生工®
Sangon Biotech

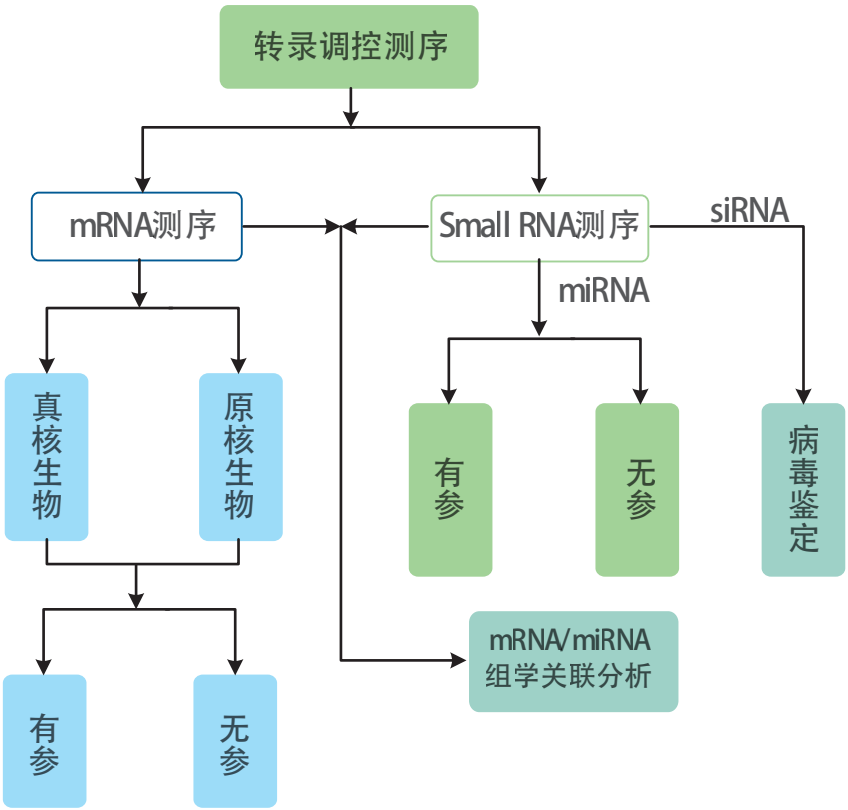
生工生物
转录调控测序服务介绍

高通量技术服务推荐

Life Biotech Future

目 录 Catalog

转录调控测序



转录组（mRNA）测序

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，用于反应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。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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。

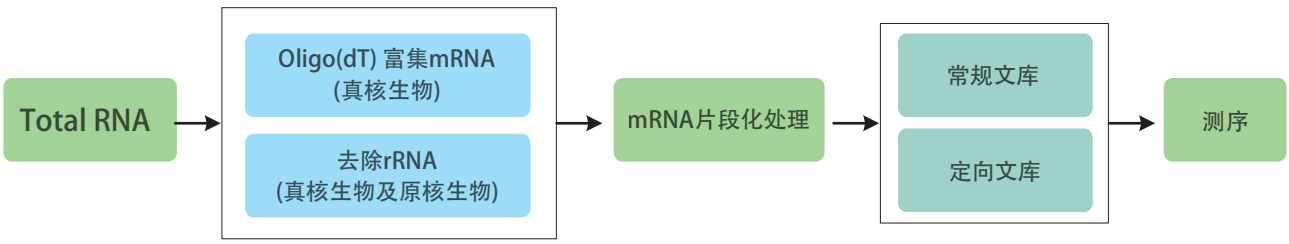
Ensembl：欧洲基因组数据库，与 NCBI，UCSC 并为三大基因组数据库，其中可下载大部分物种的基因组序列及相关注释文件；动物基因组：http://asia.ensembl.org/index.html，其它物种基因组：http://ensemblgenomes.org/。

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

3.1 实验流程图



3.2 实验流程详述

3.2.1 材料

客户提供原始样本：-70°C 冻存的新鲜动物组织、植物组织、真菌、细胞、抗凝血液等。

客户提供 RNA 样本：样本质量需达到二代测序要求，包括无降解，无凋亡，无污染，无杂带，无残留等；RNA 样本的 RIN 值≥8.0，总量≥2.0μg。

3.2.2 关键试剂

试剂/耗材	厂商	货号
Total RNA Extractor（Trizol）	上海生工	B511311
Qubit2.0 RNA 检测试剂盒	Life	Q32855
Qubit2.0 DNA 检测试剂盒	Life	Q10212
VAHTSTM mRNA-seq V2 Library Prep Kit for Illumina®	南京诺唯赞	NR601-02
Ampure XP DNA Clean Beads	南京诺唯赞	N411-03

地　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

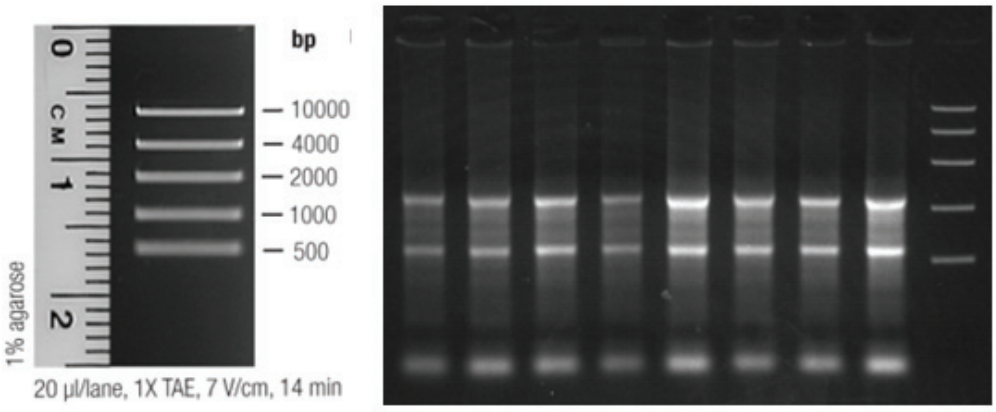
3.2.3 仪器

仪器名称	仪器型号	生产商
Qubit2.0 荧光计	Q32866	Invitrogen
微型漩涡混合仪	WH-3	上海沪西分析仪器厂有限公司
台式高速低温离心机	Thermo Scientific Sorvall Legend Micro 21R	Thermo
PCR 仪	T100™ Thermal Cycler	BIO-RAD
电泳仪	DYY-11	北京市六一仪器厂
生物电泳图像分析系统	FR-980A	复日科技

3.2.4 实验步骤

（一）Total RNA 提取方法—以 Total RNA Extractor（Trizol）提取试剂盒为例

1. 将裂解后样品或匀浆液室温放置 5-10 min，使得核蛋白与核酸完全分离。
2. 加入 0.2 ml 氯仿，剧烈震荡 15 sec，室温放置 3 min。12000 rpm 4℃离心 10 min。
3. 吸取上层水相转移至干净的离心管中，加入等体积异丙醇，混匀，室温放置 20 min。
4. 12000 rpm 4℃离心 10 min，弃上清。
5. 加入 1 ml 75%乙醇洗涤沉淀。12000 rpm 4℃离心 3 min，弃上清。室温干燥 5-10 min。
6. 加入 30-50 μL RNase-free ddH2O，充分溶解 RNA。将所得到的 RNA 溶液置于-70℃保存或立刻用于后续试验。
7. Qubit2.0 检测 RNA 浓度，琼脂糖凝胶检测 RNA 完整性以及基因组污染情况。



实验图 1：Total RNA 检测示意图

注：Marker 为 DNA Marker，其只为检测基因组污染，并不代表 RNA 条带大小。

（二）mRNA 文库构建——以 VAHTSTM mRNA-seq V2 Library Prep Kit for Illumina 试剂盒为例

利用 Qubit2.0 RNA 检测试剂盒对 Total RNA 精确定量，以确定文库构建所加入 Total RNA 的量

1. mRNA 分离/片段化

1.1. 将 mRNA Capture Beads 从 2-8 °C 取出，静置使其温度平衡至室温。

1.2. 准备 RNA 样品：在一个 Nuclease free 离心管中，将 0.1-1 μg 总 RNA 溶解于 50 μl Nuclease free 水中，冰上放置备用。

1.3. 颠倒或旋涡振荡使 mRNA Capture Beads 充分混匀，吸取 50 μl 加入到总 RNA 样品中，用移液器吹打 6 次以彻底混匀。

1.4. 将样品置于 PCR 仪中，65 °C 5 min，4 °C hold，使 RNA 变性。

1.5. 室温放置 5 分钟，使 mRNA 结合到磁珠上。

1.6. 将样品置于磁力架 5 分钟，使 mRNA 与总 RNA 分离，小心移除上清。

1.7. 将样品从磁力架上取出，用 200 µl Beads Wash Buffer 吹打 6 次以彻底混匀，在磁力架上静置 5 分钟，小心移除上清。

1.8. 将样品从磁力架上取出，50 µl Tris Buffer 重悬磁珠，用移液器吹打 6 次以彻底混匀。

1.9. 将样品置于 PCR 仪中，80 °C 2 min，25 °C hold，将 mRNA 洗脱下来。

1.10. 加入 50 µl Beads Binding Buffer，用移液器吹打 6 次以彻底混匀。

1.11. 室温放置 5 分钟，使 mRNA 结合到磁珠上。

1.12. 将样品置于磁力架 5 分钟，使 mRNA 与总 RNA 分离，小心移除上清。

1.13. 将样品从磁力架上取出，用 200 µl Beads Wash Buffer 吹打 6 次以彻底混匀，在磁力架上静置 5 分钟，吸掉全部上清(注意最后需要用 10 µl 移液器吸干净残留液体)。

1.14. 将样品从磁力架上取出，用 19.5 µl Frag/Prime Buffer 重悬磁珠，用移液器吹打 6 次以彻底混匀；将样品置于 PCR 仪中，94 °C 5 min，4 °C hold。

1.15. 将样品置于磁力架 5 分钟，转移 17 µl 上清至一个新的 nuclease free 离心管中，立刻进入第一链合成反应。

2. 双链 cDNA 合成

2.1. 将 1st Strand Buffer 从-20 °C 取出，解冻后颠倒混匀，配制第一链 cDNA 合成反应液：

Fragmented mRNA	17 µl
1st Strand Buffer	6 µl
1st Strand Enzyme Mix	2 µl
总计	25 µl

25℃温浴 10 分钟后，转入 42℃温浴 15 分钟，最后 70℃ 温浴 15 分钟。立刻进行第二链合成反应。

2.2. 将 2nd Strand Buffer 从-20 °C 取出，解冻后颠倒混匀，配制第二链 cDNA 合成反应液：

1st Strand cDNA	25 µl
2nd Strand Buffer	20 µl
2nd Strand Enzyme Mix	5 µl
总计	50 µl

16℃反应 60 分钟。加入 90 µl (1.8 ×) Ampure XP DNA Clean Beads纯化双链 cDNA，加入62.5 µl nuclease free 水溶解磁珠，小心吸取 60 µl 上清至一个新的 nuclease free 离心管中。

3. 末端修复

将 End Prep Mix 从-20 °C 取出，解冻后颠倒混匀，配制末端修复反应液：

ds cDNA	60 µl
End Prep Mix	40 µl
总计	100 µl

30℃反应 30 分钟。加入 160 µl (1.6 ×) Ampure XP DNA Clean Beads纯化末端修复产物，加入20 µl nuclease free 水溶解磁珠，小心吸取 17.5 µl 上清至一个新的 nuclease free 离心管中。

4. 末端 dA-Tailing

将 dA-Tailing Buffer Mix 从-20 °C 取出，解冻后颠倒混匀，配制末端 dA-Tailing 反应液：

纯化的末端修复产物	17.5 µl
dA-Tailing Buffer Mix	10 µl
dA-Tailing Enzyme Mix	2.5 µl
总计	30 µl

37℃温浴 30 分钟，随后 70℃温浴 5 分钟。立刻进行接头连接反应。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

5. 接头连接

将 RNA Adapter 从-20 °C 取出，解冻后颠倒混匀，配制连接反应液：

dA-Tailing 产物	30 μl
Ligation Mix	2.5 μl
RNA Adapter(with barcode,1μM)	2.5 μl
总计	35 μl

30℃反应 10 分钟。加入 5 μl Stop Ligation Mix 终止反应。

6. 连接产物纯化和片段大小分选

6.1. 用 1 x Ampure XP DNA Clean Beads 纯化连接产物

加入 40 μl (1 ×) Ampure XP DNA Clean Beads纯化连接产物，加入102.5 μl nuclease free 水溶解磁珠，小心吸取 100 μl 上清至一个新的 nuclease free 离心管中。

6.2. 用两轮 Ampure XP DNA Clean Beads 进行片段大小分选

第一轮加入 70 μl (0.7 ×) Ampure XP DNA Clean Beads 纯化连接产物，小心吸取155 μl 上清至一个新的 nuclease free 离心管中。第二轮加入 10 μl (0.1 ×) Ampure XP DNA Clean Beads 纯化连接产物，加入22.5 μl nuclease free 水溶解磁珠，小心吸取 20 μl 上清至一个新的 nuclease free 离心管中。

7. 文库扩增

将 PCR Primer Mix，Amplification Mix 1 从-20 °C 取出，解冻后颠倒混匀，配制 PCR 反应液：

纯化过的接头连接产物	20 μl
PCR Primer Mix	5 μl
Amplification Mix 1	25 μl
总计	35 μl

PCR 反应条件：

98℃，30s；

98℃，10s；

60℃，30s；　15cycles

72℃，30s；

72℃，5min；

4℃，∞；

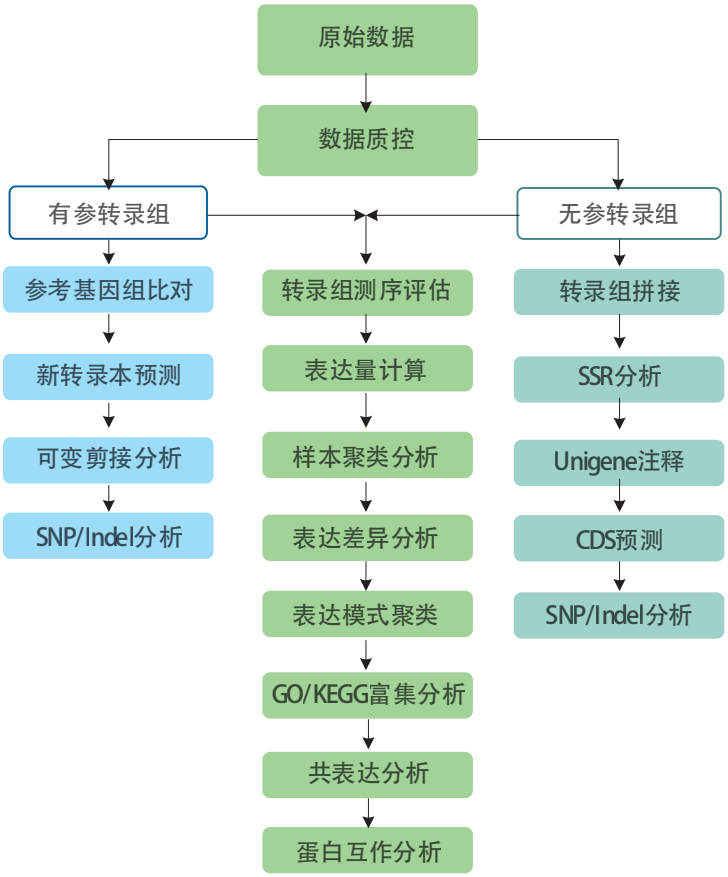
加入 50 μl (1 ×) Ampure XP DNA Clean Beads纯化连接产物，加入25 μl nuclease free 水溶解磁珠，小心吸取 22.5 μl 上清至一个新的 nuclease free 离心管中。

（三）定量混合

利用 Qubit2.0 DNA 检测试剂盒对回收的 DNA 精确定量，以方便按照 1:1 的比例等量混合后进行测序。

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	

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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. 新转录本预测	1）新转录长度统计 2）新转录本类型统计	仅有参转录组做
7. 可变剪切分析	1）可变剪切类型统计 2）可变剪切位置结果	仅有参转录组做
8. SNP/Indel 分析	1）各样本 SNP 汇总列表 2）SNP 密度分布图 3）SNP 突变图谱	无参转录组两个以上样本才做，有参转录组都做
9. 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	
10. 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
11. 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
12. 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
13. 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
14. GO/KEGG 富集分析	1）富集分析统计表 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	Pvalue 热图需比较对数目大于 1
15. 共表达分析	1）共表达基因聚类图 2）共表达基因网络图	需样本数大于10
16. 蛋白互作分析	1）蛋白互作网络图 2）候选基因列表	须有对应物种的蛋白互作数据库，网络图需客户自行绘制

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

无参转录组

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

有参转录组

2. 新转录本预测

1）将各样本 QC 之后序列比对到参考基因组，比对采用软件 tophat

2）进行带参考基因组拼接,拼接软件为 cufflinks

3）合并各样本拼接结果

4）将最终拼接结果与数据库中已知转录本比较，确定 novel 转录本

3. 可变剪切分析

1）采用 MATS 对各样本做可变剪切分析

2）统计各样本可变剪切形式数目

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 结果对原始结果进行过滤，过滤条件为：

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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：

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

2）筛选显著富集的 pathway

3）对富集的 pathway 进行上色

12. 基因共表达分析

1）采用 WGCNA 对差异表达基因做共表达分析

2）筛选出共表达基因集

3）绘制共表达网络图

13. 蛋白互作分析

1）筛选出差异表达基因，将基因导入到 String 数据库中

2）绘制蛋白互作网络，并筛选出候选基因

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 Illumina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

	C1	C2	T1	T2
Total Reads Count(#):	32756892	37281612	35017800	40611932
Total Bases Count(bp):	4913533800	5592241800	5252670000	6091789800
Average Read Length(bp):	150	150	150	150
Q30 Bases Count(bp):	4515036860	5155866430	4794125076	5598975495
Q30 Bases Ratio(%):	91.88980973	92.19677214	91.27025067	91.91018861
Q20 Bases Count(bp):	4728373189	5391114473	5036755733	5863613163
Q20 Bases Ratio(%):	96.23162029	96.40345797	95.88943781	96.254358
Q10 Bases Count(bp):	4899536179	5577034626	5236408342	6074364872
Q10 Bases Ratio(%):	99.7151211	99.72806659	99.69041158	99.71396045
N Bases Count(bp):	104747	118085	114710	132500
N Bases Ratio(%):	0.002132	0.002112	0.002184	0.002175
GC Bases Count(bp):	2430255141	2759223070	2594319691	3022005231
GC Bases Ratio(%):	49.46043398	49.34019609	49.39049457	49.60783826

地　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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：

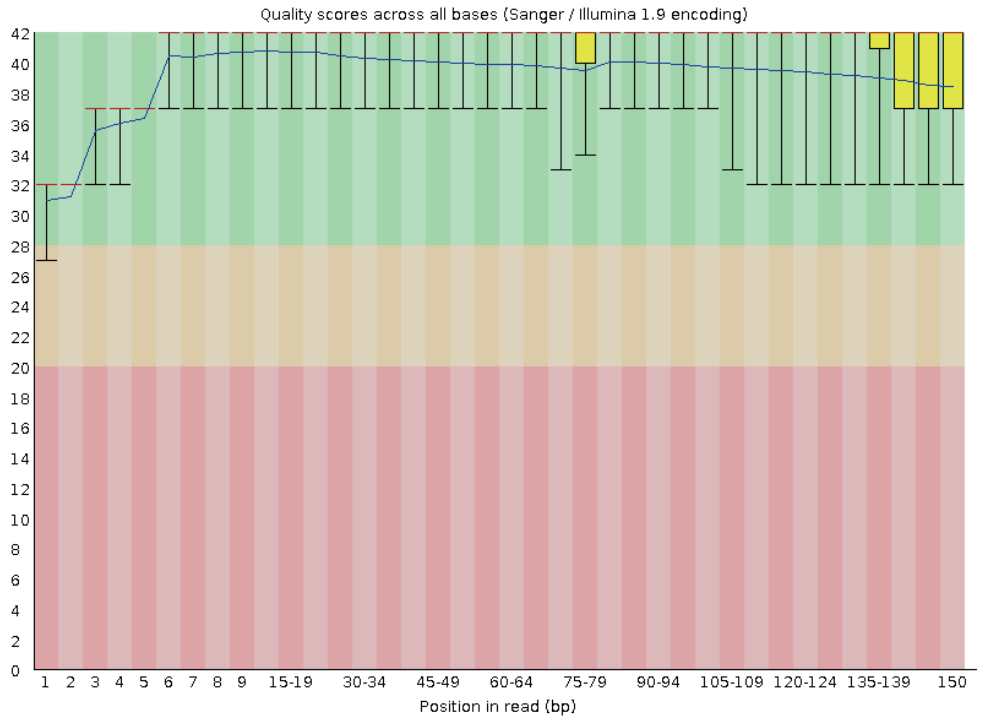


图 5.1 各位置碱基质量分布图

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 6.1 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：

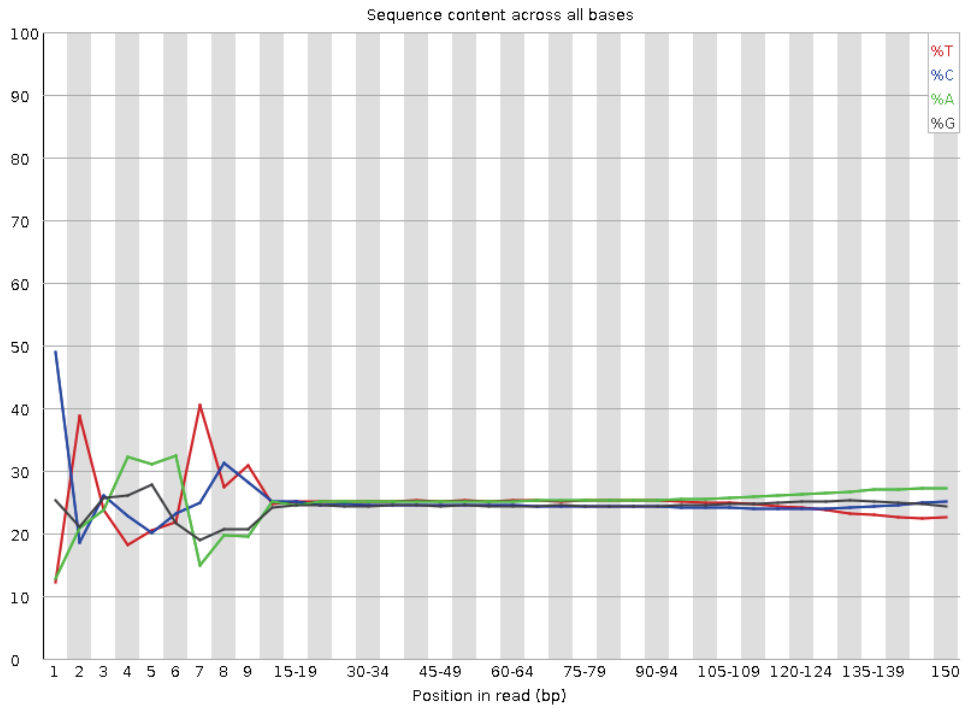


图 5.2 各位置碱基 GC 含量分布图

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

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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

	C1	C2	T1	T2
Raw_sequences	32756892	37281612	35017800	40611932
Raw_bases	4913533800	5592241800	5252670000	6091789800
Raw_mean_length	150	150	150	150
Good_sequences	32577498	37096510	34812664	40407469
Good_ratio	99.45	99.5	99.41	99.5
Good_bases	4631329208	5272541307	4945169980	5751569079
Good_mean_length	142.16	142.13	142.05	142.34

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
Populus trichocarpa	1510
Ricinus communis	1073
Hevea brasiliensis	356
Vitis vinifera	304
Bruguiera gymnorhiza	234
Fragaria vesca subsp. vesca	203
Populus tremula x Populus alba	111
Glycine max	109
Kandelia candel	60
Populus trichocarpa x Populus deltoides	50
Cucumis sativus	38
Cicer arietinum	37
Lotus japonicus	37
Solanum lycopersicum	34
Gossypium hirsutum	32
Medicago truncatula	30
Manihot esculenta	24
Sandersonia aurantiaca	22
Dianthus caryophyllus	21
Jatropha curcas	19

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：

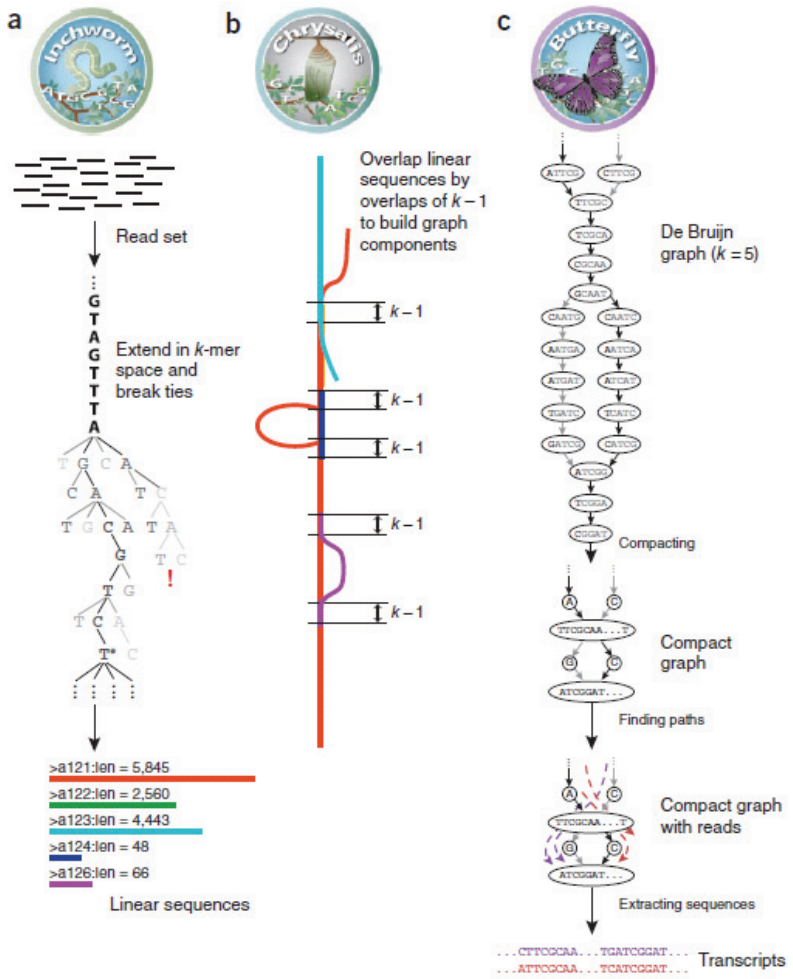


Figure 1 Overview of Trinity. (**a**) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a *k*-mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each *k*-mer present only once in the contigs. (**b**) Chrysalis pools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

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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	115194	55137	32581	1842	349	17101	201	109811956	953.28
Unigene	88131	33509	16266	1291	288	17101	201	65086731	738.52

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

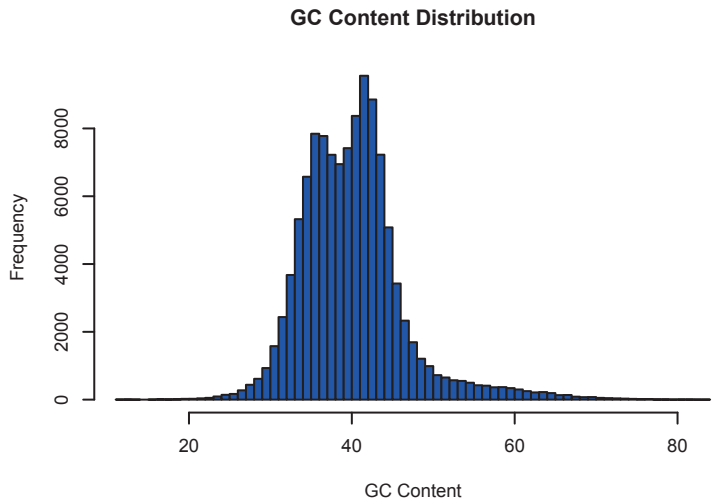


图 5.3 Transcript GC 含量分布图

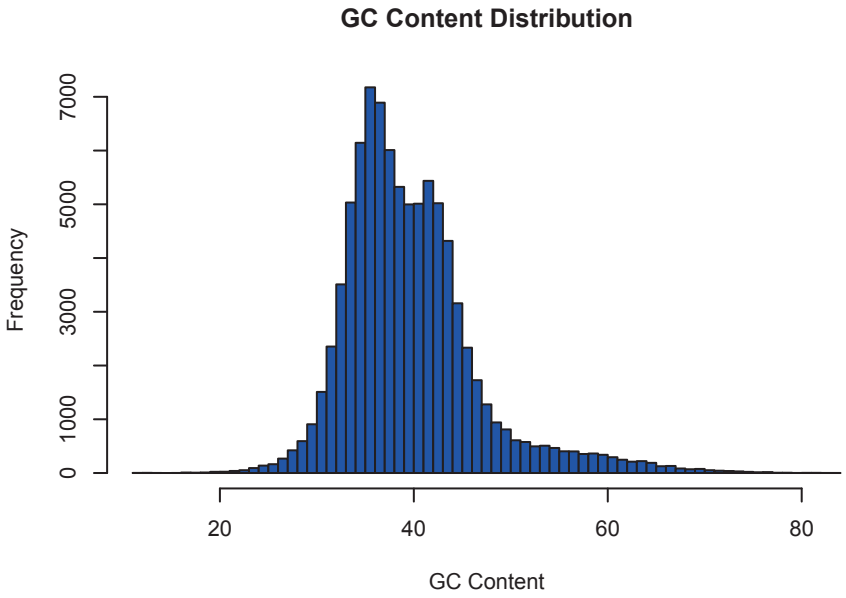


图 5.4 Unigene GC 含量分布图

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

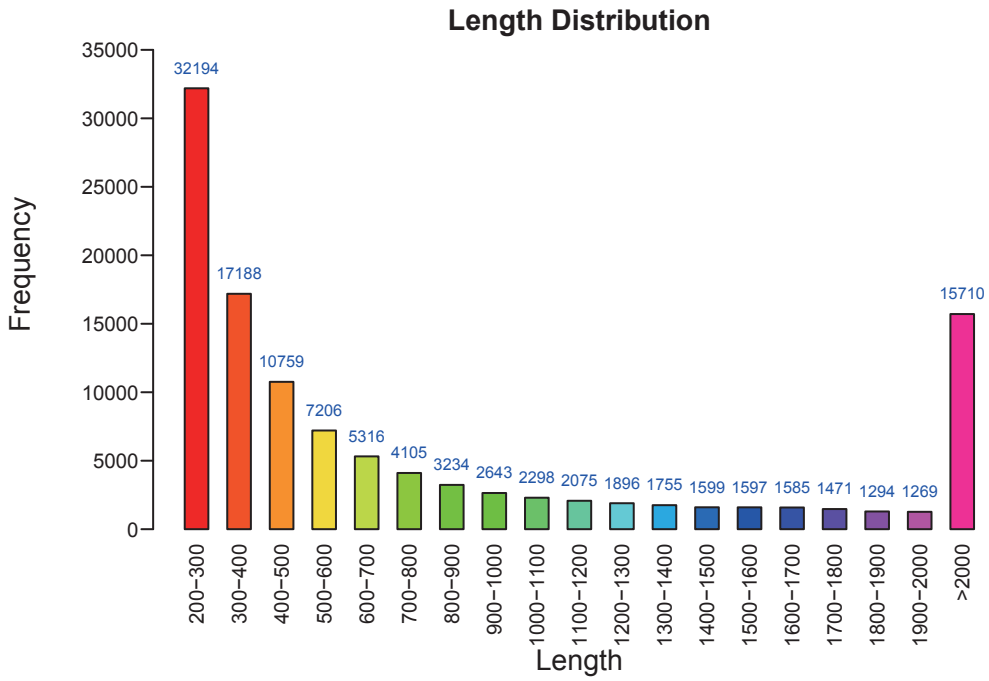


图 5.5 Transcript 长度分布图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

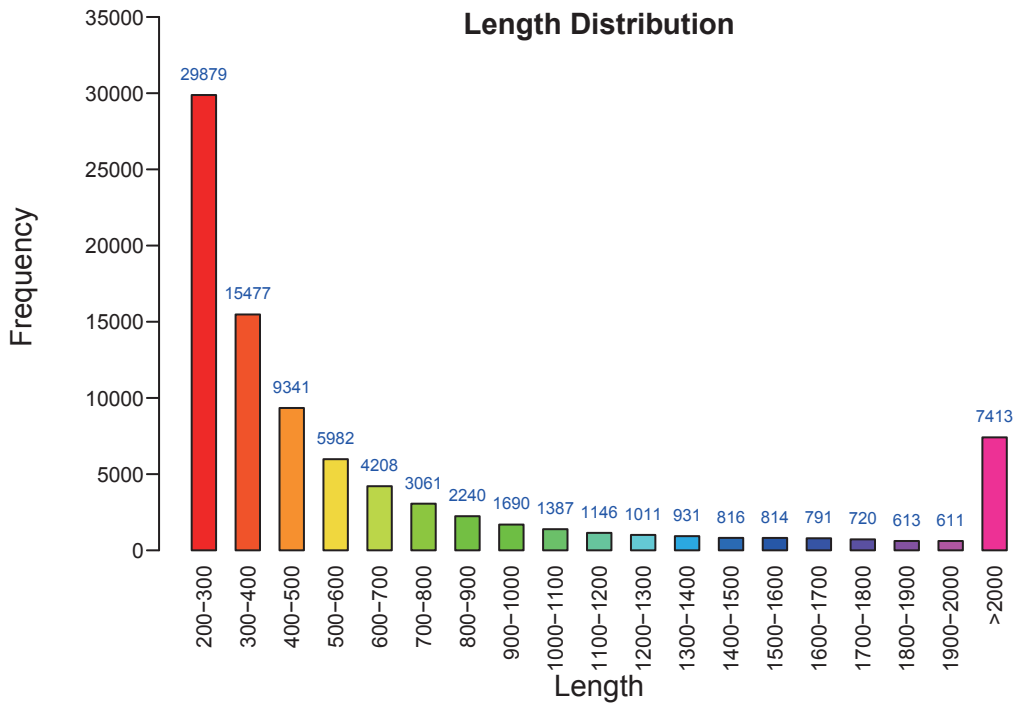


图 5.6 Unigene 长度分布图

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

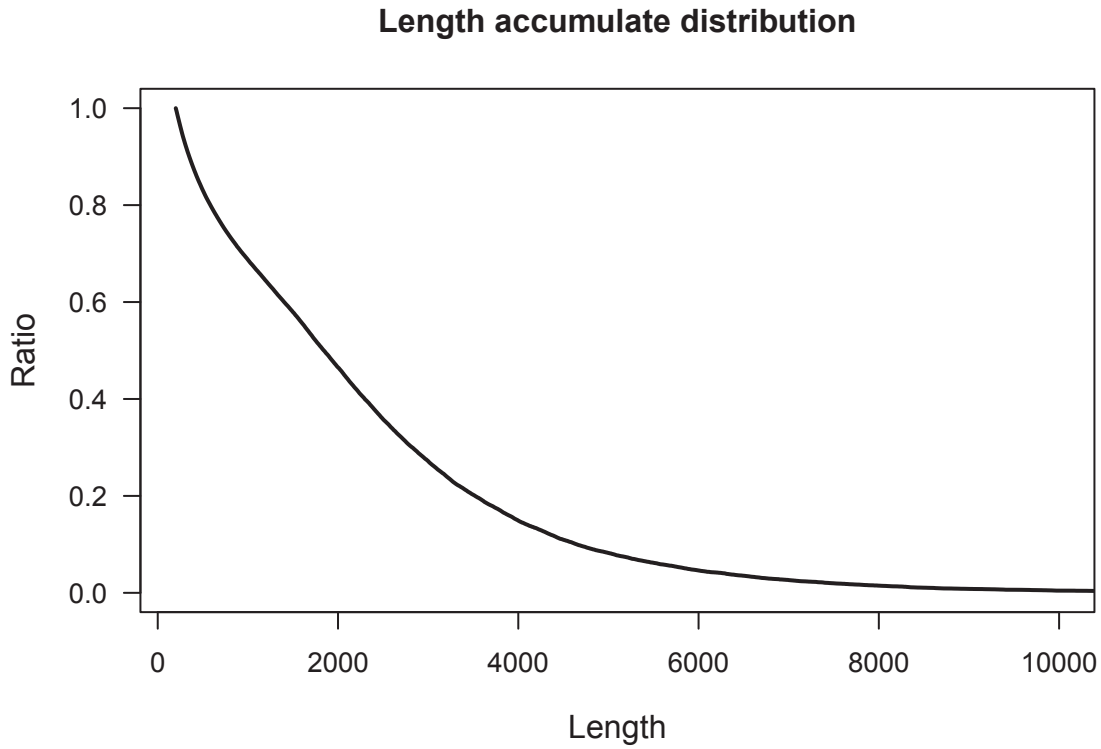


图 5.7 Transcript 长度累积分布图

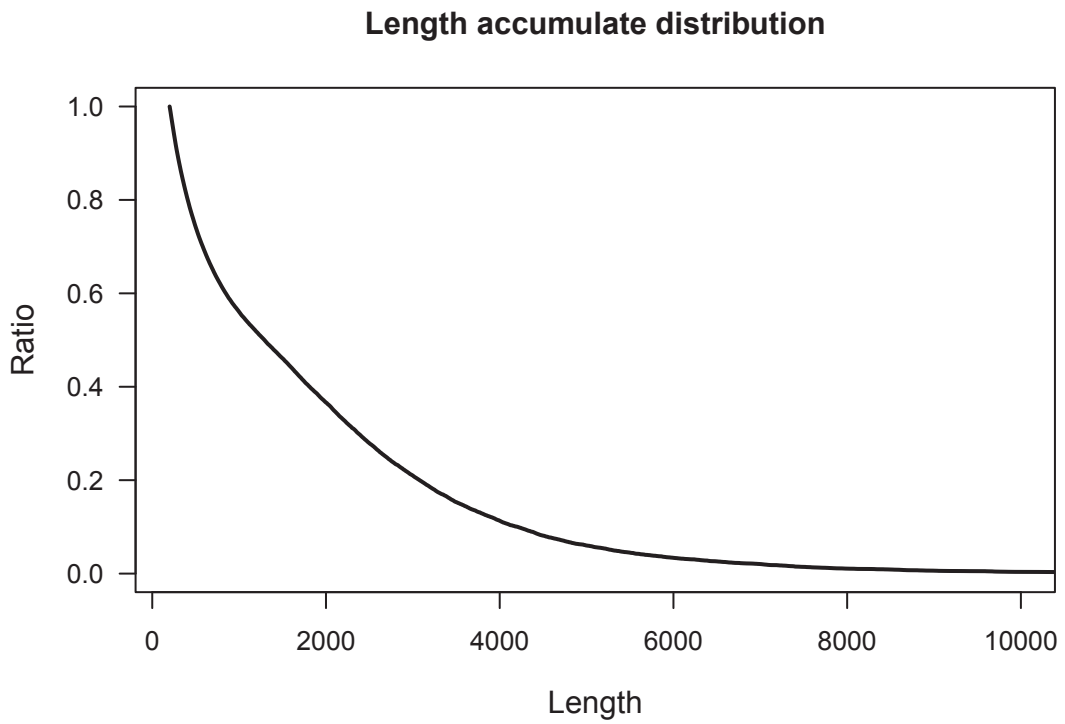


图 5.8 Unigene 长度累积分布图

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：

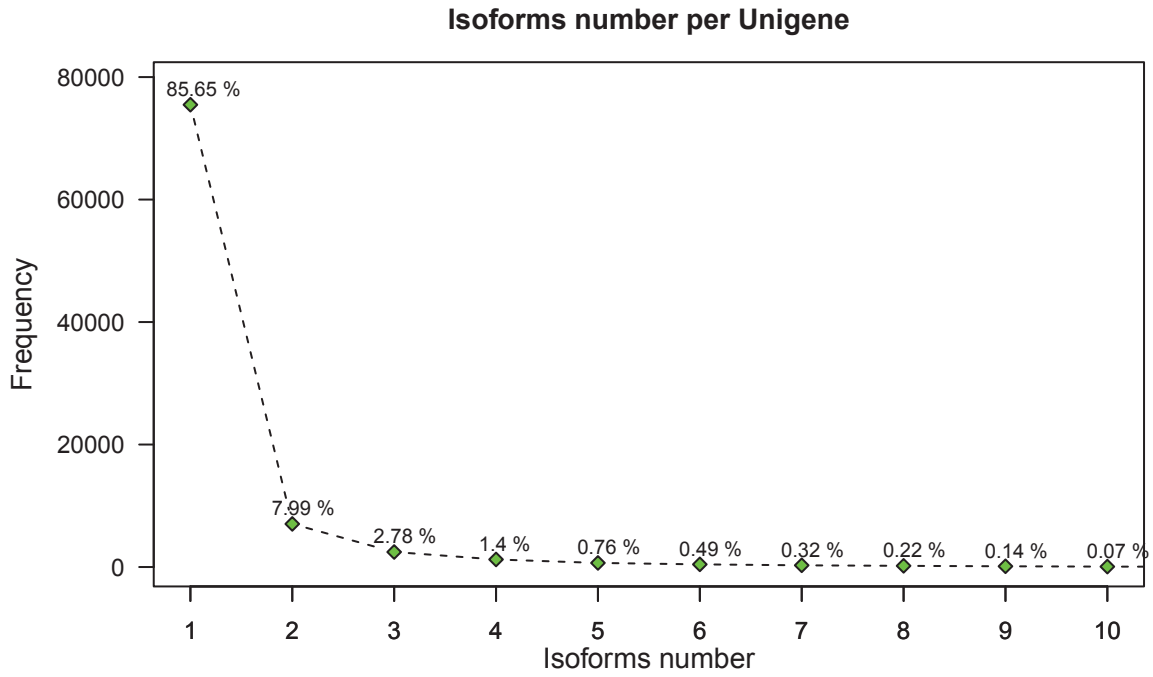


图 5.9 Unigene 下 Isoform 数目分布图

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。该图可反应出 Unigene 下可变剪切发生频率。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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	(T)16	16	6	21
c61875_g1	1	p1	(T)16	16	1	16
c16043_g1	1	p1	(A)10	10	14	23
c16043_g1	2	p1	(T)16	16	1991	2006
c48549_g1	1	p2	(GT)10	20	1	20
c25505_g2	1	p3	(ATT)7	21	335	355
c85624_g1	1	p2	(GA)9	18	1	18
c85852_g1	1	p2	(GA)8	16	1	16
c69736_g1	1	p1	(G)11	11	22	32
c69736_g1	2	p2	(CT)9	18	184	201
c20796_g1	1	p2	(AT)7	14	905	918
c7395_g2	1	p1	(T)11	11	179	189
c12790_g1	1	p1	(A)10	10	438	447
c20595_g2	1	p1	(A)17	17	1	17
c20595_g2	2	p3	(GGC)5	15	1241	1255
c71679_g1	1	p1	(T)10	10	140	149
c45016_g1	1	p3	(TGG)7	21	2	22
c45016_g1	2	p1	(A)10	10	325	334
c69952_g1	1	p1	(T)12	12	1	12
c23521_g1	1	p1	(A)13	13	11	23

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：

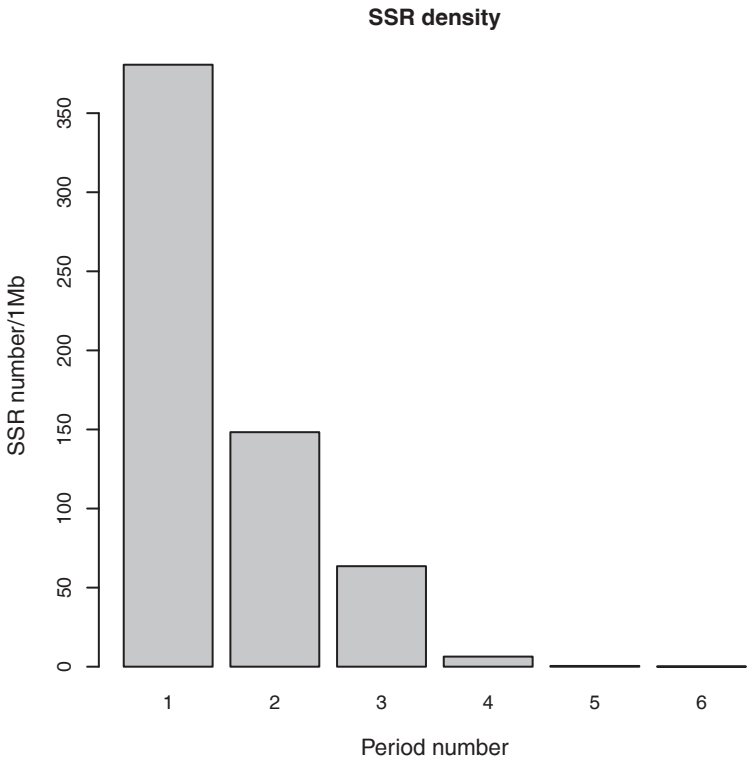


图 5.10 Unigene 下 SSR 密度分布图

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	TGAATCCGGTGGTTGTTTTT	60.206	GCAAACAGTCGTCCACCTTT	60.156
c61875_g1	GGTGTTGATGCTGTGTTTGG	60.008	CACGTGCTGATGTGACAGTG	59.931
c16043_g1	TTAGGGTTTGAAGTGGTGGC	59.971	GCAGGCAATAGAACCTCAGC	59.985
c16043_g1	TTAGGGTTTGAAGTGGTGGC	59.971	GCAGGCAATAGAACCTCAGC	59.985
c48549_g1	TGACCAAATATTCTGGGGGA	60.126	TCCACGTTTCTTGAGCTCCT	59.989
c25505_g2	TGTCCCTGACTTTCGAGCTT	59.989	TGGCAAACCACACTTGGTAA	60.004

地　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

c85624_g1	CGAAAGTTCAAGCCTCGTTC	59.993	GCACACTTTCCCCAACACTT	60.012
c85852_g1	GAGAGAGAGGTTGTTTTTGCG	59.111	TCACGGTCATGGGACATAGA	59.918
c69736_g1	GTTTTCCATAGCGTTGGGGT	61.099	TGCAGCAAAACTAGAGGGGT	59.875
c69736_g1	GTTTTCCATAGCGTTGGGGT	61.099	TGCAGCAAAACTAGAGGGGT	59.875
c20796_g1	TTCGTCAATGGTCGATGAAA	60.049	TCGATATCCGATGCAAATCA	59.997
c40970_g1	AACGTTGCCGATTTTGAAAC	59.982	GCCTTGGTTTACCAGTTGGA	59.971
c16150_g1	TTTACGAGCCAACCCAATTC	59.938	TCCAGTCTCGTCGTTTTCCT	59.844
c7395_g2	TGAAAGCCTCACTGTCATCG	59.984	AATGCAACGCCCTATACACC	59.851
c12790_g1	CCAAGATCATCATGCAAACG	60.073	CGGGGTTCCCAAACTTTTAT	60.048
c20595_g2	CAAAGAGCTTGTTGATGCCA	59.988	CGACGTAAACACAACCATCG	60.027
c20595_g2	CAAAGAGCTTGTTGATGCCA	59.988	CGACGTAAACACAACCATCG	60.027
c71679_g1	GCTTGTGAGAGAGGTCAGGG	59.986	ATCACATTTGATCTGCAGGC	58.674
c45016_g1	GTGGTGGTGGTGGTGGTATT	60.408	GTGTCCAGGGAATTGTGCTT	59.973
c45016_g1	GTGGTGGTGGTGGTGGTATT	60.408	GTGTCCAGGGAATTGTGCTT	59.973

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 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。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	21369	24.25
Annotated in KOG	16173	18.35
Annotated in NR	30258	34.33
Annotated in NT	20567	23.34
Annotated in PFAM	18762	21.29
Annotated in Swissprot	22241	25.24
Annotated in TrEMBL	30008	34.05
Annotated in GO	24070	27.31
Annotated in KEGG	5592	6.35
Annotated in at least one database	32443	36.81
Annotated in all database	4102	4.65
Total Unigenes	88131	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：

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

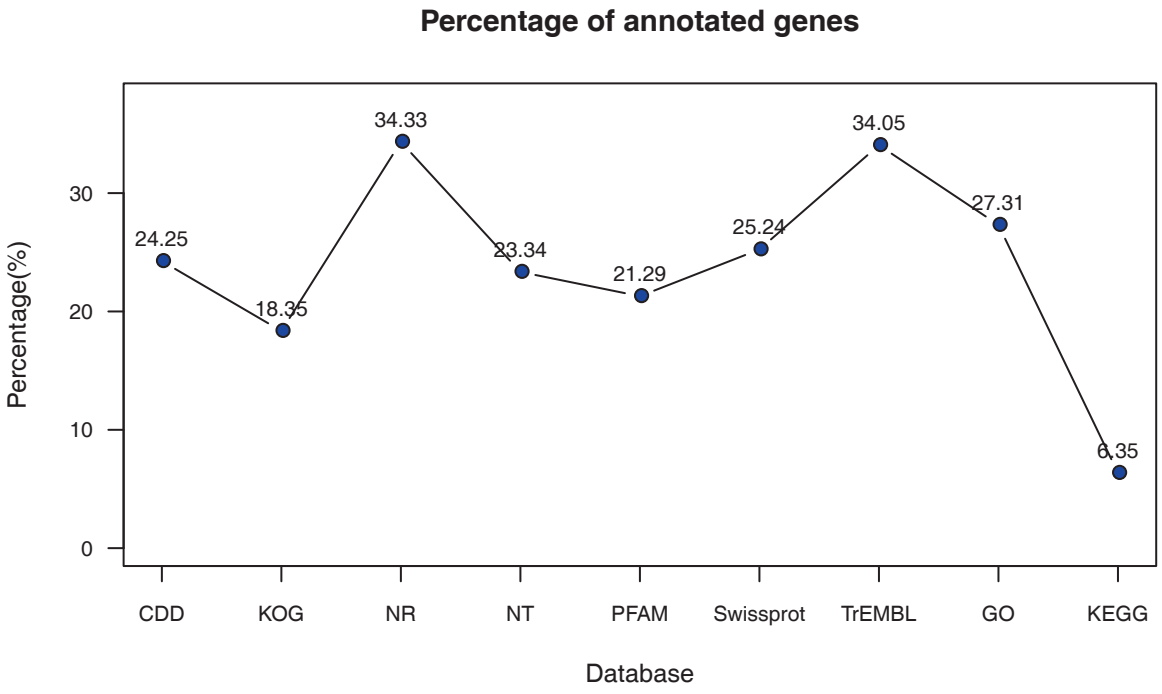


图 5.11 数据库注释比例折线图

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：

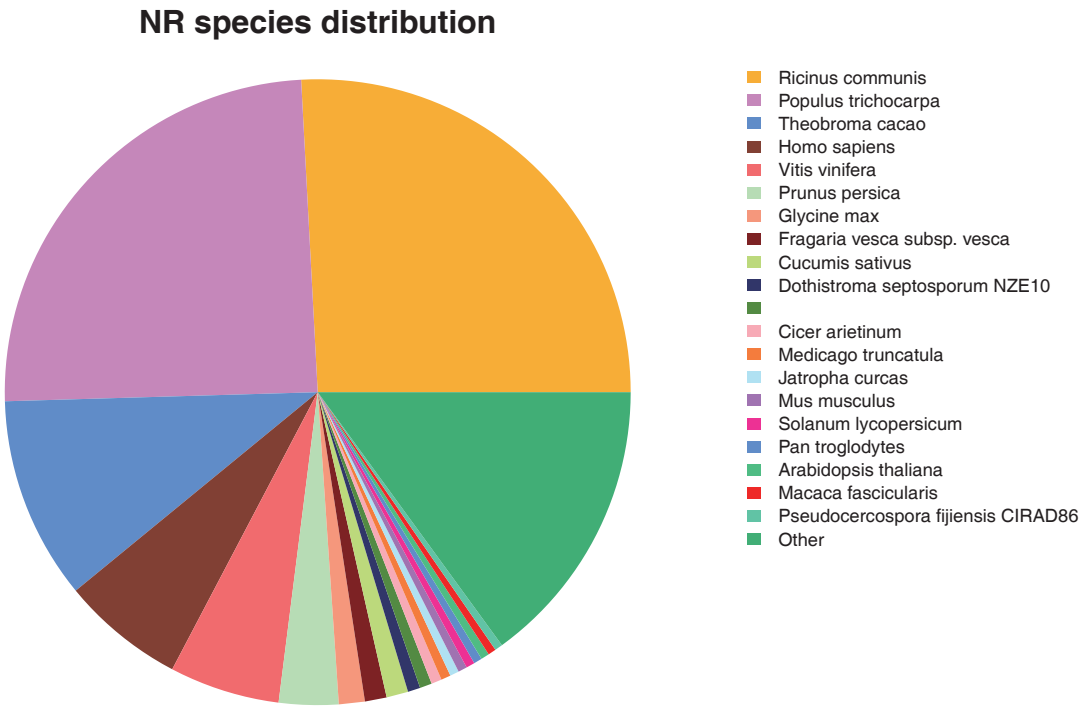


图 5.12 NR 数据库物种分布饼图

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：

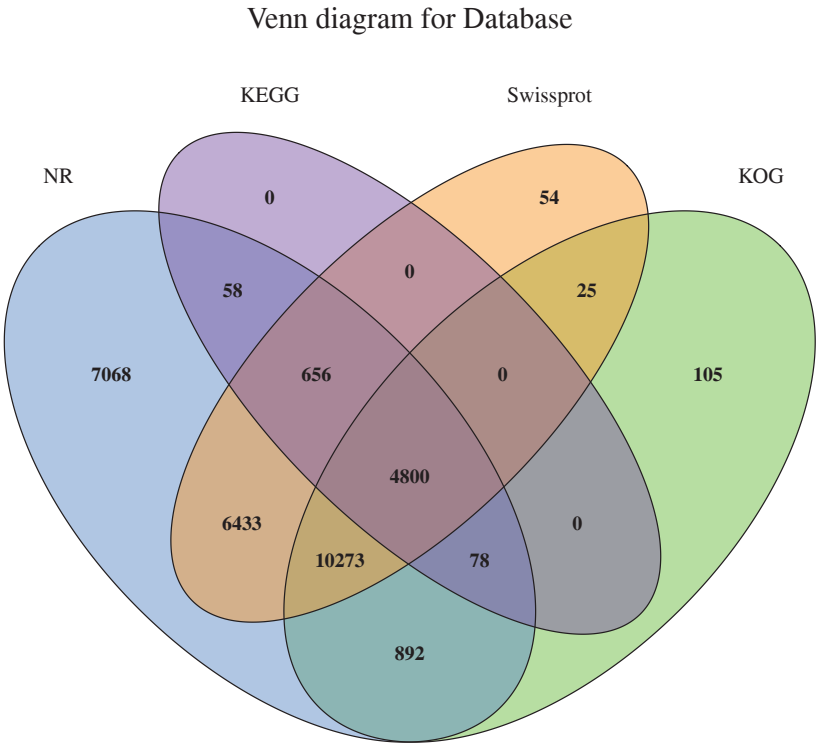


图 5.13 各数据库注释韦恩图

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
R	General function prediction only	2069	12.79
T	Signal transduction mechanisms	1889	11.68
O	Posttranslational modification, protein turnover, chaperones	1849	11.43
J	Translation, ribosomal structure and biogenesis	1036	6.41
S	Function unknown	1008	6.23
U	Intracellular trafficking, secretion, and vesicular transport	932	5.76
K	Transcription	887	5.48
C	Energy production and conversion	744	4.6
A	RNA processing and modification	672	4.16
G	Carbohydrate transport and metabolism	657	4.06

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：

地　址：上海市松江区香闵路 698 号　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

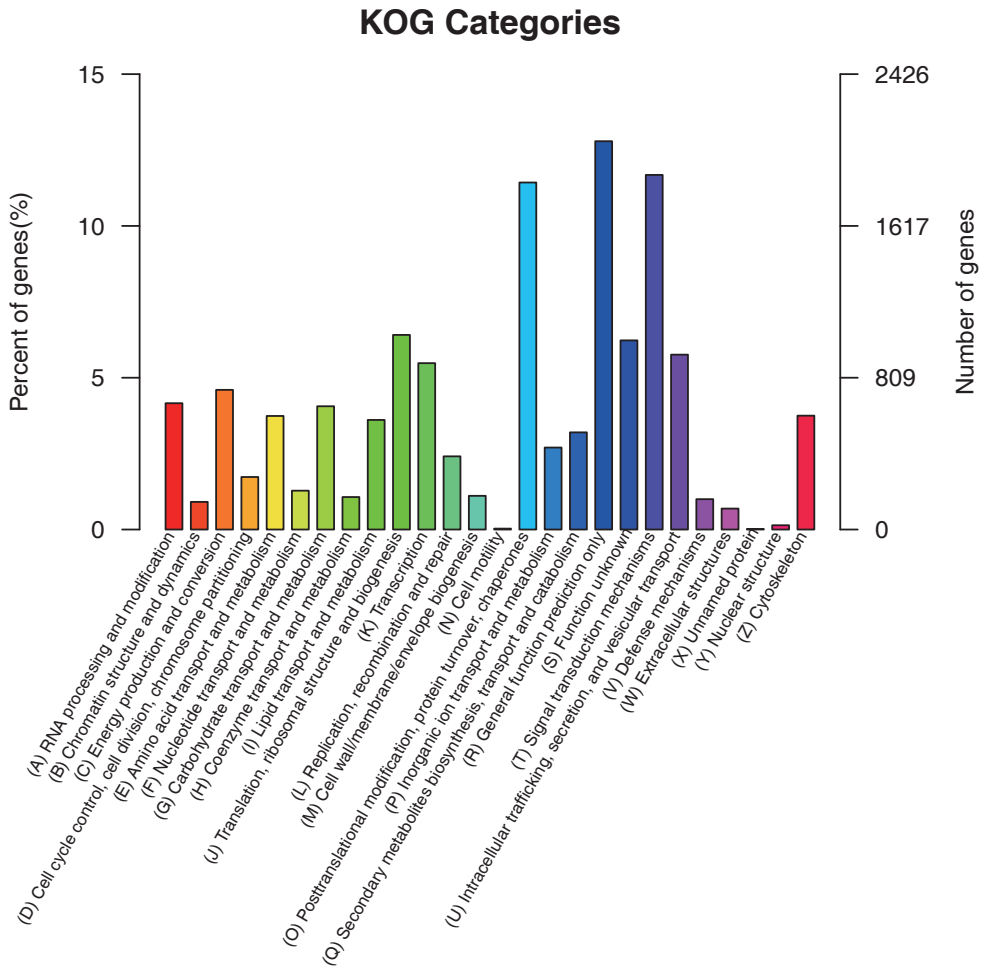


图 5.14 KOG 注释条形图

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：

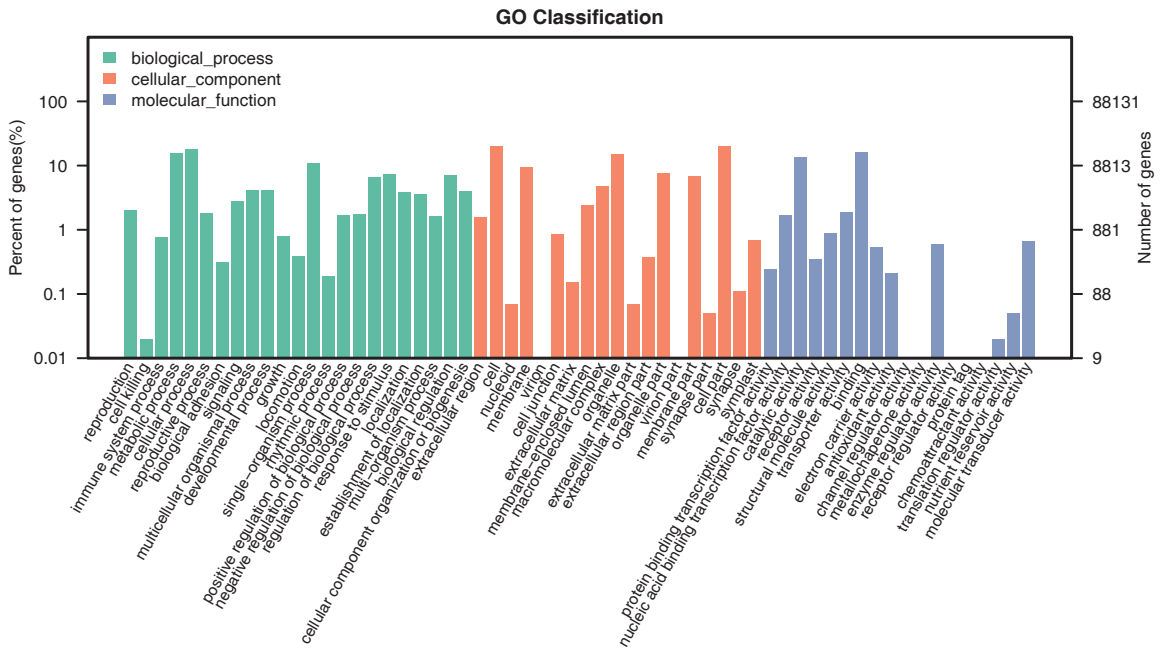


图 5.15 level2 水平 GO 注释上的基因分布

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。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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
ko03010	Ribosome	278
ko01200	Carbon metabolism	161
ko01230	Biosynthesis of amino acids	160
ko03040	Spliceosome	147
ko04075	Plant hormone signal transduction	139
ko04141	Protein processing in endoplasmic reticulum	138
ko03013	RNA transport	128
ko00230	Purine metabolism	120
ko05016	Huntington's disease	119
ko00190	Oxidative phosphorylation	118

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：

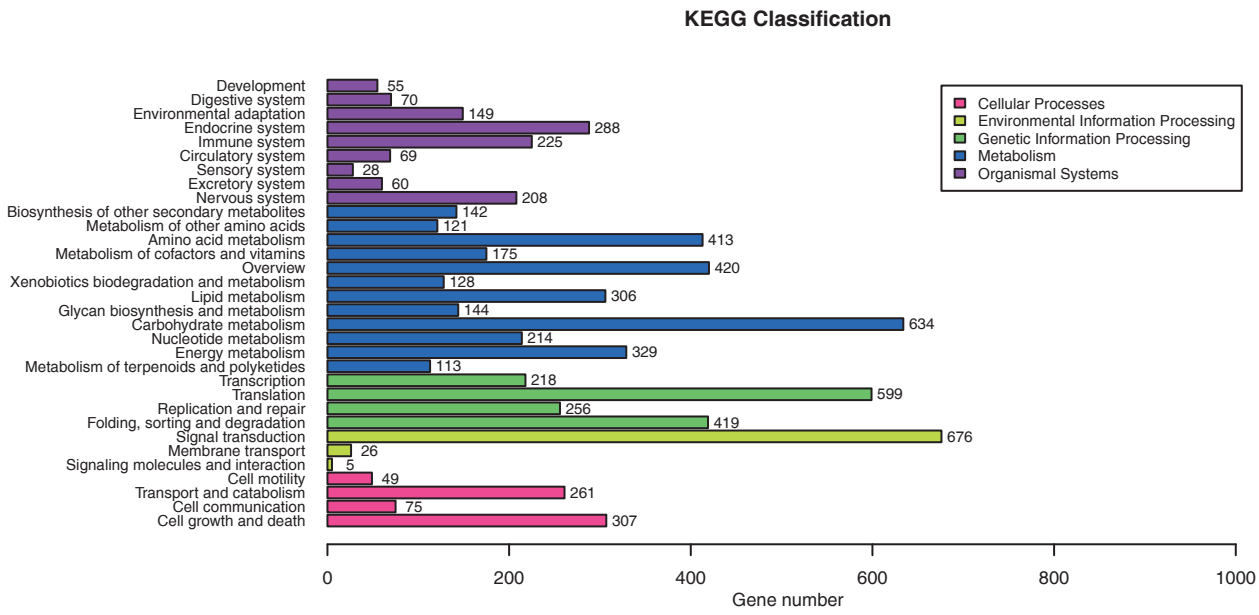


图 5.16 pathway 注释分类条形图

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 的读码框，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：

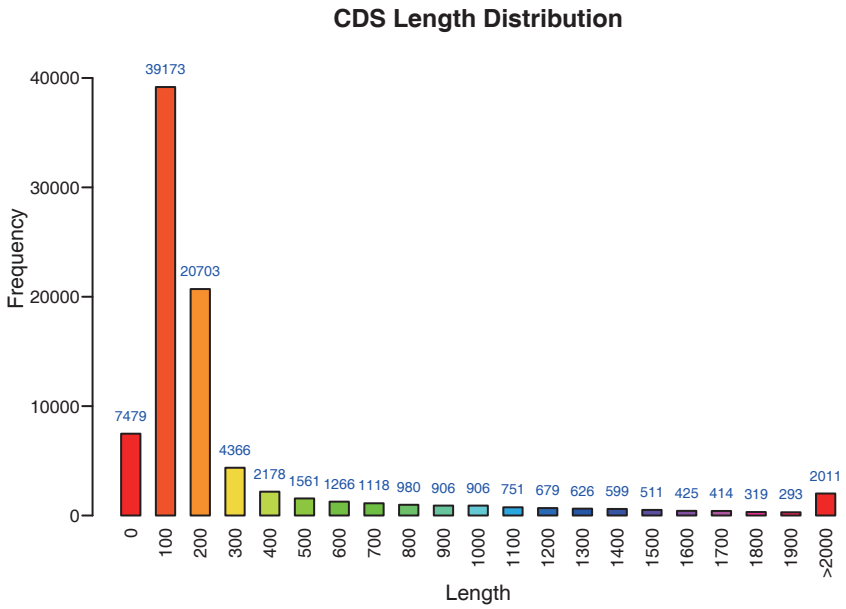


图 5.17 CDS 长度分布图

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：

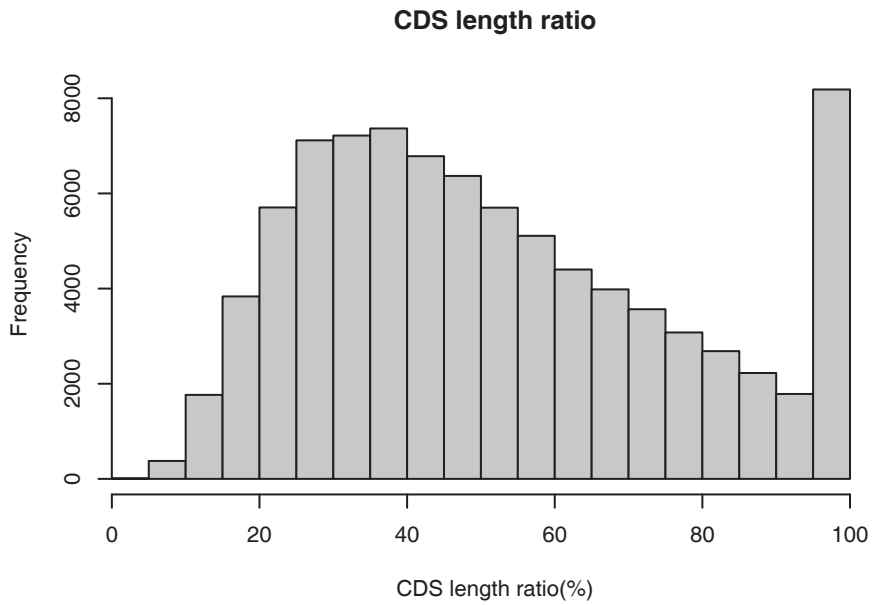


图 5.18 CDS 区域占 Unigene基因长度比例分布图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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	C1	C2	T1	T2
Total reads	32458828	36972340	34670662	40267000
Total mapped	26557207(81.82%)	30009734(81.17%)	28098092(81.04%)	32849690(81.58%)
Mutiple mapped	2895178(8.92%)	3149259(8.52%)	3166489(9.13%)	3289726(8.17%)
Unique mapped	23662029(72.90%)	26860475(72.65%)	24931603(71.91%)	29559964(73.41%)
Read1 mapped	12347541(38.04%)	13983736(37.82%)	13099863(37.78%)	15426315(38.31%)
Read2 mapped	11314488(34.86%)	12876739(34.83%)	11831740(34.13%)	14133649(35.10%)
Mapped to '+'	11845959(36.50%)	13456289(36.40%)	12487367(36.02%)	14803261(36.76%)
Mapped tp '-'	11816070(36.40%)	13404186(36.25%)	12444236(35.89%)	14756703(36.65%)
Non-splice reads	16084525(49.55%)	18680886(50.53%)	17258252(49.78%)	20309666(50.44%)
Splice reads	7577504(23.34%)	8179589(22.12%)	7673351(22.13%)	9250298(22.97%)
Reads mapped in proper pairs	21008118(64.72%)	23931742(64.73%)	21995996(63.44%)	26279516(65.26%)

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：

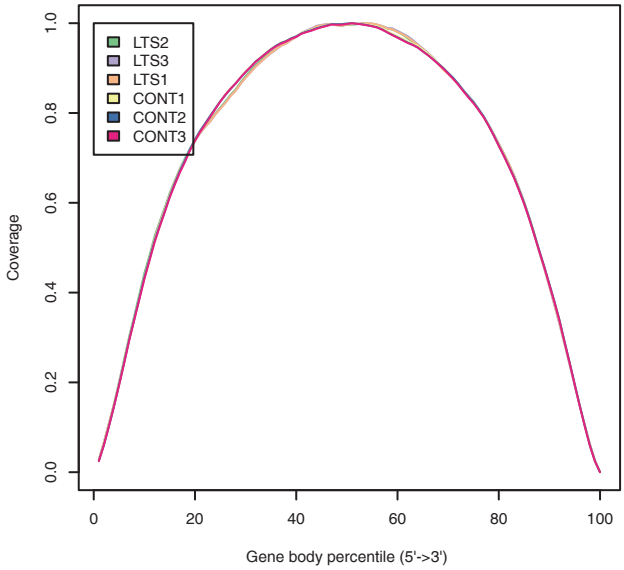


图 5.19 所有样本均一化分布曲线

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。

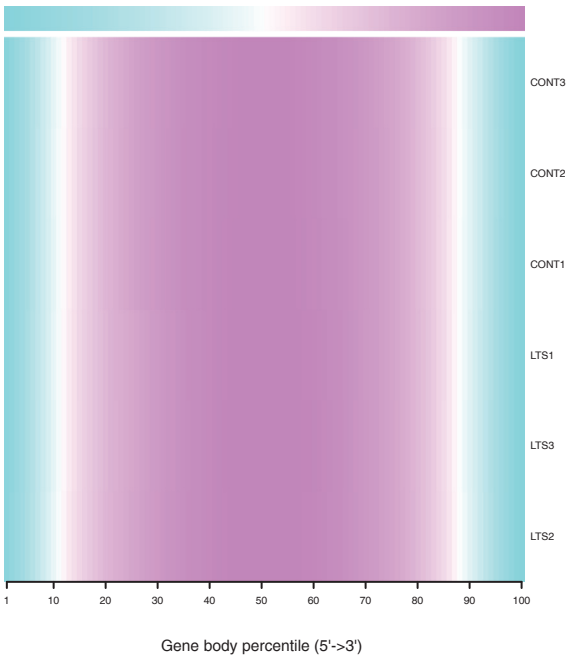


图 5.20 均一化分布热图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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。

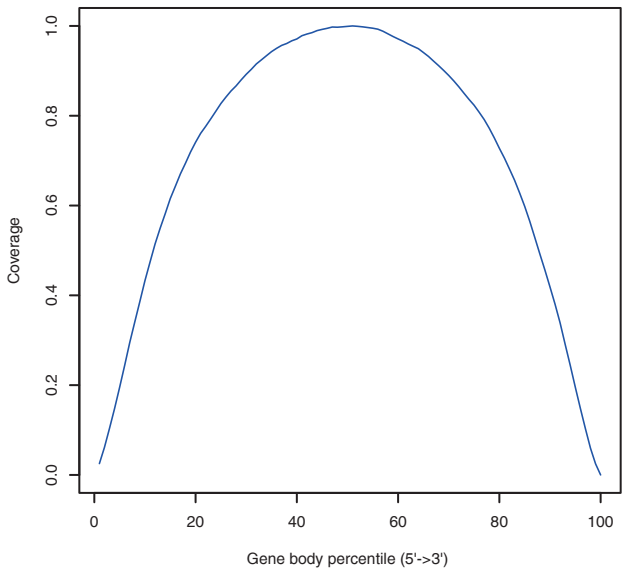


图 5.21 单样本均一化分布曲线

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：

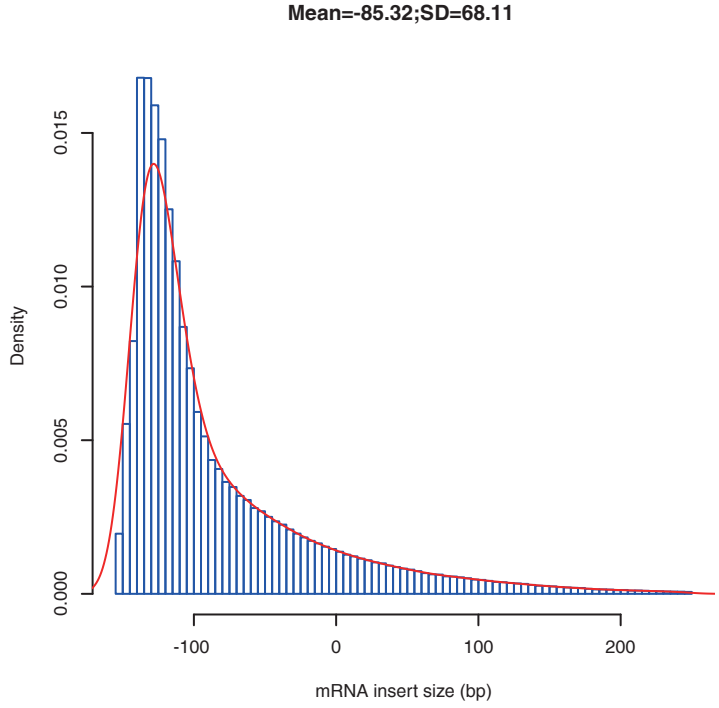


图 5.22 单样本 inner_distance 直方图

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：

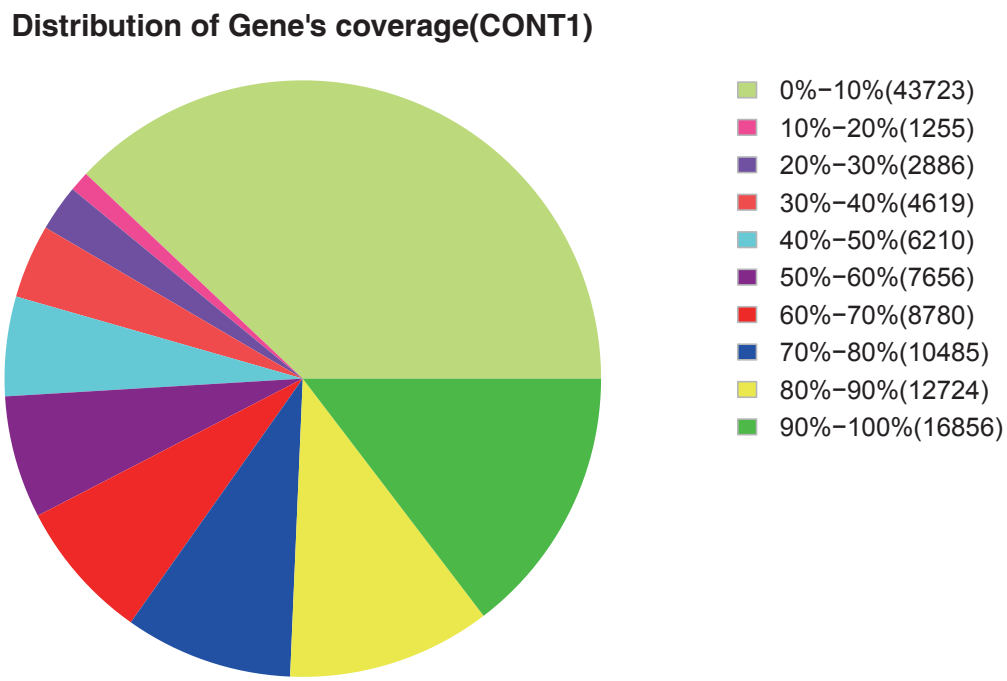


图 5.23 单样本基因覆盖度饼图

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：

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

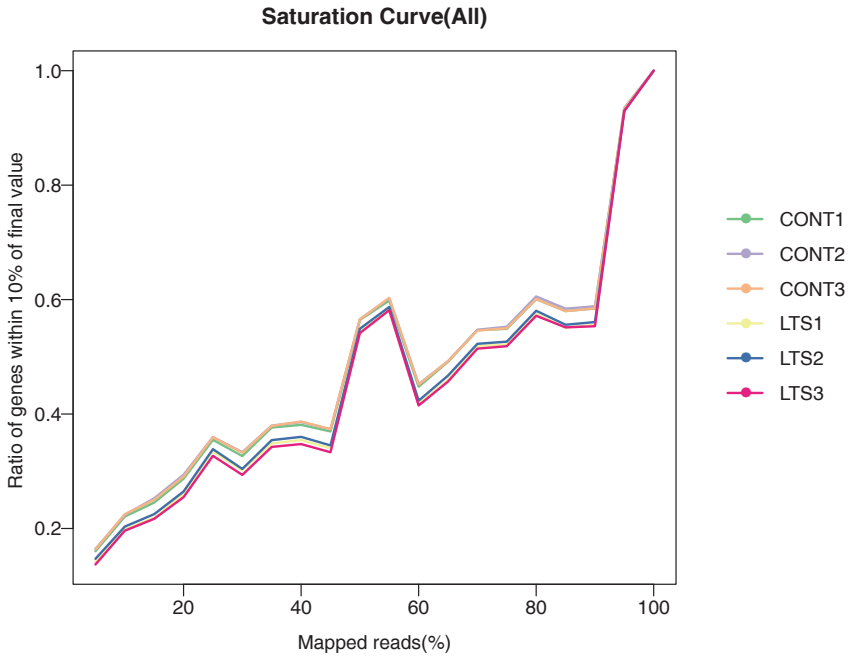


图 5.24 所有样本所有基因饱和度曲线

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：

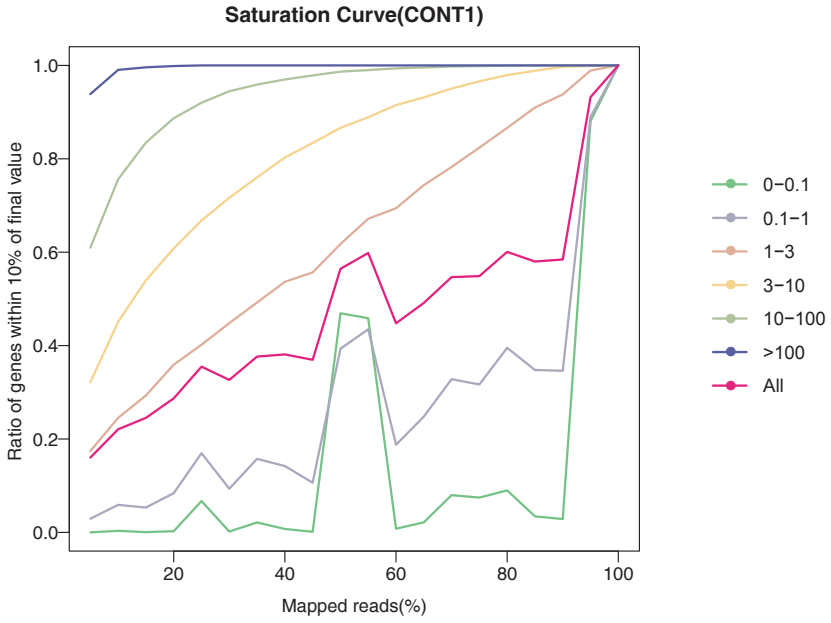


图 5.25 单样本基因饱和度曲线

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

	C1	C2	T1	T2
All	32662(100.00%)	32662(100.00%)	32662(100.00%)	32662(100.00%)
Expressed number	16097(49.28%)	16070(49.20%)	15843(48.51%)	16090(49.26%)
0-0.1	17591(53.86%)	17788(54.46%)	17927(54.89%)	17787(54.46%)
0.1-1	3057(9.36%)	3007(9.21%)	2955(9.05%)	2959(9.06%)
1-3	8208(25.13%)	8058(24.67%)	8045(24.63%)	8096(24.79%)
3-10	2924(8.95%)	2902(8.88%)	2795(8.56%)	2891(8.85%)
10-100	0(0.00%)	0(0.00%)	0(0.00%)	0(0.00%)
>100	882(2.70%)	907(2.78%)	940(2.88%)	929(2.84%)

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：

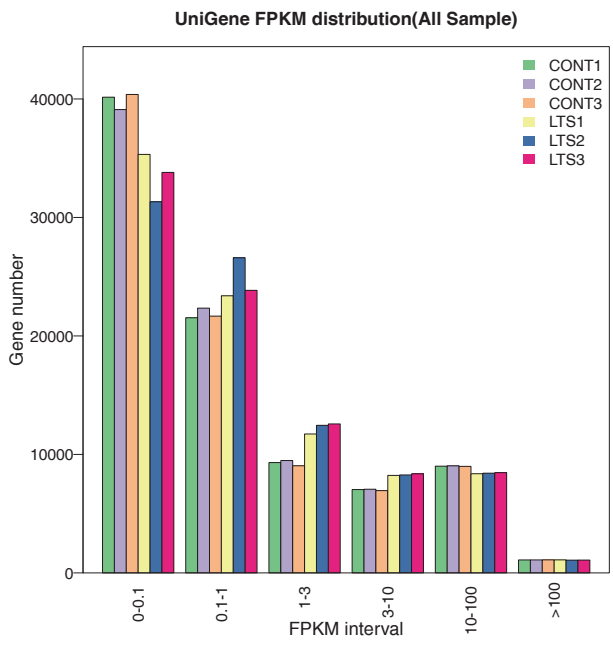


图 5.26 表达量区间条形图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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

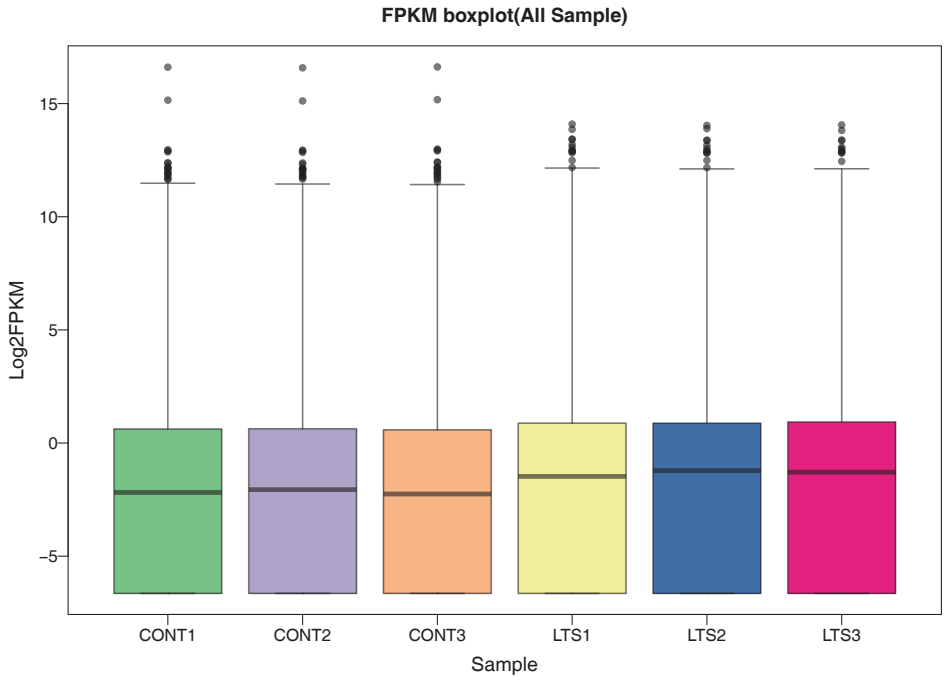


图 5.27 基因表达量盒状图

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：

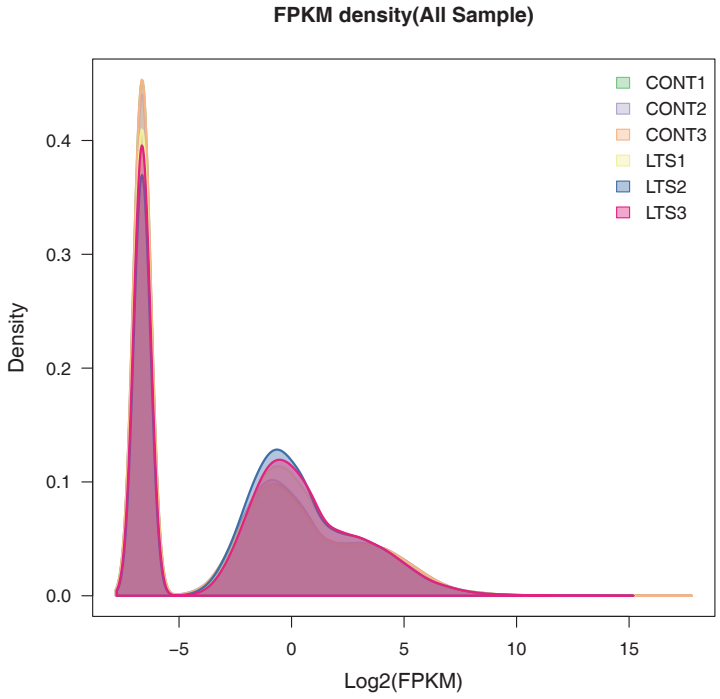


图 5.28 基因表达量密度曲线

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：

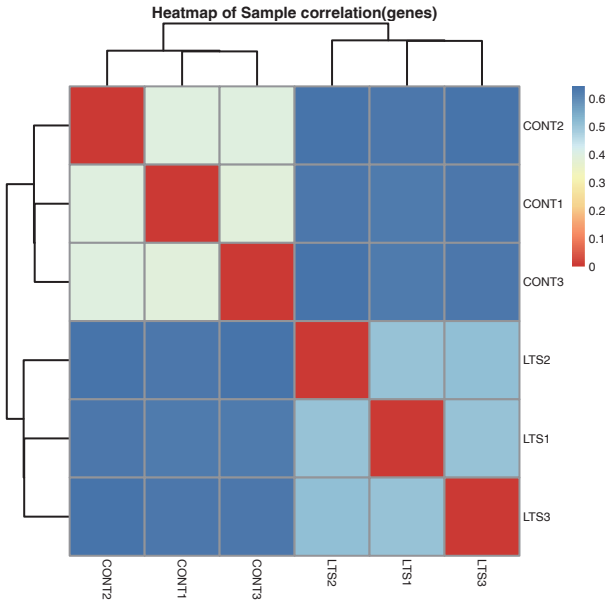


图 5.29 样本间距离热图

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：

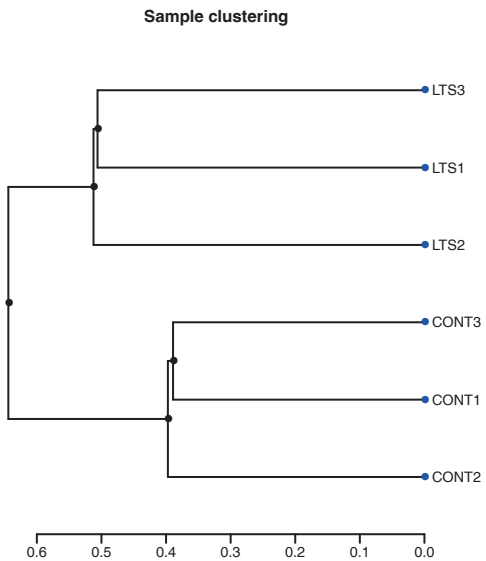


图 5.30 样本间聚类树图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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

	CONT1	CONT2	CONT3	LTS1	LTS2	LTS3
CONT1	1	0.778578	0.781552	0.611861	0.602395	0.605342
CONT2	0.778578	1	0.776422	0.605262	0.596534	0.600413
CONT3	0.781552	0.776422	1	0.611913	0.600308	0.605311
LTS1	0.611861	0.605262	0.611913	1	0.702253	0.702811
LTS2	0.602395	0.596534	0.600308	0.702253	1	0.698366
LTS3	0.605342	0.600413	0.605311	0.702811	0.698366	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：

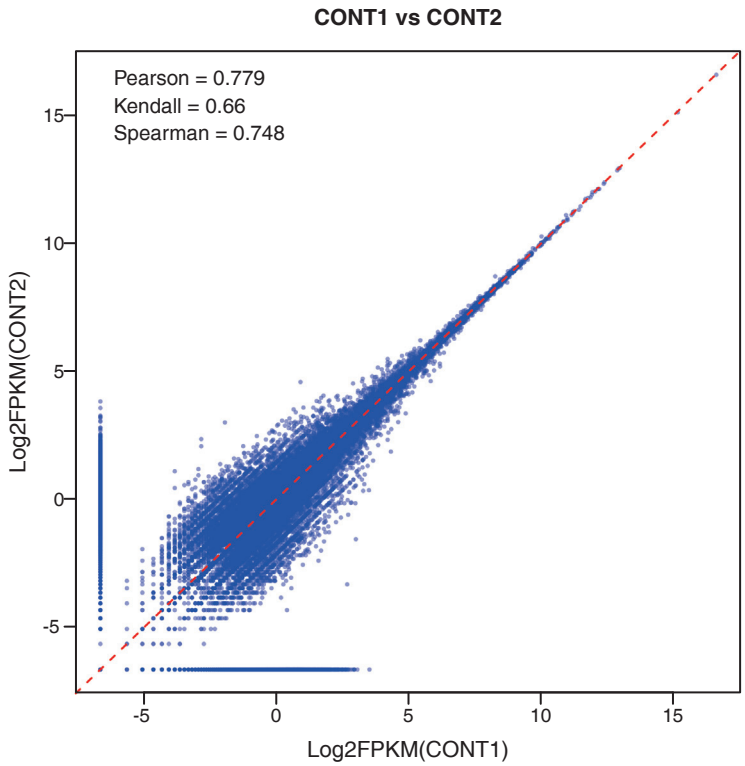


图 5.31 样本间相关性分析图

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：

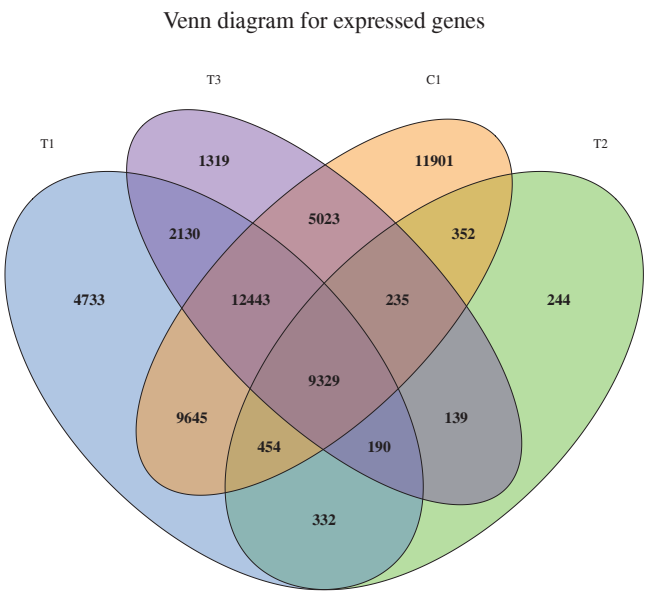


图 5.32 样本间共同表达基因韦恩图

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：

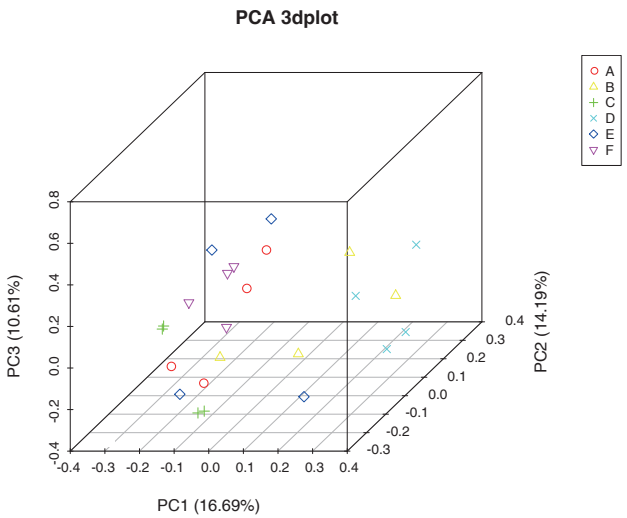


图 5.33 PCA 3Dplot

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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:

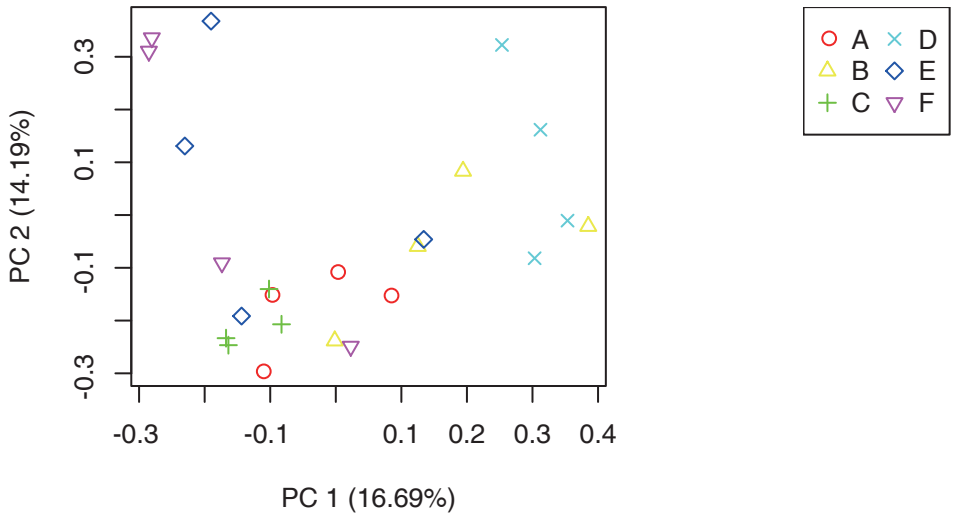


图 5.34 PCA 2Dplot (PC1 vs PC2)

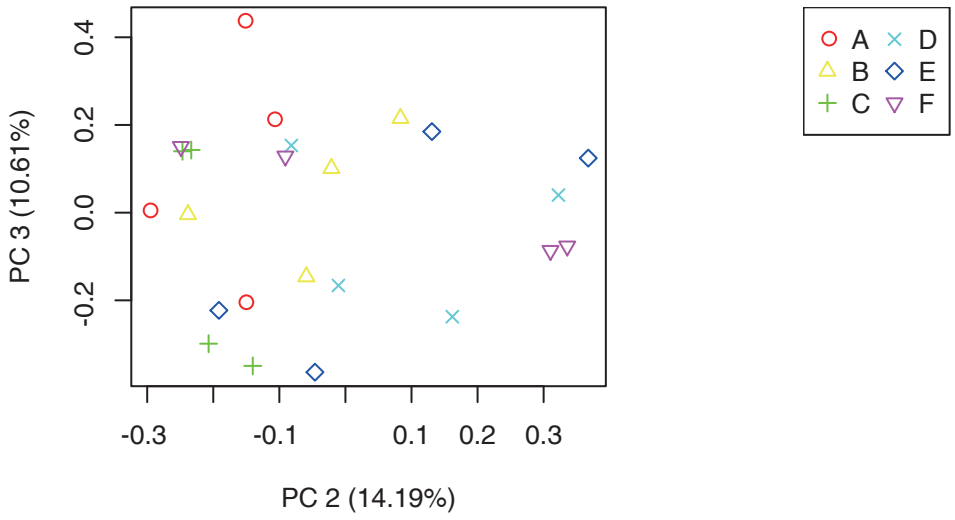


图 5.35 PCA 2Dplot (PC2 vs PC3)

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SNP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：

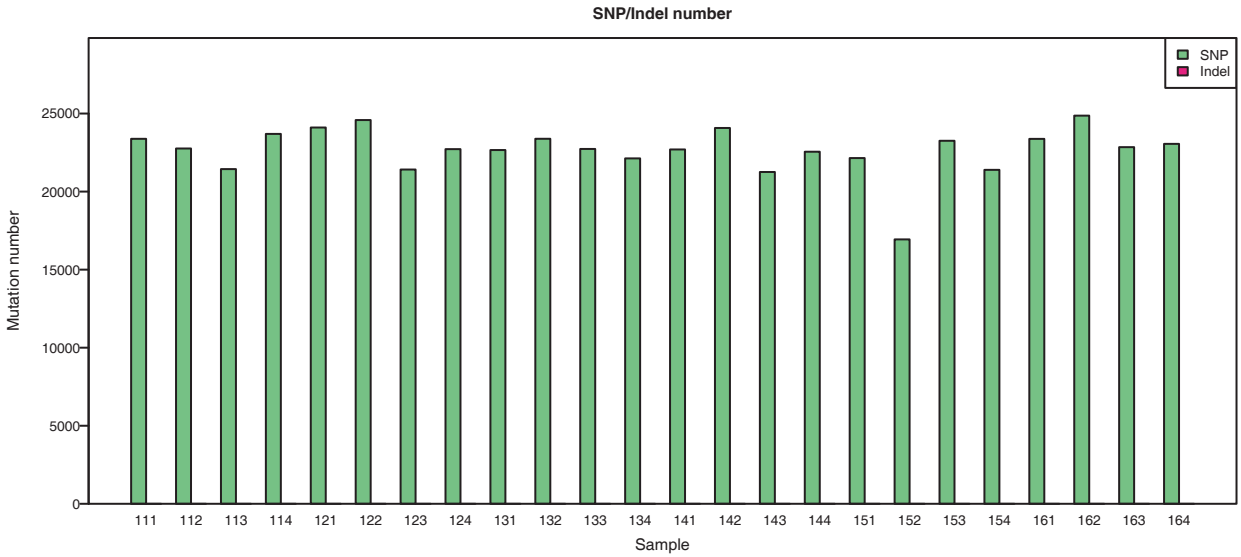


图 5.36 各样本 SNP 数目统计图

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

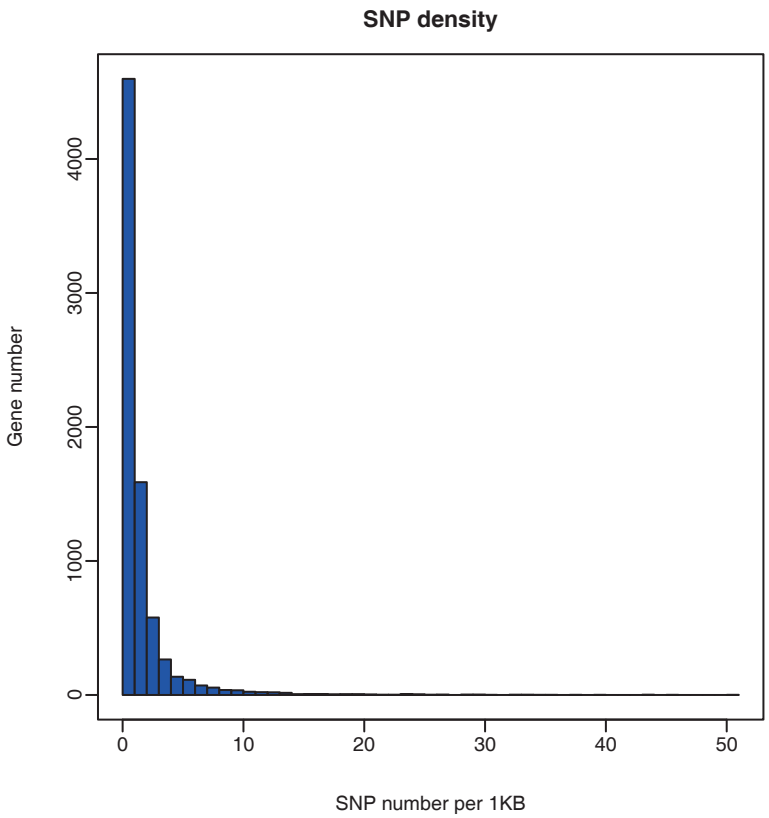


图 5.37 单样本 SNP 密度直方图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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：

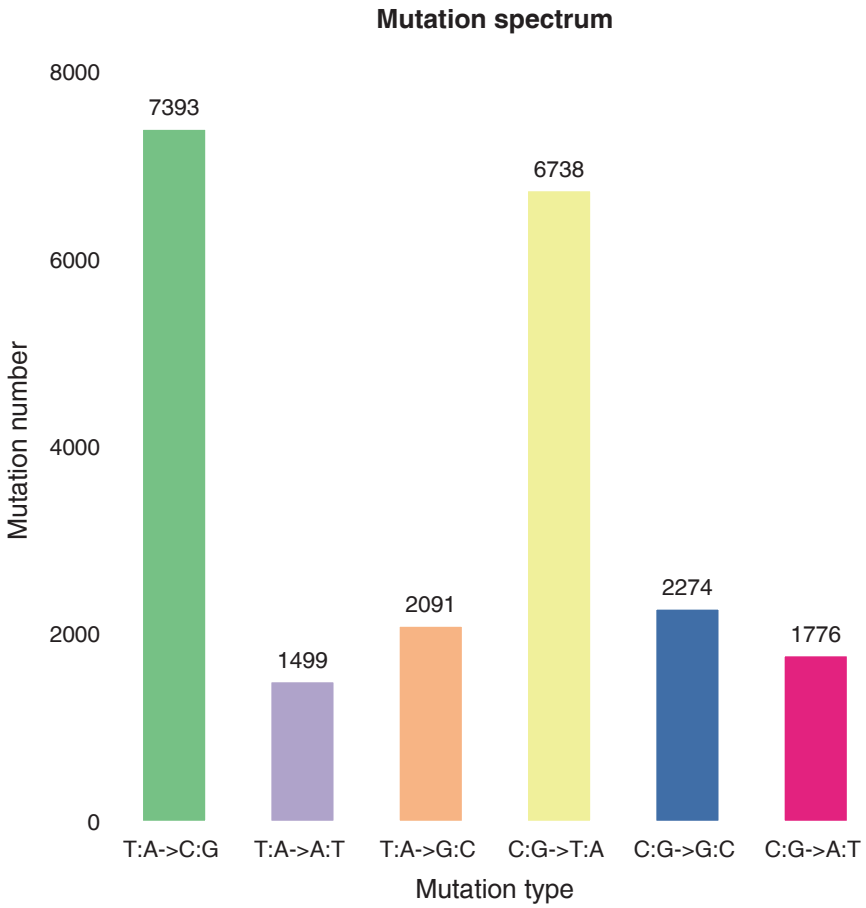


图 5.38 单样本突变谱系图

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：

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

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

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

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

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

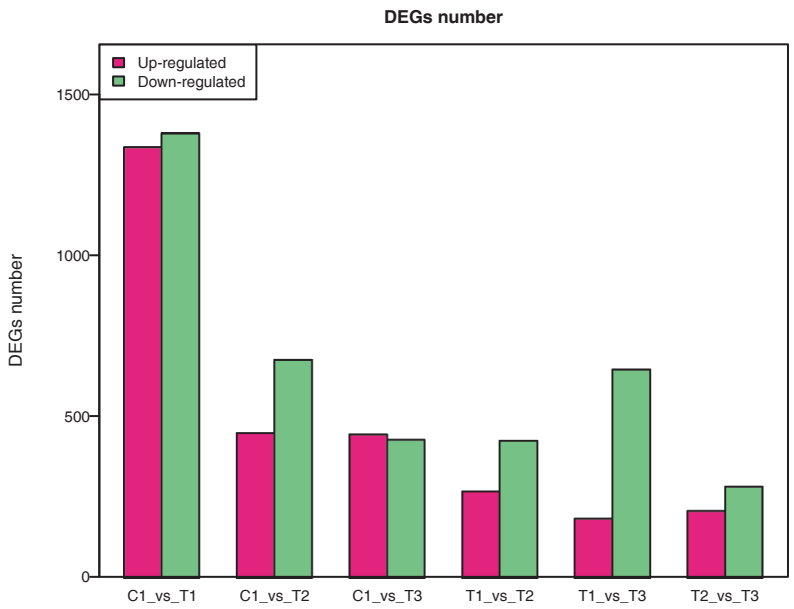


图 5.39 差异基因数目条形图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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

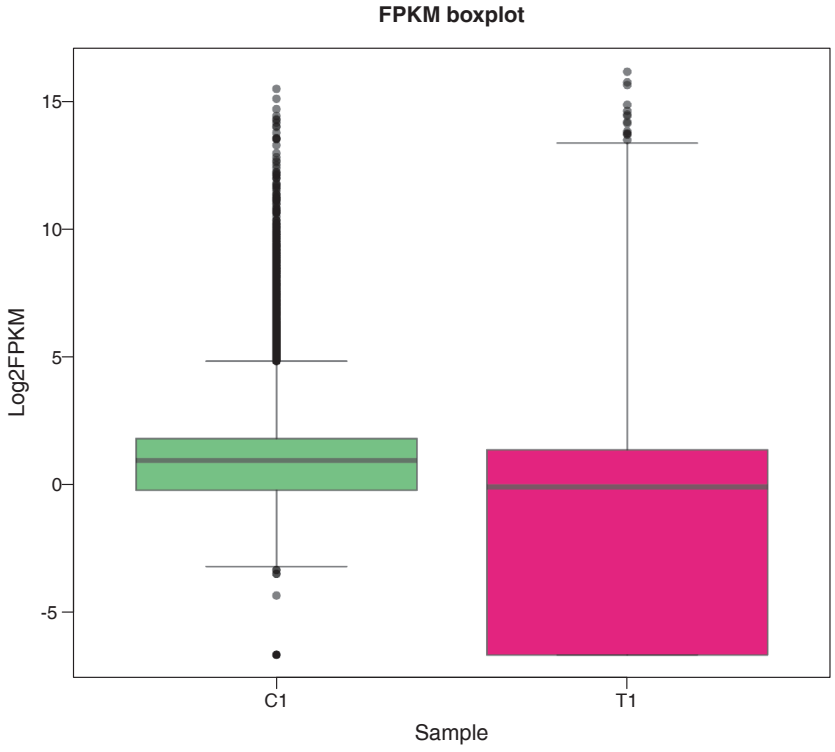


图 5.40 比较对间样本表达量盒状图

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 6.26 一样，下同。

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

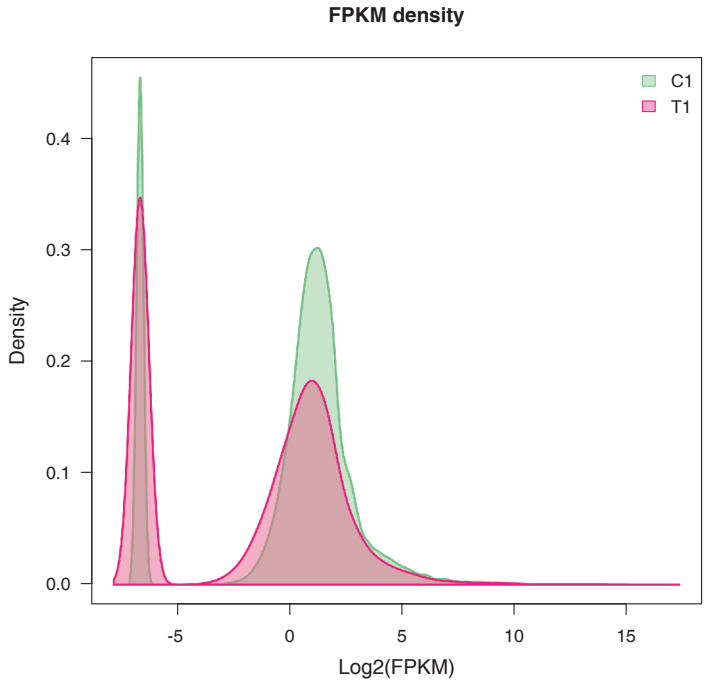


图 5.41 比较对间样本表达密度曲线

A_vs_B/.genes.FPKM.Scatter.plot.pdf：比较对样本间表达量散点图，结果如下图：

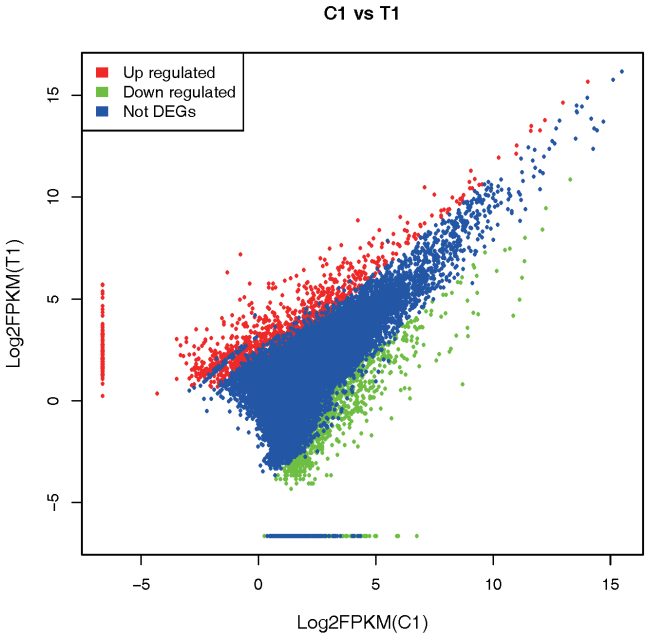


图 5.42 比较对样本件表达量散点图

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.FPKM.MA.plot.pdf：比较对样本件表达量 MA 图，结果展示如下：

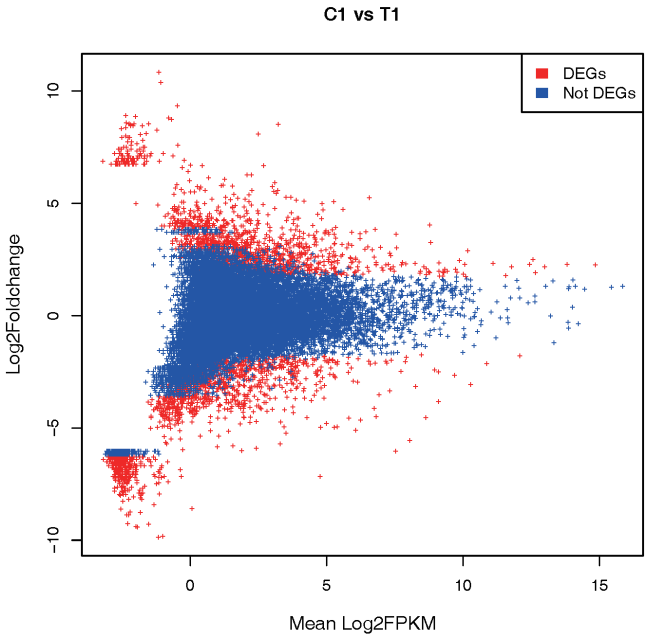


图 5.43 比较对样本件表达量 MA 图

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：

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

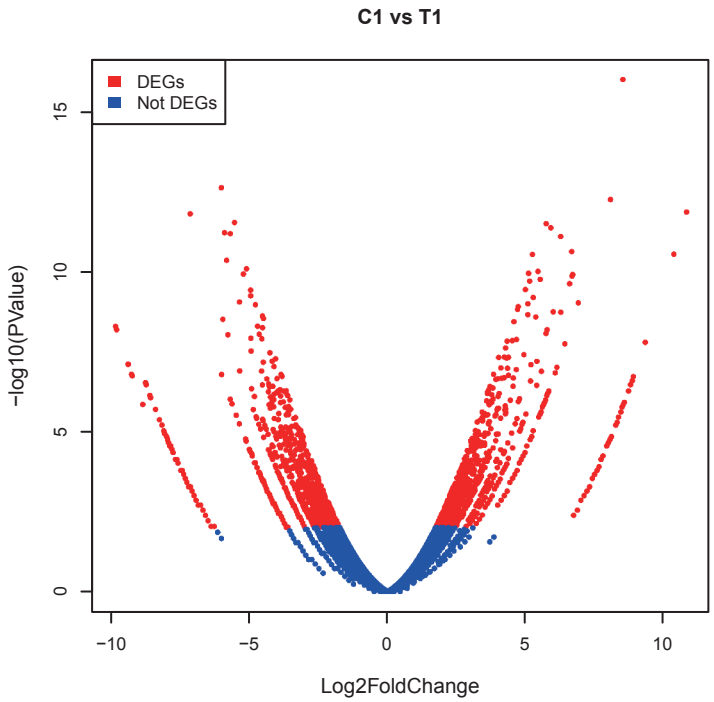


图 5.44 差异分析火山图

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

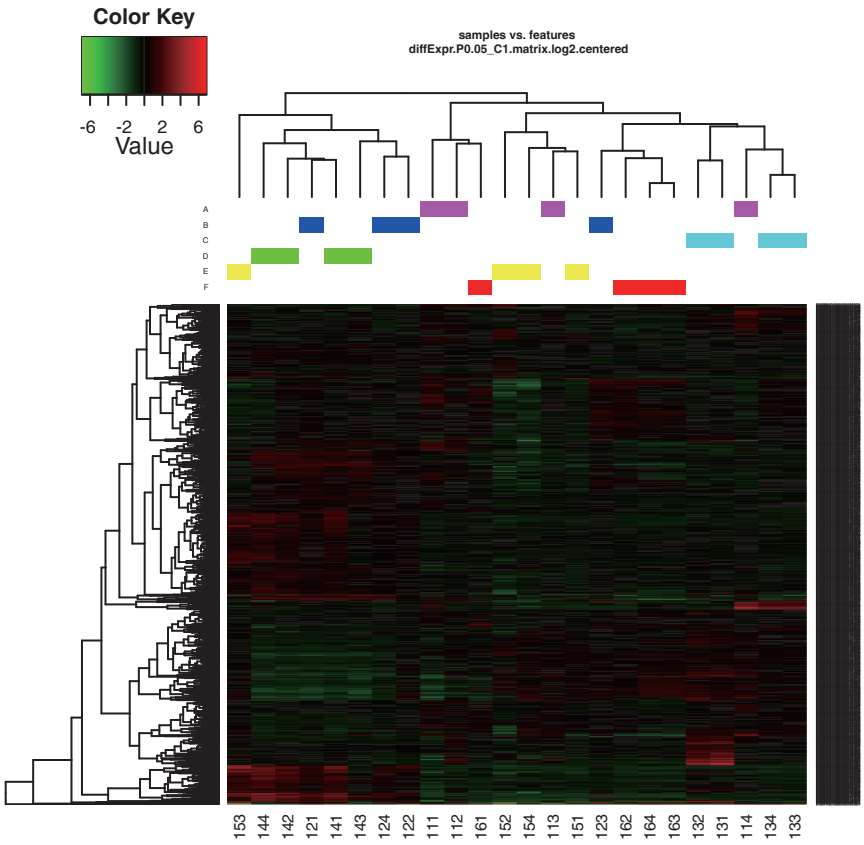


图 5.46 表达模式聚类热图

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未同一 group 或者为生物学重复。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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：

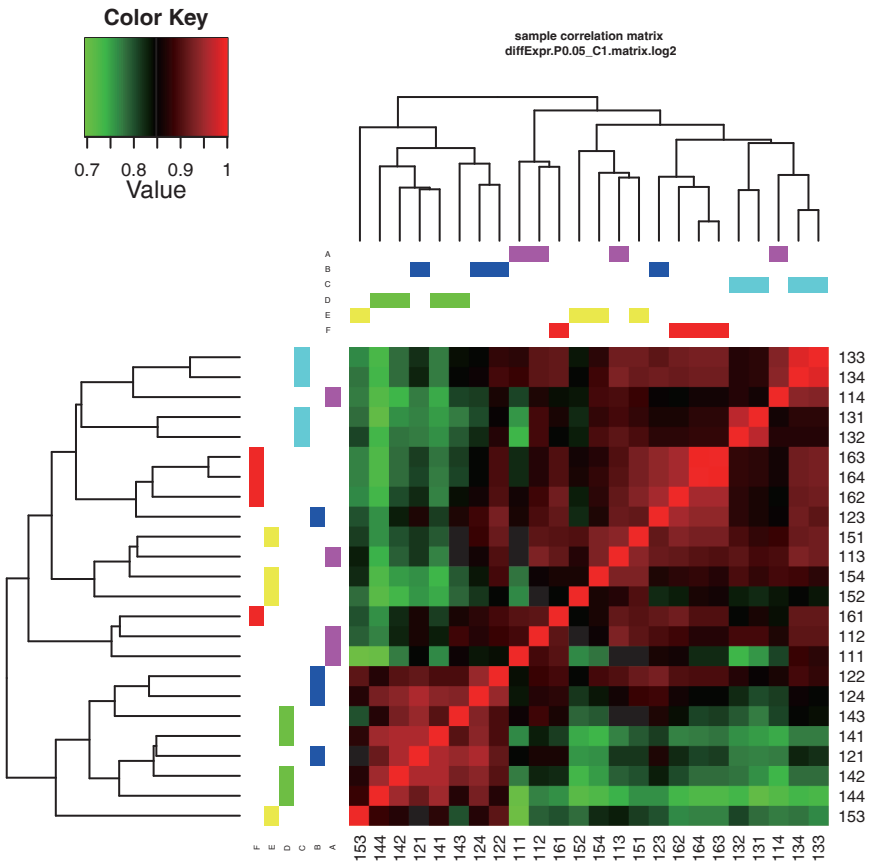
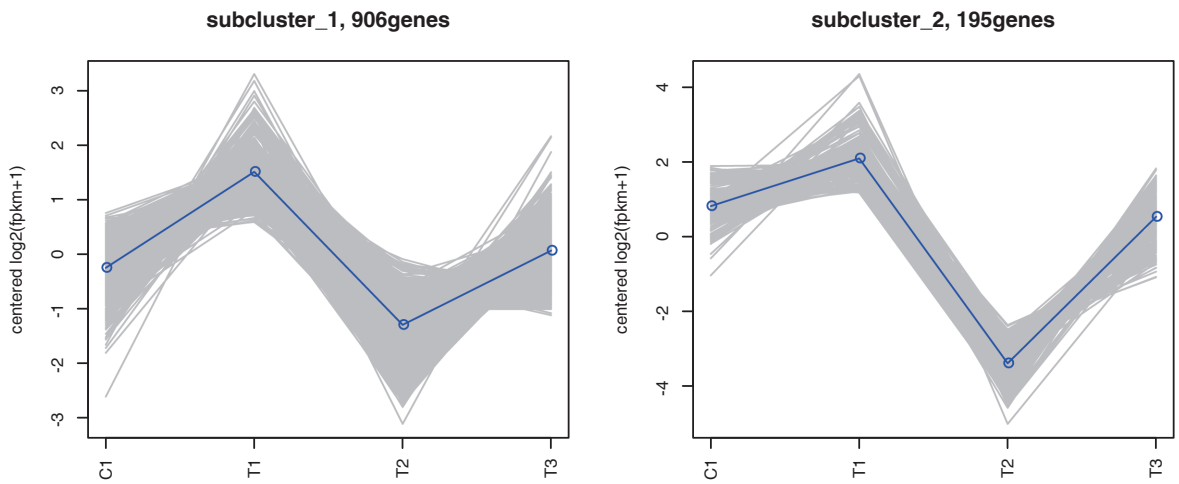


图 5.47 表达相关性热图

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未同一 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：



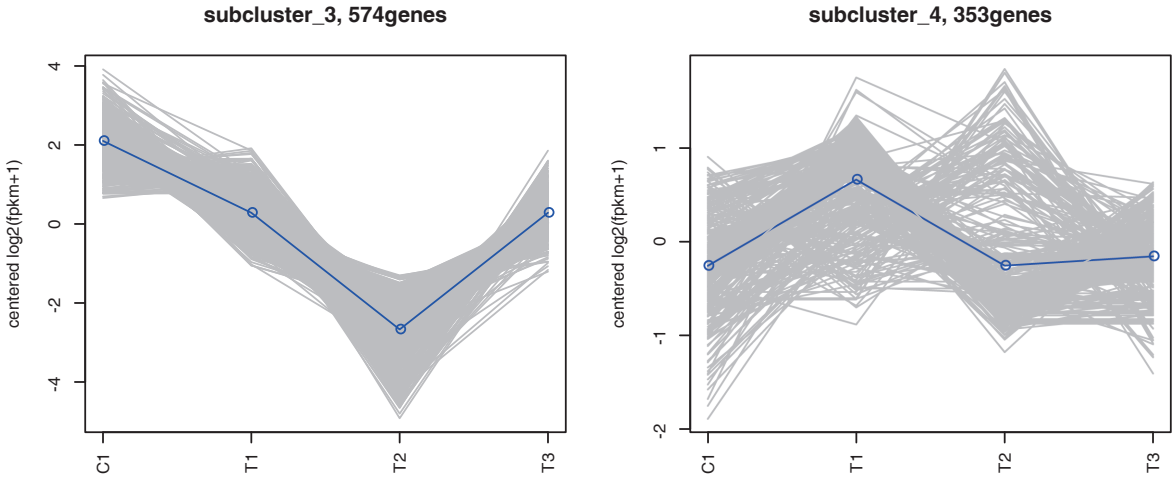


图 5.48 前 4 个 cluster 中基因在各样本中表达量折线图

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：

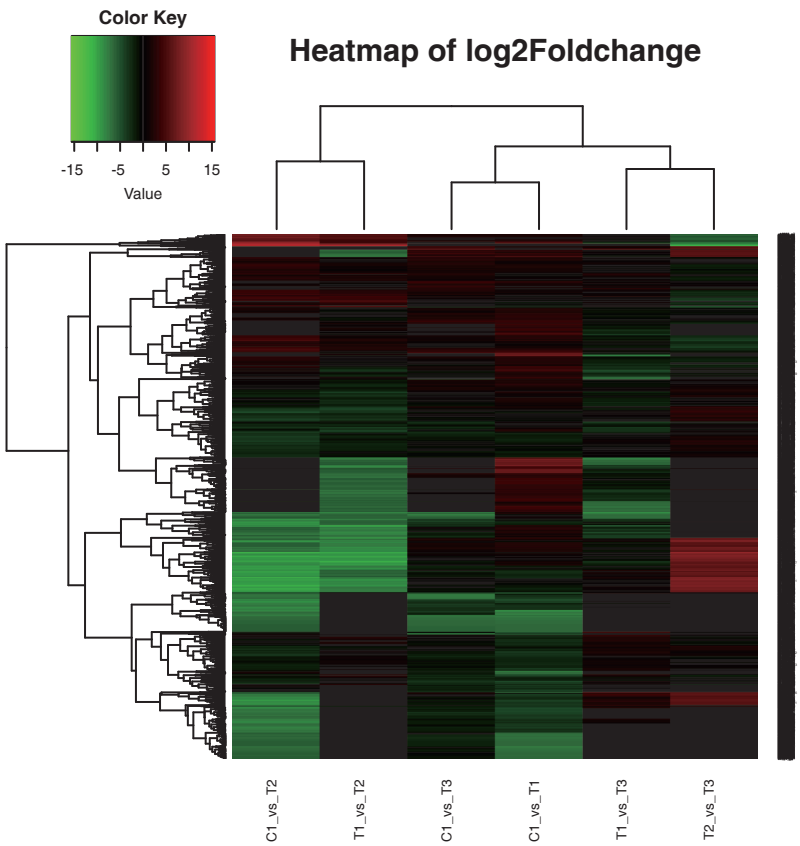


图 5.49 foldchange 热图

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。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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

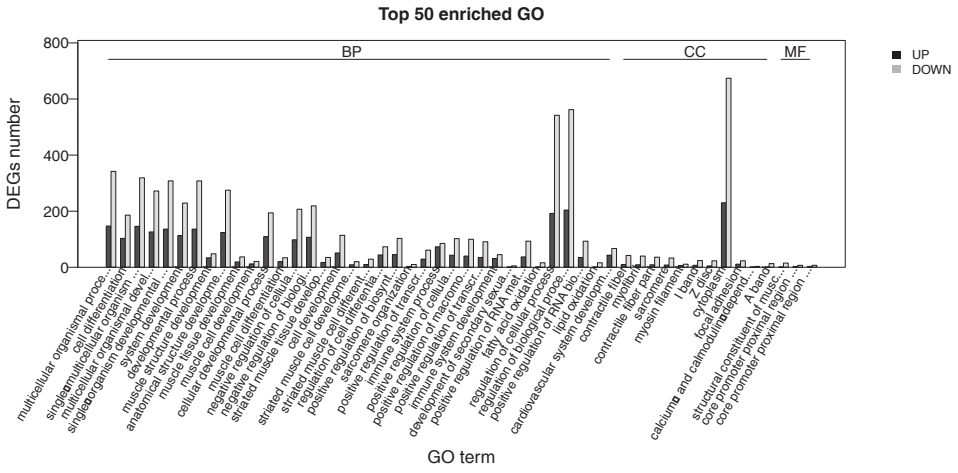


图 5.50 前 50 个 GO 差异基因数目条形图

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：

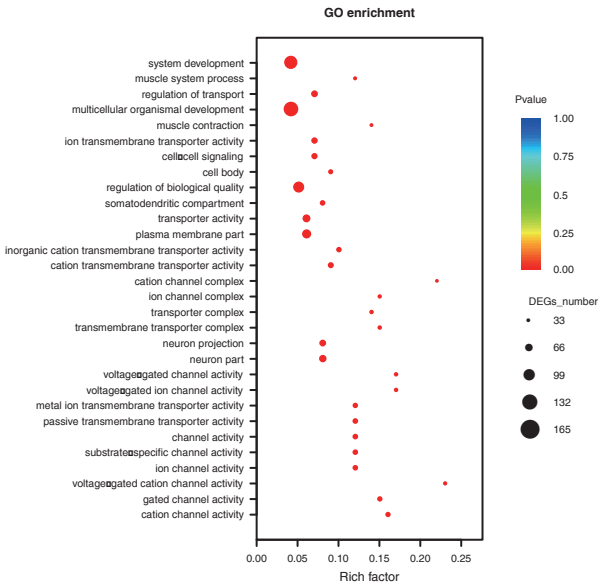


图 5.51 GO 富集分析前 30 个 GO 富集散点图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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

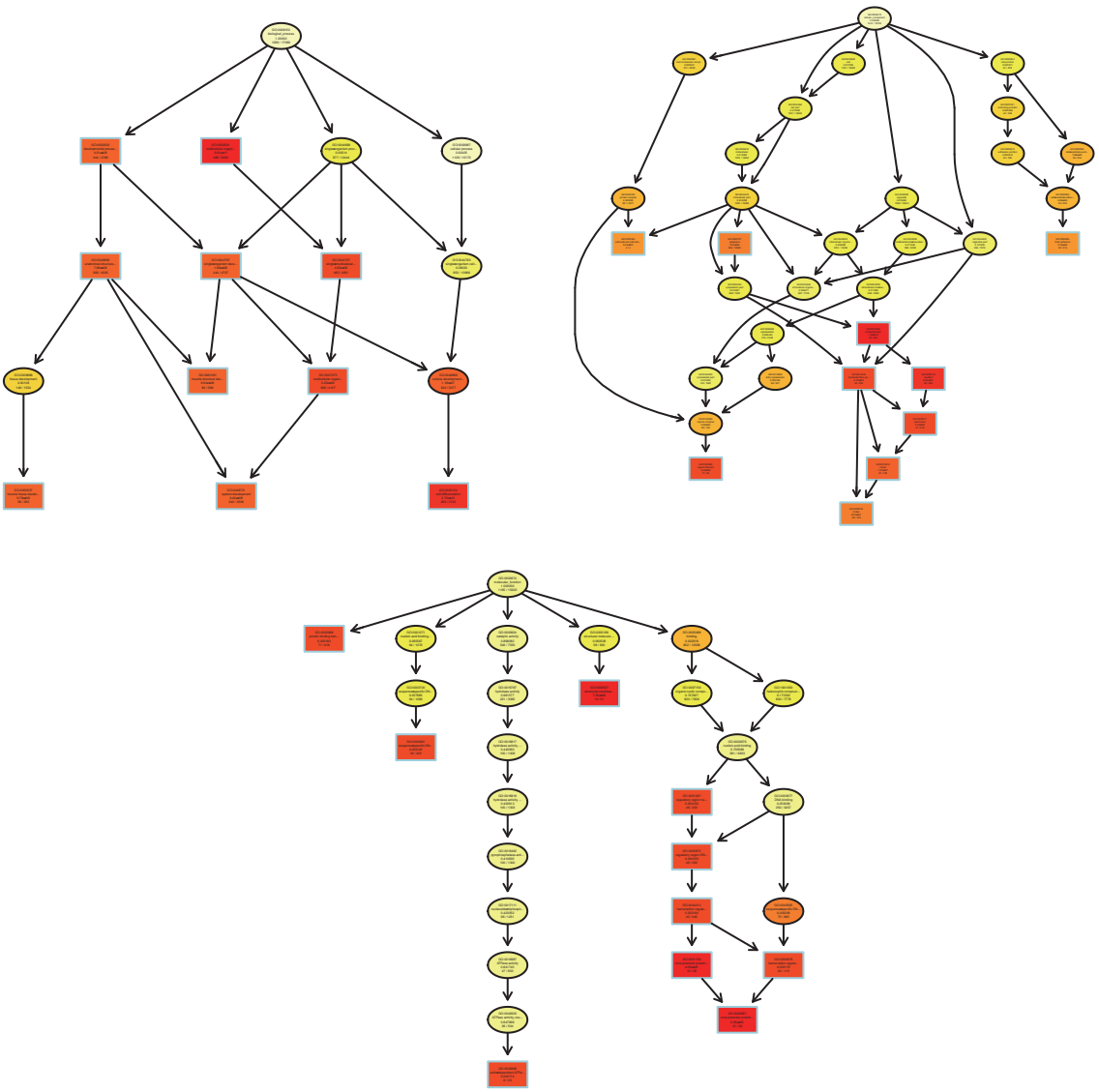


图 5.52 GO 富集有向无环图

说明：topGO 有向无环图（图 19）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：

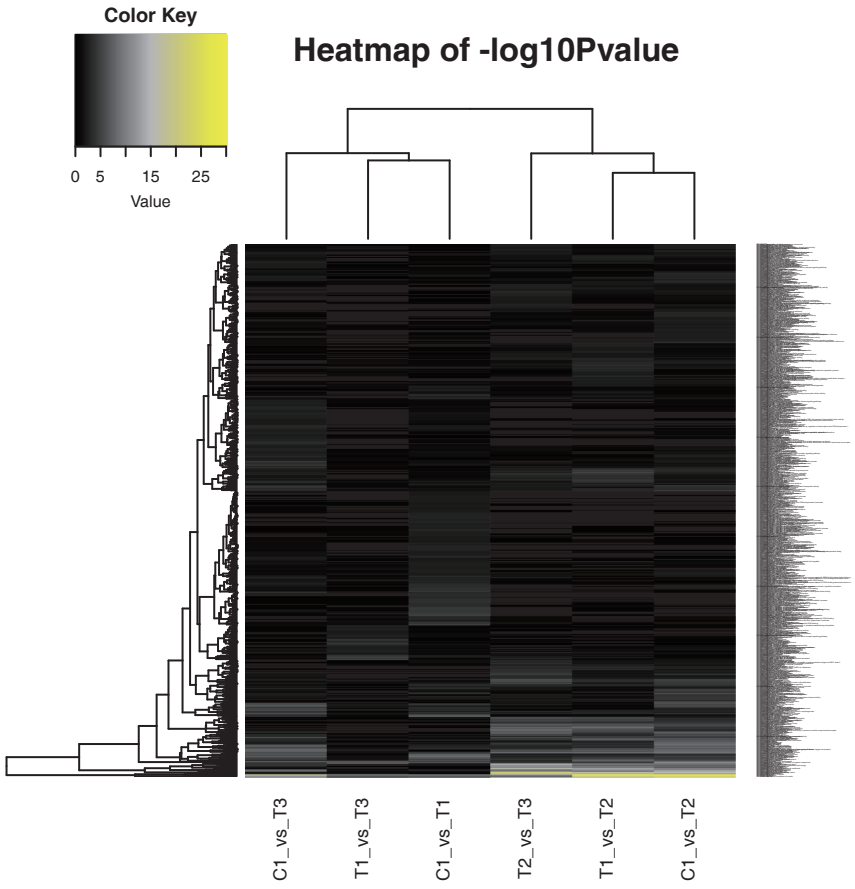


图 5.53 所有显著富集的 GO pvalue 热图

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，上图中内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：

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

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

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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

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

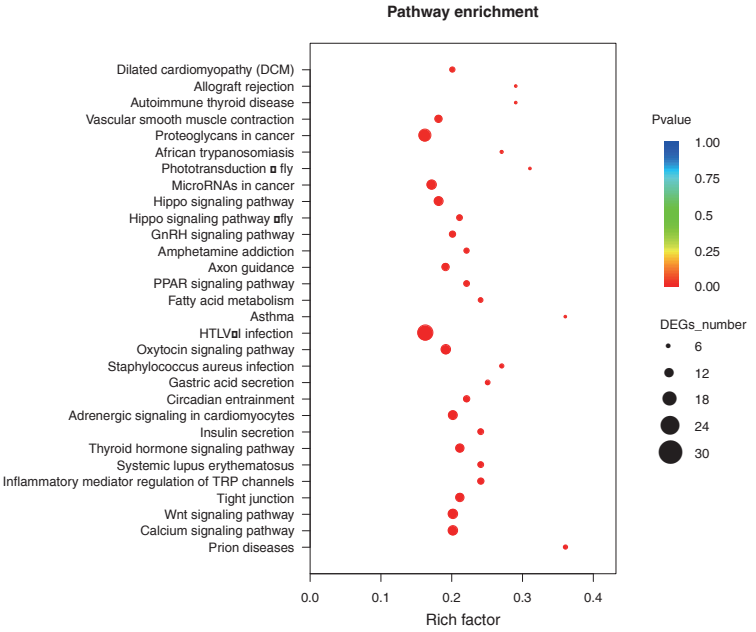


图 5.54 pathway 富集分析前 30 个富集散点图

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：

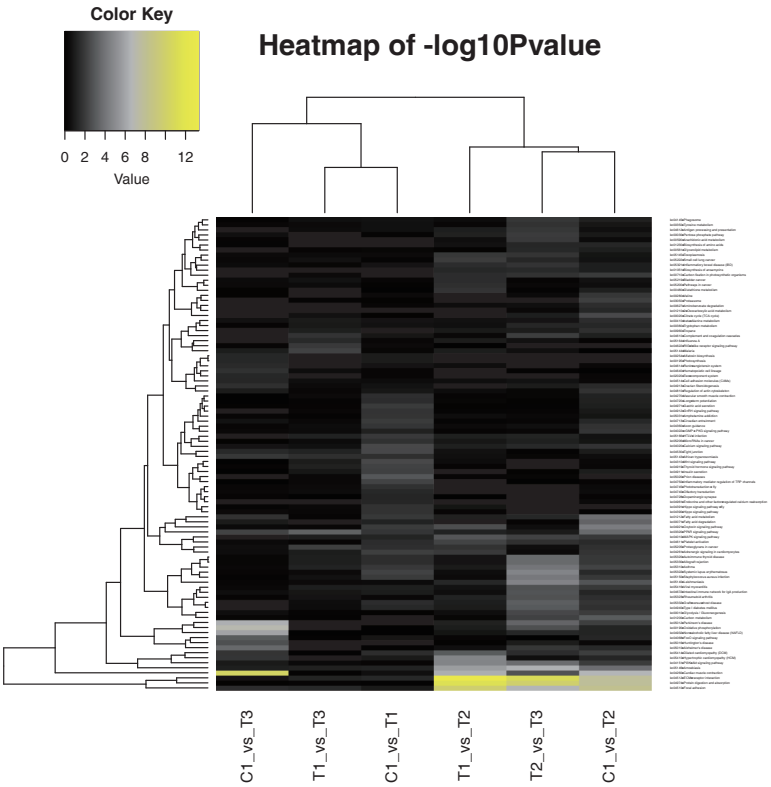


图 5.55 所有显著富集的 pathway pvalue 热图

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。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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，上图中内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：

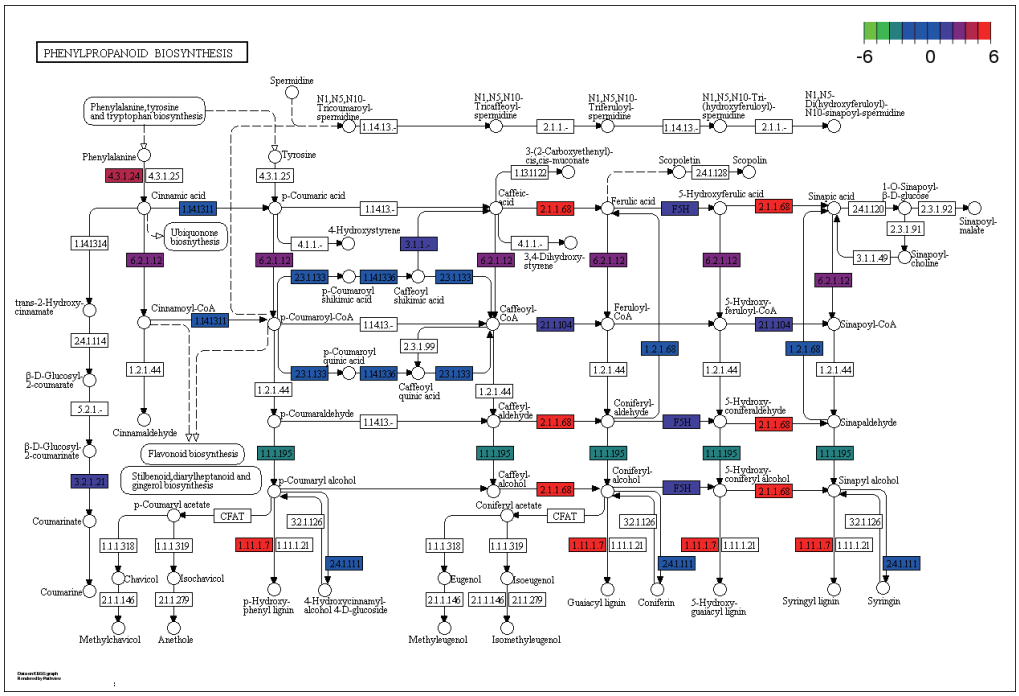


图 5.56 代谢通路上色图

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。

5.12 共表达网络分析

5.12.1 方法说明

基因共表达分析可以揭示转录调控的机制，选定一组基因，通过分析在不同样品中基因间表达量的相关性系数，构建基因间的共表达网络，从而可以明确其中的相互作用关系。分析采用目前较为权威的加权共表达网络构建方法 WGCNA（Weighted correlation networkAnalysis，

http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA）针对所选目标基因进行共表达网络的构建。大致流程如下所述：

（1）针对任意两个基因，采用斯皮尔曼方法计算二者相关系数

（2）选取合适的阈值对相关性进行 cutoff

（3）模拟 B 参数，为网络加权。最终，得到一个符合生物学意义的加权共表达网络。

5.12.2 结果展示

结果目录：11_Co-expression/

***co-expression_cluster.pdf：**共表达基因聚类图，展示如下：

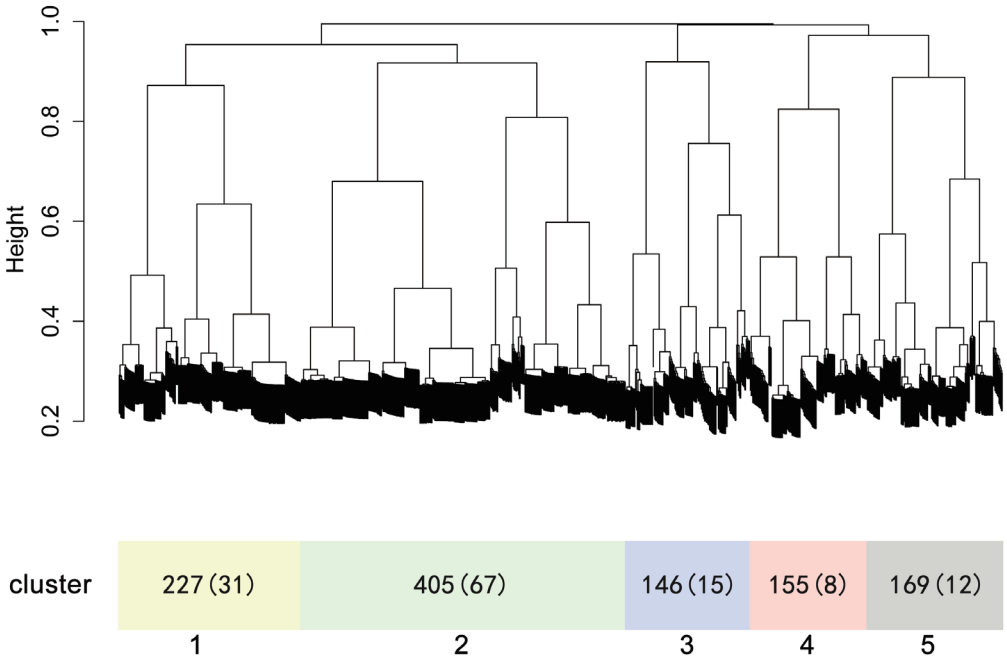


图 5.57 共表达基因聚类图

说明：上图中每一条分支代表一个差异表达基因，存在相同共表达关系的基因被聚类再一起。

***co-expression_network.pdf：**共表达网络图，展示如下：

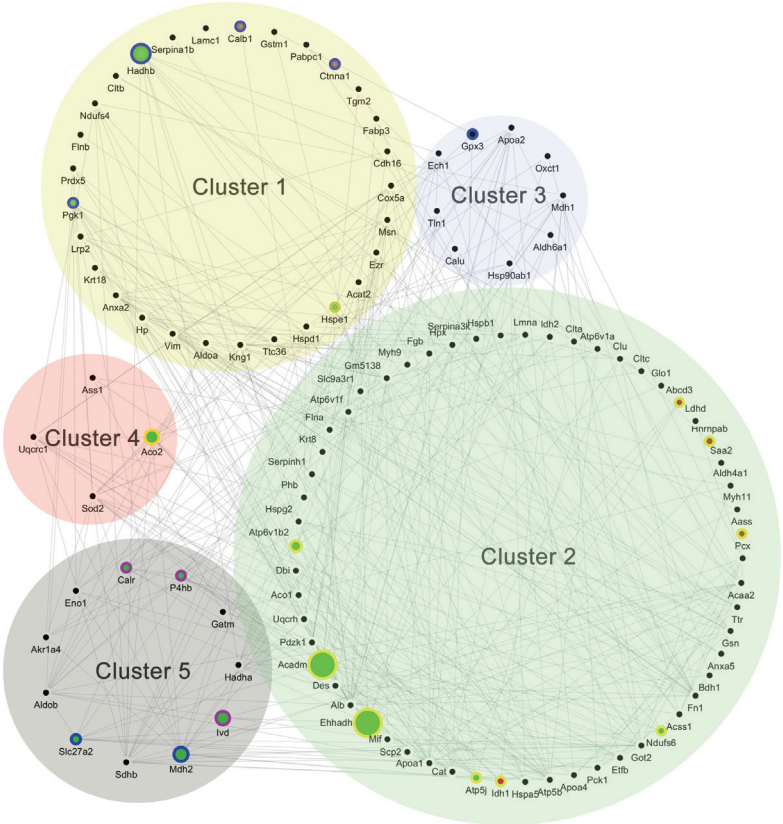


图 5.58 基因共表达网络图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

5.13 蛋白互作分析

5.13.1 方法说明

将差异基因映射至 HPRD（http://www.hprd.org/，Release 9）、biogrid（http://thebiogrid.org/)等蛋白关系网络数据库获得差异基因的相互作用关系，用 Cytoscape 软件（http://www.cytoscape.org/）对差异基因进行网络可视化。统计差异基因互作关系网络的各种拓扑性质，如网络中各个节点的联通性(Degree)，介数中心性(Betweeness)，亲密系数(Closeness)以及聚类系数(Cluster Coefficient)，可以获得互作网络中关键节点。

5.13.2 结果展示

结果目录：12_PPI_result/

***_PPI_network.pdf：**蛋白互作网络图，展示如下：

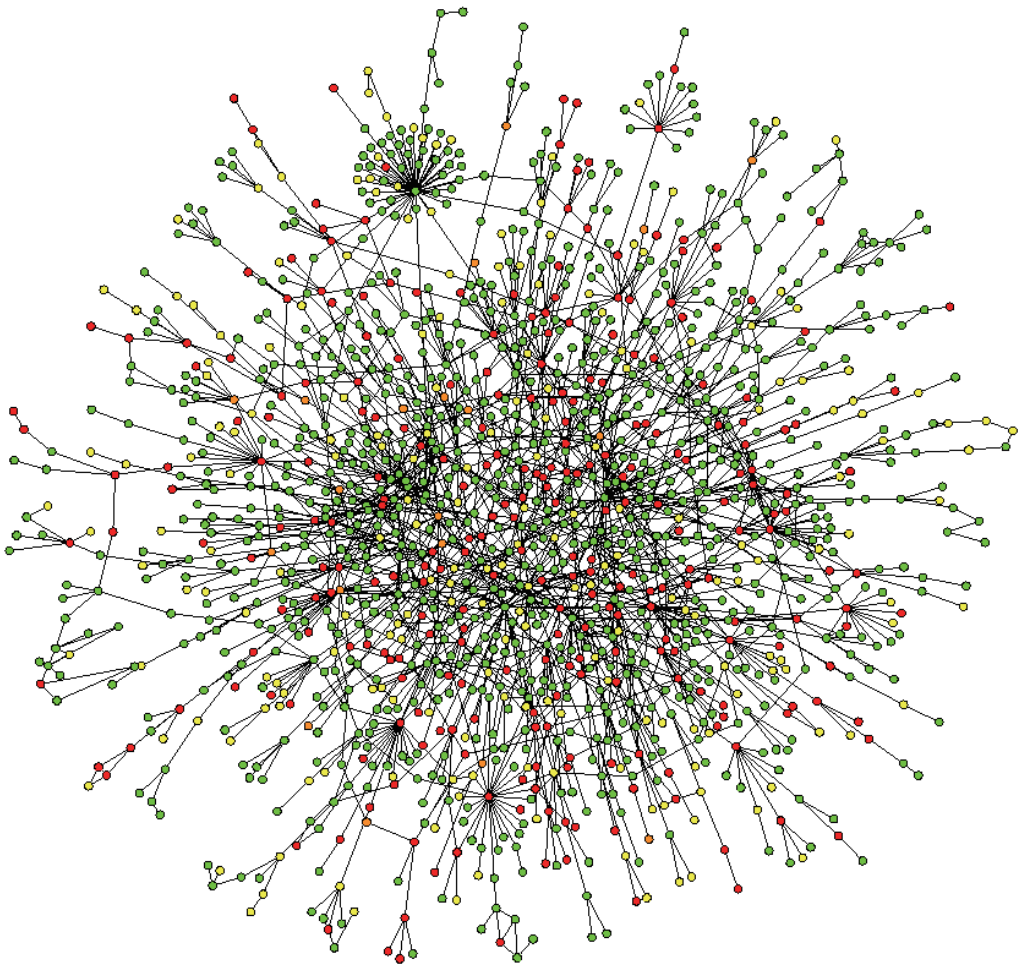


图 5.59 蛋白互作网络图

6. 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.

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

[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.

Small RNA 测序

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

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，改位判定为 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。

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。

Unique 序列：测序中相同序列的集合称之为 Unique 序列

Novel miRNA：数据库中未有的 miRNA，一般为预测的新 miRNA

TPM：TPM（Tag per million）是每百万 reads 中来自某一 miRNA 的 reads 数目，用于评估 miRNA 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。经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，用于反应不同数据集合的共性及特异性。

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，通过此分析可找出差异主要集中在哪些功能下面。

地 址：上海市松江区香闵路 698 号 Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

2. 相关软件及数据库

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。

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

mirDeep2 ：

https://www.mdc-berlin.de/8551903/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep 版本 2.0。

Miranda：

R： https://www.r-project.org 版本 3.1.2。

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

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

2.2 数据库

Pfam 数据库： http://pfam.janelia.org/

Gene Ontology ： http://www.geneontology.org/

MirBase ： http://www.mirbase.org/

NONCODE： http://www.noncode.org/NONCODERv3/

silva rRNA ： http://www.arb-silva.de/

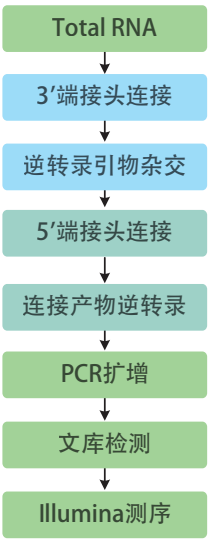
tRNAdb： http://trna.bioinf.uni-leipzig.de/DataOutput/Search

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。

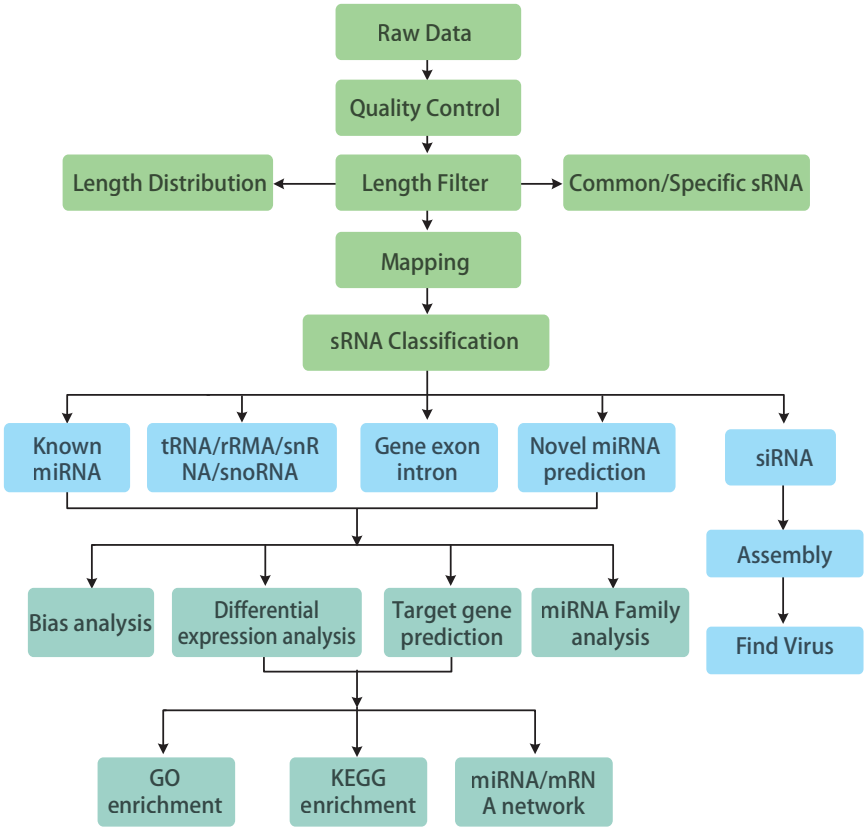
Ensembl ： 欧洲基因组数据库，与 NCBI，UCSC 并为三大基因组数据库，其中可下载大部分物种的基因组序列及相关注释文件；动物基因组：http://asia.ensembl.org/index.html，其它物种基因组：http://ensemblgenomes.org/。

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

3. 实验流程



4. 分析流程



5. 结果展示

5.1 序列预处理

5.1.1 方法说明

应用 cutadapt-1.2.1 和质量分数预处理方法对测序原始 reads 进行预处理，包括去除接头序列预处理和低质量序列预处理， 步骤如下：

1) 去除 3’端接头序列

2) 去除掉低质量序列，平均质量值小于 20 的去掉

3) 去除含 N 的序列

4) 去除含 polyA 尾巴的序列

5) 去除长度过短序列及过长序列（小于 17，大于 35 的序列过滤掉）

6) 做序列长度分布图。

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

5.1.2 结果展示

结果目录：1_QC/

All_sample_data_infor.xlsx：所有样本 QC 结果统计，结果如下：

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

表 5.1 数据统计结果

Sample	Raw Reads	Raw Base	Clean Reads	Clean Base	Average Length	Uniq_num	Clean ratio
0h	20944743	3141711450	7286184	153739482	21.1	419606	34.787651
24h	24521450	3678217500	12018170	259409360	21.58	1268384	49.010846
30h	24308244	3646236600	13825713	311154771	22.51	1326899	56.876642
48h	28478169	4271725350	16520594	380923470	23.06	1847731	58.011433

结论：所有样本原始序列条数均大于 10M，达到分析要求。

***/Length_Distribution.pdf：**各样本小 RNA 长度分布图，展示如下：

各样本长度分布图如下：

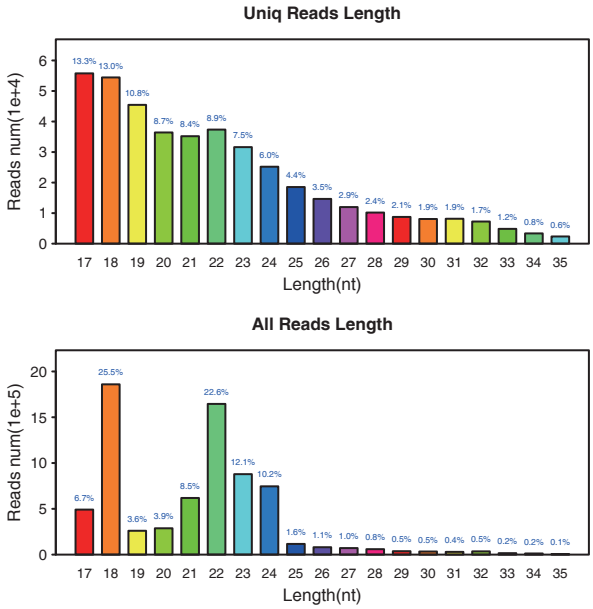


图 5.1 长度分布图

注：上图中只展示了 1 个样本结果，其他结果见 1_QC/对应样本文件夹中。

说明：图中 Total reads 表示总的 reads 数统计结果，Uniq Reads 表示每种拷贝只取一条统计结果。所有样本小 RNA 长度均分布在 18-24 之间，无异常，其它样本结果见 1_QC/中各对应样本文件夹。

Common-Specific_sRNA：样本间 sRNA 种类韦恩图，展示如下：

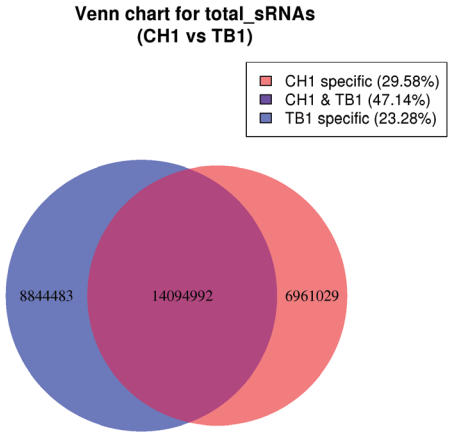


图 5.2 公共及特有序列统计

注：结题报告中只显示一对样本间的公共及特有结果，其他样本间的信息包含在结果文件夹

(1)样本 1 specific：样本 1 的特有序列。

(2)样本 1 & 样本 2：样本间的公共序列。

(3)样本 2 specific：样本 2 的特有序列。

5.2 序列注释和鉴定已知 microRNA

5.2.1 方法说明

将 clean reads 分别比对到人的 tRNA、rRNA、snoRNA、mRNA 等数据库中，允许一个错配，之后将比对到该些数据库的序列过滤；将过滤后序列比对到 mirBase 数据库中人的 miRNA，详细注释结果见表 4.1。

比对软件： bowtie（http://bowtie-bio.sourceforge.net/index.shtml ）bowtie 主要用于测序短序列比对，其特点是速度快，精度高，被大量用于短序列比对。

5.2.2 结果展示

结果目录：2_Annotation/

sRNA_Annotation_result.xlsx ：所有样本 sRNA 注释结果统计表，结果如下：

表 5.2 部分样本序列注释结果

		All	cds	lincRNA	macro_lnc	Mt_rRNA	Mt_tRNA	rRNA	scaRNA	snoRNA	snRNA	TEC	vaultRNA	miRNA	Unannot
0h	Uniq	419606	93034.338	18801.571	388.8829	2740.3667	1655.833	2727.5	117.25	3693.83564	1597.2031	84.72407	7	29596	265161.5
	All	7286184	1063763.5	446110.81	936.1164	8638.3667	9377.167	26809.62	640.41667	54801.1657	7003.8095	182.4532	10	5002001	665909.6
24h	Uniq	1268384	324289.04	79073.086	2376.217	37026.183	14372.33	13310.68	419.33333	9008.6371	7657.3357	502.2505	48.5	30567	749733.4
	All	12018170	3452647.8	847055.52	4940.349	361325.38	179433.7	414241.2	1470.6667	49953.9882	138688.17	1371.785	320	2722075	3844646
30h	Uniq	1326899	403664.61	111772.37	3185.104	26107.95	9060.833	19881.85	357.66667	7997.85477	5561.0674	663.4334	26.5	29475	709144.8
	All	13825713	3057214.5	1537904.8	6490.119	220194.12	73122	914700.4	940	43951.7639	50187.341	1858.322	115.5	1651986	6267048
48h	Uniq	1847731	601037.18	136027.41	4301.375	16850.367	5683.5	22894.49	499.16667	11377.9521	5230.831	938.1451	28.75	34363	1008499
	All	16520594	3017005.7	1722578.2	9476.78	136667.12	30287.33	1072484	993.83333	60809.4932	38485.534	2845.487	88.75	1124728	9304144

*_sRNA_Annotation_result.pdf ：各样本注释结果饼图，详细注释信息分布图如下：

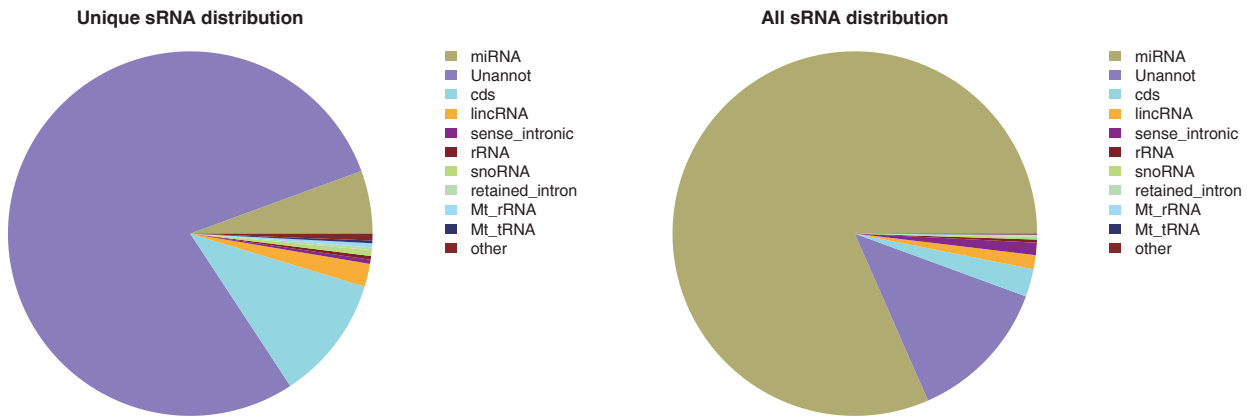


图 5.3 sRNA 注释结果饼图

注：上图展示的为部分样本结果，其他样本结果见 2_Annotation/文件夹。

说明：图中 All reads 表示总的 reads 数统计结果，Uniq Reads 表示每种拷贝只取一条统计结果。

5.3 已知 miRNA 分析

5.3.1 方法说明

过滤掉 tRNA、rRNA、snRNA、snoRNA，mRNA 等非 miRNA 序列后，将过滤之后的序列比对到 mirBaseV21 中人的已知 miRNA 中，计算各 miRNA 的表达量。数据库中人有 2588 个成熟 miRNA，总共鉴定到 1303 个。

5.3.2 结果展示

结果目录：3_known_miRNA_analysis/

All_sample_known_miRNAs_expression.xls：所有样本各已知 miRNA 表达量统计，结果如下：

地　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

表 5.3 表达丰度值前 10 位的 miRNA

miRNA	precursor	total	0h_RPM	24h_RPM	30h_RPM	48h_RPM
hsa-miR-10a-5p	hsa-mir-10a	670240.5	86498.03	33985.35	45093.78	62742.52
hsa-miR-99b-5p	hsa-mir-99b	354254	48677.95	18026.99	18936.55	27038.45
hsa-miR-10b-5p	hsa-mir-10b	342672	43577.33	18189	23491.35	32342.82
hsa-miR-7-5p	hsa-mir-7-1	299195	35734.42	16735.7	20267.67	36819.44
hsa-miR-30e-5p	hsa-mir-30e	242585.5	26909.73	17681.49	16673.52	28724.18
hsa-miR-148a-3p	hsa-mir-148a	219957.5	32355.51	10229.11	9056.65	13610.36
hsa-miR-17-5p	hsa-mir-17	211529.5	27459.44	11501.24	12339.66	19990.55
hsa-miR-92a-3p	hsa-mir-92a-1	210951	21827.16	17163.56	17516.24	23217.39
hsa-miR-191-5p	hsa-mir-191	210531	28763.07	11803.45	10997.64	14544.8

注：上表展示的展示的为部分样本结果，其他见 3_known_miRNA_analysis/

Base_bias.pdf：已知 miRNA 碱基偏好性结果，展示如下：

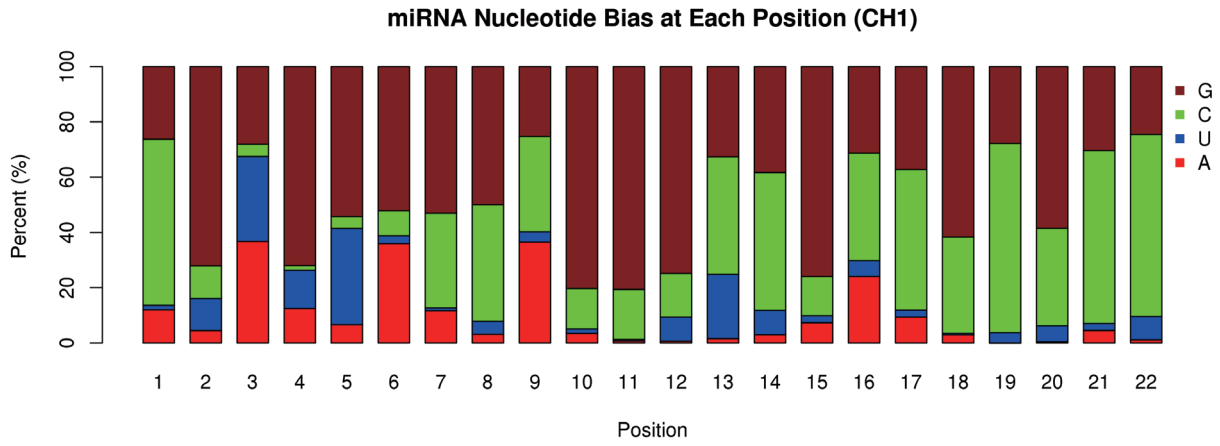


图 5.4 已知 miRNA 各位置碱基偏好性

说明：横坐标为 sRNA 的碱基位置，纵坐标为该位置 sRNA 中出现碱基 A/U/C/G 的百分率

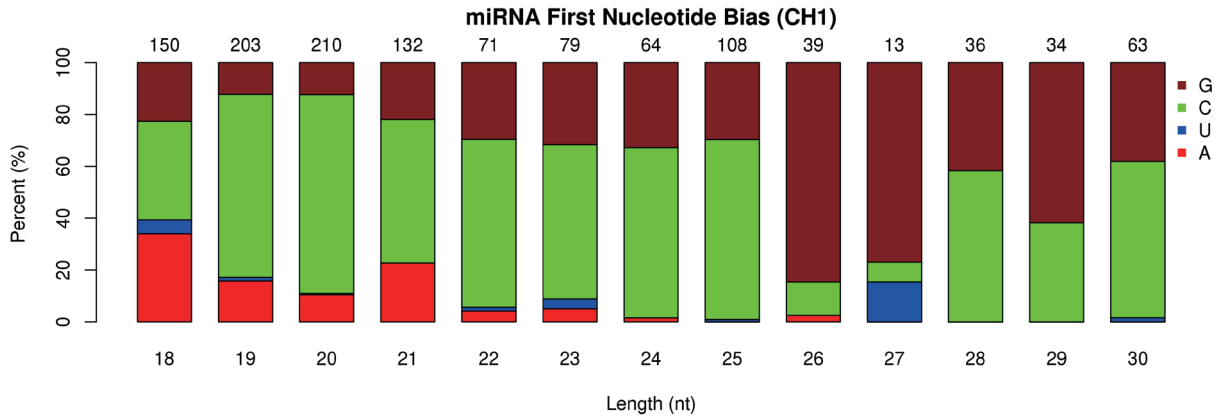


图 5.5 长度 18~30nt 的已知 miRNA 首位碱基偏好性

说明：横坐标为 sRNA 长度，纵坐标为该长度 sRNA 中首位碱基出现 A/U/C/G 的百分率（柱形图上方的数值为该长度 sRNA 的总条数）

5.4 novel miRNA 预测

5.4.1 方法说明

将各样本过滤之后的 sRNA 序列合并，比对到人基因组中，后采用 mirDeep2 做 novel miRNA 预测，参数采用默认参数。去除预测结果中人中已知的 miRN，剩余的为 novel miRNA。mirDeep2 是一款专门用于做 know 及 novel miRNA 鉴定的软件。

软件信息详见：

https://www.mdc-berlin.de/8551903/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep

5.4.2 结果展示

结果目录：4_novel_miRNA_predict/

All_novel_miRNA_expression.xlsx：各样本 novelmiRNA 表达量统计，所有样本总共预测到 354 个 novel miRNA。结果如下：

表 5.4 novel miRNA 表达量前十结果

miRNA	precursor	total
hsa-novel-214-mature	hsa-novel-214	3690831
hsa-novel-164-mature	hsa-novel-1,hsa-novel-164	316420.7
hsa-novel-1-mature	hsa-novel-1	156764.8
hsa-novel-245-mature	hsa-novel-245	109297.3
hsa-novel-2-mature	hsa-novel-2	33789
hsa-novel-154-mature	hsa-novel-154	29817
hsa-novel-10-mature	hsa-novel-10	24905
hsa-novel-267-mature	hsa-novel-267	23701
hsa-novel-148-mature	hsa-novel-148	22658
hsa-novel-7-mature	hsa-novel-7	18733

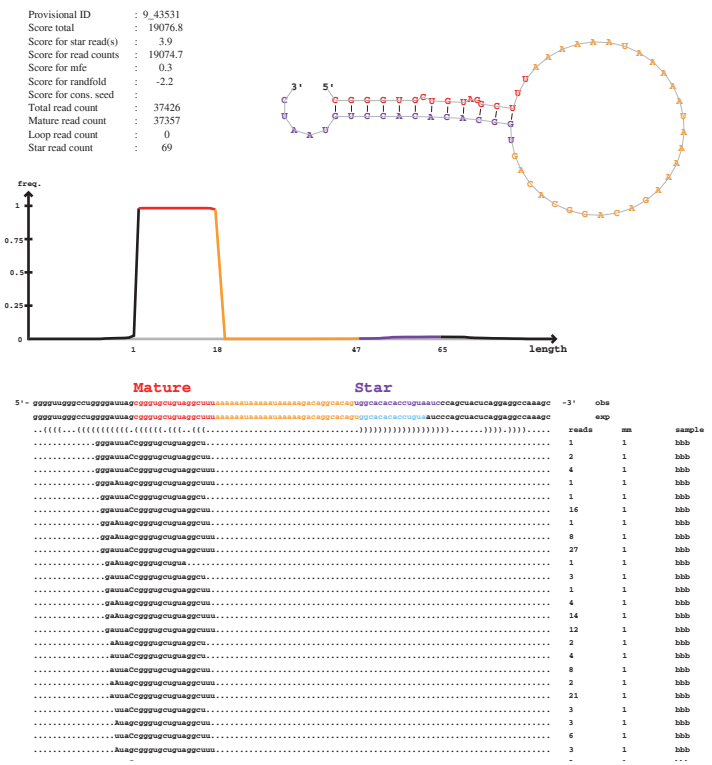


图 5.6 novel miRNA 结构示意图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

5.5 miRNA 家族分析

对检测到的已知 miRNA 和新 miRNA 进行家族分析，探索其所属的 miRNA 家族在其他物种中的存在情况。结果表格中第一列是 miRBase 数据库中收录的各个物种名，第一行是检测到的已知和新 miRNA 前体的家族名，表格中的数据项“+”表示在该物种中存在对应家族，“-”表示不存在。结果见表 4.5。

表 5.5 miRNA 家族分析结果（部分）

	mir-375	mir-15	mir-217	mir-210	mir-2190	mir-459
Xenopus tropicalis	+	+	+	+	-	-
Branchiostoma floridae	+	-	+	+	-	-
Petromyzon marinus	+	+	+	-	-	-
Macaca mulatta	+	+	+	+	-	-
Strongylocentrotus purpuratus	+	-	-	-	-	-
Tetraodon nigroviridis	+	+	+	+	-	-
Culex quinquefasciatus	+	-	-	+	-	-
Lemur catta	-	+	-	-	-	-
Ixodes scapularis	+	-	-	-	-	-

注：结题报告中只显示部分 miRNA 家族分析统计结果，更多信息见结果文件夹 miRNA_family。

5.6 样本聚类及 PCA 分析

5.6.1 方法说明

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。

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。

5.6.2 结果展示

结果目录：5_Sample_cluster/

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

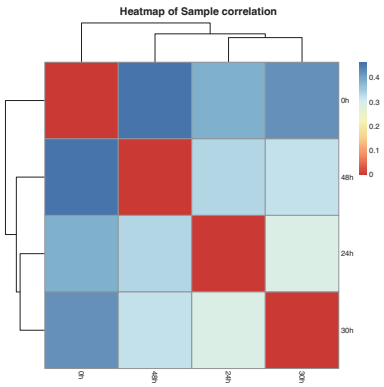


图 5.7 样本间距离热图

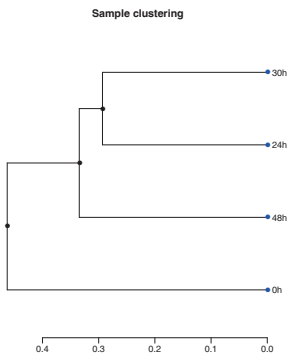


图 5.8 样本间聚类树图

说明：样本间距离热图，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。

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。

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

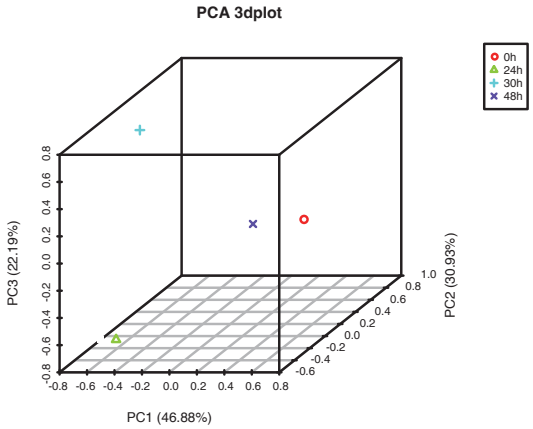


图 5.9 PCA 3Dplot

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：

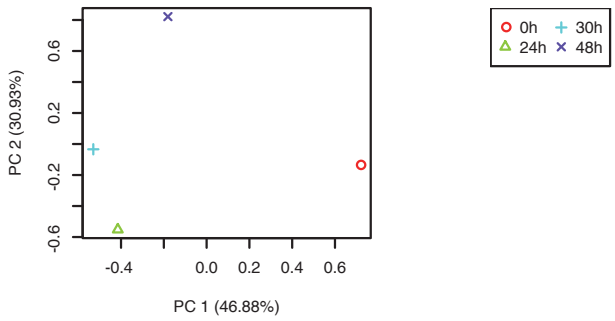


图 5.10 PCA 2Dplot (PC1 vs PC2)

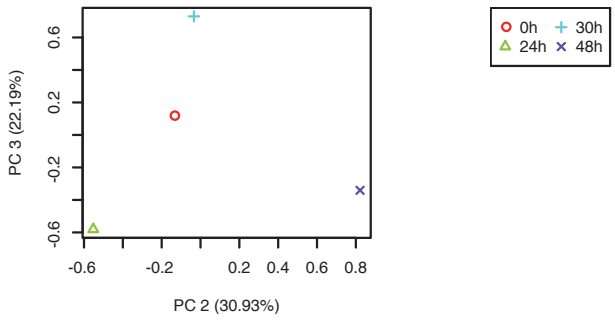


图 5.11 PCA 2Dplot (PC2 vs PC3)

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

5.7 差异表达分析

5.7.1 方法说明

无生物学重复：差异分析方法参照 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：

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

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

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

然后，我们对差异检验的 p value 作多重假设检验校正。差异基因筛选条件为： P<=0.05 且|Log2Fold Change|>=1。

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 pvalue<0.05。

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

5.7.2 结果展示

结果目录：6_DEGs_analysis

差异基因统计结果，结果如下：

表 5.5 差异基因数目统计

Compare	UP	DOWN	ALL
0h_vs_24h	184	38	222
0h_vs_30h	155	38	193
0h_vs_48h	184	62	246
24h_vs_30h	59	21	80
24h_vs_48h	73	40	113
30h_vs_48h	19	15	34

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

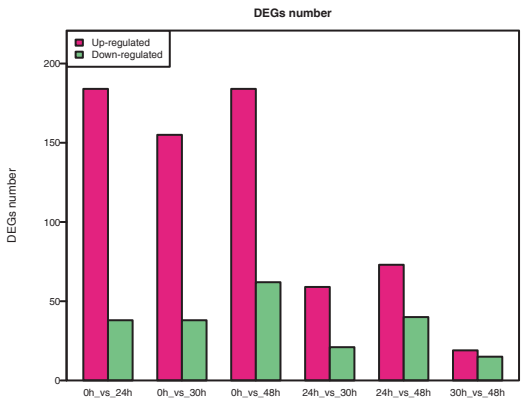


图 5.12 差异基因数目条形图

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

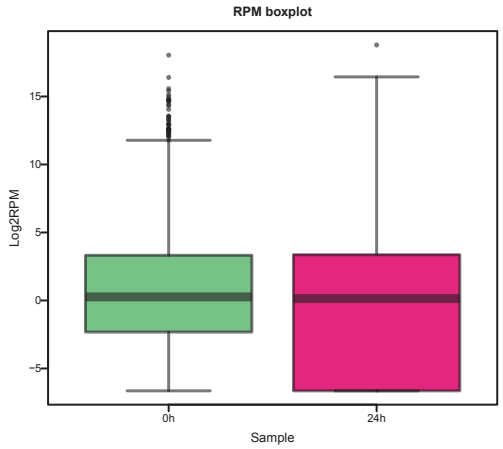


图 5.13 比较对间样本表达量盒状图

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 6.26 一样，下同。

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

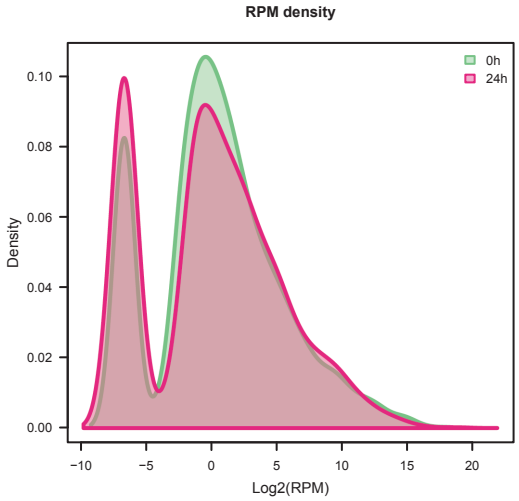


图 5.14 比较对间样本表达密度曲线

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

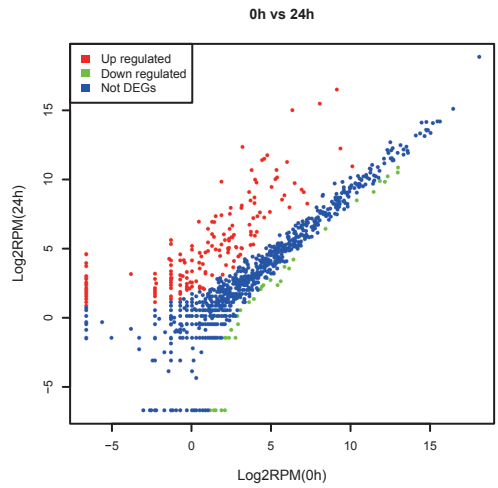


图 5.15 比较对样本件表达量散点图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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(RPM)值，若为有生物学重复样本则 X/Y 轴的值为 Log2(Mean RPM)，即为 log2(生物学重复 RPM 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.RPM.MA.plot.pdf：比较对样本件表达量 MA 图，结果展示如下：

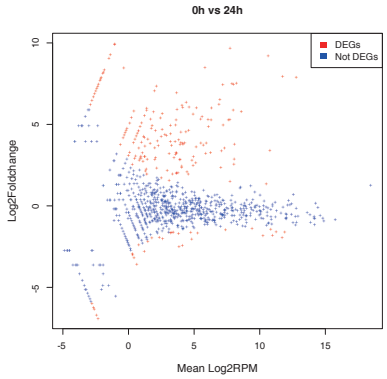


图 5.16 比较对样本件表达量 MA 图

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：

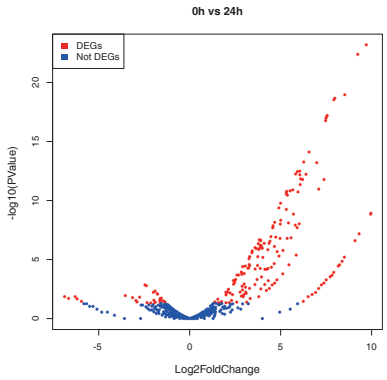


图 5.17 差异分析火山图

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：

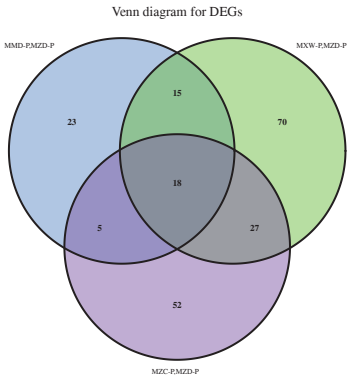


图 5.18 差异基因韦恩图

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。

5.8 差异基因表达模式聚类分析

5.8.1 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集求并集，获得该基因集在每个样品中的TPM值，做后续聚类分析，获得表达模式相近的基因集。

5.8.2 结果展示

结果目录：6_DEGs_analysis/*_DEGs_cluster

All_DEGs_samples_heatmap.pdf：所有差异基因表达聚类热图，结果展示如下

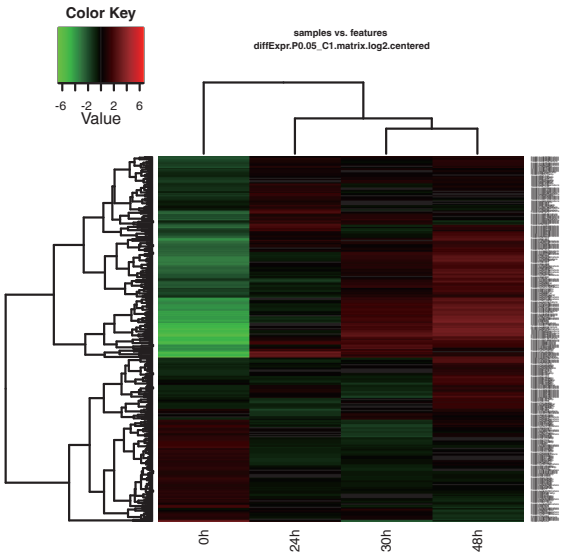


图 5.19 表达模式聚类热图

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未同一 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：

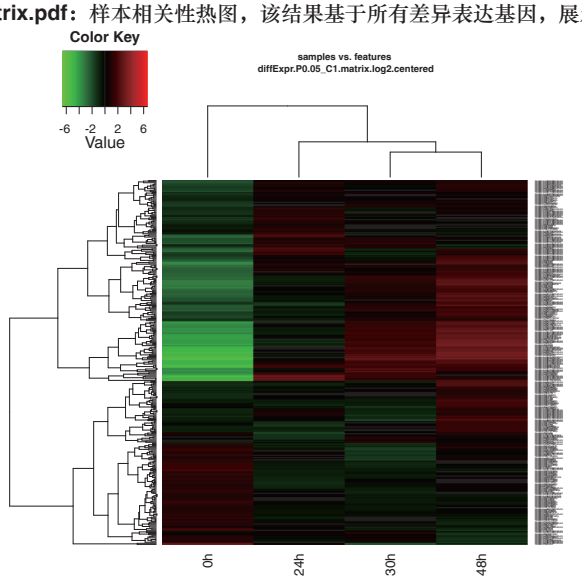


图 5.20 表达相关性热图

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

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未同一 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：

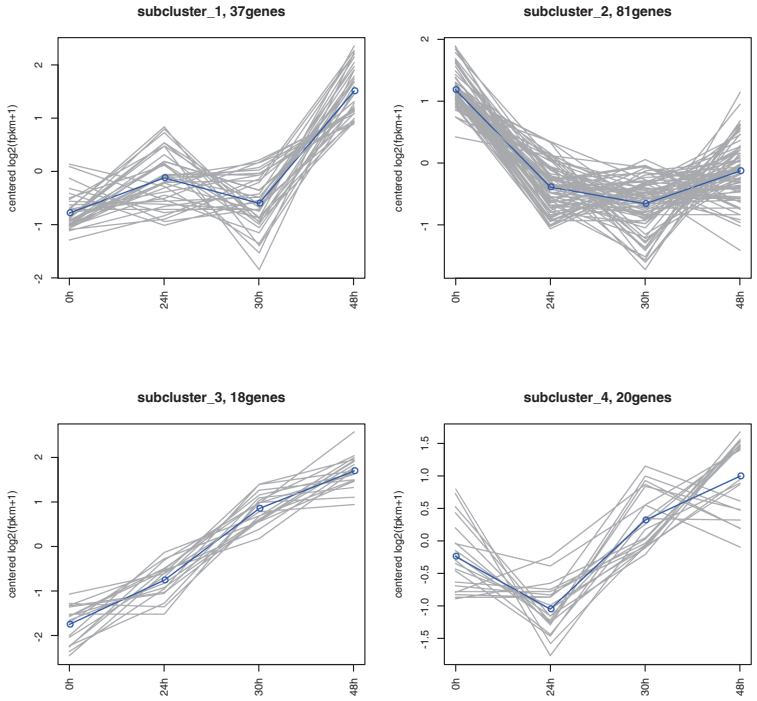


图 5.21 前 4 个 cluster 中基因在各样本中表达量折线图

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：

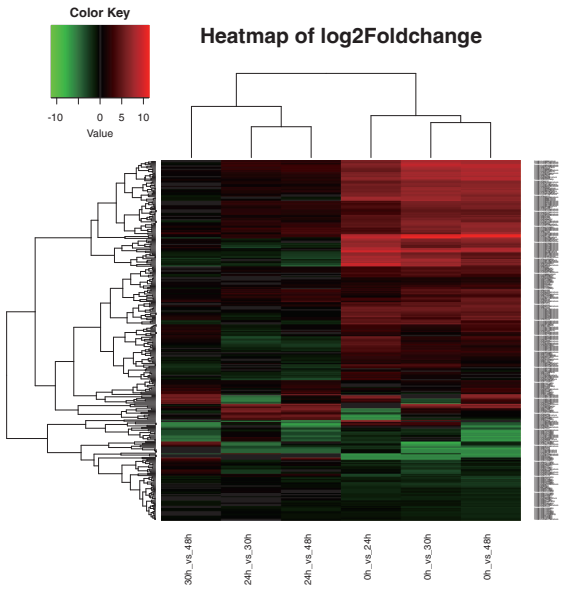


图 5.22 foldchange 热图

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。

5.9 miRNA 靶基因预测

5.9.1 方法说明

对所有已知差异表达的 miRNA 做靶基因预测，预测的数据来源为 mirDB(http://mirdb.org/miRDB/index.html)。详细结果见 7_target_prediction/

5.10 差异基因 GO 富集分析

注：一下展示内容均为第一组比较对的富集分析结果

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：8_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 5.5 GO 富集分析结果

GO_ID	Term	Type	DEGs_this_term	Expected	Pvalue	FDR
GO:0005488	binding	molecular_function	5987	5623.94	1.00E-30	2.35E-28
GO:0005622	intracellular	cellular_component	7756	7123.96	1.00E-30	2.35E-28
GO:0005634	nucleus	cellular_component	3762	3318.97	1.00E-30	2.35E-28
GO:0005737	cytoplasm	cellular_component	5596	5090.13	1.00E-30	2.35E-28
GO:0006351	transcription, DNA-templated	biological_process	2146	1792.55	1.00E-30	2.35E-28
GO:0006355	regulation of transcription, DNA-templated	biological_process	1932	1597.87	1.00E-30	2.35E-28
GO:0007154	cell communication	biological_process	3144	2735.19	1.00E-30	2.35E-28
GO:0007275	multicellular organismal development	biological_process	2634	2211.1	1.00E-30	2.35E-28
GO:0007399	nervous system development	biological_process	1244	952.37	1.00E-30	2.35E-28
GO:0009058	biosynthetic process	biological_process	3315	2947.79	1.00E-30	2.35E-28

注：上述展示的只是富集分析中前 10 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

Pvalue：富集分析 P 值，P 值越小越显著

FDR：P 值校正值

地　址：上海市松江区香闵路 698 号　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

***.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：

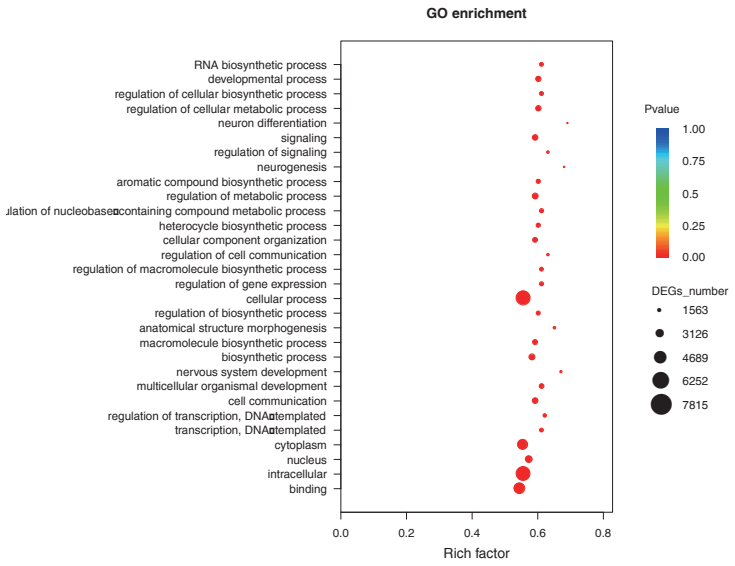


图 5.23 GO 富集分析前 30 个 GO 富集散点图

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

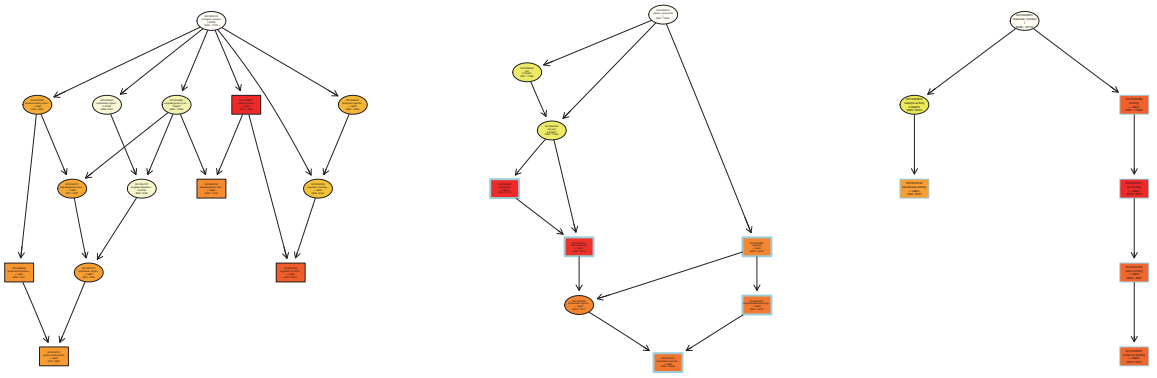


图 5.24 GO 富集工人村有向无环图

说明：topGO 有向无环图（图 19）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：

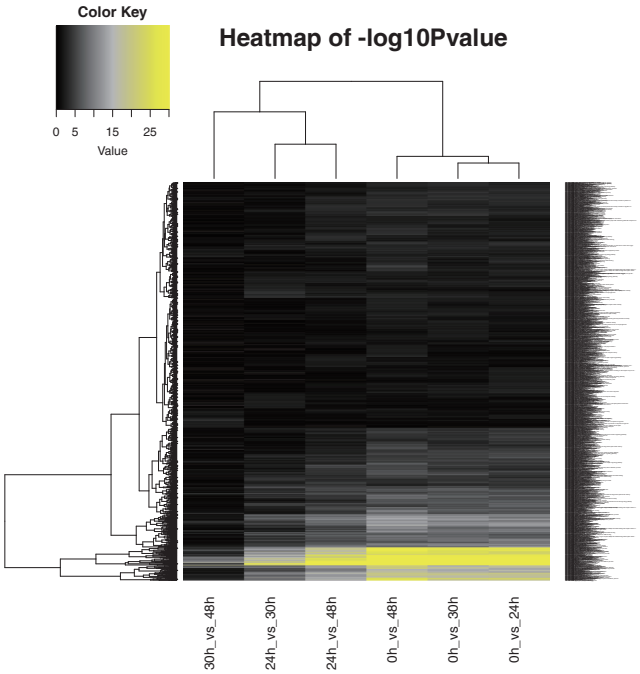


图 5.25 所有显著富集的 GO pvalue 热图

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，上图中内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 靶基因 KEGG 富集分析

注：一下展示内容均为第一组比较对的富集分析结果

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：9_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：

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

表 5.6 pathway 富集分析结果

KO_ID	Term	Type	DEGs_this_term	Pvalue	FDR
ko04724	Glutamatergic synapse	Organismal Systems	97	2.59E-18	8.42E-16
ko05200	Pathways in cancer	Human Diseases	229	1.38E-15	2.25E-13
ko04723	Retrograde endocannabinoid signaling	Organismal Systems	83	4.00E-13	4.33E-11
ko05206	MicroRNAs in cancer	Human Diseases	115	1.24E-12	9.62E-11
ko04360	Axon guidance	Organismal Systems	101	1.48E-12	9.62E-11
ko04014	Ras signaling pathway	Environmental Information Processing	165	3.20E-12	1.74E-10
ko05205	Proteoglycans in cancer	Human Diseases	161	4.98E-12	2.31E-10
ko04151	PI3K-Akt signaling pathway	Environmental Information Processing	238	6.32E-12	2.57E-10
ko04068	FoxO signaling pathway	Environmental Information Processing	104	4.60E-11	1.66E-09
ko04725	Cholinergic synapse	Organismal Systems	88	3.69E-10	1.20E-08

注：上述展示的只是富集分析中前 10 的 pathway，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。

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：位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：

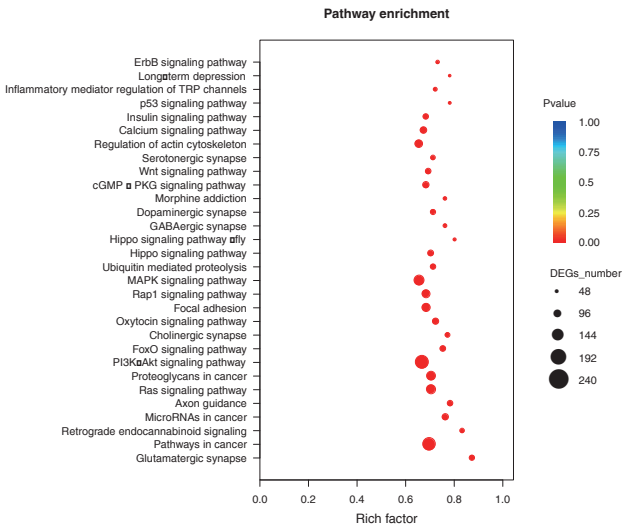


图 5.26 pathway 富集分析前 30 个富集散点图

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：

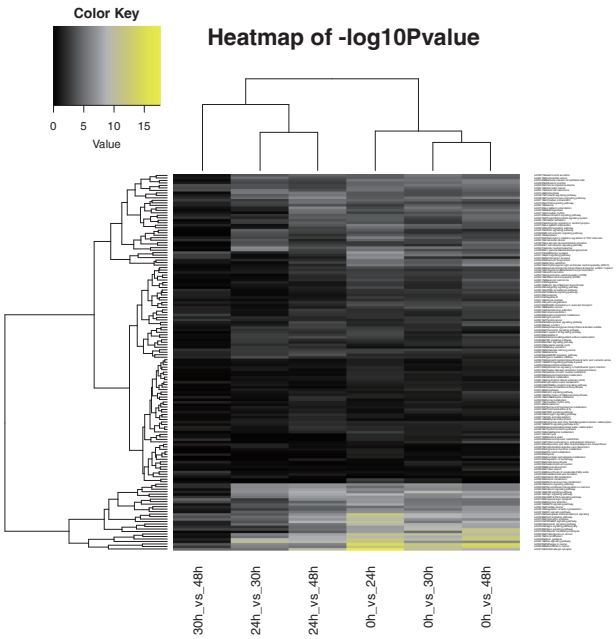


图 5.27 所有显著富集的 pathway pvalue 热图

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，上图中内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。

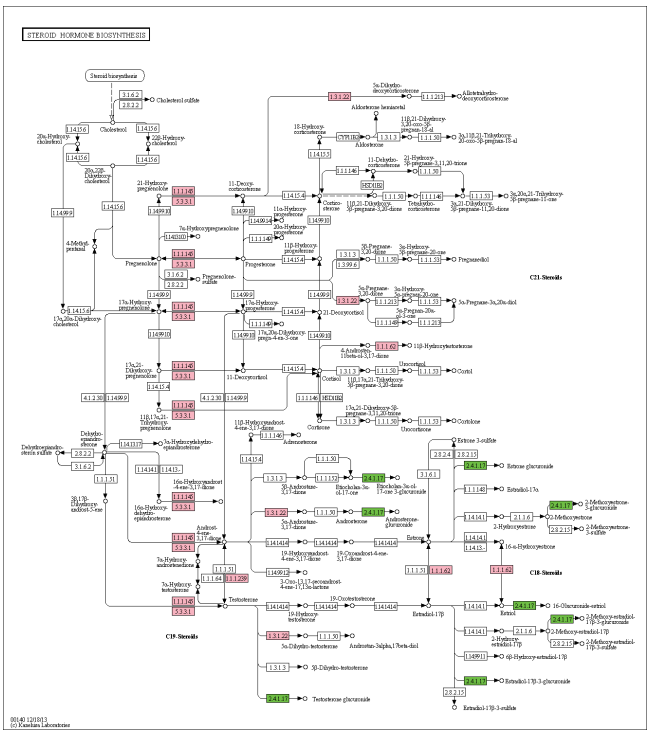


图 5.28 靶基因代谢通路图

说明：上图中绿色基因表示是差异表达 miRNA 的靶基因，红色为非差异表达 miRNA 的靶基因。

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

5.12 mRNA/miRNA 关联分析

5.12.1 方法说明

根据 miRNA 与靶标基因关系对构建 miRNA 与靶标基因网络，根据网络中节点度排序及 miRNA 与靶基因表达关系筛选关键 miRNA，并发现多种 miRNA 集中调控的核心基因功能。

另外结合基因信号通路数据库与 miRNA-mRNA 靶向序列分析技术，可发现 miRNA 调控的多种代谢通路，并通过量化计算从目标 miRNA 群中分离出核心调控作用的 miRNA 及 mRNA，并发现多种 miRNA 集中调控的核心代谢通路。

构建网络采用的是 Cytoscape 软件。此处分析采用的均为 known miRNA，novel miRNA 未做分析。

5.12.2 方法说明

结果目录：10_miRNA_mRNA_network/

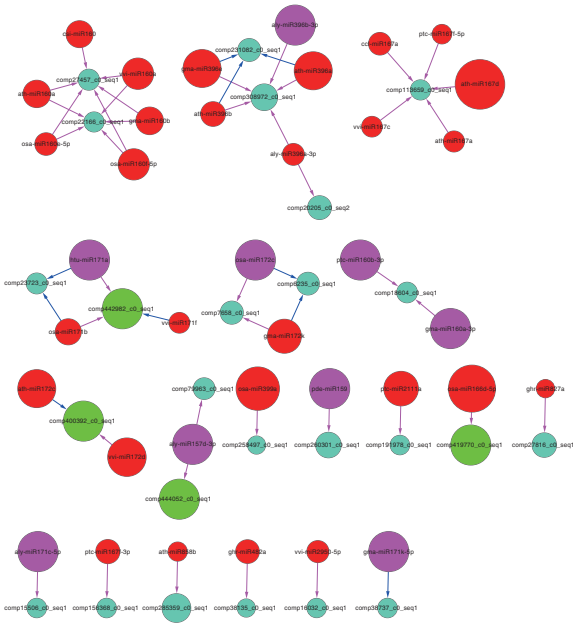


图 5.29 所有差异表达 miRNA 与靶基因网络关系图

说明：每个圈的大小表示 foldchange 的大小圈的颜色，红色与紫色表示 miRNA，红色表示上调 miRNA,紫色表示下调 mIRNA。绿色与青色表示 mRNA,绿色表示下调 mRNA，青色表示上调 mRNA。

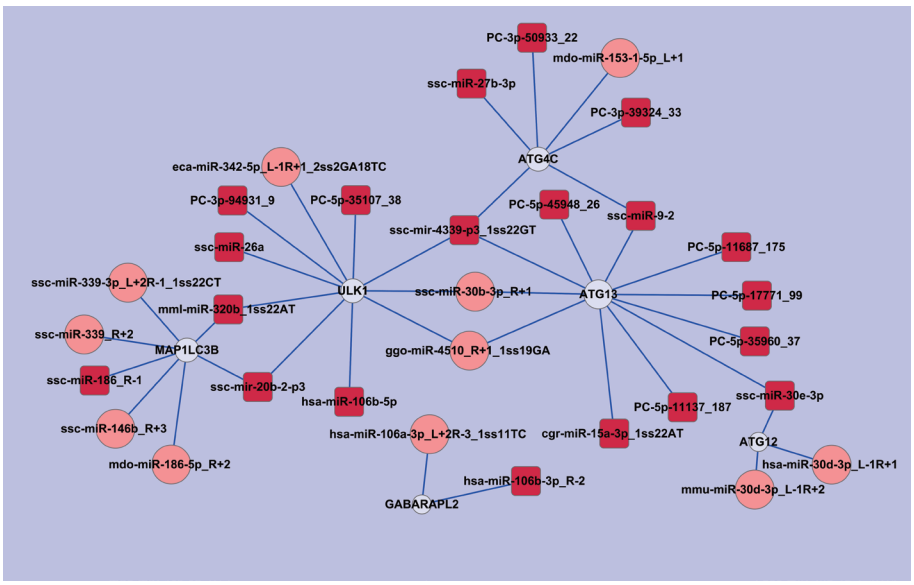


图 5.30 特定富集 GO miRNA 与靶基因网络关系图

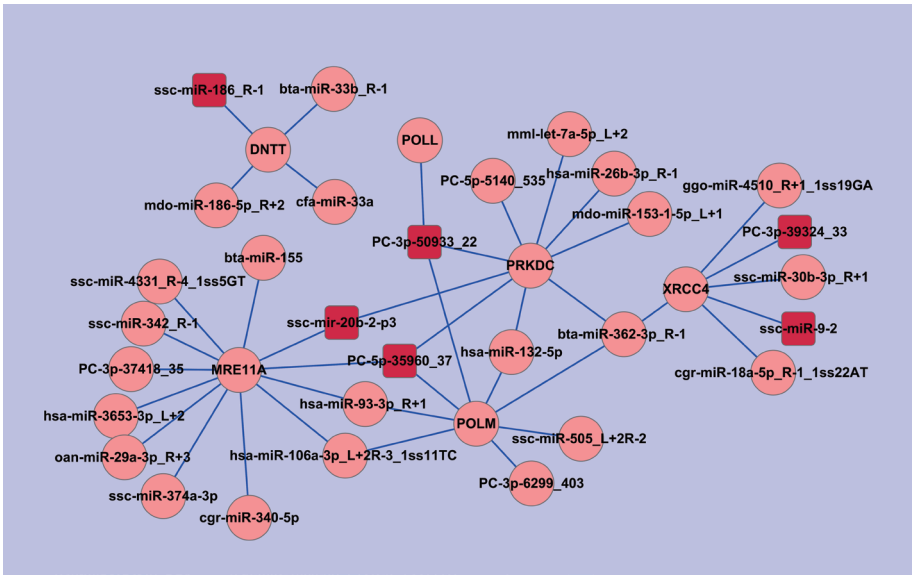


图 5.31 特定富集 pathway miRNA 与靶基因网络关系图

说明：上图中圆圈表示 miRNA，正方形表示靶基因

6. 参考文献

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

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

【3】Erlich, Y., and Mitra, P.P. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nature methods 5, 679-682.

【4】Friedlander M.R., Mackowiak S.D., Li N., Chen W., Rajewsky N. (2011). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40:37-52. (miRDeep2)

【5】Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B.(2011). Synthetic spike-in standards for RNA-seq experiments. Genome research 21, 1543-1551.

【6】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. (KEGG)

【7】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.

【8】Wen M., Shen Y., Shi S., and Tang T. (2010). miREvo: An Integrative microRNA Evolutionary Analysis Platform for Next-generation Sequencing Experiments. BMC Bioinformatics 13:140. (miREvo)

【9】Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. goseq: Gene Ontology testing for RNA-seq datasets.(goseq)

【10】Zhou L., Chen J., Li Z., Li X., Hu X., et al. (2010). Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. PLoS One 5: e15224. (TPM)

地　　址：上海市松江区香闵路 698 号　　Add: 698 Xiang Min Road SongJiang Shanghai China
技术支持：021-37772413/57072083/57072061

生工® Sangon Biotech

生工生物工程（上海）股份有限公司
Sangon Biotech (Shanghai) Co., Ltd.
地址：上海市松江区香闵路698号　邮编：201611
咨询热线：800-820-1016; 400-821-0268
电话（总机）：021-37772168
传真：021-37772170
Email：sales@sangon.com
http://www.sangon.com

扫描生工微博、微信二维码
了解更多优惠信息

微博　微信

Life·Biotech·Future
www.sangon.com