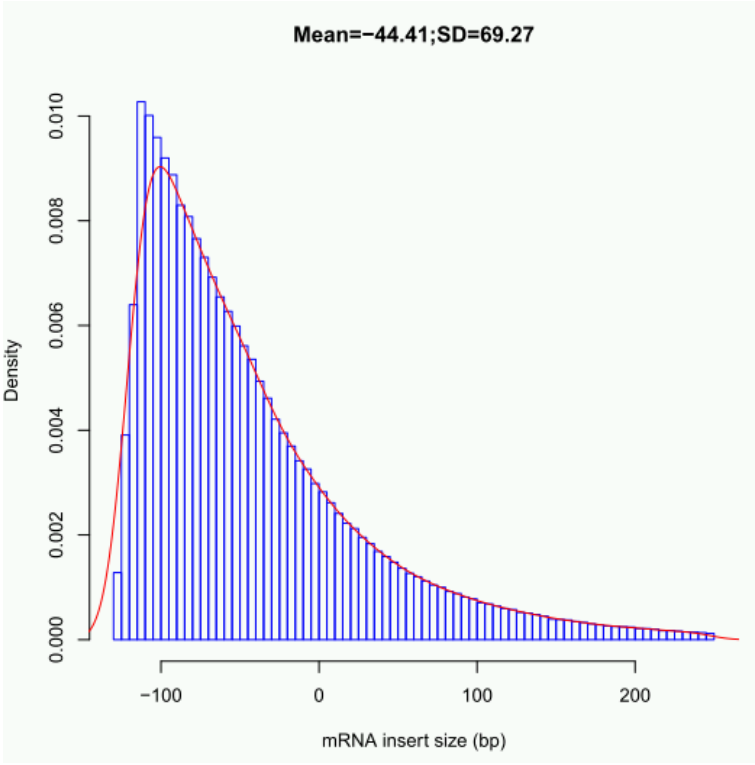


图 5.22 单样本 inner_distance 直方图

注：每个样本均会有一个该文件，在对应的样本文件夹中，上面展示的只是其中一个样本的结果，其他样本的见对应的样本名文件夹。

说明：横坐标表示建库片段的 inner_distance，纵坐标表示片段的密度，该图可以反应出转录组测序的建库片段大小分布，一般转录组测序文库大小为 180-200bp。通过上图可以估算出建库片段大小，计算方式为（2*测序长度+Mean）。

5.5.3 基因覆盖度分析

通过比对结果，计算各转录本各位置的覆盖度，并计算各转录本的覆盖率，通过计算每个样本各基因的覆盖比率可以看出该次测序各样本被完全测到以及未被测到基因的比例，亦可看出样本间是否存在较多特异性表达的基因。

覆盖度分析软件：samtools（http://samtools.sourceforge.net/）

软件参数设置：samtools depth -b

结果目录：5_RNASeq_evaluation/

***/*.coverage.interval_plot.pdf：**基因覆盖分布饼图，绘图源文件为*.coverage.interval.xls，展示结果如下：

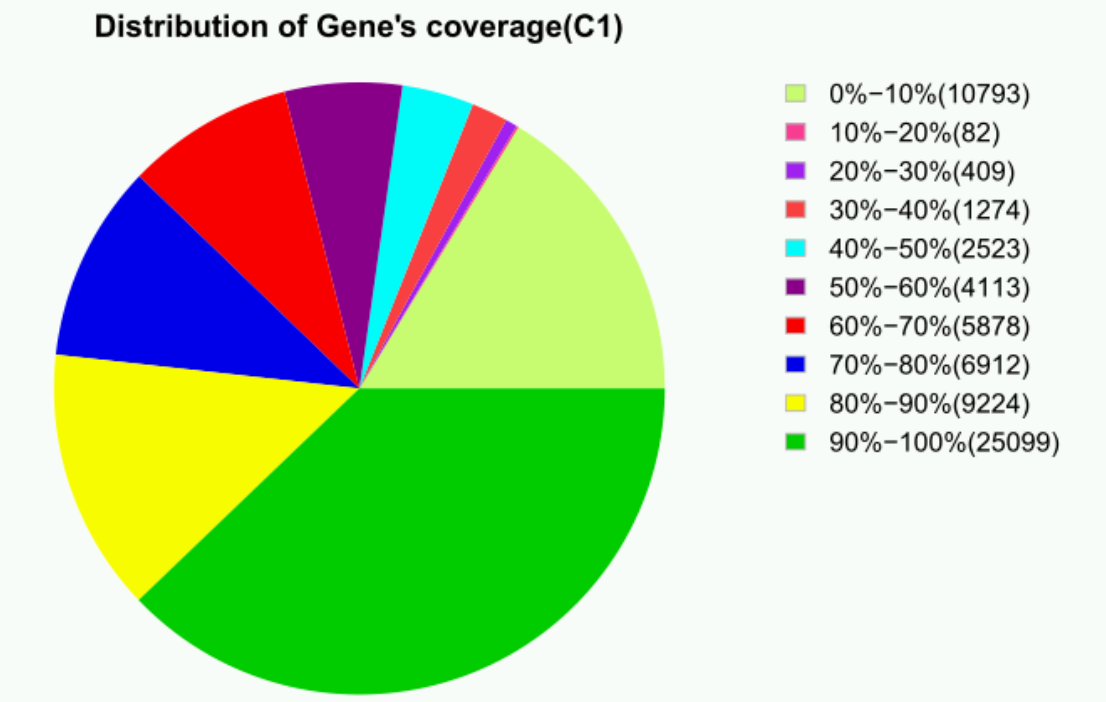


图 5.23 单样本基因覆盖度饼图

注：每个样本均会有一个该文件，在对应的样本文件夹中，上面展示的只是其中一个样本的结果，其他样本的见对应的样本名文件夹。

5.5.4 测序饱和度分析

测序饱和度曲线反映了基因表达水平定量对数据量的要求。表达量越高的基因，就越容易被准确定量；反之，表达量低的基因，需要较大的测序数据量才能被准确定量。当曲线达到饱和，说明测序数据量已满足定量要求。表达水平的饱和曲线的具体算法描述如下：分别对 10%、15%，20%、25%……90%的总体 mapped reads 单独进行基因定量分析，把 100%mapped reads 数据条件下得到的基因表达水平作为最终数值。用每个百分比条件下求出的单个基因的 RPKM 数值和最终对应基因的表达水平数值进行比较，如果差异小于 10%，则认为这个基因在这个条件下定量是准确的。

饱和度分析软件：RSeQC RPKM_saturation.py 模块，参数默认。

结果目录：5_RNASeq_evaluation/

All_saturation_curve_plot.pdf：所有样本所有基因饱和度曲线，绘图源文件为 All.saturation.xls，详细展示如下图：

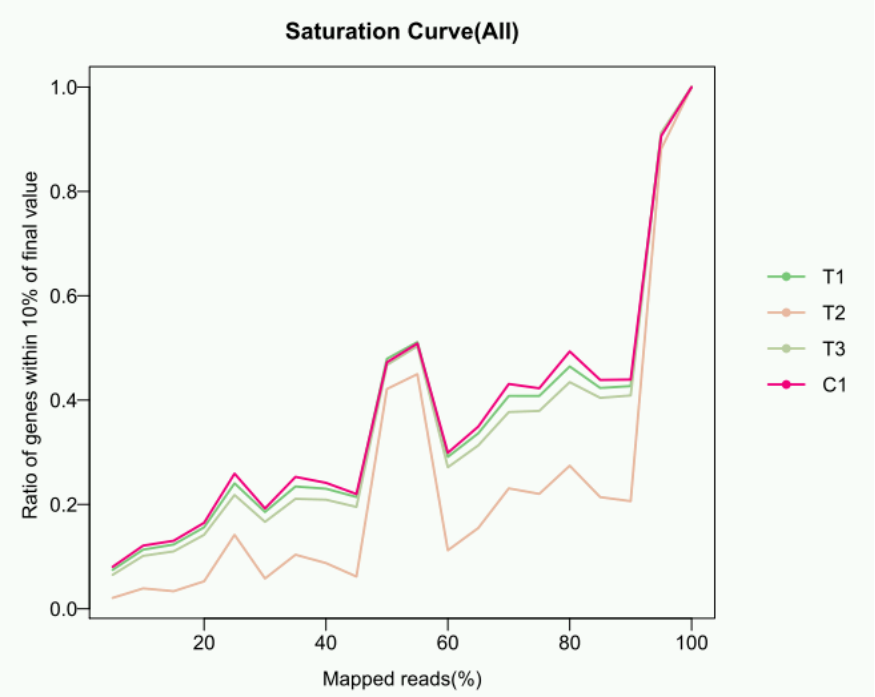


图 5.24 所有样本所有基因饱和度曲线

说明：横坐标代表定位到基因组上的 reads 数占总 mapped reads 数的百分比，纵坐标代表定量误差在 10%以内的基因占总基因数的比例。

***/*_saturation_curve_plot.pdf：**单样本饱和度曲线图，结果展示如下：

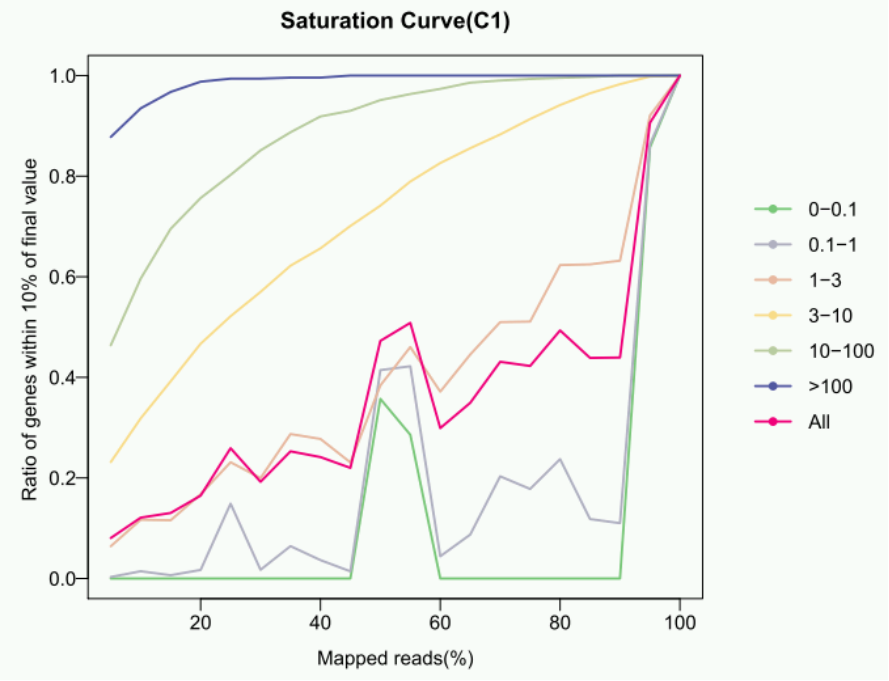


图 5.25 单样本基因饱和度曲线

注：每个样本均会有一个该文件，在对应的样本文件夹中，上面展示的只是其中一个样本的结果，其他样本的见对应的样本名文件夹。

说明：横坐标代表定位到基因组上的 reads 数占总 mapped reads 数的百分比，纵坐标代表定量误差在 10%以内的基因占总基因数的比例。不同颜色的线条代表不同 RPKM 区间。图例方框中为不同颜色对应的 100%mapped reads 时的 RPKM 区间。上图中可反应出表达量位于哪些区间已达到饱和，表达量较低的基因尚未达到饱和。

5.6 表达量统计及样本间聚类分析

注：以下全部展示的均为基因层面，转录本层面的结果在对应的文件夹里面均有，带 isoforms 命名的均为转录本层面结果。

5.6.1 表达量统计及绘图

采用 RSEM（Li et al., 2011）对 bowtie 的比对结果进行统计，进一步得到每个样品比对到每个基因上的 read count 数目。在 RNA-seq 技术中，FPKM（Fragment Per Kilo bases per Million mapped Reads）是每百万 reads 中来自某一基因每千碱基长度的 reads 数目，FPKM 同时考虑了测序深度和基因长度对 reads 计数的影响，是目前最为常用的基因表达水平估算方法（Mortazavi et al., 2008）。所以我们将 read count 数进行了 FPKM 转换。FPKM 计算公式如下：

$$FPKM = \frac{totalexonFragments}{mappedFragments(millions) * exonlength(KB)}$$

计算表达量软件：RSEM（http://deweylab.biostat.wisc.edu/rsem/），参数采用默认参数。

结果目录：6_expression_profile/

All.genes.FPKM.interval.xls：各样本表达量区间统计表，结果如下：

表 5.11 基因表达量区间统计

	T1	T2	T3	C1
All	59774(100.00%)	59774(100.00%)	59774(100.00%)	59774(100.00%)
Expressed number	39256(65.67%)	11275(18.86%)	30808(51.54%)	49382(82.61%)
0-0.1	20587(34.44%)	50136(83.88%)	29013(48.54%)	10397(17.39%)
0.1-1	9892(16.55%)	6898(11.54%)	10402(17.40%)	6203(10.38%)
1-3	16613(27.79%)	1709(2.86%)	12079(20.21%)	24403(40.83%)
3-10	9334(15.62%)	760(1.27%)	5924(9.91%)	14460(24.19%)
10-100	2894(4.84%)	236(0.39%)	1989(3.33%)	3724(6.23%)
>100	454(0.76%)	35(0.06%)	367(0.61%)	587(0.98%)

All.genes.FPKM.interval.barplot.pdf：各样本表达量区间条形图，绘图源文件为 All.genes.FPKM.interval.xls，结果展示如下：

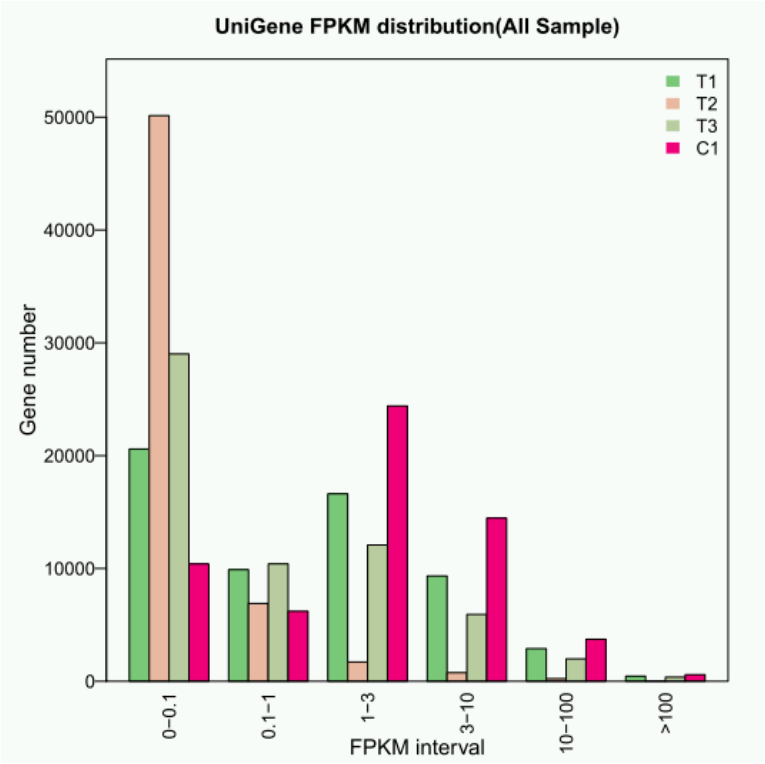


图 5.26 表达量区间条形图

All.genes.FPKM.boxplot.pdf：各样本基因表达量盒状图，展示如下图：

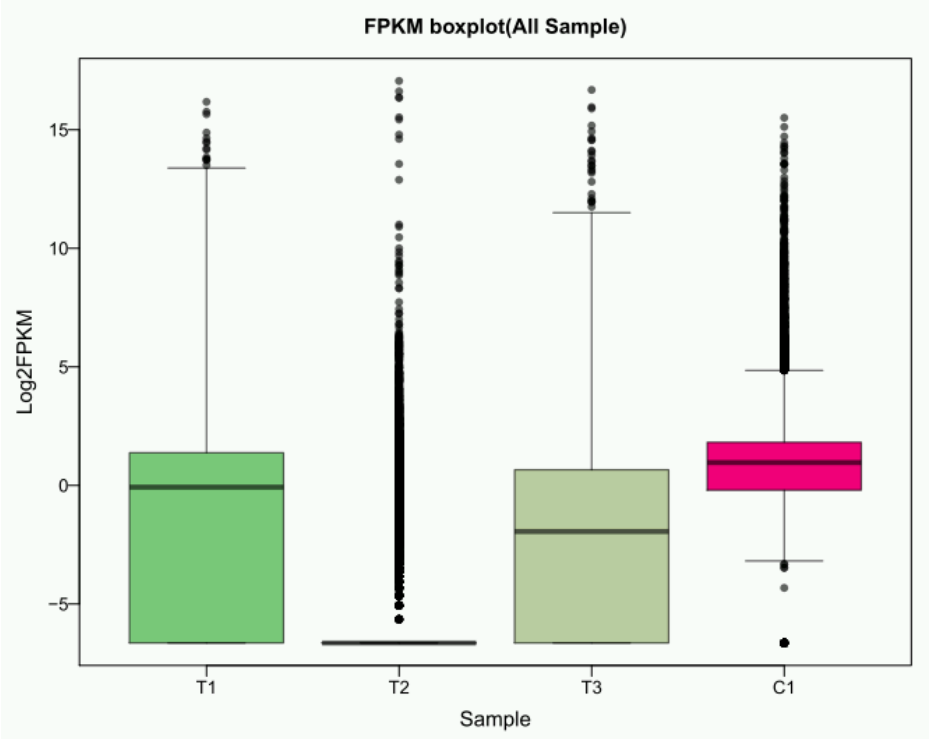


图 5.27 基因表达量盒状图

说明：横坐标为样本，不同颜色代表不同样本，纵坐标为 Log2(FPKM)值，触须的范围表示表达量的最大与最下值，盒子里面区域为 25%-75%区域，盒子里面的黑线为中位数。

All.genes.FPKM.density.pdf：各样本基因表达量密度分布图，展示如下图：

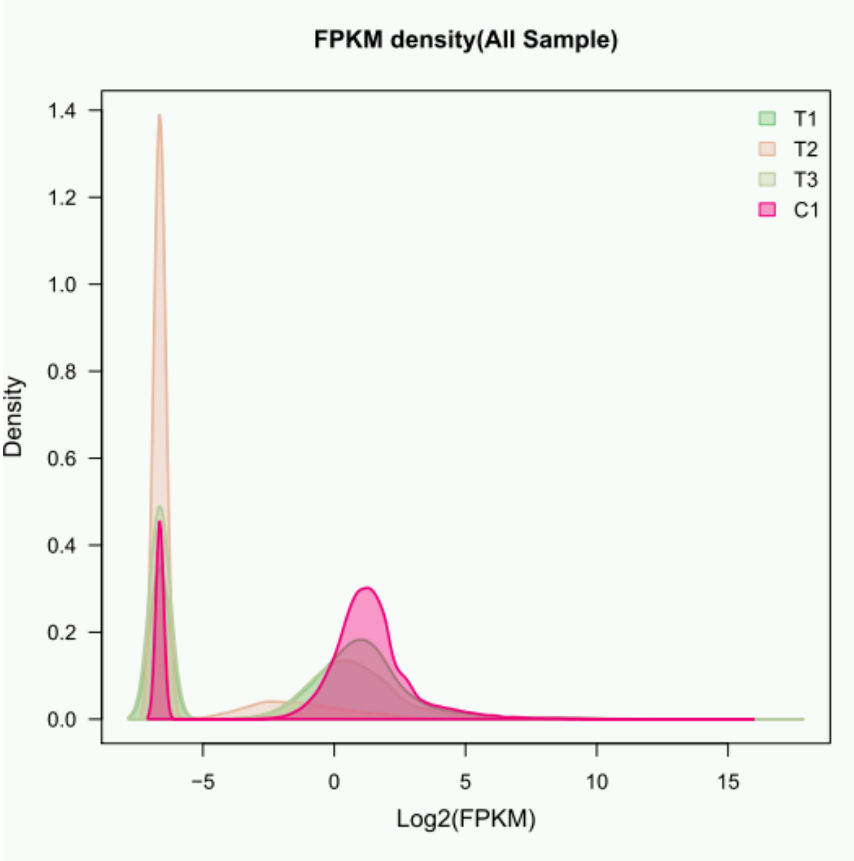


图 5.28 基因表达量密度曲线

说明：横坐标为表达量 Log2（FPKM）值，纵坐标为对应 Log2(FPKM)的相对密度值，最左边的峰值为未表达的基因。

5.6.2 样本聚类分析

通过计算样本间距离可以获取样本间相似度，表达模式越接近的样本在聚类分析的时候会越靠近，样本间距离计算方式为 1-R^2，其中 R 为皮尔森相关系数。样本间聚类方式为 Hierarchical clustering。

聚类软件：R

结果目录：6_expression_profile/

All.genes.correlation.heatmap.pdf：样本间距离热图，结果展示如下图：

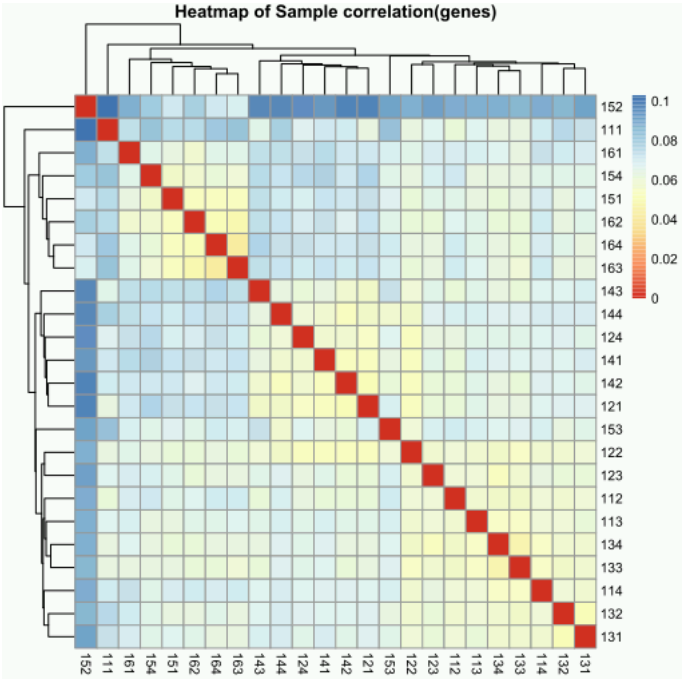


图 5.29 样本间距离热图

说明：上图中每个颜色方块表示两两样本间距离，聚类最大值为 1，最小值为 0，距离值越大颜色越蓝，反之距离越小颜色越红，且越相似的样本在聚类时会越靠近。上图可反应出所有样本间的相似度情况。

All.genes.Sample.clustering.pdf：样本间聚类树图，如下图：

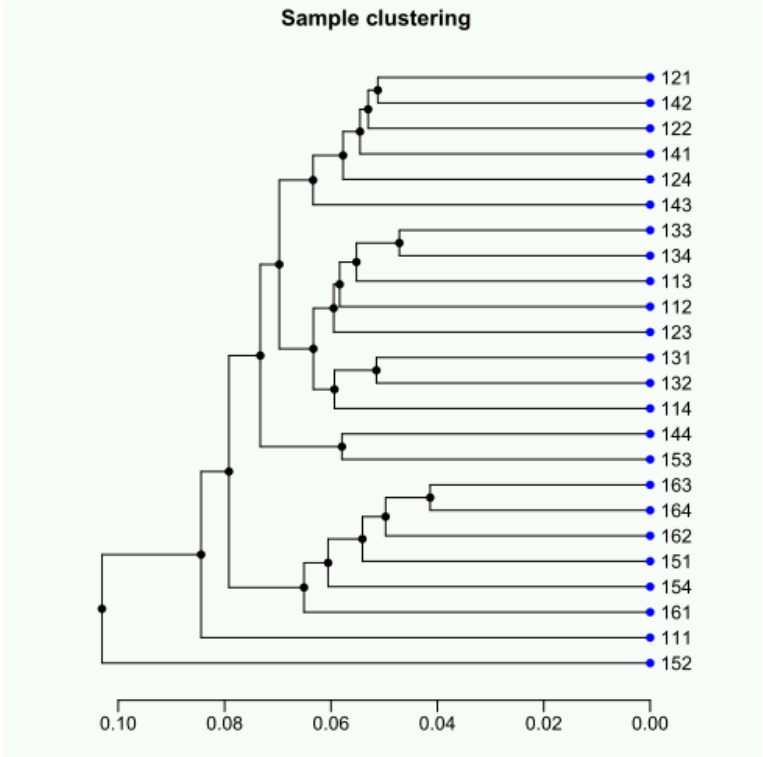


图 5.30 样本间聚类树图

说明：样本聚类图，图中每一个分支代表一个样本，长度值表示样本间的距离，样本间相似度越高，则在树图中越靠近。

5.6.3 样本间相关性分析

样品间基因表达水平相关性是检验实验可靠性和样本选择是合理性的重要指标。相关系数越接近 1，表明样品之间表达模式的相似度越高。若样品中有生物学重复，通常生物重复间相关系数要求较高。

结果目录：6_expression_profile/correlation_analysis/

All.genes.pearson.correlation.matrix.csv：各样本间 pearson 相关系数矩阵，结果如下表：

表 5.12 pearson 相关性系数矩阵

	162	164	163	161	152	153	151	154
162	1	0.974815997	0.974978148	0.9705087	0.959163516	0.966993748	0.973253823	0.970435462
164	0.974815997	1	0.979102	0.966917442	0.963540797	0.963803336	0.972577933	0.969251416
163	0.974978148	0.979102	1	0.96728142	0.965220528	0.965163643	0.973161403	0.969825448
161	0.9705087	0.966917442	0.96728142	1	0.954710482	0.96522001	0.967814442	0.967265068
152	0.959163516	0.963540797	0.965220528	0.954710482	1	0.953225691	0.963719539	0.9587678
153	0.966993748	0.963803336	0.965163643	0.96522001	0.953225691	1	0.965564544	0.967375381
151	0.973253823	0.972577933	0.973161403	0.967814442	0.963719539	0.965564544	1	0.971669752
154	0.970435462	0.969251416	0.969825448	0.967265068	0.9587678	0.967375381	0.971669752	1

注：上面只展示了 pearson 相关性系数矩阵结果，其它两类系数结果见相关文件夹。

A_vs_B.genes.correlation.pdf：样本间相关性分析图，结果如下：

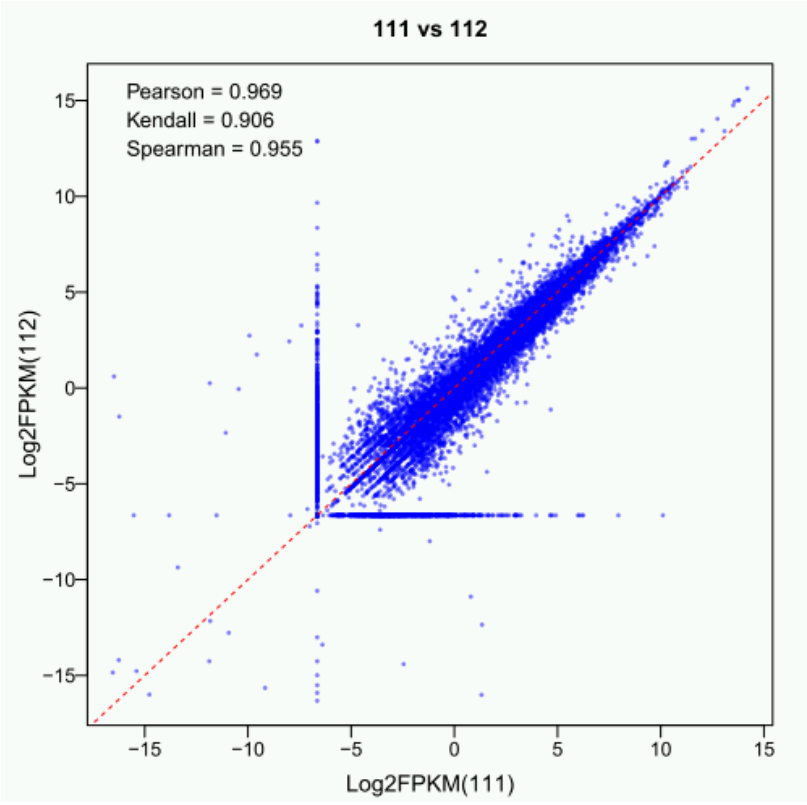


图 5.31 样本间相关性分析图

注：此处只展示了两样本间的结果，其它样本间分析见对应的文件夹。

说明：上图横坐标为样品 1 的 log2FPKM， 纵坐标为样品 2 的 log2FPKM，并计算了三个相关性系数，分别为 pearson、kendall、spearman。样本越相似，则上图中的大部分的点应集中在对角线附近。

5.6.4 样本间共同表达基因韦恩图

结果目录：6_expression_profile/VENN/

***.genes.venn.pdf：**样本间共同表达基因韦恩图，结果如下：

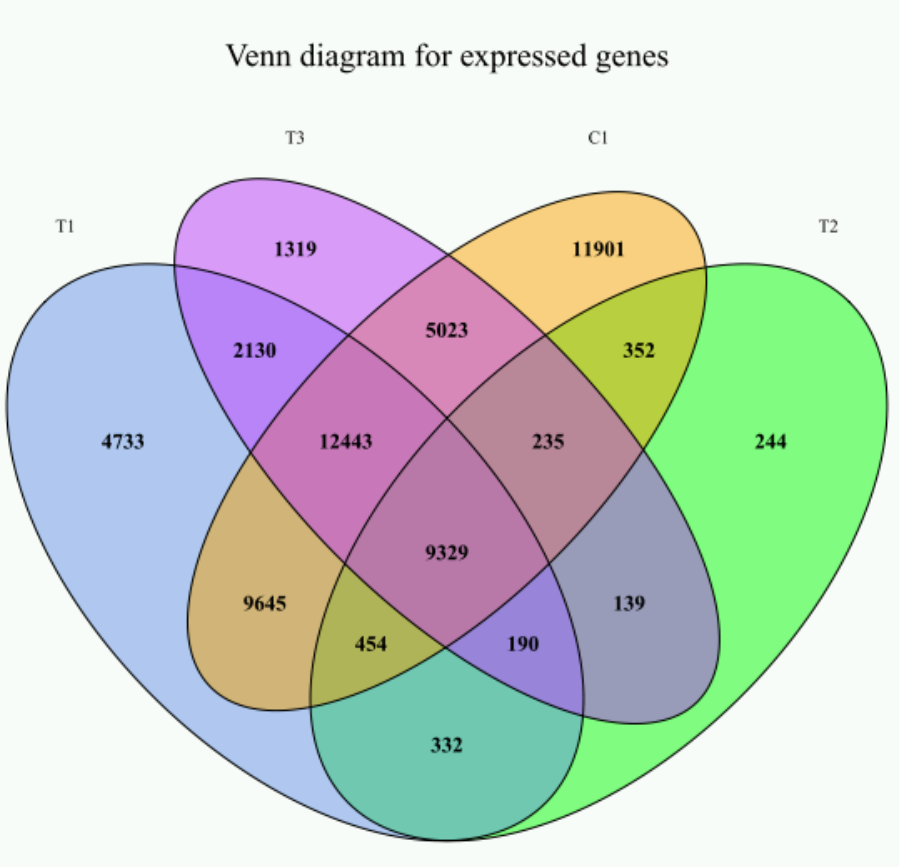


图 5.32 样本间共同表达基因韦恩图

注：此处只展示一组韦恩图，且最多只会做五个样本间的韦恩图，若样本数目超过 5 个则不做，其它韦恩图结果见对应文件夹。

5.6.5 PCA 分析

PCA 分析（Principal Component Analysis）即主成分分析，是一种对数据进行简化分析的技术，这种方法可以有效的找出数据中最“主要”的元素和结构，去除噪音和冗余，将原有的复杂数据降维，揭示隐藏在复杂数据背后的简单结构。通过 PCA 分析可以较好的找出样本间的关系以及主要影响样本间差异的一些基因。

结果目录：6_expression_profile/PCA/

All.genes.PCA.3dplot.pdf：前 3 主成分 3D 图，结果展示如下：

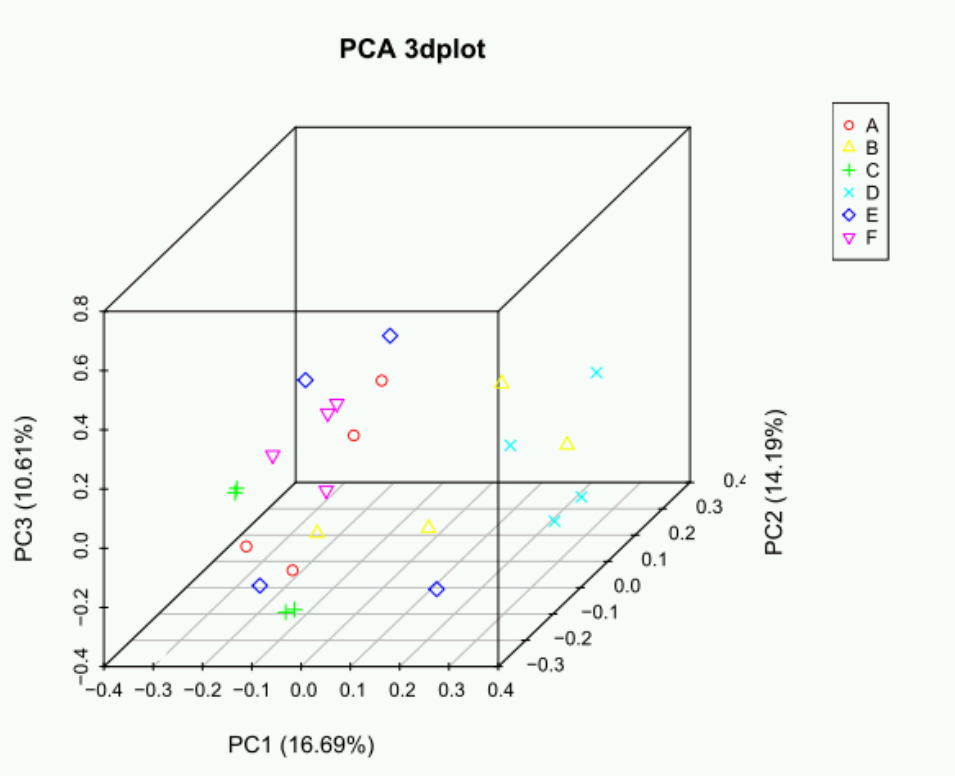


图 5.33 PCA 3Dplot

说明：PCA 三维散点图，图中不同颜色代表不同样本或者不同 group 中的样本，样本间相似度越高则在图中越聚集，反之样本间相似度越低则空间距离越远。

All.genes.PCA.2dplot.heatmap.pdf：前 3 主成分 2D 图，结果展示如下:

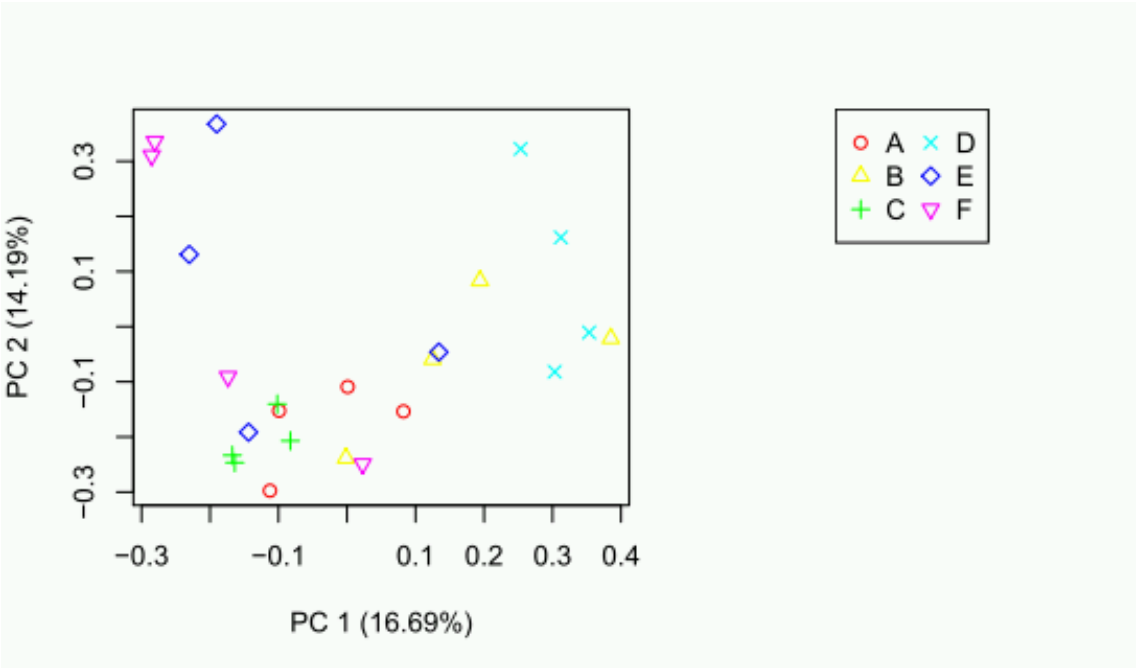


图 5.34 PCA 2Dplot (PC1 vs PC2)

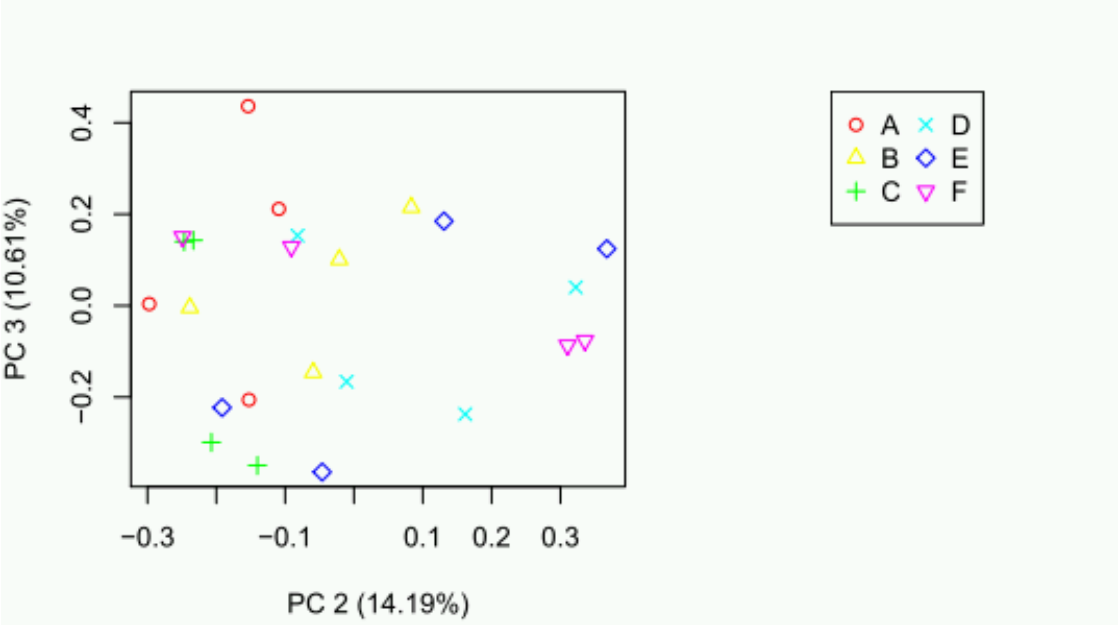


图 5.35 PCA 2Dplot (PC2 vs PC3)

5.7 SNP 分析（样本数大于等于 2 时才做）

5.7.1 方法说明

基于比对结果对各样本做 SNP/INDEL calling，采用软件为 samtools，得到 SN 及 INDEL 结果后对原始结果进行过滤，过滤条件为：1）QUAL 值大于 20；2）覆盖度大于 2。通过 SNP 结果可以找出不同品系间样本在 mRNA 层面的基因型差异，进而与表达量及表型关联起来。

采用软件：Samtools，bcftools（http://samtools.sourceforge.net/）。

参数设置：samtools mpileup -uD

5.7.2 结果展示

结果目录：7_SNP/

All.bam.filtered.vcf：各样本原始 SNP 结果 VCF 文件，vcf 文件格式详细介绍见 http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40。

All.SNP.Indel.count.pdf：所有样本 SNP、INDEL 数目条形图，绘图源文件为 All.snp.indel.count.xls，结果展示如下图：

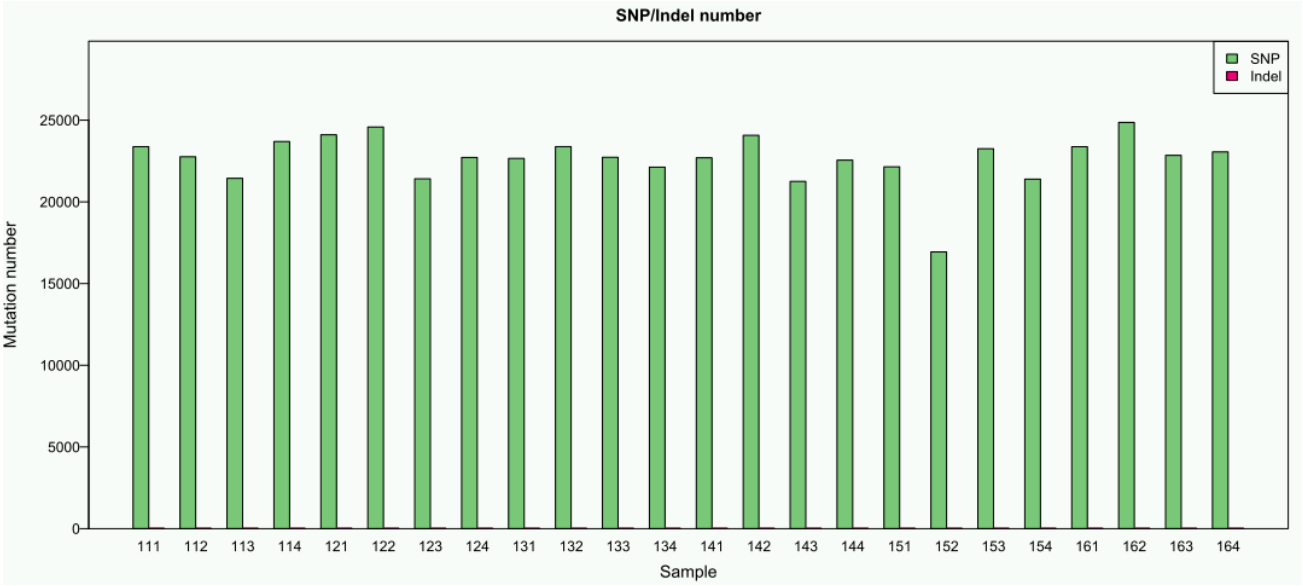


图 5.36 各样本 SNP 数目统计图

***/*.snp.density.pdf：**某样本 SNP 密度度，结果展示如下：

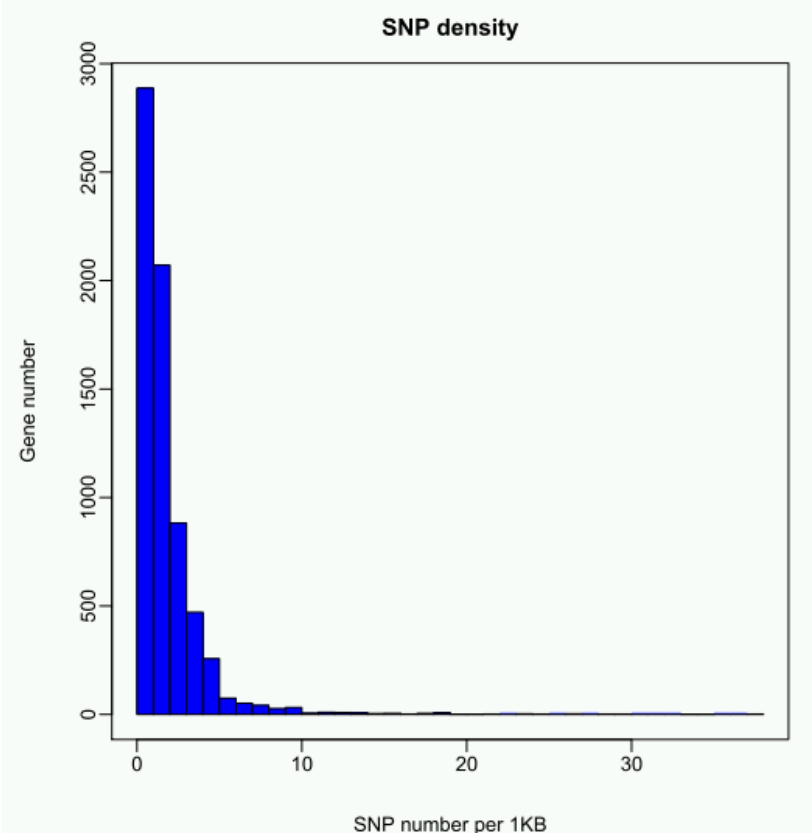


图 5.37 单样本 SNP 密度直方图

说明：上图横坐标表示每 1KB 里面 SNP 的数目，纵坐标为对应的基因数目

***/*.mutation.spectrum.pdf：**各样本突变谱系图，结果展示如下：

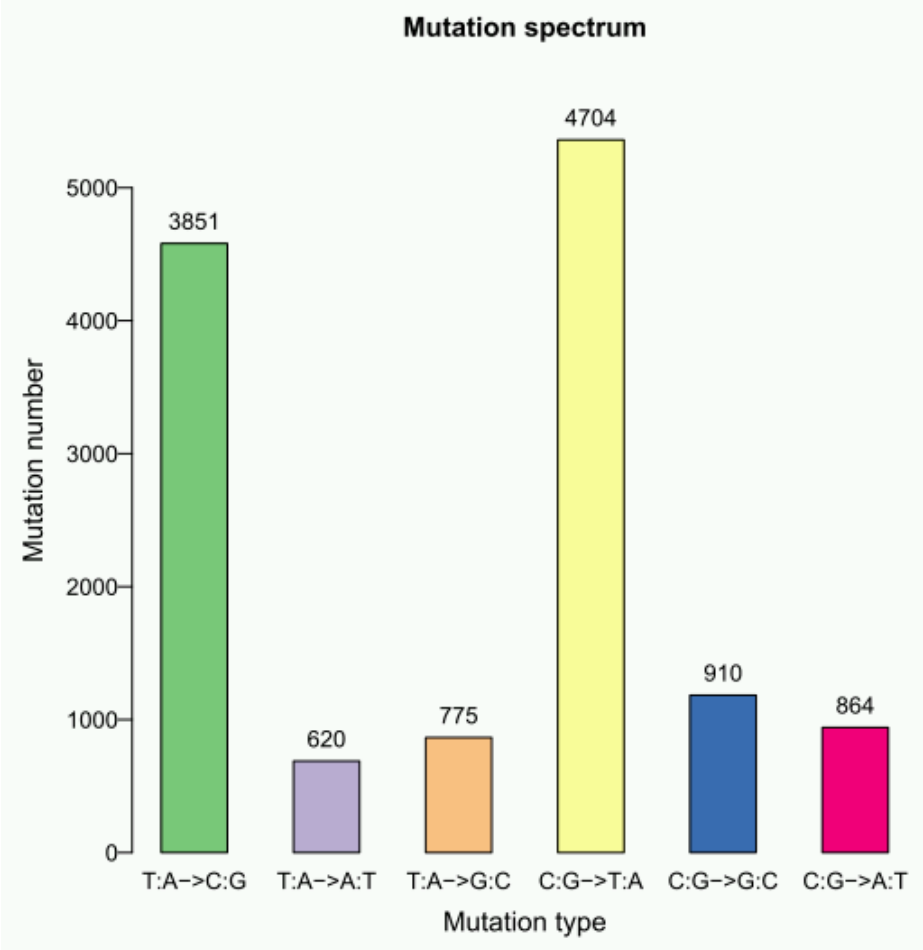


图 5.38 单样本突变谱系图

注：在 SNP 结果文件夹中每个样本分别会有自身样本名对应的文件夹，上面展示的只是其中一个样本的结果，其他样本结果见对应的文件夹。

5.8 差异表达分析

注：以下全部展示的均为基因层面，转录本层面的结果在对应的文件夹里面均有，带 isoforms 命名的均为转录本层面结果。

5.8.1 方法说明

无生物学重复样本分析方法如下：

参照 Audic S.等人发表在 Genome Research 上的基于测序的差异基因检测方法{Audic, 1997 #8}（该文献已被引用超过五百次）。假设观测到基因 A 对应的 reads 数为 x，已知在一个大文库中，每个基因的表达量只占所有基因表达量的一小部分，在这种情况下，p(x)的分布服从泊松分布：

$$p(x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$$ （λ 为基因 A 的真实转录数）

已知，样本一中唯一比对上总 reads 数为 N1，样本二中比对上的总 reads 数为 N2，样本一中比对到基因 A 的总 reads 数为 x，样本二中比对到基因 A 的总 reads 数为 y，则基因 A 在两样本中表达量相等的概率可由以下公式计算：

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

$$或 2\times(1-\sum_{i=0}^{i=y} p(i|x)) \quad (如果 \sum_{i=0}^{i=y} p(i|x)>0.5)$$

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

然后，我们对差异检验的 p value 作多重假设检验校正，采用的方法为 FDR，在我们的分析中，差异表达基因定义为 p≤0.01 且倍数差异在 2 倍以上的基因。

有生物学重复样本筛选方法如下：

采用 DEseq 进行差异分析,筛选阈值为 qvalue<0.001 且|FoldChange|>2。

差异分析软件：DESeq、edgeR（http://www.bioconductor.org/）

5.8.2 结果展示

结果目录：8_DEGs_analysis/

genes.DEGs.num.xls：差异基因统计结果，结果如下：

表 5.13 差异基因数目统计

Compare	UP	DOWN	ALL
C1_vs_T1	1337	1380	2717
C1_vs_T2	450	678	1128
C1_vs_T3	446	429	875
T1_vs_T2	269	426	695
T1_vs_T3	184	648	832
T2_vs_T3	208	284	492

All.genes.DEGs.count.pdf：差异基因数目统计条形图，结果展示如下：

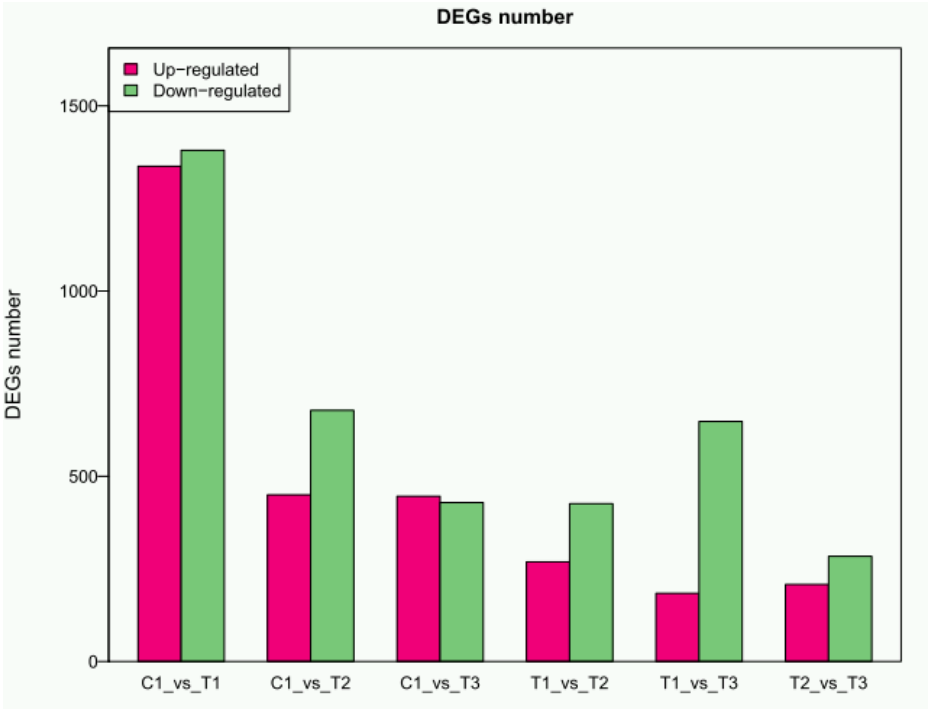


图 5.39 差异基因数目条形图

A_vs_B/*.genes.FPKM.boxplot.pdf：比较对样本表达盒状图，结果如下：

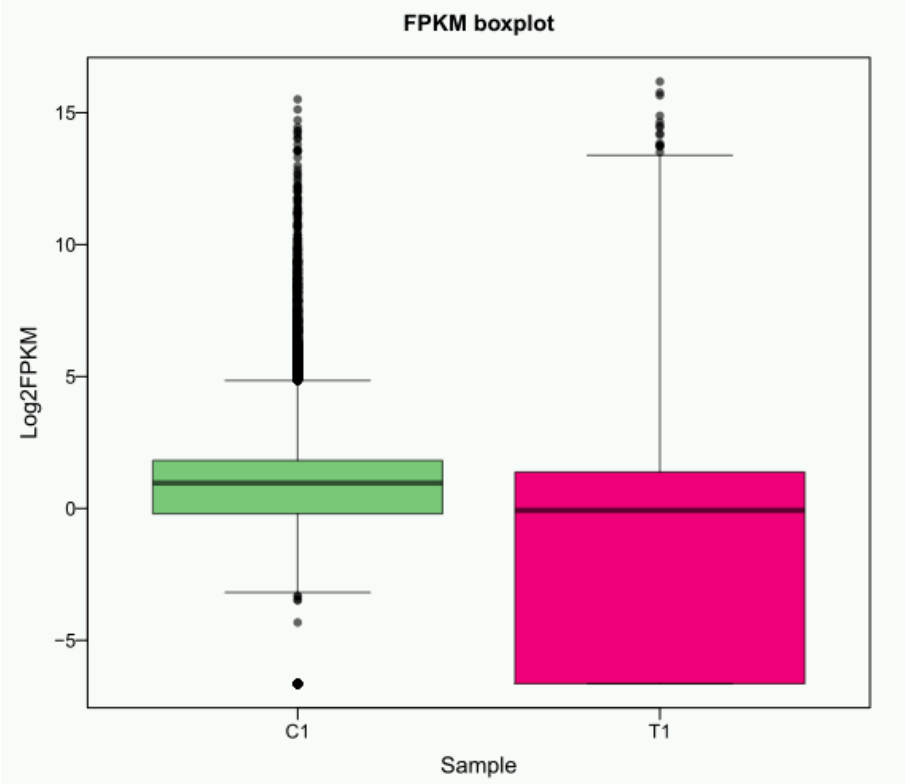


图 5.40 比较对间样本表达量盒状图

注：上述展示的只是一组比较对间的结果，若有多组比较，到对应的比较对文件夹中可以找到相应的结果，若只有一组比较此处展示的结果与图 5.40 一样，下同。

A_vs_B/*.genes.FPKM.density.pdf：比较对间样本表达密度曲线图，结果如下：

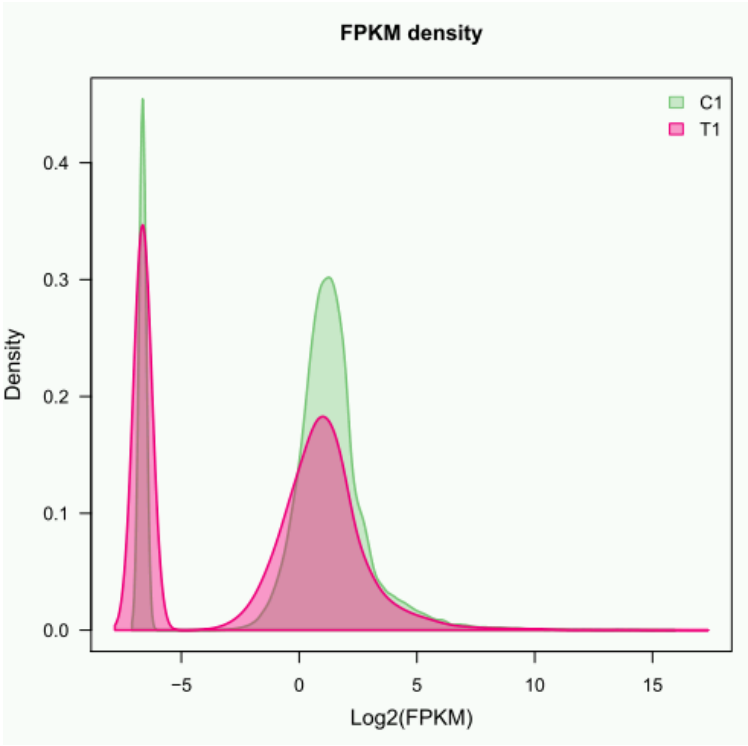


图 5.41 比较对间样本表达密度曲线

A_vs_B/.genes.FPKM.Scatter.plot.pdf：比较对间样本件表达量散点图，结果如下图：

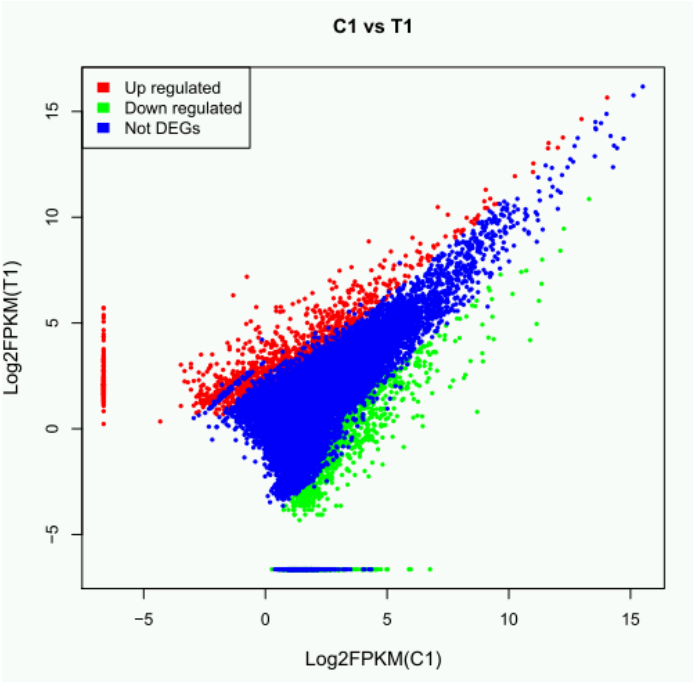


图 5.42 比较对间样本件表达量散点图

说明：上图为样本间表达量散点图，每一个点代表一个基因，横纵坐标分别表示 log2(FPKM)值，若为有生物学重复样本则 X/Y 轴的值为 Log2(Mean FPKM)，即为 log2(生物学重复 FPKM 的均值)。其中红色表示上调基因，绿色表示下调基因，蓝色表示非差异表达基因，上调/下调均是 Y 轴样本相对于 X 轴样本。

A_vs_B/.genes. MA.plot.pdf：比较对样本间表达量 MA 图，结果展示如下：

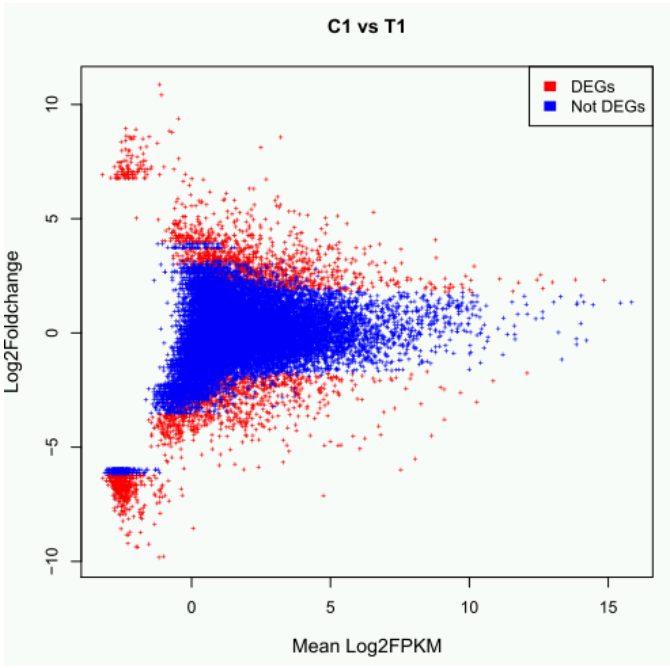


图 5.43 比较对样本间表达量 MA 图

说明：横坐标 X 轴表示 log 均值，即（log2(A)+log2(B)）/2，纵坐标为代表 log（Foldchange），即 log2(B/ A)，各个数据点红色代表筛选出的差异基因，蓝色代表非差异基因。

A_vs_B/*.genes.volcano.plot.pdf：火山图，结果展示如下：

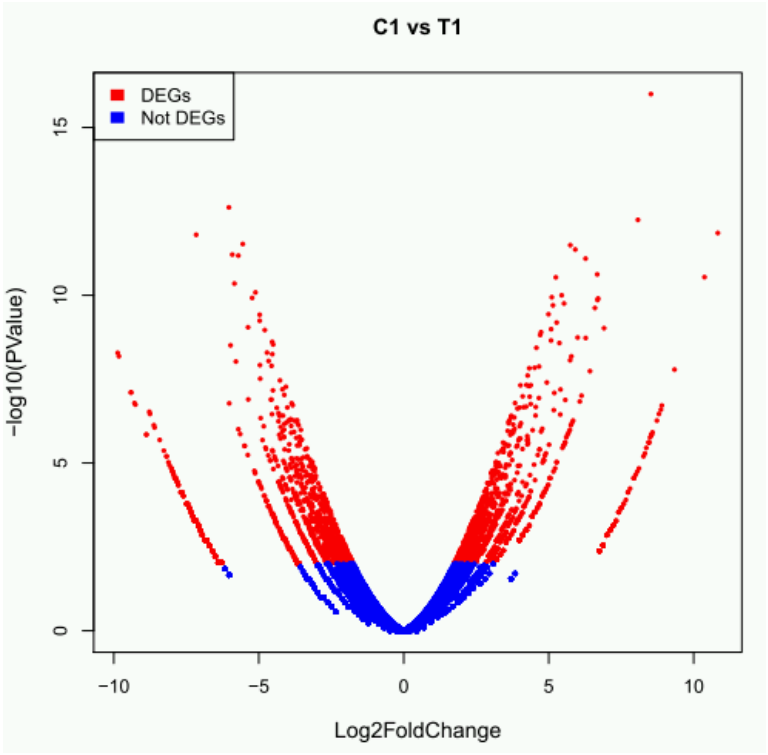


图 5.44 差异分析火山图

说明：横坐标代表基因在不同实验组中/不同样品中表达倍数变化；纵坐标代表基因表达量变化的统计学显著程度，p-value 越小，-log10(p-value)越大，即差异越显著。图中的散点代表各个基因，蓝色圆点表示无显著性差异的基因，红色圆点表示有显著性差异的基因，火山图可以直观展现 pvalue 与 log2（foldchange）的关系。

VENN/*.genes.all.venn.pdf：指定的比较对间差异表达基因韦恩图，结果展示如下：

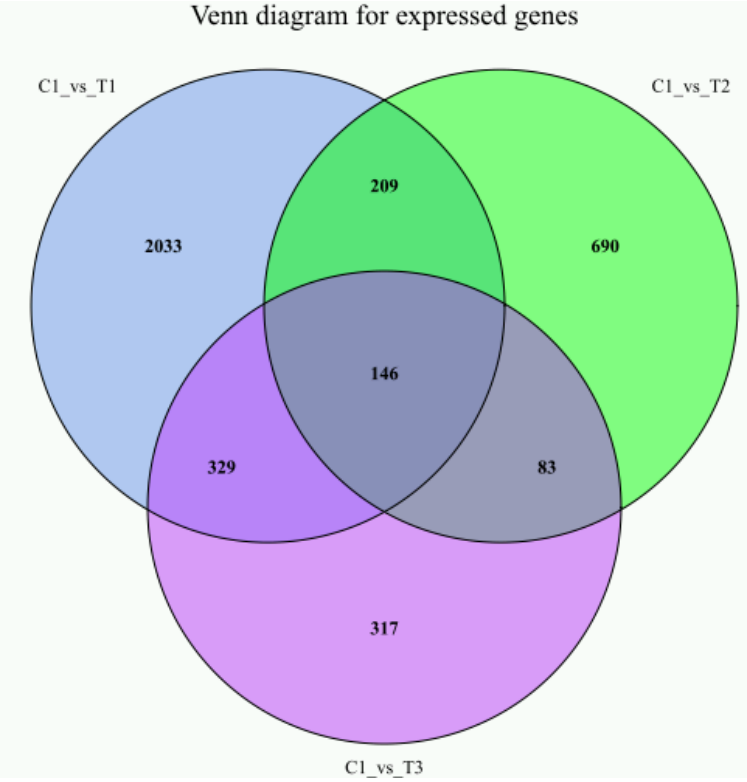


图 5.45 差异基因韦恩图

注：默认为所有比较对间做韦恩图，当比较对数目大于五时或未指定，则该项分析不会做，上图展示的为所有差异基因，上调与下调差异基因韦恩图见相关文件夹。

说明：上图展示的为特定比较对间差异基因的韦恩图，通过该图可以看出不同比较对间差异基因的异同。

5.9 差异基因表达模式聚类分析

5.9.1 方法说明

差异基因聚类分析用于判断不同实验条件下差异基因表达量的聚类模式。每个比较组合都会得到一个差异基因集，将所有比较组合的差异基因求并集，获得该基因集在每个样品中的 FPKM 值，做后续聚类分析，获得表达模式相近的基因集。

5.9.2 结果展示

结果目录：8_DEGs_analysis/*_DEGs_cluster/genes

All_DEGs_samples_heatmap.pdf：所有差异基因表达聚类热图(指定比较组别)，结果展示如下：

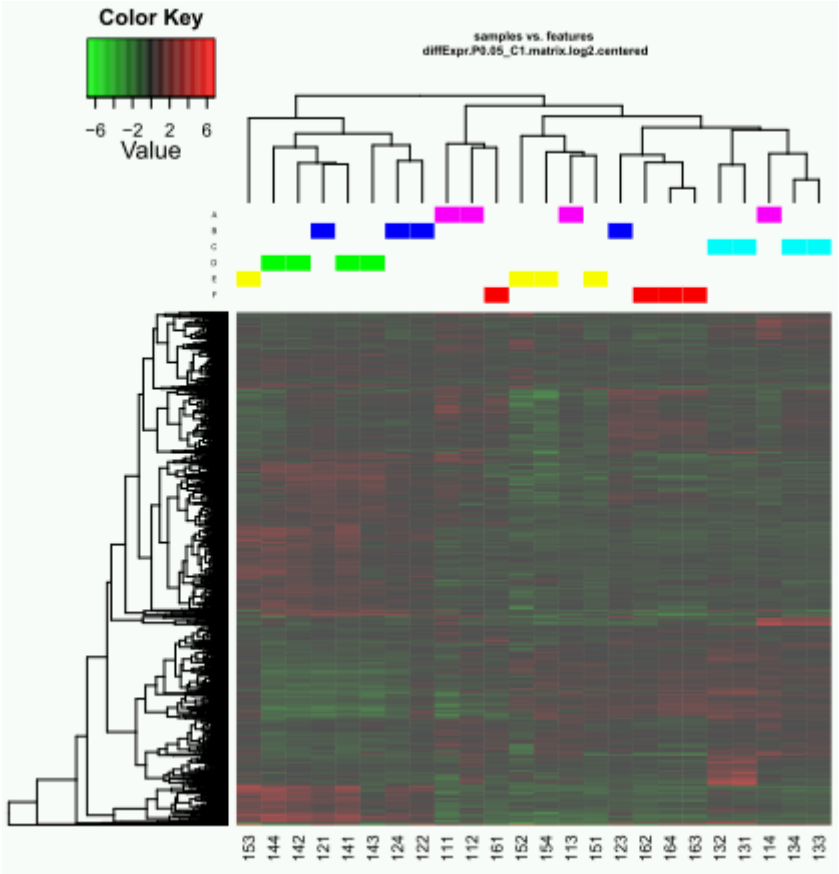


图 5.46 表达模式聚类热图

说明：热图，所有差异表达基因表达量热图，图中每行代表一个基因，每列代表一个样本，颜色表示表达量高低，越红表示表达量越高，反之越绿表示表达量越低。图中分别对样本及基因做聚类，相似的样本会聚在一起，另表达模式相近的基因亦会聚在一起，如图左侧的距离结果。聚类树下面的颜色块表示 group，颜色相同说明这些样本为同一 group 或者为生物学重复。

All_DEGs_sample_cor_matrix.pdf：样本相关性热图，该结果基于所有差异表达基因，展示如下：

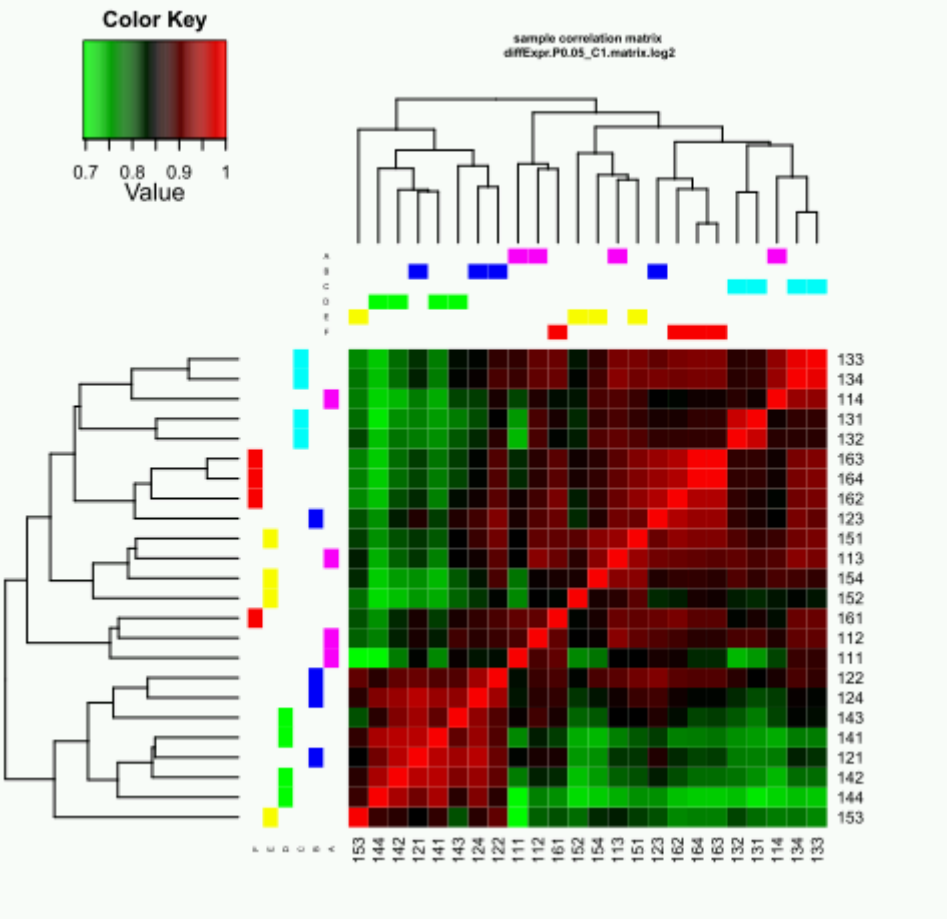
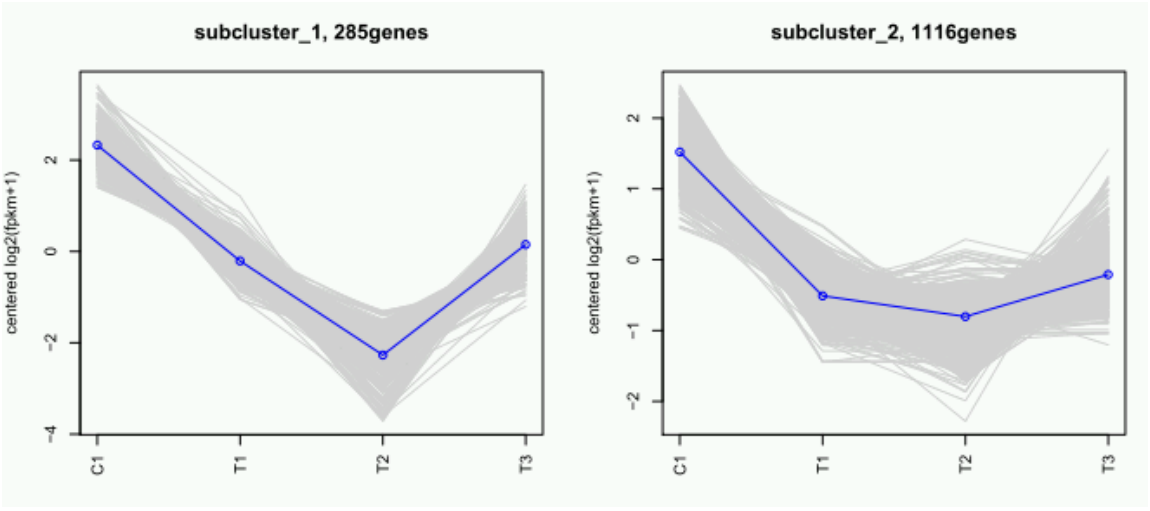


图 5.47 表达相关性热图

说明：样本相关性热图，图中行列代表样本，每一格表示两样本间的相关性，颜色越红表示样本间相关性越高，越相似，反之越绿表示相关性越低。聚类树旁边的颜色块表示 group，颜色相同说明这些样本为同一 group 或者为生物学重复。

DEGs_cluster_plot.pdf：基因集表达量散点图，表达模式相近的基因聚为一类，归为一个 cluster，结果展示如下：



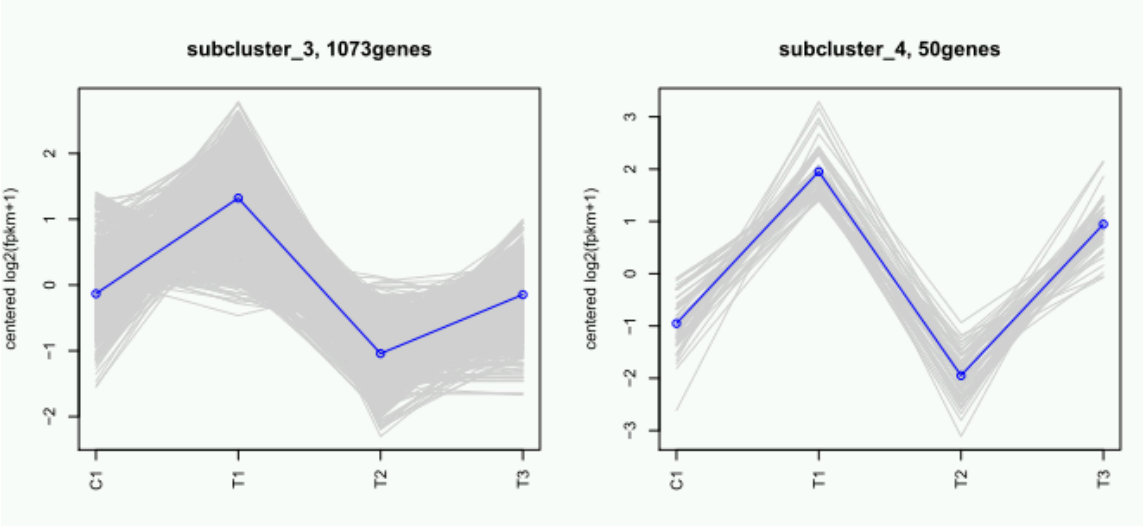


图 5.48 前 4 个 cluster 中基因在各样本中表达量折线图

说明：图中一条折线表示一个基因在不同样本中的表达量值，图中看出每个 cluster 下面的所有基因在所有样本中表达模式均类似。

All_DEGs_genes_foldchange_heatmap.pdf：所有差异表达基因 log2(foldchange)热图，结果展示如下：

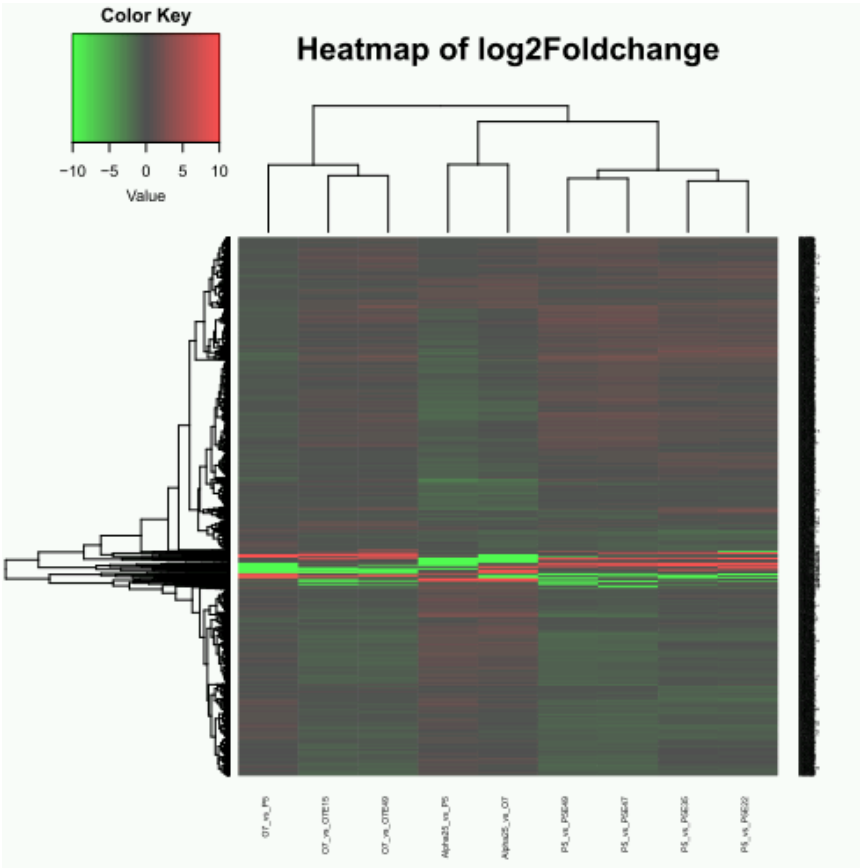


图 5.49 foldchange 热图

注：当比较对大于两组时此图才会生成，只有一组比较时此图没有。

说明：上图中红色表示上调表达，绿色表示下调表达，颜色越红表示上调倍数越高，颜色越绿表示下调倍数越高，每一行代表一个基因，每一列代表一组比较对。

5.10 差异基因 GO 富集分析

5.10.1 方法说明

Gene Ontology（简称 GO, http://www.geneontology.org/）是基因功能国际标准分类体系。根据实验目的筛选差异基因后，研究差异基因在 Gene Ontology 中的分布状况将阐明实验中样本差异在基因功能上的体现。GO 富集分析方法为 GOseq（Young et al, 2010）,此方法基于 Wallenius non-central hyper-geometric distribution 。相对于普通的 Hyper-geometric distribution，此分布的特点是从某个类别中抽取个体的概率与从某个类别之外抽取一个个体的概率是不同的，这种概率的不同是通过对基因长度的偏好性进行估计得到的，从而能更为准确地计算出 GOterm 被差异基因富集的概率。

5.10.2 结果展示

结果目录：9_GO_enrichment/，每个比较对在这里面均会有对应的文件夹

A_vs_B/*.genes.all_GO_enrichment.xls：所有差异表达基因 GO 富集分析列表，结果如下

表 5.14 GO 富集分析结果

GO_ID	Term	Type	DEGs_thi	UP	Down	Expected	Pvalue	FDR
GO:0032501	multicellular organismal process	biological_process	489	147	342	383.79	3.90E-11	2.90E-07
GO:0043292	contractile fiber	cellular_component	52	10	42	19.94	1.30E-10	4.83E-07
GO:0030154	cell differentiation	biological_process	289	103	186	206.59	2.70E-10	6.69E-07
GO:0030016	myofibril	cellular_component	49	9	40	19.03	6.90E-10	1.28E-06
GO:0044449	contractile fiber part	cellular_component	45	9	36	17.6	4.30E-09	5.63E-06
GO:0044707	single-multicellular organism process	biological_process	465	146	319	373.01	4.80E-09	5.63E-06
GO:0007275	multicellular organismal development	biological_process	398	126	272	310.93	5.30E-09	5.63E-06
GO:0044767	single-organism developmental process	biological_process	444	136	308	356.89	1.90E-08	1.77E-05
GO:0030017	sarcomere	cellular_component	41	8	33	16.32	3.40E-08	2.30E-05
GO:0032982	myosin filament	cellular_component	17	6	11	3.63	3.40E-08	2.30E-05
GO:0048731	system development	biological_process	342	113	229	264.15	3.40E-08	2.30E-05
GO:0032502	developmental process	biological_process	444	136	308	359.23	4.50E-08	2.79E-05
GO:0061061	muscle structure development	biological_process	82	34	48	44.9	6.60E-08	3.77E-05
GO:0048856	anatomical structure development	biological_process	399	124	275	318.62	7.70E-08	4.09E-05
GO:0060537	muscle tissue development	biological_process	56	19	37	26.75	9.80E-08	4.65E-05
GO:0055001	muscle cell development	biological_process	33	12	21	12.13	1.00E-07	4.65E-05
GO:0048869	cellular developmental process	biological_process	303	109	194	231.83	1.20E-07	5.25E-05
GO:0042692	muscle cell differentiation	biological_process	54	20	34	25.77	1.60E-07	6.61E-05
GO:0031674	I band	cellular_component	31	7	24	11.25	1.90E-07	7.43E-05
GO:0048523	negative regulation of cellular process	biological_process	305	98	207	235.67	2.70E-07	0.0001

注：上述展示的只是富集分析中前 20 的 GO，且为所有差异基因富集分析结果，UP/Down 基因分别的富集分析结果见对应的文件夹。

GO_ID：GO ID

Term：GO 名字

Type：GO 功能类

DEGs_this_term：该功能类下的差异基因数目

UP：该功能类下上调基因数目

Down：该功能类下下调基因数目

Pvalue：富集分析 P 值，P 值越小越显著

FDR：P 值校正值

***.genes.all_GO_enrichment_barplot.pdf：**GO 富集分析前 50 个 GO 差异基因数目条形图，结果展示如下图：

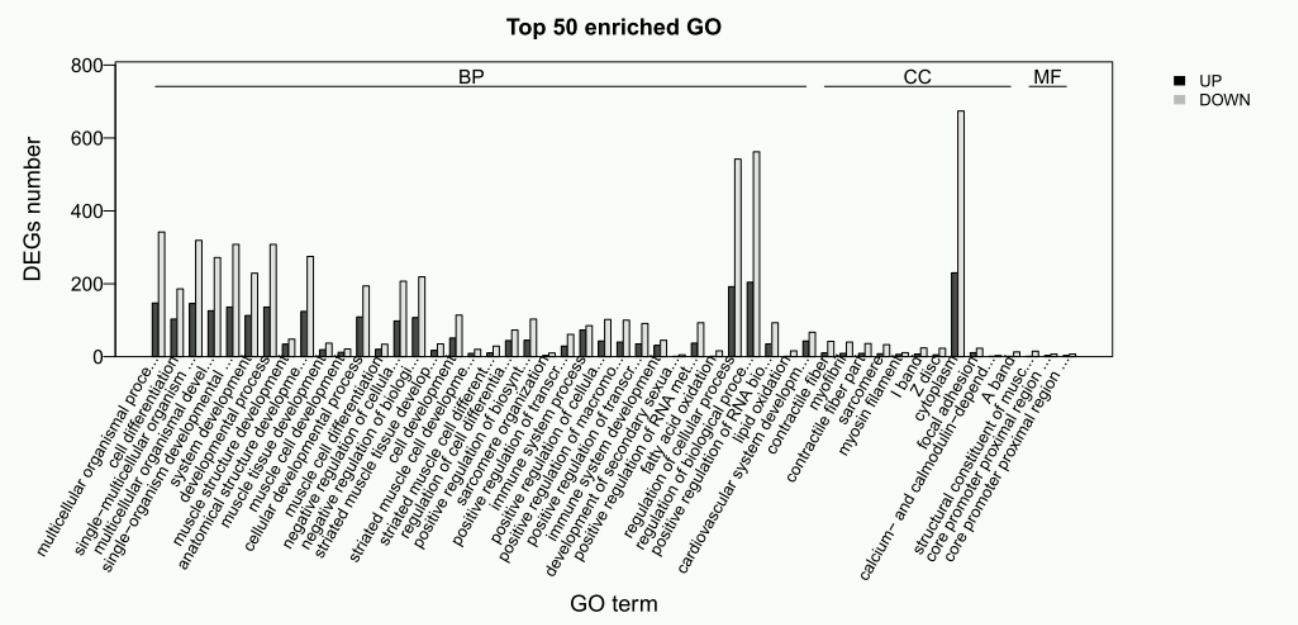


图 5.50 前 50 个 GO 差异基因数目条形图

说明：横轴为前 50 个显著富集的 GO 名称，纵坐标为差异基因数目，黑色条为上调基因数目，灰色条为下调基因数目，该图可以反应出所有差异基因在富集的 GO 中的分布。从左到右分别表示 BP,CC,MF。

***.genes.all_GO_enrichment_scatterPlot.pdf：**所有差异基因 GO 富集分析前 30 个 GO 富集散点图，结果如下：

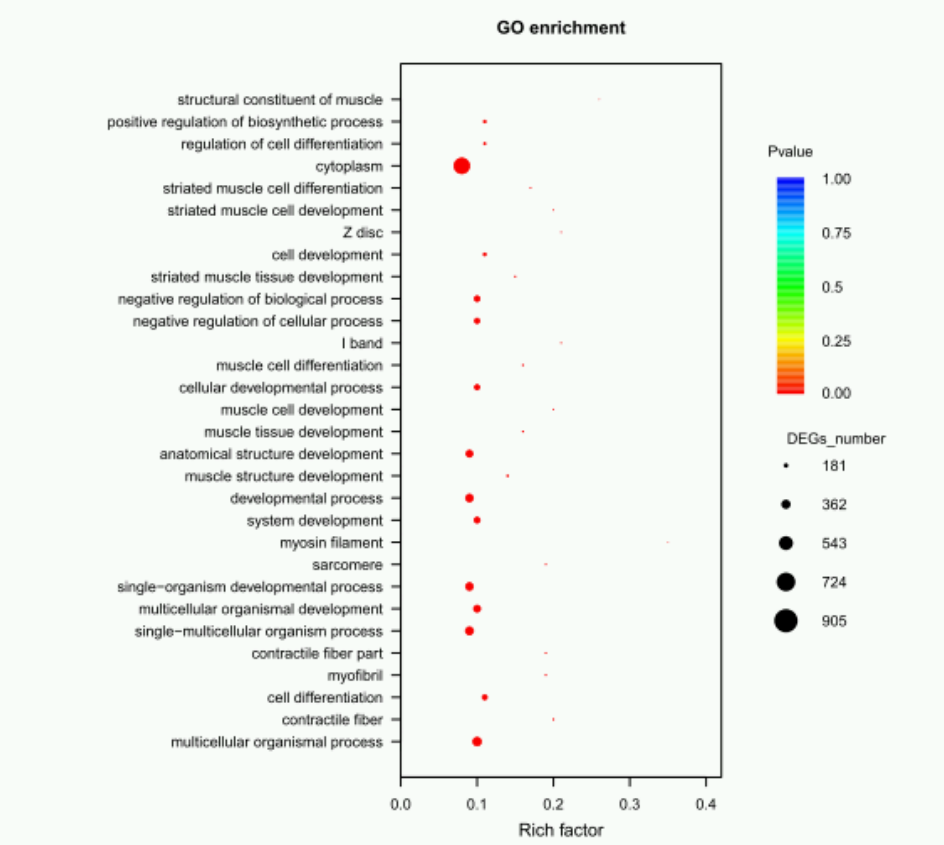
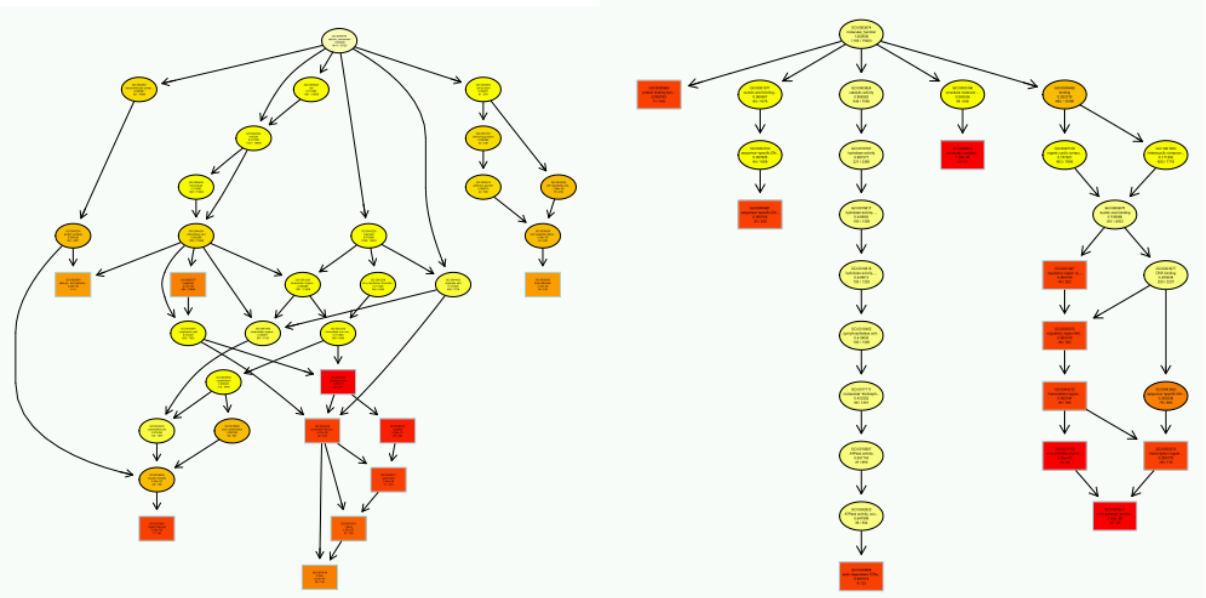


图 5.51 GO 富集分析前 30 个 GO 富集散点图

说明：纵轴表示 GO 名称，横轴表示 GO 对应的 Rich factor，Pvalue 的大小用点的颜色来表示，Pvalue 越小则颜色越接近红色，每个 GO 下包含的差异基因的多少用点的大小来表示。

***.genes.all_*_classic_5_all.pdf：**topGO 有向无环图



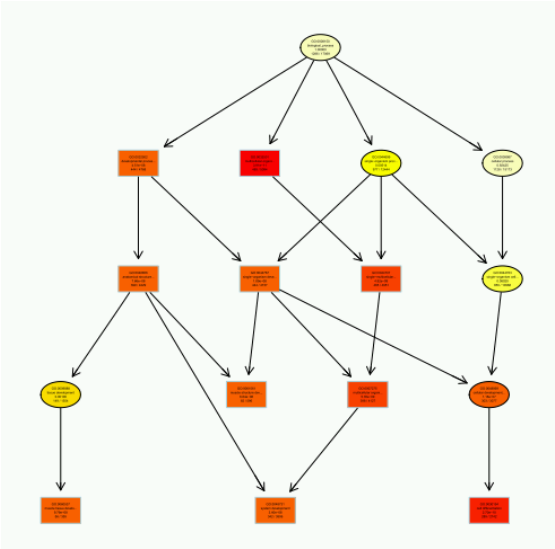


图 5.52 GO 富集有向无环图

说明：topGO 有向无环图能直观展示差异基因富集的 GO term 及其层级关系。有向无环图为差异基因 GO 富集分析的结果图形化展示方式，分支代表包含关系，从上至下所定义的功能范围越来越具体。对 GO 三大分类（CC 细胞成分，MF 分子功能，BP 生物学过程）的每一类都取富集程度最高的前 5 位作为有向无环图的主节点，用方框表示，并通过包含关系将相关联的 GO Term 一起展示，颜色的深浅代表富集程度，颜色越深代表富集程度越高。每个方框或圆圈代表一个 GO term，放大方框中内容从上到下代表的含义依次为:GO term 的 id、GO 的描述、GO 富集的 Pvalue、该 GO 下差异基因的数目/该 GO 下背景基因的数目。每组比较三张图（BP,CC,MF）。

***genes.enriched.GO.heatmap2.pdf**：所有比较组 GO 富集 Pvalue 热图，该图通过对所有比较组（或指定比较组）显著富集的 GO 的 P 值做热图（默认为 p<0.01，可调整），结果如下：

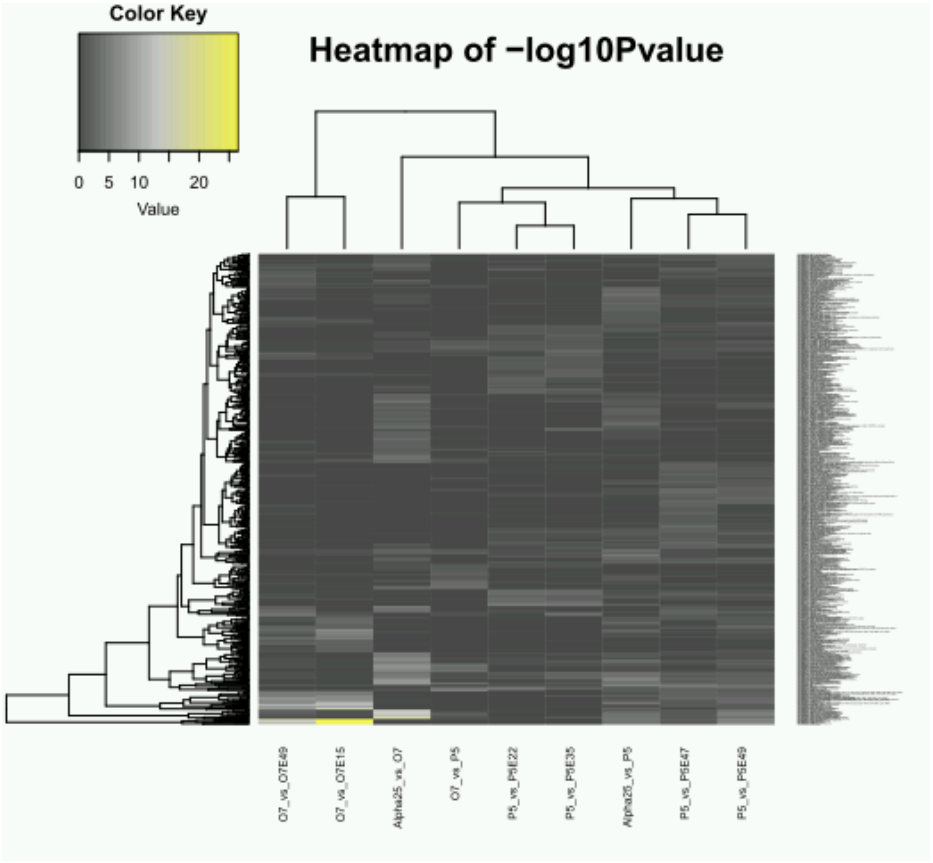


图 5.53 所有显著富集的 GO pvalue 热图

注：当比较对大于两组时此图才会生成，只有一组比较时此图没有。

说明：上图中每一行代表一个 GO term，每一列为一组比较组，颜色越黄表示越显著，即 P 值越小，上图反应出在不同比较对间富集的 GO 差异，尤其当样本为时间序列样本时可以很好的看出在不同时间段差异表达基因功能的差异。

5.11 差异基因 KEGG 富集分析

5.11.1 方法说明

在生物体内，不同基因相互协调行使其生物学功能，通过 Pathway 显著性富集能确定差异表达基因参与的最主要生化代谢途径和信号转导途径。KEGG（Kyoto Encyclopedia of Genes and Genomes）是有关 Pathway 的主要公共数据库（Kanehisa,2008）。Pathway 显著性富集分析以 KEGG Pathway 为单位，应用超几何检验，找出与整个基因组背景相比，在差异表达基因中显著性富集的 Pathway。该分析的计算公式如下：

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

在这里 N 为所有基因中具有 Pathway 注释的基因数目；n 为 N 中差异表达基因的数目；M 为所有基因中注释为某特定 Pathway 的基因数目；m 为注释为某特定 Pathway 的差异表达基因数目。P≤0.05 的 Pathway 定义为在差异表达基因中显著富集的 Pathway。

5.11.2 结果展示

结果目录：10_kegg_enrichment/ 每个比较对在这里面均会有对应的文件夹

*.genes.all_kegg_enrichment.xls：所有差异基因 KEGG 富集分析结果，结果如下：

表 5.15 pathway 富集分析结果

KO_ID	Term	DEGs_this_term	UP	Down	Pvalue	FDR
ko05020	Prion diseases	8	5	3	0.001034	0.180191
ko04020	Calcium signaling pathway	18	5	13	0.003128	0.180191
ko04310	Wnt signaling pathway	18	5	13	0.003128	0.180191
ko04530	Tight junction	16	4	12	0.003363	0.180191
ko04750	Inflammatory mediator regulation of TRP channels	12	1	11	0.004629	0.180191
ko05322	Systemic lupus erythematosus	11	11	0	0.004847	0.180191
ko04919	Thyroid hormone signaling pathway	16	3	13	0.005062	0.180191
ko04911	Insulin secretion	11	1	10	0.005796	0.180191
ko04261	Adrenergic signaling in cardiomyocytes	17	5	12	0.006556	0.180191
ko04713	Circadian entrainment	12	4	8	0.008752	0.180191
ko04971	Gastric acid secretion	9	1	8	0.008998	0.180191
ko05150	Staphylococcus aureus infection	8	8	0	0.009041	0.180191
ko04921	Oxytocin signaling pathway	18	5	13	0.00907	0.180191
ko05166	HTLV-I infection	29	13	16	0.009696	0.180191
ko05310	Asthma	5	5	0	0.010334	0.180191
ko01212	Fatty acid metabolism	9	1	8	0.010839	0.180191
ko03320	PPAR signaling pathway	11	7	4	0.012922	0.198418
ko04360	Axon guidance	14	6	8	0.013596	0.198418
ko05031	Amphetamine addiction	10	1	9	0.014173	0.198418
ko04912	GnRH signaling pathway	12	1	11	0.0153	0.203487

注：上述展示的只是富集分析中前 20 的 pathway，且为所有差异基因富集分析结果，UP/Down 基因的富集分析结果见对应的文件夹。

KO_ID：KO ID

Term：pathway 名称

All_num_this_term：注释到该通路上的所有基因

DEGs_this_term：该功能类下的差异基因数目

UP：该功能类下上调基因数目

Down：该功能类下下调基因数目

Pvalue：富集分析 P 值，P 值越小越显著

FDR：P 值校正值

***.genes.all_kegg_enrichment_scatterPlot.pdf**：所有差异基因 pathway 富集分析前 30 个富集散点图，结果如下：

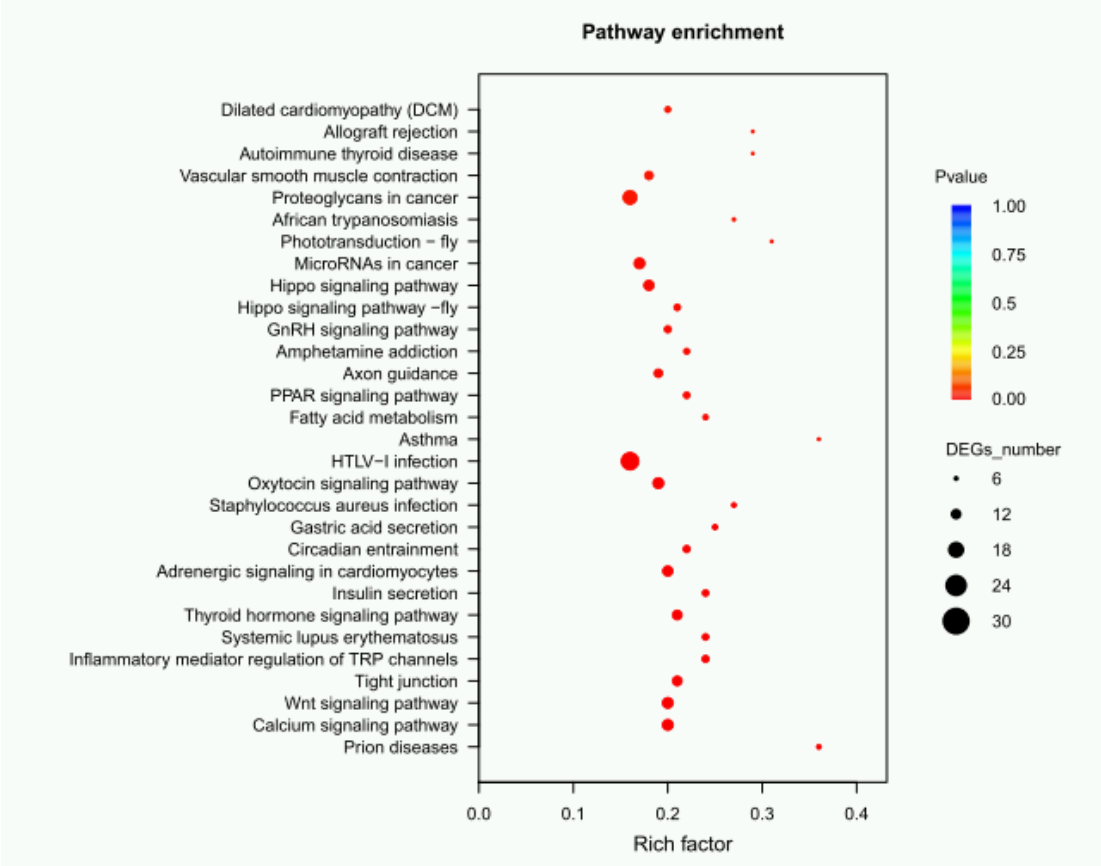


图 5.54 pathway 富集分析前 30 个富集散点图

说明：纵轴表示 pathway 名称，横轴表示 pathway 对应的 Rich factor，Pvalue 的大小用点的颜色来表示，Pvalue 越小则颜色越接近红色，每个 pathway 下包含的差异基因的多少用点的大小来表示。

***genes.enriched.kegg.heatmap2.pdf**：所有比较组 kegg 富集 Pvalue 热图，该图通过对所有比较组显著富集的 kegg 的 P 值做热图（默认为 p<0.05，可调整），结果如下：

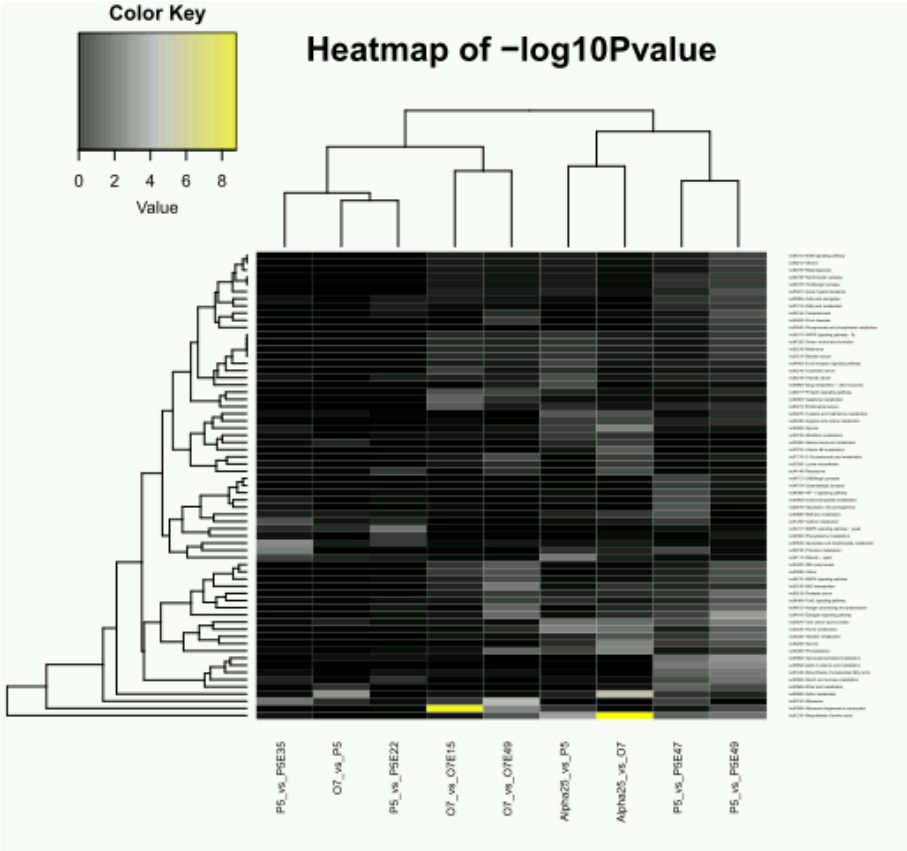


图 5.55 所有显著富集的 pathway pvalue 热图

注：当比较对大于两组时此图才会生成，只有一组比较时此图没有。

说明：上图中每一行代表一个 pathway，每一列为一组比较组，颜色越黄表示越显著，即 P 值越小，上图反应出在不同比较对间富集的 pathway 差异，尤其当样本为时间序列样本时可以很好的看出在不同时间段差异表达基因功能的差异。

***.pathway.tgz：**差异基因代谢通路图，里面为所有代谢通路图，并对差异基因进行上色，如下图：

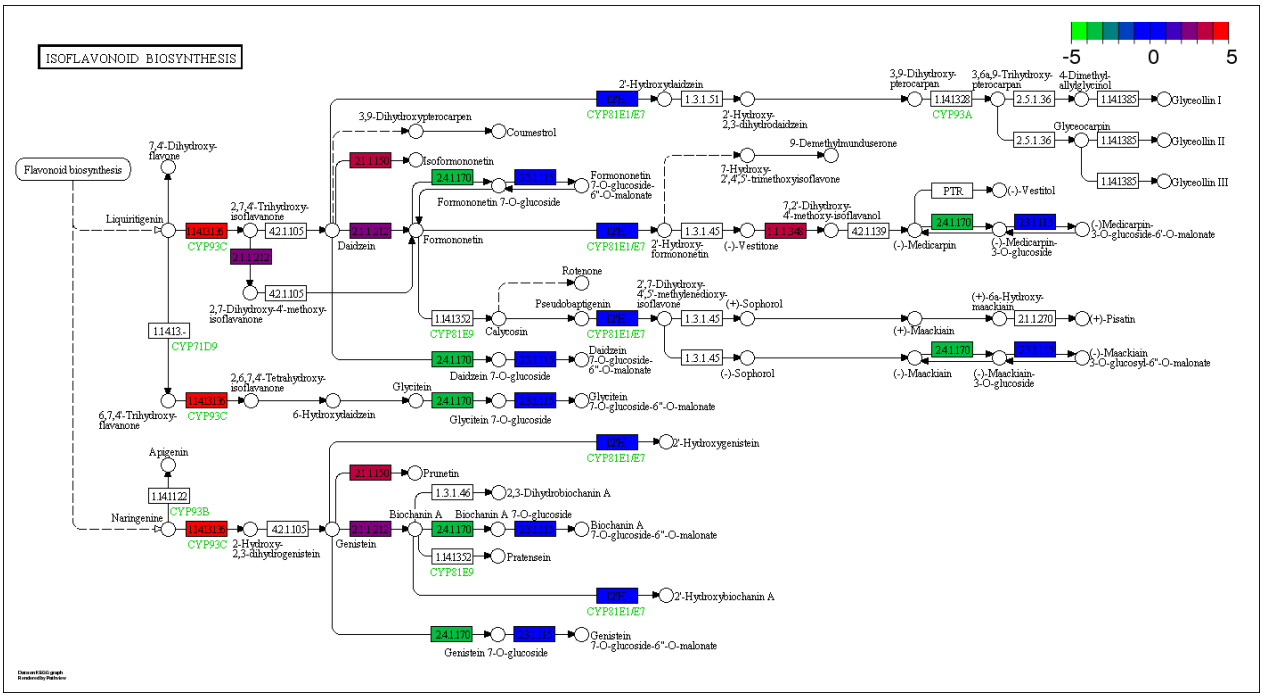


图 5.56 代谢通路上色图

注：上图只展示某一比较对中最显著富集的代谢通路

说明：上图中所有有颜色的表示该物种在该通路上注释上的基因，颜色越红表示上调差异倍数越大，颜色越绿表示下调差异倍数越大，蓝色表示非差异表达基因。

6. 结果说明

1_QC/ 原始数据及 QC 结果，此文件夹中会有对应样本的子文件夹，里面存放了各样本原始数据统计及 QC 结果。

All_sample_raw_data_infor.xls：所有样本原始数据统计结果，格式说明见上报告 5.1

All_sample_QC_infor.xls：所有样本 QC 之后结果统计， 格式说明见上报告 5.1

*/PE_trimed_infor.xls：样本对应的 QC 之后结果统计，格式同上

*/Raw_data_infor.xls：样本对应原始数据统计，格式同 All_sample_raw_data_infor.xls

*/*_rand_100000.fa_blast_out.best_species_count.xls：样本污染比对结果，为每个样本随机挑选 100000 条序列 NT 数据库对应的物种统计结果

*/*_R1_fastqc.zip：样本 Read1 序列对应的 FASTQC 分析结果

*/*_R2_fastqc.zip：样本 Read2 序列对应的 FASTQC 分析结果，详细 FASTQC 结果解释见 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

2_assembly/ denovo 拼接结果，里面包含 Transcript 和 Unigene 结果

assembly_result.xls：拼接结果统计，格式说明见报告 5.4

Unigene.fa：所有 Unigene 序列

Transcript.fa：所有转录本序列文件

Unigene_GC_content.pdf：Unigene GC 含量分布图

Unigene_GC_content.xls：Unigene GC 含量统计表

Unigene_Len_Dis.pdf：Unigene 长度分布图

Unigene_len_distribution.xls：Unigene 长度分布统计

Unigene_Len_accumulate.pdf：Unigene 长度累积分布图

Unigene_len_accumulate.xls：Unigene 长度累积分布统计

Unigene_isoforms_num_count.pdf：Unigene 下转录本数目分布图

Unigene_isoforms_num.xls：Unigene 下转录本数目统计表

转录本结果解释同上。

3_SSR/ SSR 分析结果，里面包含 Transcript 和 Unigene 结果

Unigene.fa.ssr.xls：Unigene SSR 结果，格式说明见报告 5.5

Unigene.fa.statistics.txt：Unigene SSR 结果统计

Unigene.fa_density.pdf：Unigene SSR 密度分布图

Unigene.fa_density.xls：Unigene SSR 密度统计，格式说明如下：

Unit_size	SSR 数目大小，单碱基或者其他数目碱基数
SSR_number	该类型 SSR 总数目
SSR_num_per_M	每百万个碱基该类型 SSR 数目

Unigene.fa.ssr.primer.xls：Unigene SSR 引物设计结果，没个 SSR 设计了三对引物

转录本 SSR 结果格式同 Unigene 结果

4_Annotation/ Unigene 注释结果

Annotation_ratio.pdf：各数据库注释比例图

Annotation_statistics.xls：各数据库注释统计表

nr_species_count.pdf：NR 数据库注释物种分布图

Unigene_annotation.xls：所有注释结果整合文件

nr_species_count.xls：NR 注释物种统计表

Venn_diagram_for_annotation.pdf：注释上的基因在各数据库韦恩图

blast_best_hit/ 各数据库最佳比对结果，各列说明如下：

Query_ID	比对的蛋白 ID
Query_len	Query 序列长度
Sbjct	比对上的数据中序列 ID
Sbjct_len	Sbjct 序列长度
Bitscore	分数，越高比对结果越好
Evalue	E 值，越小比对结果越好
Identity	相似度
Align_len	比对上的序列长度
Query_ratio	比对上的长度占 Query 序列总长度比例
Sbjct_ratio	比对上的长度占 Sbjct 序列总长度比例

KOG/ 目录下是注释 KOG 分类结果文件

gene_to_KOG.xls：各基因注释上的 KOG 详细信息

KOG_Categories.pdf：各 KOG code 注释上的基因数条形图

KOG_code_count.xls：各 KOG code 注释上的基因数统计表

KOG_blast.out.infor.best.xls：KOG 注释最佳比对结果

KEGG/ 目录下是注释 KEGG 结果文件：包括基因注释信息、ko 注释基因列表、ko 图

kegg_annot1.xls：基因 pathway 注释结果

kegg_annot2.xls：注释上的 pathway 统计结果

kegg_categories.xlsx：pathway 注释分类结果

KEGG_Categories.pdf：pathway 注释分类条形图

GO/ 目录下是注释 GO 功能分类的结果文件

full_GO_annot.xls：基因注释上 GO 详细信息

part_GO_annot.xls：GO 注释信息，每一行代表一条基因注释上的所有 GO 信息

full_GO_annot _all_level.txt：所有 GO 树上被注释到的 GO 统计结果

full_GO_annot _level2.xlsx：level2 水平上 GO 统计结果

GO_classification_level2.pdf：level2 水平上 GO 注释结果分布图

WEGO_annot.xls：WEGO 格式 GO 注释结果

CDS/ CDS 预测结果

CDS_Len_Dis.pdf：CDS 长度分布图

cds_length_distribution.xls：CDS 长度统计表

CDS_length_ratio.pdf：CDS 长度占 Unigene 长度比例分布图

cds_length_ratio.xls：各 Unigene CDS 长度比例统计表

cds_nucl.fasta：CDS 序列

cds_pep.fasta：CDS 蛋白序列

noORF_id.txt：没有预测到 CDS 的 Unigene

ORF6frame.txt：各 Unigene6 种编码框结果

5_RNASeq_evaluation/ RNASeq 质量评估结果

All.coverage.interval.xls：所有样本基因覆盖比例统计表，计算的为各样本在基因覆盖为某一比例的基因数目占比。

All.geneBodyCoverage.curves.pdf：所有样本均一化分析图

All.geneBodyCoverage.heatMap.pdf：所有样本均一化分析热图

All.geneBodyCoverage.txt：所有样本均一化分析统计表，每一行表示样本在各均一化位置对于的平均覆盖深度，总共 100 个位置

All.saturation.xls：饱和度拟合结果，各百分比抽样序列下，FPKM 误差范围在 10%以内的基因占比数目，可用于绘制饱和度曲线

All_saturation_curve_plot.pdf：饱和度曲线

All_sample_mapping_stastics.xls：所有样本 Mappping 统计，说明见 5.10

此文件夹还包含各样本对应的质量评估结果，说明如下：

*.coverage.interval.xls：单个样本基因覆盖统计表，格式如下：

Coveage_ratio	基因覆盖度比例
Transcript_num	在该比例的转录本数目
Transcript_Ratio	在该比例下转录本数目占比

*.coverage.interval_plot.pdf：单个样本基因覆盖分布饼图

*.coverage.xls：各转录本覆盖度详细列表，格式说明如下：

Transcript_ID	转录本 ID
Length	长度
Mead_depth	平均覆盖深度

Coverage(%)	覆盖度，及被覆盖到的区域比率

*.geneBodyCoverage.curves.pdf：单个样本均一化分析图

*.geneBodyCoverage.txt：单样本均一化分析结果统计

*.inner_distance_plot.pdf：样本片段长度分布图

*.saturation.pdf：单样分区间饱和度盒状图

*.saturation.xls：单样本饱和度拟合结果

*_saturation_curve_plot.pdf：单样本拟合度曲线图

6_expression_profile/ 表达量计算结果

All.genes.correlation.heatmap.pdf：样本间相关性热图

All.genes.counts.xls：各基因比对上的 reads 数统计

All.genes.FPKM.boxplot.pdf：各样本表达量盒状图

All.genes.FPKM.density.pdf：各基因表达量密度曲线

All.genes.FPKM.interval.barplot.pdf：各样本分区间基因表达量条形图

All.genes.FPKM.interval.xls：各样本分区间基因表达量基因数目统计

All.genes.FPKM.xls：所有样本表达量矩阵

All.genes.Sample.clustering.pdf：样本聚类图

转录本层面结果解释同上，各样本均有对应文件，文件说明如下：

*.FPKM.interval.xls：单样本表达量区间统计

*.genes.FPKM.boxplot.pdf：单样本表达量盒状图

*.genes.FPKM.density.pdf：单样本表达量密度曲线

*.genes.FPKM.interval.barplot.pdf：单样本表达量区间条形图

*.genes.FPKM.xls：单样本基因表达量

*isoforms*为对应的转录本层面结果，解释同上

correlation_analysis/ 样本间相关性分析

All.genes.kendall.correlation.matrix.csv：样本间 kendall 相关系数矩阵

All.genes.pearson.correlation.matrix.csv：样本间 pearson 相关性系数矩阵

All.genes.spearman.correlation.matrix.csv：样本间 spearman 相关性系数矩阵

A_vs_B.genes.correlation.pdf：样本间相关性分析散点图

*.Isoforms*为转录本层面结果，解释同上

PCA/ PCA 分析结果

All.genes.PC1.extreme50.xls：第一主成分前 50 基因表达量矩阵

All.genes.PC2.extreme50.xls：第二主成分前 50 基因表达量矩阵

All.genes.PC3.extreme50.xls：第三主成分前 50 基因表达量矩阵

All.genes.PC_all.extreme50.xls：所有主成分前 50 基因表达量矩阵

All.genes.PCA.2dplot.heatmap.pdf：PCA 2d plot 图

All.genes.PCA.3dplot.pdf：PCA 3d plot 图

All.genes.PCA.matrix.xls：各样本对应的主成分矩阵

*.isoforms*为转录本层面结果，解释同上

VENN/ 各样本共同表达韦恩图结果

*.genes.venn.pdf：共同表达基因韦恩图

7_SNP/ SNP 分析结果

All.bam.filtered.vcf：所有样本原始 SNP 结果。VCF 格式

All.mutation.spectrum.xls：所有样本突变谱系统计

All.SNP.Indel.count.pdf：各样本 SNP、INDEL 数目条形图

All.snp.indel.count.xls：各样本 SNP/INDEL 数目条形图

All_snp_result.xls：各样本过滤后 SNP 结果

各样本均有对应的文件夹，文件说明如下：

*.mutation.spectrum.pdf：突变谱系条形图

*.snp.density.pdf：SNP 密度分布图

*.snp.density.txt：SNP 密度统计

*.SNP.result.xls：SNP 结果

8_DEGs_analysis/ 差异表达基因分析结果

genes.DEGs.num.xls：差异表达基因数目统计

isoforms.DEGs.num.xls：差异表达转录本数目统计

各比较对有对应的文件夹，各文件说明如下：

A_vs_B.genes.DEGs.count.pdf：样本对间差异表达基因数目条形图

A_vs_B.genes.dif.test.all.xls：所有差异表达基因结果，包括上调及下调基因

A_vs_B.genes.dif.test.up.xls：上调基因结果，为 B 相对于 A

A_vs_B.genes.dif.test.down.xls：下调基因结果，为 B 相对于 A

A_vs_B.genes.dif.test.xls：所有基因差异表达检验结果，结果说明如下：

Type	是否为差异基因，分为 Up/Down/-
Id	基因 ID
baseMean	样本间平均表达量值，为校正后的序列数目，下同
baseMeadA	A group 中的平均表达量
baseMeadB	Bgroup 中的平均表达量
Foldchange	差异倍数
Log2fold change	差异表达倍数 log2 值
Pval	P 值
Padj	校正后的 P 值

A_vs_B.genes.FPKM.boxplot.pdf：样本对间表达量盒状图

A_vs_B.genes.FPKM.density.pdf：样本对间表达量密度曲线

A_vs_B.genes.FPKM.Scatter.plot.pdf：样本间表达量散点图

A_vs_B.genes.MA.plot.pdf：样本间表达量 MA 图

A_vs_B.genes.volcano.plot.pdf：样本间火山图

VENN/ 差异基因韦恩图结果

*.all.venn.pdf：所有差异基因韦恩图

*.up.venn.pdf：上调差异基因韦恩图

*.down.venn.pdf：下调差异基因韦恩图

*isoforms*为转录本层面结果，格式同上

*_DEGs_cluster/ 差异基因表达模式聚类结果，分类 gene 层面及转录本层面，文件说明如下：

All_DEGs_genes_FPKM_matrix.xls：所有差异表达基因表达量矩阵

All_DEGs_sample_cor_matrix.pdf：样本间相关系数热图

All_DEGs_samples_heatmap.pdf：所有差异基因表达量热图

DEGs_cluster_plot.pdf：所有差异基因各类别基因表达量折线图

subcluster_*_log2_medianCentered_fpkm.matrix：各类基因表达量矩阵

Isoforms 层面结果同上。

9_GO_enrichment/GO 富集分析结果

*.genes.enriched.GO.combine.xls：所有比较对显著富集 GO 汇总

*genes.enriched.GO.heatmap2.pdf：GO 富集热图 2

*genes.enriched.GO.heatmap.pdf：GO 富集热图 1

genes.all.GO_enrichment_count.xls：所有差异基因富集分析结果统计

genes.down.GO_enrichment_count.xls：下调基因富集分析结果统计

genes.up.GO_enrichment_count.xls：上调基因富集分析结果统计

各比较对会有对应的文件夹，文件说明如下：

A_vs_B.genes.all_BP_classic_5_all.pdf：所有差异基因 BP 类上的 topGO 有向无环图

A_vs_B.genes.all_CC_classic_5_all.pdf：所有差异基因 CC 类上的 topGO 有向无环图

A_vs_B.genes.all_MF_classic_5_all.pdf：所有差异基因 MF 类上的 topGO 有向无环图

A_vs_B.genes.all_GO_enrichment.xls：所有差异基因 GO 富集分析结果，格式说明如下：

GO_ID	GO ID
Term	GO 名字
Type	GO 功能类
All_annotated_num	所有注释上 GO 的基因数目
All_num_this_term	所有注释上该 Term 的基因数目
All_DEGs_num	左边的数字为所有差异基因数目，右边为差异基因中有多少注释上 GO
DEGs_this_term	差异基因中注释上该 Term 的基因数目
UP	上调表达基因注释上该 Term 的基因数目
DOWN	下调表达基因注释上该 Term 的基因数目
Expected	正常情况下注释上该 term 的差异基因数目
Pvalue	P 值，P 值越小越显著
FDR	FDR 值，P 值校正值

A_vs_B.genes.all_GO_enrichment_barplot.pdf：前 50 个富集的 GO 基因数目条形图

A_vs_B.genes.all_GO_enrichment_pvalue.pdf：所有差异基因富集分析 P 值分别直方图

A_vs_B.genes.all_GO_enrichment_scatterPlot.pdf：前三十各富集的 GO 散点图

up：均为上调基因富集结果，解释同上

down：均为下调基因富集结果，解释同上

10_kegg_enrichment/ KEGG 富集分析结果

*.genes.enriched.kegg.combine.xls：所有比较对显著富集 kegg 汇总

*genes.enriched.kegg.heatmap2.pdf：kegg 富集热图 2

*genes.enriched.kegg.heatmap.pdf：kegg 富集热图 1

genes.all.kegg_enrichment_count.xls：所有差异基因富集分析结果统计

genes.down.kegg_enrichment_count.xls：下调基因富集分析结果统计

genes.up.kegg_enrichment_count.xls：上调基因富集分析结果统计

各比较对会有对应的文件夹，文件说明如下：

A_vs_B.genes.all_kegg_enrichment.xls：所有差异基因 KEGG 富集分析结果，格式同 GO

A_vs_B.genes.all_kegg_enrichment_pvalue.pdf：所有差异基因富集分析 P 值分别直方图

A_vs_B.genes.all_kegg_enrichment_scatterPlot.pdf：前三十各富集的 pathway 散点图

up：均为上调基因富集结果，解释同上

down：均为下调基因富集结果，解释同上

7. 参考文献

[1] A Franceschini.STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res.2013 Jan;41(Database issue).

[2] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A.Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52. doi: 10.1038/nbt.1883.

[3] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM,Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, WolfYI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioin-formatics. 2003 Sep 11;4:41. Epub 2003 Sep 11.

[4] Anders, S., and Huber, W. Differential expression analysis for sequence count data. Genome Biol 11, R106.

[5] Chepelev, I., Wei, G., Tang, Q., and Zhao, K. (2009). Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic acids research 37, e106-e106.

[6] Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic acids research 35, W297-W299.

[7] Mamanova, L., Andrews, R.M., James, K.D., Sheridan, E.M., Ellis, P.D., Langford, C.F., Ost, T.W.B., Collins, J.E., and Turner, D.J. FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nature methods 7, 130-132.

[8] Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5, 621-628.

[9] Sammeth, M., Foissac, S., and Guigo, R. (2008). A general definition and nomenclature for alternative splicing events. PLoS computational biology 4, e1000147.

[10] Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.

[11] Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28, 511-515.

[12] Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476.

[13] Wang L., Feng Z., Wang X., Wang X., Zhang X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-8.

[14] Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, S.H., Robles, M., Talón, M., Dopazo, J., Conesa, A. (2008).High-throughput functional annotation and data mining with the Blast2GO suite.Nucleic Acids Research 36, 3420-3435.

[15] Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing.Genome Research 19, 1124-1132.

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

[17] Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010).Gene ontology analysis for RNA-seq: accounting for selection bias. GenomeBiology, doi:10.1186/gb-2010-11-2-r14.

[18] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids research36:D480–484.

[19] Mao, X., Cai, T., Olyarchuk, J.G., Wei, L. (1995). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as acontrolled vocabulary.bioinformatics 21, 3787–3793.

[20] Cole Trapnell , Adam Roberts , Loyal Goff , Geo Pertea , Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn and Lior Pachter(2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.Natrual protocol doi:10.1038/nprot.2012.016.