原核转录组学分析报告

项目名称：　天津科技大学_梭菌_转录组学测试

委托单位：　天津科技大学

项目负责人：

核验人员：

技术负责人：

项目编号：　P20191101913

报告时间：　2020-03-11

1 项目信息

基本思想

原核转录组测序，基于 Illumina 测序平台，通过对特定状态或者特定时期下的原核生物的样本的总 RNA 去除 rRNA 后进行测序，针对实际样品信息采用灵活的差异分析策略可以鉴定到生物体不同时期或不同组织或不同个体或不同实验处理间差异表达的 mRNA，再利用软件与特定数据库进行功能注释，最终可以得到 mRNA 在生物体中参与生命活动的清晰生物信息图谱。

1.2 实验流程

1.2.1 样本检测

我们对总 RNA 的样本检测包括以下 3 种方法：

（1）1%的琼脂糖电泳检测 RNA 样品是否有降解以及杂质；

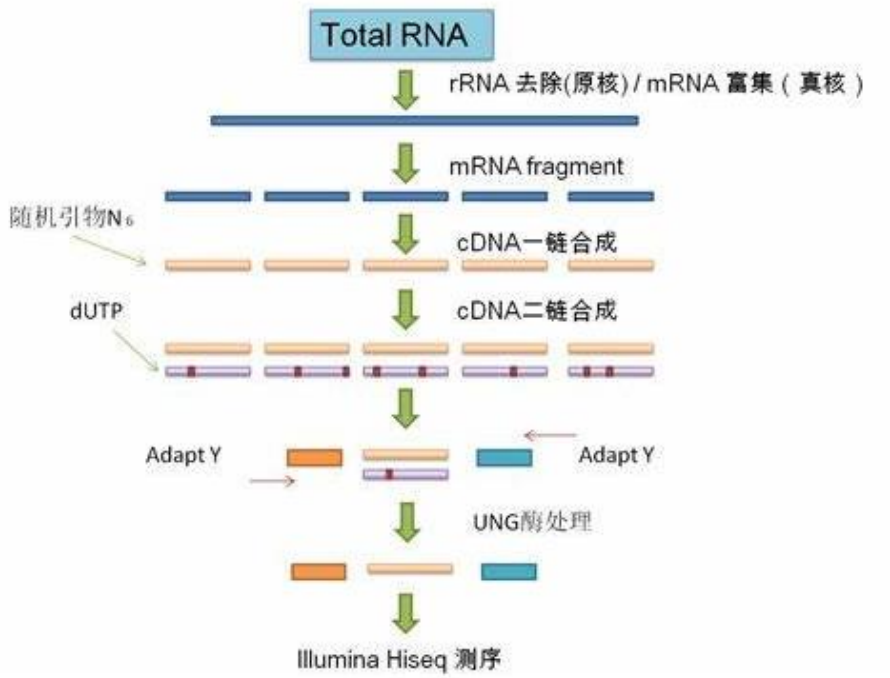
（2）凯奥 K5500 分光光度计检测样品纯度（凯奥，北京）；

（3）安捷伦 2100 RNA Nano 6000 Assay Kit（Agilent Technologies, CA, USA）检测 RNA 样品的完整性和浓度。

1.2.2 文库构建和上机测序

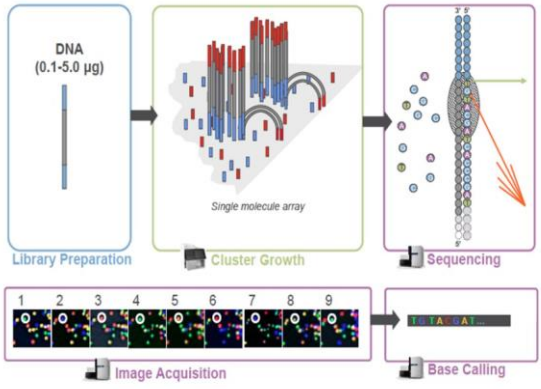
总 RNA 样本检测合格后，对于真核生物，用带有 Oligo（dT）的磁珠富集 mRNA，对于原核生物，用试剂盒去除 rRNA，向得到的 RNA 中加入 Fragmentation Buffer 使其片断成为短片段，再以片断后的 RNA 为模板，用六碱基随机引物合成 cDNA 第一链，并加入缓冲液、dNTPs、RNaseH 和 DNA Polymerase I 合成 cDNA 第二链，经过 QIAQuick PCR 试剂盒纯化并加 EB 缓冲液洗脱。洗脱纯化后的双链 cDNA 再进行末端修复、加碱基 A、加测序接头处理，然后经琼脂糖凝胶电泳回收目的大小片段并进行 PCR 扩增，从而完成整个文库制备工作。

构建好的文库用 Illumina HiSeq *X Ten* 进行测序。测序策略为 PE150（pair end 150bp）。其实验流程如下：



1.2.3 上机测序

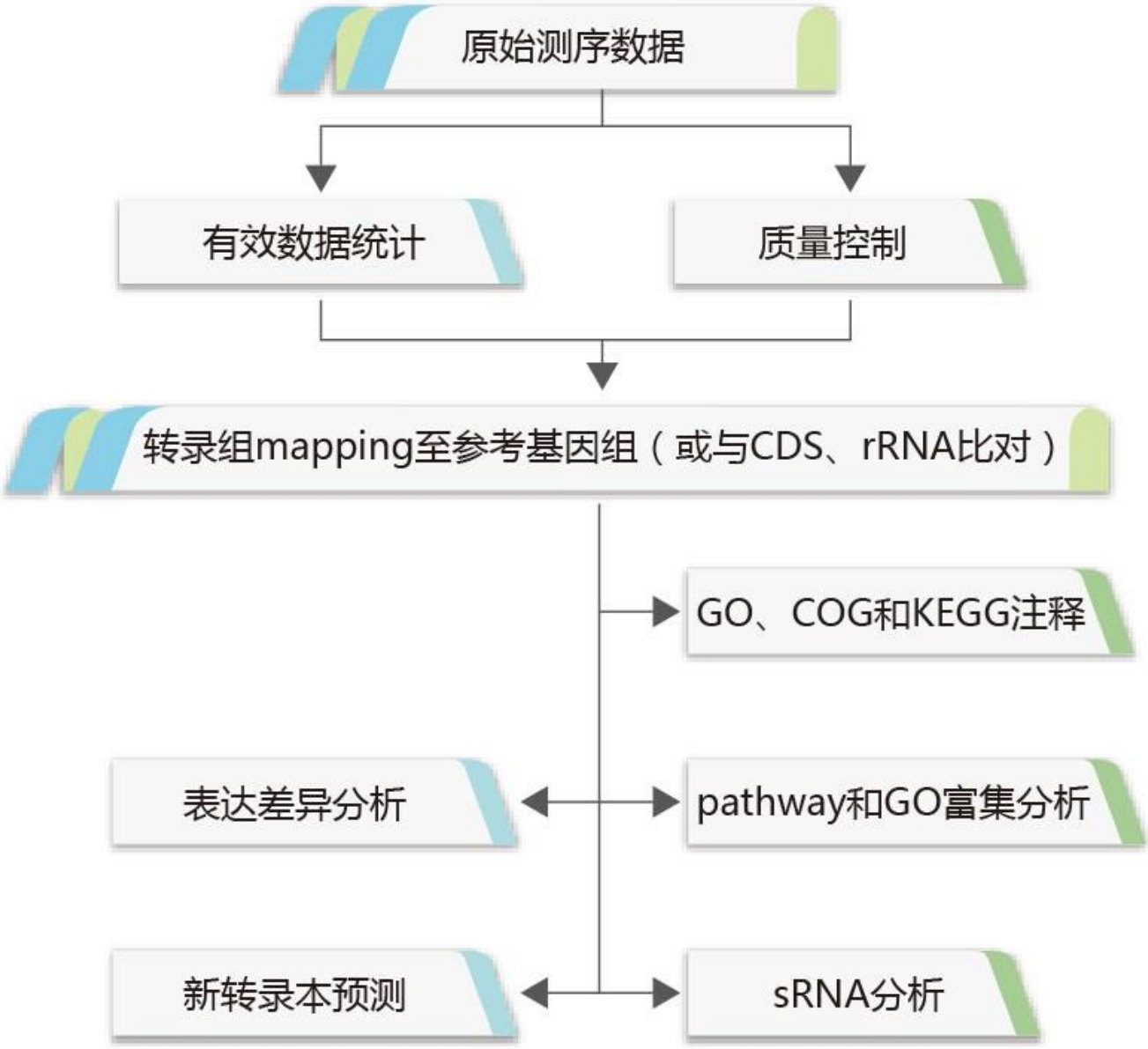
库检合格后，把不同文库按照有效浓度及目标下机数据量的需求 pooling 后进行 Illumina 测序。测序的基本原理是边合成边测序（Sequencing by Synthesis）。在测序的 flow cell 中加入四种荧光标记的 dNTP、DNA 聚合酶以及接头引物进行扩增，在每一个测序簇延伸互补链时，每加入一个被荧光标记的 dNTP 就能释放出相对应的荧光，测序仪通过捕获荧光信号，并通过计算机软件将光信号转化为测序峰，从而获得待测片段的序列信息。测序过程如下图所示。



1.3 信息分析流程

illumina 测序仪测序所得原始下机序列（Raw Reads），通过去低质量序列、去接头污染等过程完成数据处理得到高质量的序列（Clean Reads），后续所有分析都是基于 Clean Reads。

我们的转录组测序信息分析流程主要分为三部分：测序数据质控（QC）、数据比对分析(mapping)和转录组深层分析。其中，测序数据质控包括过滤测序所得序列、评估测序数据质量以及计算序列长度分布等；数据比对分析主要是针对比对到基因组中的序列，根据不同的基因组注释信息依次进行分类和特征分析，并计算相应的表达量；转录组深层分析包括差异表达分析、新转录本预测和变异分析等其他个性化分析。具体的信息分析流程图如下：



1.4 样品信息

本项目共 12 个样本，样品信息如下表所示。

表 1.1 样品信息

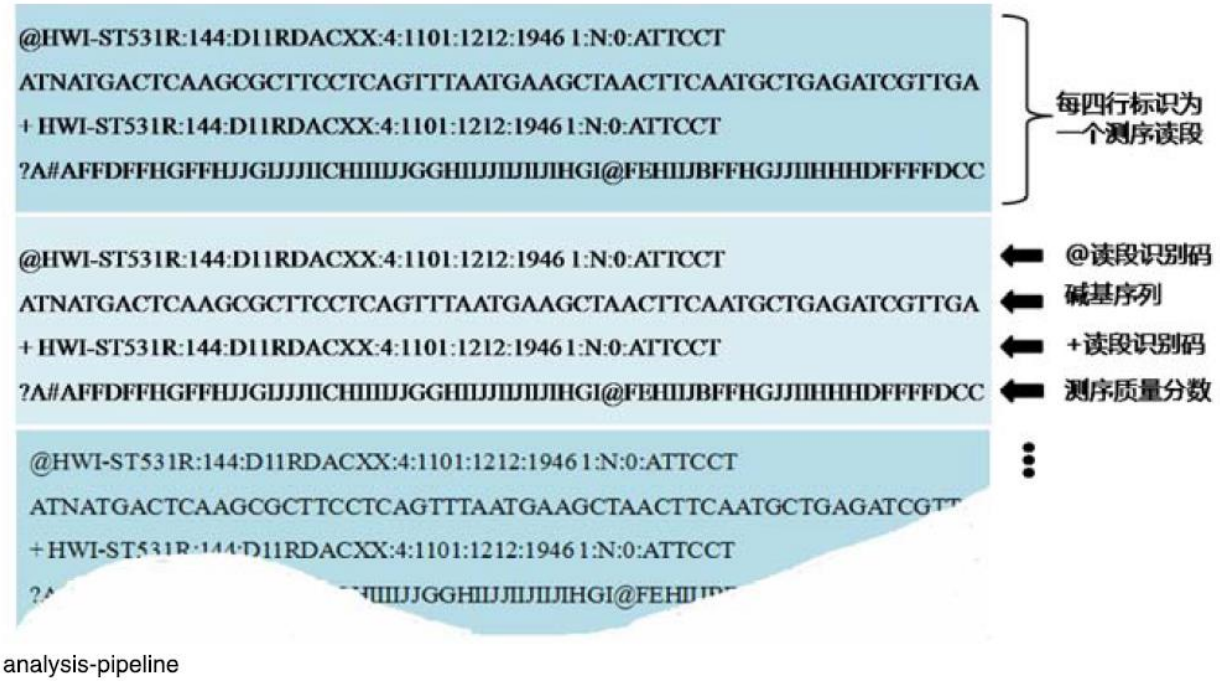
sample	group
F_3_6h_1	F3_6h
F_3_6h_2	F3_6h
F_3_6h_3	F3_6h
824_6h_1	824_6h
824_6h_2	824_6h
824_6h_3	824_6h
F_3_24h_1	F3_24h
F_3_24h_2	F3_24h
F_3_24h_3	F3_24h
824_24h_1	824_24h
824_24h_2	824_24h
824_24h_3	824_24h

2 数据过滤

2.1 原始数据

Illumina 高通量测序结果最初以原始图像数据文件存在，经 CASAVA 软件进行碱基识别（Base Calling）后转化为原始测序序列（Sequenced Reads），我们称之为 Raw Data，其结果以 FASTQ（简称为 fq）文件格式存储。FASTQ 文件包含每条测序序列（Read）的名称、碱基序列以及其对应的测序质量信息。在 FASTQ 格式文件中，每个碱基对应一个碱基质量字符，每个碱基质量字符对应的 ASCII 码值减去 33（Sanger 质量值体系），即为该碱基的测序质量得分。不同 Score 代表不同的碱基测序错误率，如 Score 值为 20 和 30 分别表示碱基测序错误率为 1%和 0.1%。

其中 FASTQ 格式示例如下：



analysis-pipeline

图 2.1 FASTQ 文件格式示例

（1）第一行以“@”开头，随后为 Illumina 测序标识别符（Sequence Identifiers）和描述文字（选择性部分）；

（2）第二行是碱基序列；

（3）第三行以“+”开头，随后为 Illumina 测序标识别符（选择性部分）；

（4）第四行是对应碱基的测序质量，该行中每个字符对应的 ASCII 值减去 33，即为对应第二行碱基的测序质量值。

2.2 数据过滤统计

测序得到的某些原始下机序列，会含有测序接头序列以及低质量序列，为了保证信息分析数据的质量，我们对原始序列进行过滤，得到高质量的 Clean Reads，再进行后续分析，后续分析都基于 Clean Reads。所用的软件为 trimmomatic 和 FastQC。

数据处理步骤如下:

（1）去除接头污染的 Reads（Reads 中接头污染的碱基数大于 5bp。对于双端测序，若一端受到接头污染，则去掉两端的 Reads）；

（2）去除低质量的 Reads（Reads 中质量值 Q≤19 的碱基占总碱基的 15%以上，对于双端测序，若一端为低质量 Reads，则会去掉两端 Reads）；

（3）去除含 N 比例大于 5%的 Reads（对于双端测序，若一端含 N 比例大于 5%，则会去掉两端 Reads）。数据过滤统计结果见下表：

sample	ReadsNum_raw	BaseNum_raw	ReadsNum_clean	BaseNum_cleanGC%	>Q30	>Q20	clean_rate
824_24h_1	21890178	3283526700	13868520	199680782843.54%	94.11%	97.98%	60.81%
824_24h_2	25228342	3784251300	13983702	201592289043.49%	94.4%	98.12%	53.27%
824_24h_3	18945184	2841777600	13928044	202693772144.42%	92.88%	97.24%	71.33%
824_6h_1	22767336	3415100400	14415204	207952242242.34%	93.1%	97.51%	60.89%
824_6h_2	21853126	3277968900	16474694	236827501741.72%	93.18%	97.58%	72.25%
824_6h_3	46721596	7008239400	37139596	535660590041.01%	93.02%	97.5%	76.43%
F_3_24h_1	19271288	2890693200	13870550	200980795740.95%	94.64%	98.3%	69.53%
F_3_24h_2	17090318	2563547700	13939874	203722461041.76%	94.41%	98.14%	79.47%
F_3_24h_3	20626942	3094041300	14407800	209608966940.86%	94.08%	97.99%	67.75%
F_3_6h_1	44344902	6651735300	33893140	486181752841.63%	93.22%	97.58%	73.09%
F_3_6h_2	22161232	3324184800	17869096	258808517340.6%	93.66%	97.84%	77.86%
F_3_6h_3	20656804	3098520600	15918974	229731131341.8%	93.39%	97.66%	74.14%

表 2.1 数据过滤与数据产出统计表

（1）total reads：原始下机序列的总序列数；

（2）total base：原始下机序列的总碱基数；

（3）GC%：每个样本序列的 GC 含量；

（4）Q30：总序列中质量值大于 30（错误率小于 0.1%）的碱基数的比例。该值越大说明测序质量越好。

（5）Q20：总序列中质量值大于 20（错误率小于 1%）的碱基数的比例。该值越大说明测序质量越好。

2.3 质控分析 Q&A

问：测序错误率会随着测序序列长度的增加而升高，错误率在多少是可以接受的范围？

答：我们的测序会进行严格的数据质量把控。一般情况下，单个碱基位置的测序错误率应该低于 1%，最高在 6%左右可以接受。

问：我们质控的标准是什么？是否严格？

答：为保证后续分析的质量，我们会严格把控 cleandata 的筛选标准，具体标准如下：

(1) 去除带接头(adapter)的 reads；

(2) 去除含 N(N 表示无法确定碱基信息)的 reads；

(3) 去除低质量 reads(质量值 Qphred <= 20 的碱基数占整个 read 的 50%以上的 reads)。

问：相关名词解释

adapter：接头，用于上机测序。建库时引入的接头序列与测序芯片（flow cell）上固定的接头相互识别。

index：测序的标签，用于测定混合样本，通过每个样本添加的不同标签进行数据区分，鉴别测序样品。

Q20,Q30：Phred 数值大于 20、30 的碱基占总体碱基的百分比，其中 Phred=-10log10(e).

raw data/raw reads：测序下机的原始数据。

clean data/clean reads：对原始数据进行过滤后，剔除了低质量数据的剩余数据。后续分析均基于 clean data。

3 比对分析(mapping)

3.1 比对率分析

将各样品过滤后的测序序列与参考基因组进行比对(mapping)，使其定位到基因组。 本项目分析中，我们使用 bowtie2 软件进行 mapping(http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)。参考基因组信息由客户提供。

在参考基因组选择合适并且组装完整，样品无外源物种污染的情况下，比对率通常都会在80%左右。由于基因组中会存在重复区域，在比对中，会出现一条序列比对到基因组多个位置的情况（MultiMap Reads）。同时，因不同物种基因组中的重复区域比例不同，这种比对到多个位置序列的比例会随着物种的变化而有差异。

下表为比对率统计结果表：

表 3.1 各个样品的比对率统计表

Items	824_24h_1	824_24h_2	824_24h_3	824_6h_1	824_6h_2	824_6h_3	F_3_24h_1	F_3_24h_2	F_3_24h_3	F_3_6h_1	F_3_6h_2	F_3_6h_3
Total records	8040894(100%)	8270666(100%)	6227022(100%)	12661034(100%)	13144962(100%)	30750380(100%)	10976112(100%)	10195240(100%)	12606908(100%)	27221698(100%)	15120786(100%)	13162770(100%)
QC failed	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Optical/PCR duplicate	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Non primary hits	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Unmapped reads	201241(2.503%)	153370(1.854%)	169155(2.716%)	1552864(12.26%)	513583(3.907%)	855080(2.781%)	551647(5.026%)	858429(8.42%)	1167444(9.26%)	1054228(3.873%)	345905(2.288%)	462958(3.517%)
mapq < mapq_cut (non-unique)	1622045(20.17%)	1478551(17.88%)	1192806(19.16%)	2824456(22.31%)	1538084(11.7%)	3263743(10.61%)	1307511(11.91%)	1092893(10.72%)	1292411(10.25%)	3059140(11.24%)	1271287(8.408%)	1444026(10.97%)
mapq >= mapq_cut (unique)	6217608(77.32%)	6638745(80.27%)	4865061(78.13%)	8283714(65.43%)	11093295(84.39%)	26631557(86.61%)	9116954(83.06%)	8243918(80.86%)	10147053(80.49%)	23108330(84.89%)	13503594(89.3%)	11255786(85.51%)
Read-1	3101716(38.57%)	3314180(40.07%)	2431510(39.05%)	4162383(32.88%)	5564586(42.33%)	13328751(43.34%)	4553310(41.48%)	4121680(40.43%)	5073143(40.24%)	11567320(42.49%)	6753867(44.67%)	5638595(42.84%)
Read-2	3115892(38.75%)	3324565(40.2%)	2433551(39.08%)	4121331(32.55%)	5528709(42.06%)	13302806(43.26%)	4563644(41.58%)	4122238(40.43%)	5073910(40.25%)	11541010(42.4%)	6749727(44.64%)	5617191(42.67%)
Reads map to '+'	3108255(38.66%)	3319193(40.13%)	2431042(39.04%)	4135485(32.66%)	5546967(42.2%)	13317610(43.31%)	4556669(41.51%)	4121707(40.43%)	5073156(40.24%)	11549678(42.43%)	6748384(44.63%)	5625475(42.74%)
Reads map to '-'	3109353(38.67%)	3319552(40.14%)	2434019(39.09%)	4148229(32.76%)	5546328(42.19%)	13313947(43.3%)	4560285(41.55%)	4122211(40.43%)	5073897(40.25%)	11558652(42.46%)	6755210(44.67%)	5630311(42.77%)
Non-splice reads	6217608(77.32%)	6638745(80.27%)	4865061(78.13%)	8283714(65.43%)	11093295(84.39%)	26631557(86.61%)	9116954(83.06%)	8243918(80.86%)	10147053(80.49%)	23108330(84.89%)	13503594(89.3%)	11255786(85.51%)
Splice reads	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
Reads mapped in proper pairs	5841868(72.65%)	6381962(77.16%)	4643670(74.57%)	7890884(62.32%)	10584206(80.52%)	25483066(82.87%)	8525326(77.67%)	7846264(76.96%)	9641796(76.48%)	22125100(81.28%)	12940926(85.58%)	10770940(81.83%)
Proper-paired reads map to different chrom	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)
total_mapping	7839653(97.5%)	8117296(98.15%)	6057867(97.28%)	11108170(87.74%)	12631379(96.09%)	29895300(97.22%)	10424465(94.97%)	9336811(91.58%)	11439464(90.74%)	26167470(96.13%)	14774881(97.71%)	12699812(96.48%)

其中比较重要的几个参数：

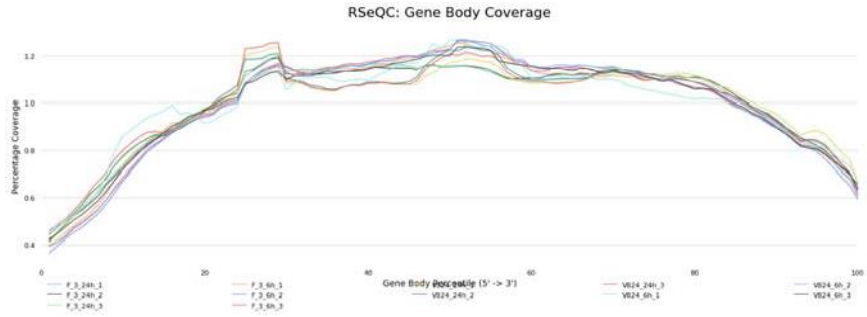
（1）Total reads：过滤后总的序列数；

（2）unique mapped：唯一比对的 reads；

比对结果 BAM 文件查看起来较为困难，可通过 IGV 软件进行可视化，显示 reads 在各染色体上的分布及在基因组中注释的外显子、内含子、基因间区等功能区域的分布，如下图所示：具体的使用方法可参考提供的使用说明文档。

3.2 比对区域分布

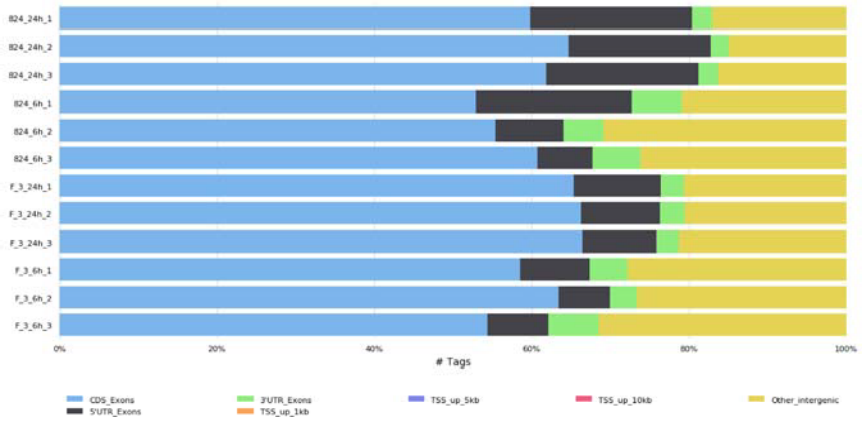
根据比对结果，分别统计 reads 在基因组外显子区域，内含子区域以及基因间区所占的比例。一般模式物种的基因注释较为完善（如人和小鼠），其比对到外显子区域的比例最高。比对到内含子区域的 reads 可能来源于前体 mRNA 或可变剪接事件滞留的内含子。比对到基因间区的 reads，可能来源于 ncRNA 或少许 DNA 片段污染，也可能是基因注释还不够完善。所有样本的测序 reads 在基因组区域分布情况如下图所示:



All.geneBodyCoverage.curves: reads gene body 各个区域上覆盖的

All.geneBodyCoverage.curves:显示了 reads 在 gene body 各个区域上覆盖的密度图。

Reads_distribution_across_genomic_regions:显示了 reads 在 gene body 上分布百分比统计。本次测序为去除 rRNA 测序，理论上在外显子区域的 reads 比例占比最多。



3.3 比对分析 Q&A

问：有参分析都需要什么文件？

答：相应的参考基因组及基因结构注释文件（gtf/gff/gff3/bed 等格式，推荐 gtf，gff）、基因的 GO 注释文件的直接下载链接以及基因功能描述文件。

问：造成 mapping rate 较低的原因可能有哪些？

答：当 mapping rate 较低时主要可能有 2 个原因：（1）由于 reference 组装不好，或者所测物种与 reference 的亲缘关系较远；（2）由于样品的特殊前处理或者相对于参考基因组此样品本身的变异太大，导致 mapping rate 相对较低。

问：mapping 时用的是 read 全长，还是头尾有处理？

答：实验方面，我们使用标准的 RNA-seq 试剂盒，其 index 处于 Adapter 中间，在测序中由 Index read 完成，由此测序得到的 Read 1 和 Read 2 的各个碱基全都是样本的序列，因此 mapping 时，头尾不处理。信息分析方面，我们会将过滤得到的 clean reads 的全长进行 mapping。

问：Read-1,Read-2 的具体含义是什么？

答：双端测序会在 cDNA 片段的两端先后读取一定长度的碱基（如 pe150，就是分别测 150bp）得到一对 reads，其中一条称为 Read-1，另一条称为 Read-2。大量文献和实际项目都表明，转录组分析中使用双端 reads 对于序列拼接和定量都大有裨益。

4 表达量分析

4.1 表达量估计

4.1.1 FPKM 值和 counts 值

基因表达水平一般是通过该基因转录的 mRNA 的多少来衡量的。每个基因转录产生的 mRNA 的量，是受到时空等多种因素调控的，个体在不同的生长发育阶段，或者不同的组织水平,或者响应不同的实验处理，基因转录出 mRNA 的量都是不一样的。我们通过 featurecounts(Liao Y, Smyth GK and Shi W., 2014)软件统计基因的表达量。

FPKM(Wagner, et al., 2012)是利用 RNA-Seq 技术用来定量估计基因表达值的一个非常有效的工具。FPKM 是 Reads per Kilobase Millon Mapped Reads 的缩写，由 Mortazavi 于 2008 年第一次提出。其计算公式为:

RPKM (FPKM)

- Reads (fragments) per Kilobase Per Million

$$\text{RPKM} = \frac{\text{raw number of reads}}{\text{exon length}} \times \frac{1{,}000{,}000}{\text{Number of reads mapped in the sample}}$$

- In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:
  - The number of fragments is biased towards larger genes
  - Total number of fragments is related to total library depth
- Use of FPKM/RPKM normalizes for gene size and library depth

FPKM

图 4.1 RPKM 计算公式

设 FPKM(A)为基因 A 的表达量，则分子为唯一比对到该基因的 Reads 数，分母为唯一比对到参考基因的总 Reads 数乘以基因 A 的外显子区域的长度。FPKM 法能消除基因长度和测序量差异对计算基因表达的影响，计算得到的基因表达量可直接用于比较不同样品间的基因表达差异。

```
## [1] "├── genes_count.txt"
## [2] "├── genes_count.txt.summary"
## [3] "├── multiqc_data"
## [4] "│   ├── mqc_featureCounts_assignment_plot_1.txt"
## [5] "│   ├── multiqc_data.json"
## [6] "│   ├── multiqc_featureCounts.txt"
## [7] "│   ├── multiqc_general_stats.txt"
## [8] "│   ├── multiqc.log"
## [9] "│   └── multiqc_sources.txt"
## [10] "├── multiqc_plots"
## [11] "│   ├── pdf"
## [12] "│   │   ├── mqc_featureCounts_assignment_plot_1_pc.pdf"
## [13] "│   │   └── mqc_featureCounts_assignment_plot_1.pdf"
## [14] "│   ├── png"
## [15] "│   │   ├── mqc_featureCounts_assignment_plot_1_pc.png"
## [16] "│   │   └── mqc_featureCounts_assignment_plot_1.png"
## [17] "│   └── svg"
## [18] "│       ├── mqc_featureCounts_assignment_plot_1_pc.svg"
## [19] "│       └── mqc_featureCounts_assignment_plot_1.svg"
## [20] "├── multiqc_report.html"
## [21] "├── protein_count.txt"
## [22] "├── protein_count.txt.summary"
## [23] "├── Sample_FPKM_boxplot.pdf"
## [24] "├── Sample_FPKM_boxplot.png"
## [25] "├── Sample_FPKM_density.pdf"
## [26] "├── Sample_FPKM_density.png"
## [27] "├── Sample_normalized_counts_correlation.pdf"
```

```
## [28] "├─── Sample_normalized_counts_correlation.png"

## [29] "├─── Sample_TMM_correlation.pdf"

## [30] "├─── Sample_TMM_correlation.png"

##  [ reached getOption("max.print") -- omitted 4 entries ]
```

其中 FPKM.xlsx 和 counts.xlsx 分别为用 FPKM 值和 counts 表示的表达值。

ID	824_24h_1	824_24h_2	824_24h_3	824_6h_1	824_6h_2	824_6h_3	F_3_24h_1	F_3_24h_2	F_3_24h_3	F_3_6h_1	F_3_6h_2	F_3_6h_3	seqnames	start	end	width	strand	source	type	Name	gene_biotype	locus_tag	product	protein_id
gene0	18.31595	28.09917	35.79180	358.23432	282.100161	316.67231	421.86721	424.92288	402.56929	251.068506	259.92509	291.923347	NC_017295.1	467	1807	1341	+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosomal replication initiator protein DnaA	WP_010963330.1
gene1	12.10356	13.00525	23.33191	59.60318	51.219856	56.40512	127.08542	116.82640	94.84225	43.161708	48.14479	45.373547	NC_017295.1	2064	3164	1101	+	RefSeq	gene	CEA_RS00010	protein_coding	CEA_RS00010	DNA polymerase III subunit beta	WP_010963331.1
gene10	0.00000	0.00000	0.00000	0.00000	0.000000	0.00000	0.00000	0.00000	0.00000	1.069595	0.00000	2.260567	NC_017295.1	14497	14573	77	+	RefSeq	gene	CEA_RS00055	tRNA	CEA_RS00055	NA	NA
gene100	14.80100	20.24977	18.40990	56.00001	29.874943	26.06816	138.16438	109.74067	143.03253	28.350096	31.83628	34.130133	NC_017295.1	100676	101134	459	+	RefSeq	gene	CEA_RS00505	protein_coding	CEA_RS00505	DUF441 domain-containing protein	WP_010963419.1
gene1000	13.14690	20.14393	18.06949	16.22727	8.893785	13.04238	29.12252	19.44634	19.55262	6.863235	10.58707	11.494771	NC_017295.1	1086261	1086896	636	+	RefSeq	gene	CEA_RS04980	protein_coding	CEA_RS04980	hemolysin D	WP_010964265.1
gene1001	16.06066	24.94250	26.36924	14.96132	14.182652	13.96983	20.46569	16.21406	14.69913	8.956279	14.56341	19.134660	NC_017295.1	1087584	1088425	846	+	RefSeq	gene	CEA_RS04985	protein_coding	CEA_RS04985	fatty acid-binding protein DegV	WP_010964266.1
gene1002	60.79337	55.07438	41.58907	50.96889	44.098122	49.92310	231.05112	156.95652	163.23694	97.467152	110.89282	83.721903	NC_017295.1	1088431	1088877	447	+	RefSeq	gene	CEA_RS04990	protein_coding	CEA_RS04990	PadR family transcriptional regulator	WP_010964267.1
gene1003	28.40159	40.04743	39.80140	63.85080	56.514786	57.04975	166.67688	128.73015	141.46267	71.914769	87.46327	102.483143	NC_017295.1	1088991	1089542	552	+	RefSeq	gene	CEA_RS04995	protein_coding	CEA_RS04995	membrane protein	WP_010964268.1
gene1004	334.48127	409.37373	406.07629	638.27279	493.193291	352.25004	371.89246	393.52319	494.86169	354.791132	402.92796	531.861234	NC_017295.1	1089719	1090150	432	+	RefSeq	gene	CEA_RS05000	protein_coding	CEA_RS05000	transcriptional repressor	WP_010964269.1
gene1005	536.97932	645.85583	609.38998	983.79614	845.113418	587.46610	638.40526	632.89610	829.03091	554.310684	646.13813	913.203637	NC_017295.1	1090152	1090496	345	-	RefSeq	gene	CEA_RS05005	protein_coding	CEA_RS05005	hypothetical protein	WP_010964270.1

ID:注释文件中 Gene 的 ID；

后续多列:为该基因在各个样本中的 FPKM 值表达值；

seqnames：染色体的编号；start：起始坐标；end：终止坐标；width：基因长度；strand：正义链/反义链；source：数据库来源；

type：类型；Dbxref：其他数据库中的 ID；

Names：基因的缩写,适合于在文献中检索；gene_biotype：gene 的 bio 类型比如 蛋白编码、非编码的；

product：基因的编码产物；protein_id：编码的蛋白编号。

counts 值表

ID	824_24h_1	824_24h_2	824_24h_3	824_6h_1	824_6h_2	824_6h_3	F_3_24h_1	F_3_24h_2	F_3_24h_3	F_3_6h_1	F_3_6h_2	F_3_6h_3	seqnames	start	end	width	strand	source	type	Name	gene_biotype	locus_tag	product	protein_id
gene0	94	150	142	2467	2207	5785	2810	2534	2952	4088	2433	2249	NC_017295.1	467	1807	1341	+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosomal replication initiator protein DnaA	WP_010963330.1
gene1	51	57	76	337	329	846	695	572	571	577	370	287	NC_017295.1	2064	3164	1101	+	RefSeq	gene	CEA_RS00010	protein_coding	CEA_RS00010	DNA polymerase III subunit beta	WP_010963331.1
gene10	0	0	0	0	0	0	0	0	0	1	0	1	NC_017295.1	14497	14573	77	+	RefSeq	gene	CEA_RS00055	tRNA	CEA_RS00055	NA	NA
gene100	26	37	25	132	80	163	315	224	359	158	102	90	NC_017295.1	100676	101134	459	+	RefSeq	gene	CEA_RS00505	protein_coding	CEA_RS00505	DUF441 domain-containing protein	WP_010963419.1
gene1000	32	51	34	53	33	113	92	55	68	53	47	42	NC_017295.1	1086261	1086896	636	+	RefSeq	gene	CEA_RS04980	protein_coding	CEA_RS04980	hemolysin D	WP_010964265.1
gene1001	52	84	66	65	70	161	86	61	68	92	86	93	NC_017295.1	1087584	1088429	846	+	RefSeq	gene	CEA_RS04985	protein_coding	CEA_RS04985	fatty acid-binding protein DegV	WP_010964266.1
gene1002	104	98	55	117	115	304	513	312	399	529	346	215	NC_017295.1	1088431	1088877	447	+	RefSeq	gene	CEA_RS04990	protein_coding	CEA_RS04990	PadR family transcriptional regulator	WP_010964267.1
gene1003	60	88	65	181	182	429	457	316	427	482	337	325	NC_017295.1	1088991	1089542	552	+	RefSeq	gene	CEA_RS04995	protein_coding	CEA_RS04995	membrane protein	WP_010964268.1
gene1004	553	704	519	1416	1243	2073	798	756	1169	1861	1215	1320	NC_017295.1	1089719	1090150	432	+	RefSeq	gene	CEA_RS05000	protein_coding	CEA_RS05000	transcriptional repressor	WP_010964269.1
gene1005	709	887	622	1743	1701	2761	1094	971	1564	2322	1556	1810	NC_017295.1	1090152	1090496	345	-	RefSeq	gene	CEA_RS05005	protein_coding	CEA_RS05005	hypothetical protein	WP_010964270.1

counts 值表表头与 FPKM 表一致，其中的值为 counts。后续差异表达基因的分析基于 counts 表进行。

4.1.2 表达量分布统计

一般而言，差异表达基因的数量只占整体基因的小部分，因此少量的差异表达基因对样品的表达量分布没有太大影响，因此所有样品应该具有类似的表达量分布情况。根据所有样品的基因表达量，得到该样品的表达量密度图如下：

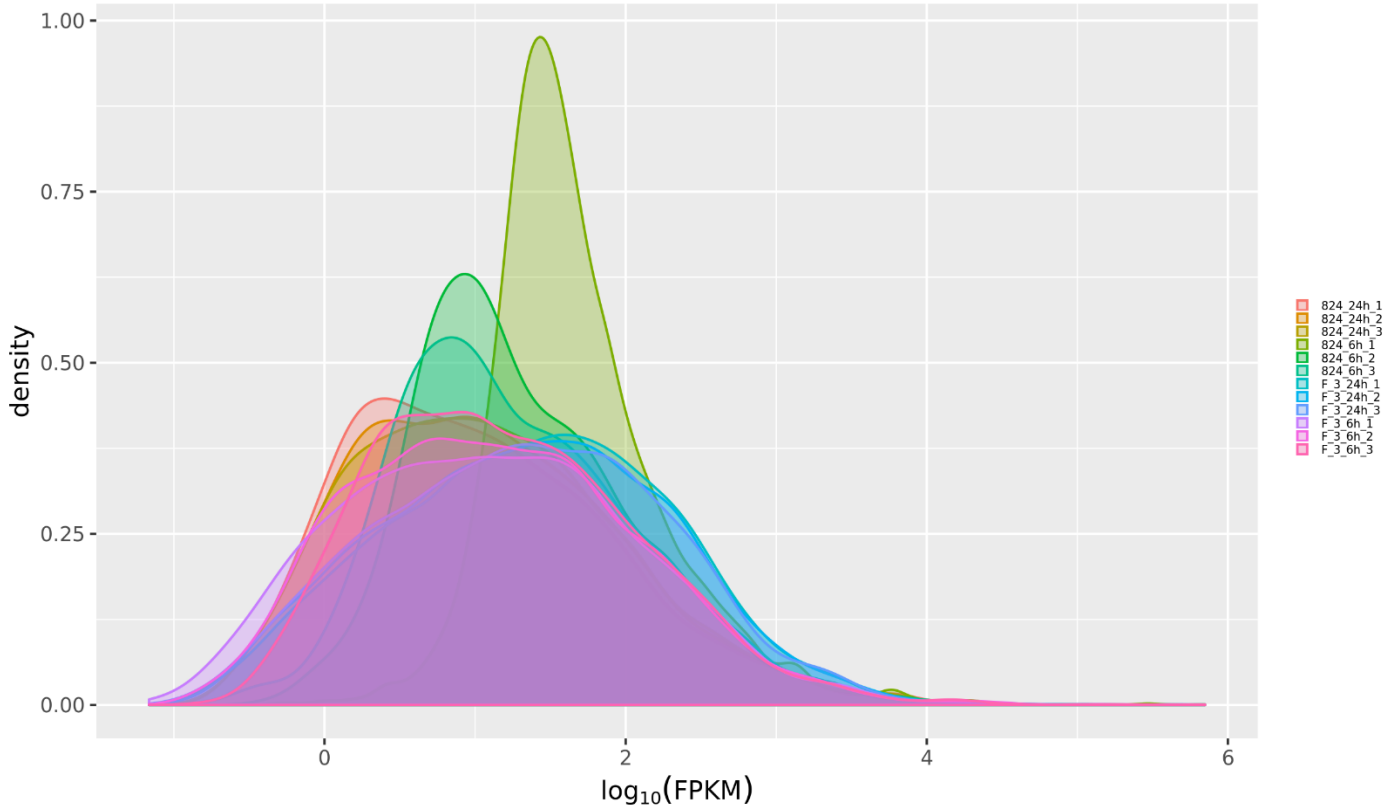


图 4.1 表达量 FPKM 分布图

对每组样品的基因表达量，取以 10 为底的对数后，做出密度分布图。横坐标为 log10(RPKM)，纵坐标为基因的密度。不同颜色代表不同样品。

根据每个样品的表达量，对每个样品进行绘制箱子图，查看样品的表达量整体分布趋势，得到所有样品的表达量的分布箱式图如下：

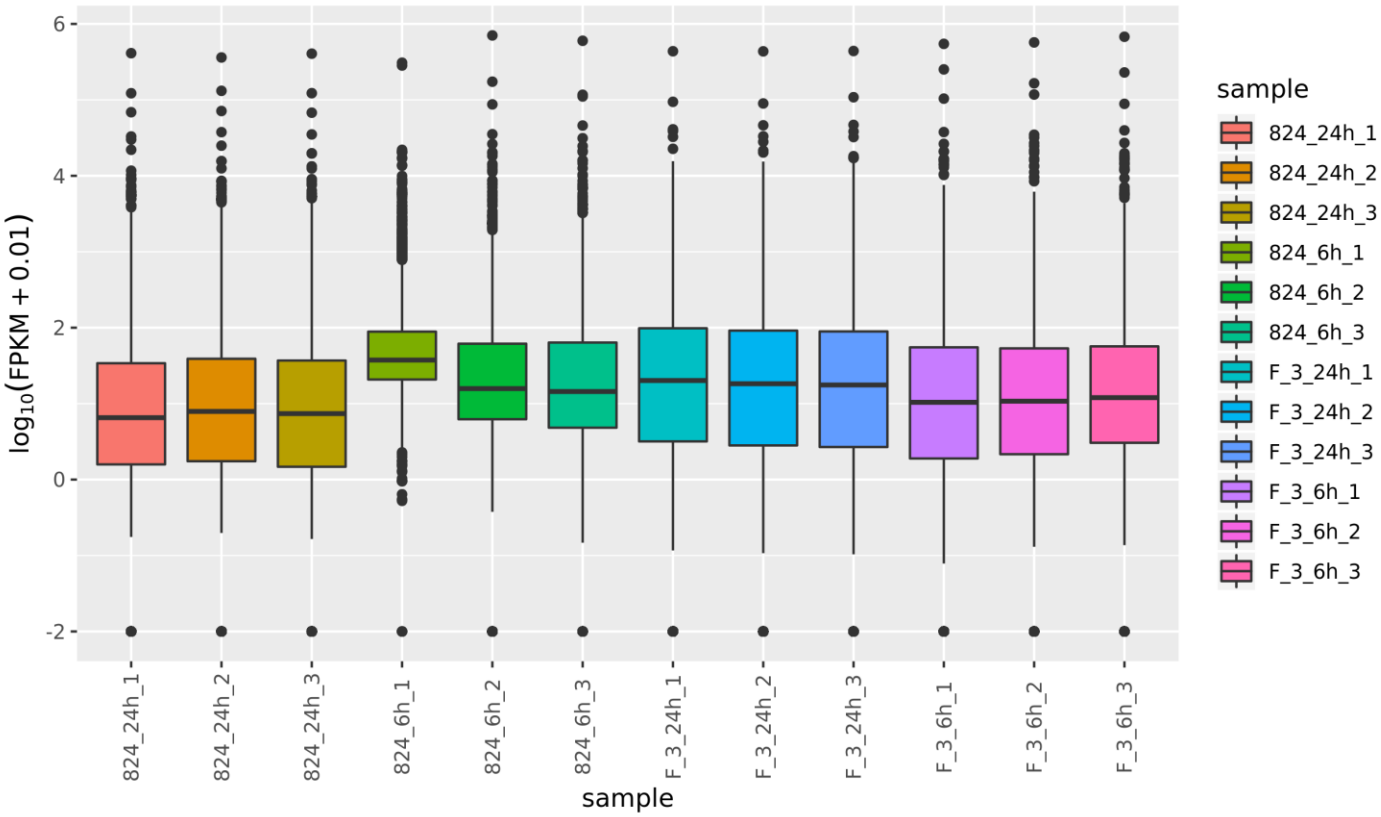


图 4.2 表达量箱式分布图

4.1.3 样品的聚类

生物学重复是任何生物学实验所必须的，高通量测序技术也不例外。生物学重复主要有两个用途：一个是证明所涉及的生物学实验操作是可以重复的且变异不大，另一个是为了确保后续的差异基因分析得到更可靠的结果。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否合理的重要指标。相关系数越接近 1，表明样品之间表达模式的相似度越高。Encode 计划建议皮尔逊相关系数的平方(R2)大于 0.92(理想的取样和实验条件下)。具体的项目操作中，我们要求生物学重复样品间 R2 至少要大于 0.8，否则需要对样品做出合适的解释，或者重新进行实验。根据各样本所有基因的表达值：RPKM，计算组内及组间样本的相关性系数，绘制成热图，可直观显示组间样本差异及组内样本重复情况。样本间相关性系数越高，其表达模式越为接近，样本相关性热图如下图所示：

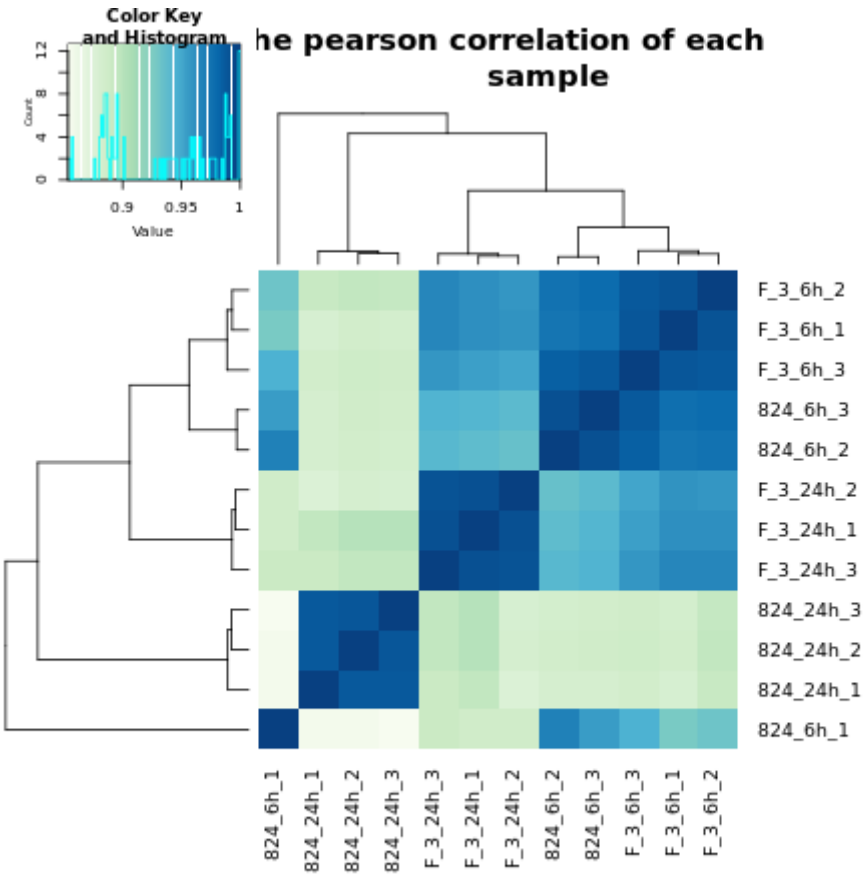
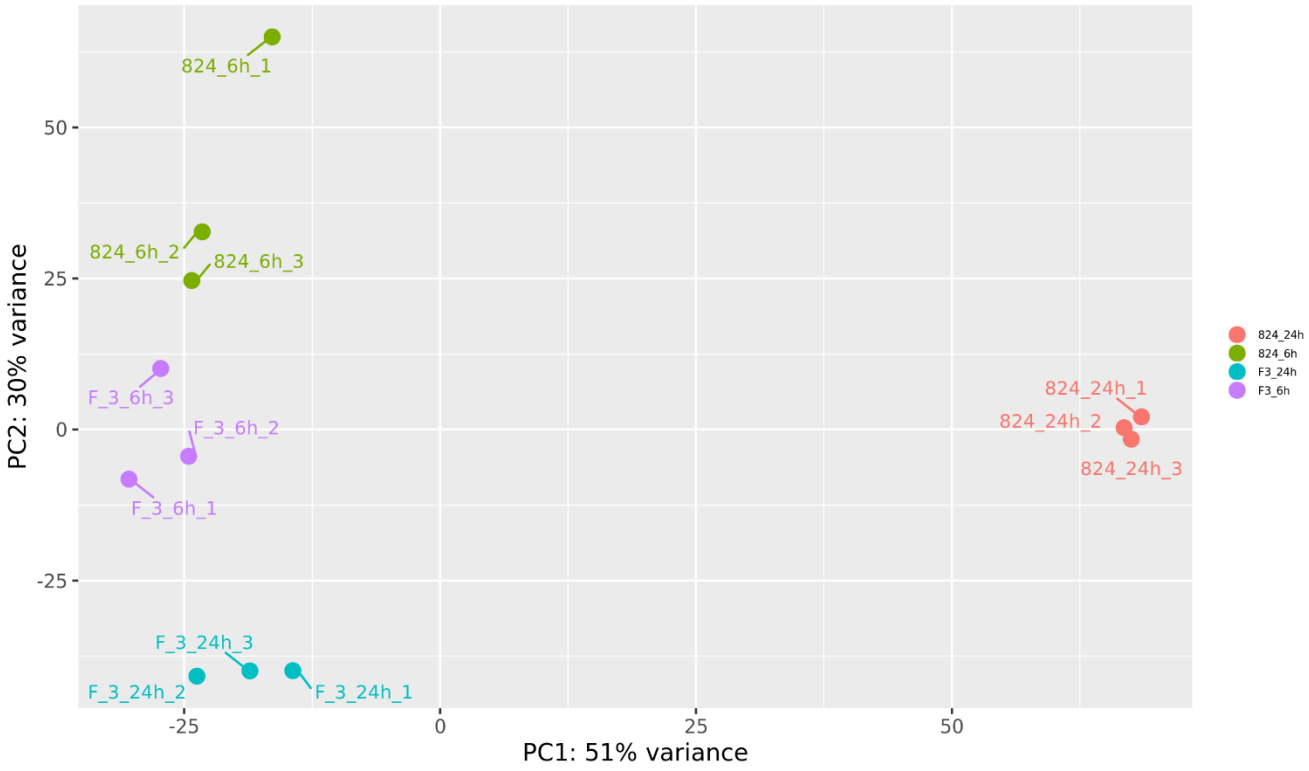


图 4.3 样本两两之间的相关性热图

计算样本两两之间的基因表达值的皮尔逊相关系数（Pearson Correlation Efficiency），再利用系统聚类法（Hierarchical Cluster）将相似度高的样品归为一类，以此类推，最终得到样品的整体聚类结果。

主成分分析（PCA）也常用来评估组间差异及组内样本重复情况，PCA 采用线性代数的计算方法，对数以万计的基因变量进行降维及主成分提取。理想条件下，PCA 图中，组间样本应该分散，组内样本应该聚在一起。



PCA

图 4.4 主成分分析图

主成分分析图，这里列出了两个主成分，分别为 PC1 和 PC2，这两个主成分对样本分类的解释度分别见坐标轴上的百分比。

4.2 差异表达分析

基因差异表达分析的输入数据为基因表达水平分析中得到的 readcount 数据,分析主要分为三部分：

1)首先对 readcount 进行标准化（normalization）；

2)然后根据模型进行假设检验概率（pvalue）的计算；

3)最后进行多重假设检验校正，得到 FDR 值（错误发现率)。

针对不同情况，会采用不同的软件进行基因差异表达的分析。分析方法如下表：

类型	软件	标准化方法	pvalue计算模型	FDR计算方法	差异基因筛选标准
有生物学重复	DESeq2(Anders et al, 2014)	DESeq2	负二项分布	BH	|log2(FoldChange)| > log2(1.5)&padj < 0.05
无生物学重复	DEGseq(Wang L & Wang. X, 2019)	TMM	负二项分布	BH	|log2(FoldChange)| > log2(1.5)&qvalue < 0.001

4.2.1 差异表达分析统计结果

```
## [1] "./4.DEG"
## [2] "├── DEG_stat.xlsx"
## [3] "├── DESeq2.normalized.rlog.pearson.pdf"
## [4] "├── DESeq2.normalized.rlog.pearson.png"
## [5] "├── induced_VS_ctrl"
## [6] "|      ├── cluster.xlsx"
## [7] "|      ├── heatmap.pdf"
## [8] "|      ├── heatmap.png"
## [9] "|      ├── induced_VS_ctrl-DEseq2_DEG_results.xlsx"
## [10] "|      ├── induced_VS_ctrl-DEseq2_results.xlsx"
```

```
## [11] "|      ├── induced_VS_ctrl-volcano.pdf"
## [12] "|      └── induced_VS_ctrl-volcano.png"
## [13] "├── normalized_counts.xlsx"
## [14] "├── PCA.pdf"
## [15] "└── PCA.png"
## [16] ""
## [17] "1 directory, 13 files"
```

如上述文件夹结构示意图所示，其中的子文件夹名即为差异比较组的组名。以前者为实验组，后者为对照组，实验组_VS_对照组。其中 DEGse2_results.xlsx 的表格表示所有参与比较分析的基因的结果。而 DEG_DEseq2_results.xlsx 的结果则是差异基因筛选标准筛选出来的差异基因结果。

我们采用 Deseq2 进行基因差异表达分析，比较处理组与参考组，并选取|log2FoldChange|≥log21.5 和 padj< 0.05 的基因作为差异表达基因筛选标准，得到上下调基因。本项目所有组别差异表达基因结果见：

表 4.5 组间比较得到的差异表达基因的数量统计表

compare	up	down
824_24h_VS_824_6h	822	1134
F_3_6h_VS_824_6h	151	714
F3_24h_VS_824_24h	876	963
F3_24h_VS_F3_6h	367	584

如上表所示，一共进行了 4 组比较，分别为 824_24h_VS_824_6h, F3_24h_VS_824_24h, F3_24h_VS_F3_6h, F_3_6h_VS_824_6h。每一组比较的差异基因数目如上表所示。

表 4.6 表头信息展示

ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	F_3_6h_1	F_3_6h_2	F_3_6h_3	824_6h_1	824_6h_2	824_6h_3	F_3_24h_1	F_3_24h_2	F_3_24h_3	824_24h_1	824_24h_2	824_24h_3	seqnames	start	end	width	strand	sourcetypeName	gene_biotype	locus_tag	product	protein_id
gene3928	56354.800	9.2155810	.249300336	.96578	0	0	542.850852866	.76164504	.297970312	.85119488	.18811331	.24415	425.35688	441.71806	484.62840239347	.132237132	.320195380	.250	NC_017296.1	80971	82002	1032	*	RefSequ	CEA_RS19555	protein_coding	CEA_RS19555	epoxyqueuosine reductase	WP_010890762.1
gene3929	19318.134	8.7633730	.269885532	.47071	0	0	174.508688301	.31217141	.245282208	.39844162	.46775159	.43555	70.63688	44.63677	60.67705	80042.975	78405.310	72046.002	NC_017296.1	82040	83218	1179	*	RefSequ	CEA_RS19560	protein_coding	CEA_RS19560	acetyl-CoA C-acetyltransferase	WP_014519101.1
gene3927	10530.642	7.3233030	.242127930	.24559	0	0	318.566196293	.48588295	.045701296	.62552292	.75589184	.53516	285.61870	380.34250	382.18663	45809.236	46548.813	31280.495	NC_017296.1	80094	80870	777	*	RefSequ	CEA_RS19550	protein_coding	CEA_RS19550	penicillin-binding protein	WP_010890761.1
gene209	2355.663	7.8208680	.282927327	.64268	0	0	33.964778	51.84917	34.526624	54.76163	34.53421	32.52344	152.02286	115.31166	195.42739	11044.207	7880.803	8638.020	NC_017295.1	212460	213185	726	*	RefSequ	nagB	protein_coding	CEA_RS01050	glucosamine-6-phosphate deaminase	WP_010963511.1
gene3620	2099.528	6.0963310	.226334026	.93511	0	0	177.436686229	.89727145	.430328	86.19887126	.36380136	.10352	145.88052	130.19059	153.66266	8131.752	7885.432	7845.986	NC_017295.1	37165393717219	681	*	RefSequ	CEA_RS18020	protein_coding	CEA_RS18020	ABC transporter permease	WP_077836360.1
gene741	1142.426	5.0797780	.194591126	.10488	0	0	168.652692165	.33038140	.199021103	.94569124	.00920137	.51758	168.91428	136.70011	185.18321	3900.708	4281.787	4196.164	NC_017295.1	798627	800030	1404	*	RefSequ	CEA_RS03700	protein_coding	CEA_RS03700	uronate isomerase	WP_010964011.1
gene886	1003.175	7.9227430	.303952226	.06576	0	0	8.198395	10.76115	7.323829	24.33850	14.91250	10.25195	31.47948	21.38845	32.30856	3402.453	4469.260	4005.429	NC_017295.1	964360	965046	687	*	RefSequ	CEA_RS04410	protein_coding	CEA_RS04410	ABC transporter permease	WP_010964152.1
gene3621	1005.087	5.4394160	.215763325	.21010	0	0	110.092729122	.28578	79.515862	69.97320	92.61446	96.50977	88.29610	73.46469	82.74143	3509.812	3659.192	4076.551	NC_017295.1	37172123718156	945	*	RefSequ	CEA_RS18025	protein_coding	CEA_RS18025	ABC transporter ATP-binding protein	WP_010966786.1
gene3617	3106.122	3.9340090	.160226624	.55277	0	0	1055.250523974	.37313929	.080078722	.54934696	.96308610	.52149	403.85870	393.36155	479.11231	10686.344	10313.321	10008.724	NC_017295.1	37139033714955	1053	*	RefSequ	CEA_RS18005	protein_coding	CEA_RS18005	sensor histidine kinase	WP_010966782.1
gene1610	1526.892	6.4118500	.262884724	.39035	0	0	92.524741147	.72123	98.348567	70.98730	65.92894	66.10742	172.75325	106.01233	211.18766	5282.611	6040.792	5967.734	NC_017295.1	16921541692699	546	*	RefSequ	CEA_RS08015	protein_coding	CEA_RS08015	glutathione peroxidase	WP_010964857.1

basemean 值为在这些样本中表达值的平均值，可以以此值来过滤掉掉在所有样本中表达都不高的基因；

log2FoldChange：差异倍数取以 2 为底的对数，这里以|log2FoldChange|≥log2(1.5)为差异表达的标准；

padj 为统计显著性的 p 值的校正值，这里以该值< 0.05 为显著。

满足以上两个条件的基因则认为是差异表达基因。

DEseq2 会对基因的表达量进行标准化，见： normalized_counts.xlsx。建议后续基于表达量分析的内容基于该表格的值进行。

根据各比较组上下调基因，绘制差异表达基因火山图：

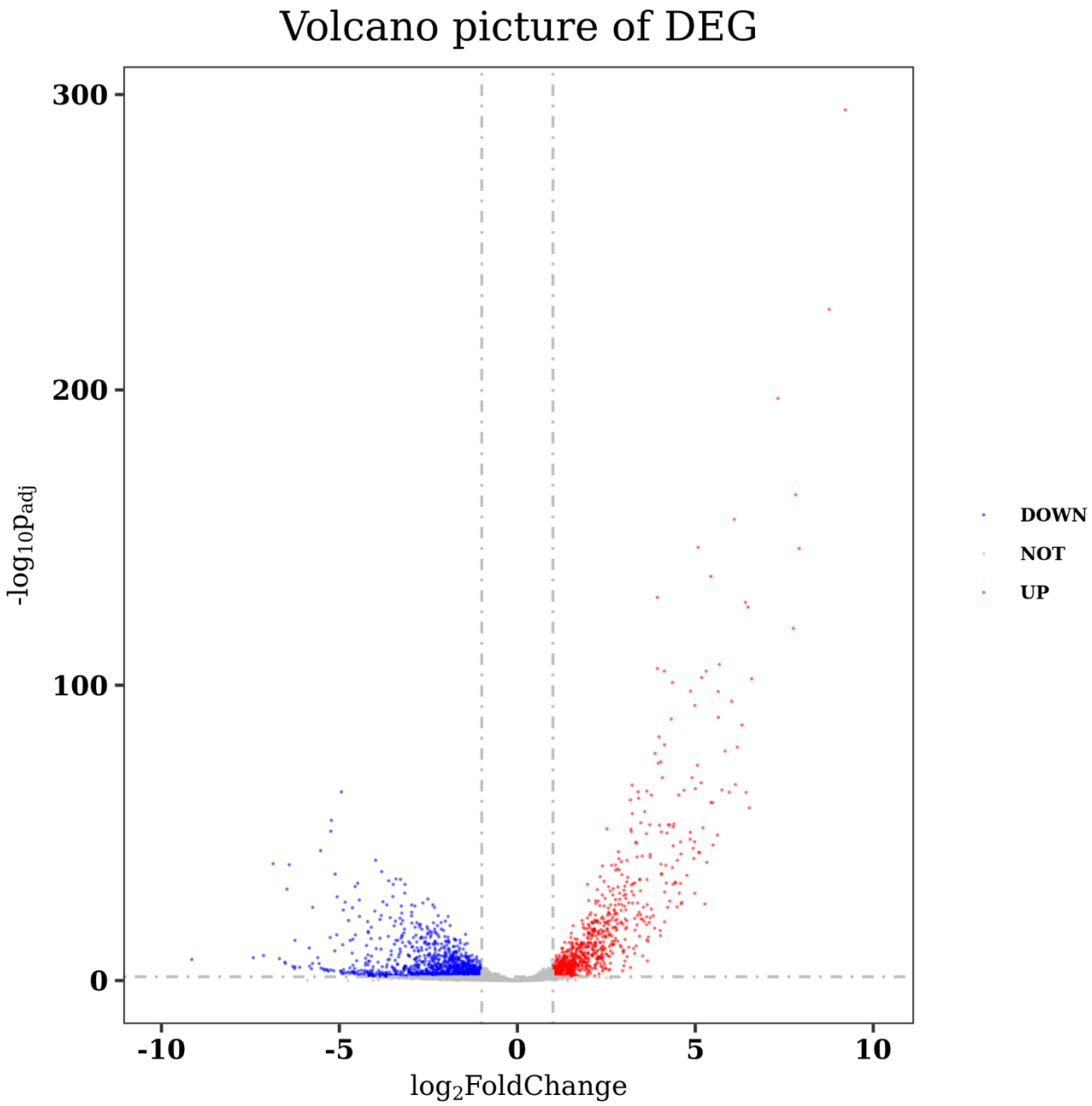
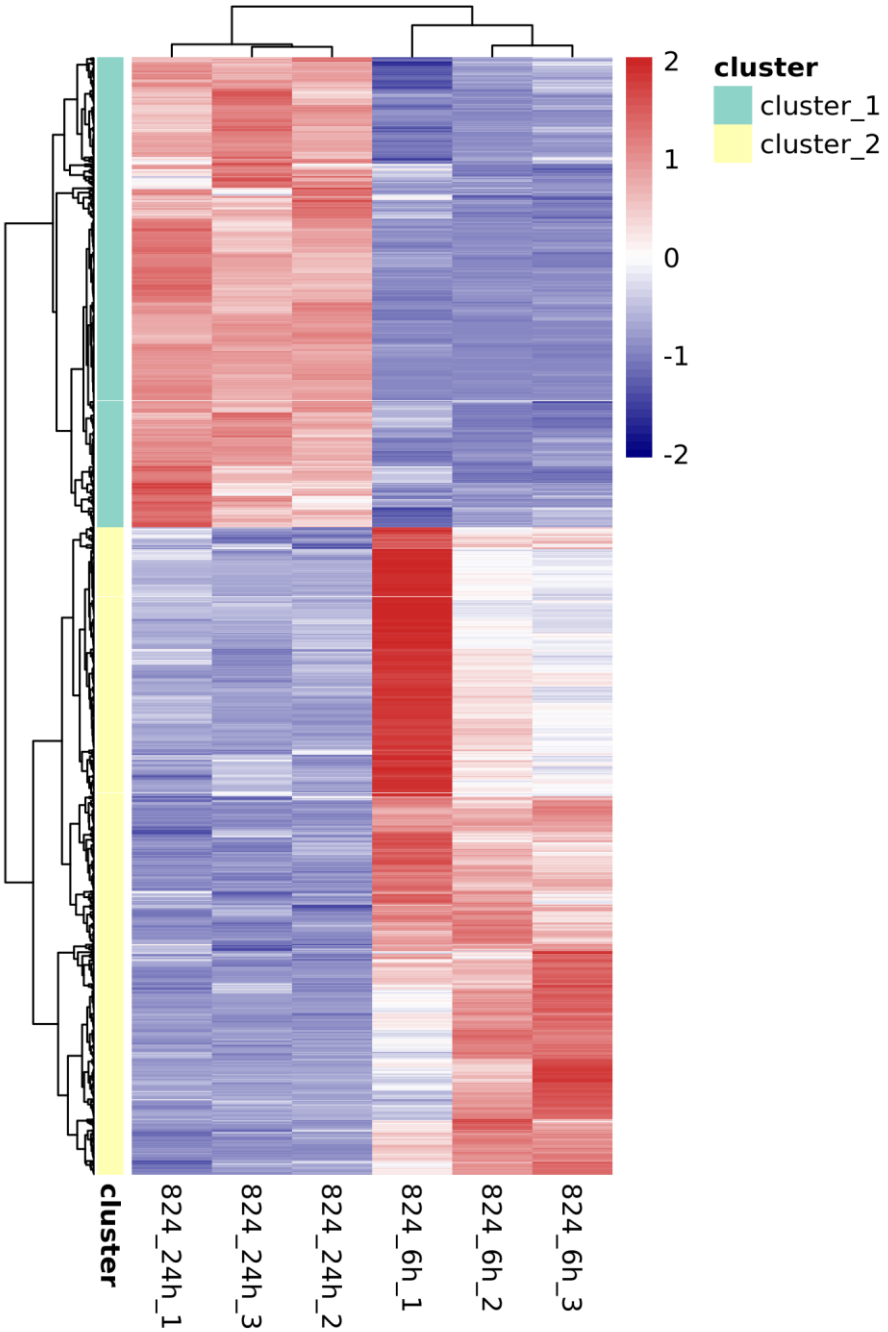


图 4.6 差异表达基因火山图

横坐标为不同实验组中/不同样品中表达倍数变化，纵坐标为表达量变化的统计学显著程度，不同颜色表示不同的分类。两条竖虚线表示 log2FoldChange 的阈值，而横虚线则表示 padj 值的阈值。这三条虚线与坐标轴框出来的区间中，左上为显著下调的基因，右上为显著上调的基因。

4.2.2 差异表达基因聚类图

通过比较处理组和参考组，对差异表达基因进行聚类分析，可以很直观反映出不同实验条件下样本差异表达基因的变化情况。我们利用 R 软件，对差异表达基因和不同样本/实验条件同时进行分层聚类分析。下图为差异表达基因聚类热图：



heatmap

图 4.7 差异基因聚类图

根据差异表达基因在样本中标准化后的表达值，计算皮尔森相关，再利用系统聚类法（Hierarchical Cluster），最终得到样品的整体聚类结果。在图中，表达量的变化用颜色的变化表示，蓝色表示表达量较低，红色表示表达量较高。

4.3 定量分析&差异表达分析 Q&A

问：认为基因表达的阈值是多少？为什么设置为这个阈值？

答：认为 FPKM ＞ 1 是基因表达的。这个阈值是主流杂志推荐的，也能够很好的反应基因的表达水平。

问：样品间的相关性有何意义？如何计算?

答：样品间的相关性反应了样品间的相似程度，即不同处理或组织的样品在表达水平方面的相似度。相关系数越接近 1，样品间的相似度越高，样品间的差异基因也越少。生物学重复间的样品的相关系数应大于生物学重复外的样品的相关系数。相关系数的计算方法有三种：A. Pearson correlation; B. Spearman rank correlation; C. Kendall’s τ。我们使用 R 语言进行 Pearson 相关系数的计算。

问：主成份分析是什么?

答：主成份分析 (Principal Component Analysis, PCA) 是一种多元分析技术。PCA 的核心思想在于，在尽可能保留数据的差异的前题下，降低数据的维度，也就是抽象出更少的互不相关的变量来描述各数据。数据集是一群在多维空间中的点，在保持这一群点的相对空间位置不变的情况下，旋转到一个新的坐标系（坐标轴就是各 PC），使得各点在新的坐标轴上的坐标（投影）的方差最大，而投影方差最大的坐标轴即为 PC1，其次为 PC2。

问：如何判断一个基因是否是差异基因？如果是差异基因，如何判断该基因的表达量是上调还是下调？

答：如果基因的 log2Foldchange>0，则认为该差异基因是上调，反之，若 log2Foldchange<0，认为该差异基因是下调。

问：能否用 FPKM 进行差异分析？

答：在做差异分析时，我们是采用 readcount 数据，通过 DESeq 或者 TMM 标准化后，进行差异分析。FPKM 实际上也是对 readcount 进行标准化处理的一种方法，各种标准化方法优劣势比较见下图(Dillies, M. A. et al, 2013)，可以看出，在进行差异分析时，DESeq 和 TMM 的标准化效果最好，FPKM 的标准化效果较差，所以，不推荐使用 FPKM 进行差异分析。

问：差异基因筛选条件最大能设的阈值是多少？很多客户希望通过调整差异基因筛选阈值来找相关基因是否有必要？

答：一般来说，等级较高的文章阈值的设置会比较严格；而在某些文章中，差异基因筛选阈值会适当放宽：如在一些无生物学重复的文章中，只将 qvalue 作为差异基因筛选标准，不考虑 log2foldchange；有的文章则将 pvalue 作为差异基因的筛选标准。

问：如何判断差异基因在两个样品间的差异大小？

答：差异的显著情况可通过|log2Foldchange|来判断，|log2Foldchange|越大，差异倍数越大。

问：差异基因筛选条件最大能设的阈值是多少？很多客户希望通过调整差异基因筛选阈值来找相关基因是否有必要？

答：差异基因的筛选是基于统计学意义的，不能直观的通过两个数值的大小判断是否为差异基因：

首先：受测序深度的影响，有些样品的测序深度较深，可能导致该样品的 readcount 数值较高，做差异分析的第一步就是要消除测序深度的影响，对原始数据进行标准化处理（我们在有重复项目中，使用 DESeq 标准化方法；无重复项目中，使用 TMM 标准化方法）

其次：在差异分析过程中，需要对 readcount 的分布进行估计，经验表明，readcount 服从负二项分布。在有重复的项目中，重复的好坏也会对差异基因与否产生影响。如果重复较差，组内差异情况会屏蔽掉部分组间的差异。在估计完参数后，需要用特定检验方法来判断差异基因与否

再次：在计算完 pvalue 以后，需要对 pvalue 进行多重假设检验校正，来减少假阳性。这个过程会使得 padj 会大于原来的 pvalue，使得部分通过 pvalue 阀值的基因，无法通过 padj 的阀值。

问：聚类分析是怎么做的？

答：聚类使用的为 R 中的聚类软件包 pheatmap，所针对的数据为 heatmap.txt（差异基因的并集），以基因的表达水平值 log10(FPKM+1)值进行聚类。其采用相应的距离算法，算出每个基因之间的距离，然后通过反复迭代，计算基因之间的相对距离，最后根据基因的相对距离远近来分成不同的 subcluster，从而实现聚类。该软件包是免费的，只需通过 R 来运行。

问：为什么要用校正后的 p 值，能直接用 p_value 吗？

答：校正后的 p 值（padj/qvalue），是对 p 值进行了多重假设检验，能更好地控制假阳性率。

5 功能分析

5.1 GO 功能分析

5.1.1 GO 富集分析

Gene Ontology（简称 GO）是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（Controlled Vocabulary）来全面描述生物体中基因和基因产物的属性。GO 总共有三个 Ontology，分别描述基因的分子功能（Molecular Function）、细胞组分（Cellular Component）、生物过程（Biological Process）。

如果研究的物种有相关 GO 注释数据库，直接该数据库进行 GO 的分析；如果没有，可以利用注释工具，得到每个基因对应的 GO 条目。

根据计算每个条目的基因数目，然后应用超几何检验，找出与整个基因组背景相比，差异表达基因显著富集的 GO 条目，其计算公式为:

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

GO_formula

图 5.1 GO 富集计算公式

其中，N 为该物种所有具有 GO 注释的基因数目；X 为 N 中差异表达基因的数目；n 为所有基因中注释到某特定 GO 条目的基因数目；x 为注释到某特定 GO 条目的差异表达基因数目。计算得到的 p-value 通过校正之后，以 q<0.05 为阈值，满足此条件的 GO 条目定义为在差异表达基因中显著富集的 GO 条目。通过 GO 功能显著性富集分析能确定差异表达基因行使的主要生物学功能。

```
##  [1] "./5-1.GO_KEGG_enrichment"
##  [2] "└── induced_VS_ctrl"
##  [3] "    ├── ALL"
##  [4] "    |    ├── GO_enrichment"
##  [5] "    |    |    ├── gene2GO_Biological Process.xlsx"
##  [6] "    |    |    ├── gene2GO_Cellular Component.xlsx"
##  [7] "    |    |    ├── gene2GO_Molecular Function.xlsx"
##  [8] "    |    |    ├── GO_Biological Process_bar.pdf"
##  [9] "    |    |    ├── GO_Biological Process_bar.png"
## [10] "    |    |    ├── GO_Biological Process_dot.pdf"
## [11] "    |    |    ├── GO_Biological Process_dot.png"
## [12] "    |    |    ├── GO_Biological Process_enrichment.xlsx"
## [13] "    |    |    ├── GO_Cellular Component_bar.pdf"
## [14] "    |    |    ├── GO_Cellular Component_bar.png"
## [15] "    |    |    ├── GO_Cellular Component_dot.pdf"
```

```
## [16] "    |      |      ├─── GO_Cellular Component_dot.png"
## [17] "    |      |      ├─── GO_Cellular Component_enrichment.xlsx"
## [18] "    |      |      ├─── GO_Molecular Function_bar.pdf"
## [19] "    |      |      ├─── GO_Molecular Function_bar.png"
## [20] "    |      |      ├─── GO_Molecular Function_cnetplot.pdf"
## [21] "    |      |      ├─── GO_Molecular Function_cnetplot.png"
## [22] "    |      |      ├─── GO_Molecular Function_dot.pdf"
## [23] "    |      |      ├─── GO_Molecular Function_dot.png"
## [24] "    |      |      └─── GO_Molecular Function_enrichment.xlsx"
## [25] "    |      └─── KEGG_enrichment"
## [26] "    |           ├─── gene2KEGG.xlsx"
## [27] "    |           ├─── kegg_bar2.pdf"
## [28] "    |           ├─── kegg_bar2.png"
## [29] "    |           ├─── kegg_bar.pdf"
## [30] "    |           ├─── kegg_bar.png"
##  [ reached getOption("max.print") -- omitted 70 entries ]
```

对 4 组比较：824_24h_VS_824_6h, F3_24h_VS_824_24h, F3_24h_VS_F3_6h, F3_6h_VS_824_6h 的差异基因分为上调 up、下调 down 和全部的差异基因 all，分别进行 GO 和 KEGG 富集分析

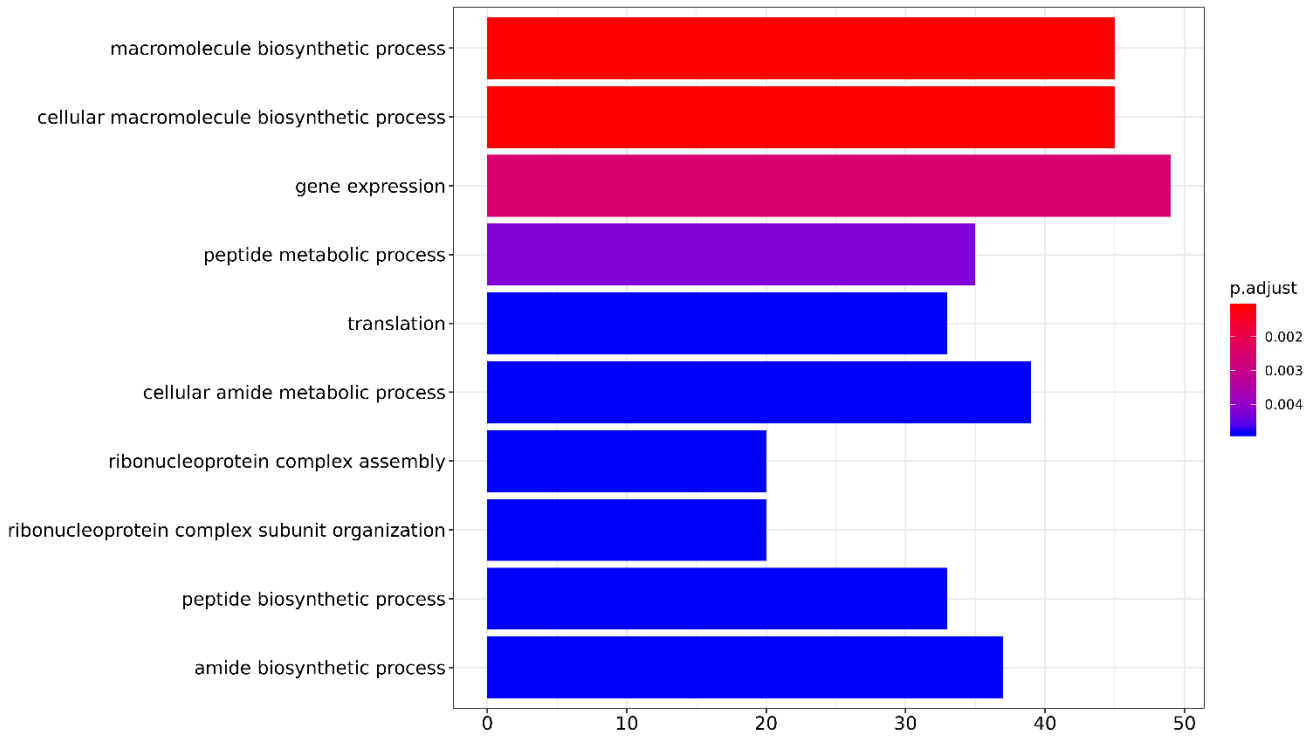
每两组比较得到的差异表达基因 GO 统计结果示例见下表：

表 5.1 GO 富集结果示例表

gene_ID	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	F_3_6h_1	F_3_6h_2	F_3_6h_3	X824_6h_1	X824_6h_2	X824_6h_3	F_3_24h_1	F_3_24h_2	F_3_24h_3	X824_24h_1	X824_24h_2	X824_24h_3	seqnames	start	end	widthstrand	source	typeName	gene_biotype	locus_tag	product	protein_id
gene0	GO:0009059	macromolecule biosynthetic process	45/132	99/484	0.000009	0.001129	0.000981	45	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0034645	cellular macromolecule biosynthetic process	45/132	99/484	0.000009	0.001129	0.000981	45	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0044260	cellular macromolecule metabolic process	56/132	159/484	0.004512	0.042690	0.037082	56	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0044249	cellular biosynthetic process	71/132	220/484	0.015783	0.122504	0.106411	71	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0009058	biosynthetic process	71/132	222/484	0.020816	0.150615	0.130830	71	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:1901576	organic substance biosynthetic process	71/132	222/484	0.020816	0.150615	0.130830	71	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0006259	DNA metabolic process	5/132	31/484	0.956782	1.000000	0.868635	5	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0043170	macromolecule metabolic process	69/132	200/484	0.001987	0.023280	0.020221	69	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0034641	cellular nitrogen compound metabolic process	77/132	242/484	0.015935	0.122504	0.106411	77	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1
gene0	GO:0006807	nitrogen compound metabolic process	90/132	299/484	0.046594	0.279666	0.242840	90	2393.931	2380.171	2353.042	1250.898	1732.204	2045.088	2157.496	2356.45	2326.216	258.7626	347.1719	459.0565	NC_017295.1	467	1807	1341+	RefSeq	gene	CEA_RS00005	protein_coding	CEA_RS00005	chromosome replication initiator protein DnaA	WP_010963330.1

GOID:在 GO 数据库中 ID；GOTerm：具体条目描述；GeneRatio 为该行 GO 条目在该 geneset 的前景值，BgRatio 为该行 GO 条目在这个物种中的背景值；Fisher 精确检验的 p 值；p.adjust p 值的校正值； qvalue q 值；geneIDs 为具体找到基因名字。

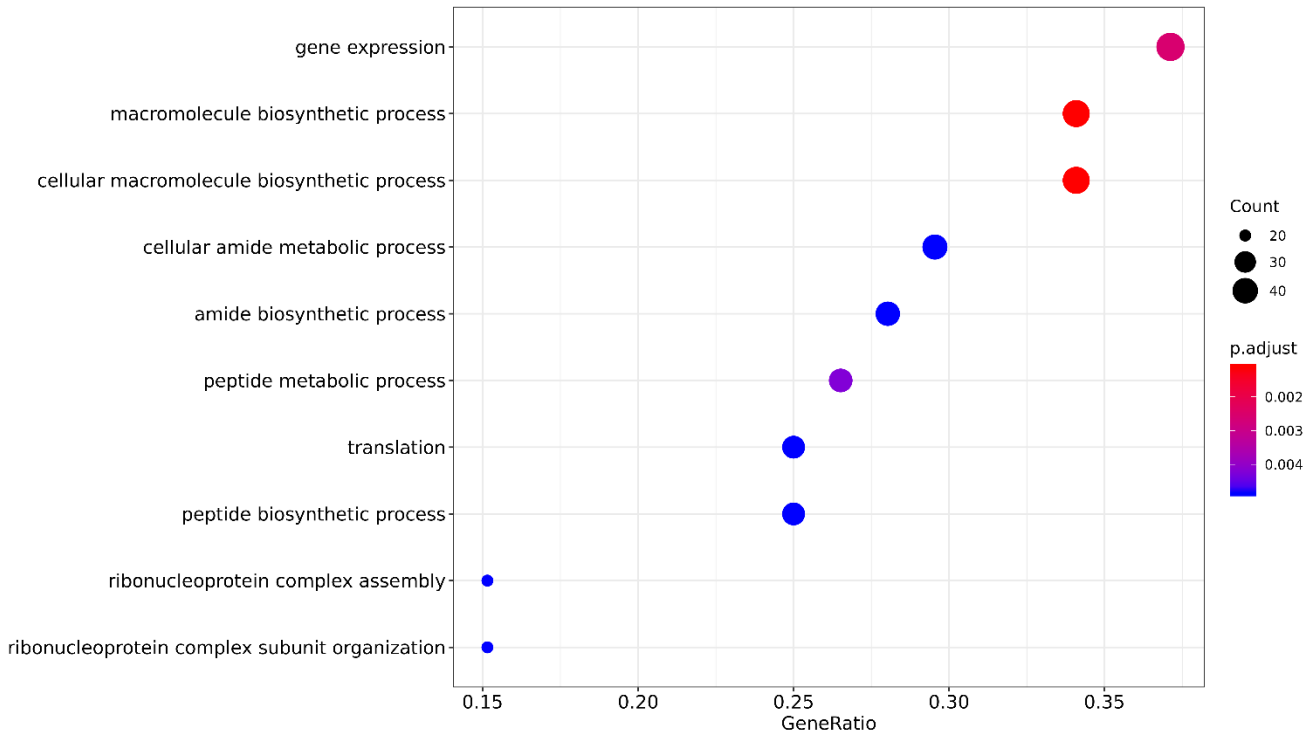
根据样品在该 GO 的富集显著性 p 值做出分布图，结果如下：



GO_BP_bar

图 5.2 富集 GO 条目 p 值分布图

取所有样品中富集的 GO 条目进行分析，纵坐标为 GO 的条目，横坐标为不同的其中找到的基因数目，不同的颜色代表富集的显著程度。



GO_BP_dot

每个点表示该 GO 条目的富集程度，颜色越趋近于红色表示富集程度越高。每个点的大小表示富集到该 GO 条目的基因的个数，点越大表示富集到该 GO 条目的基因越多，反之则越少。

5.2 KEGG 通路分析

KEGG（Kyoto Encyclopedia of Genes and Genomes，京都基因与基因组百科全书）是基因组破译方面的数据库。在给出染色体中一套完整基因的情况下，它可以对蛋白质交互（互动）网络在各种各样的细胞活动过程起的作用做出预测。KEGG 的 PATHWAY 数据库整合当前在分子互动网络（比如通路、联合体）的知识，GENES/SSDB/KO 数据库提供关于在基因组计划中发现的基因和蛋白质的相关知识，COMPOUND/GLYCAN/REACTION 数据库提供生化复合物及反应方面的知识。

其中基因数据库（GENES Database）含有所有已知的完整基因组和不完整基因组。有细菌、蓝藻、真核生物等生物体的基因序列，如人、小鼠、果蝇、拟南芥等等；通路数据库（PATHWAY Database）储存了基因功能的相关信息，通过图形来表示细胞内的生物学过程，例如代谢、膜运输、信号传导和细胞的生长周期；配体数据库（LIGAND Database）包括了细胞内的化学复合物、酶分子和酶反应的信息。

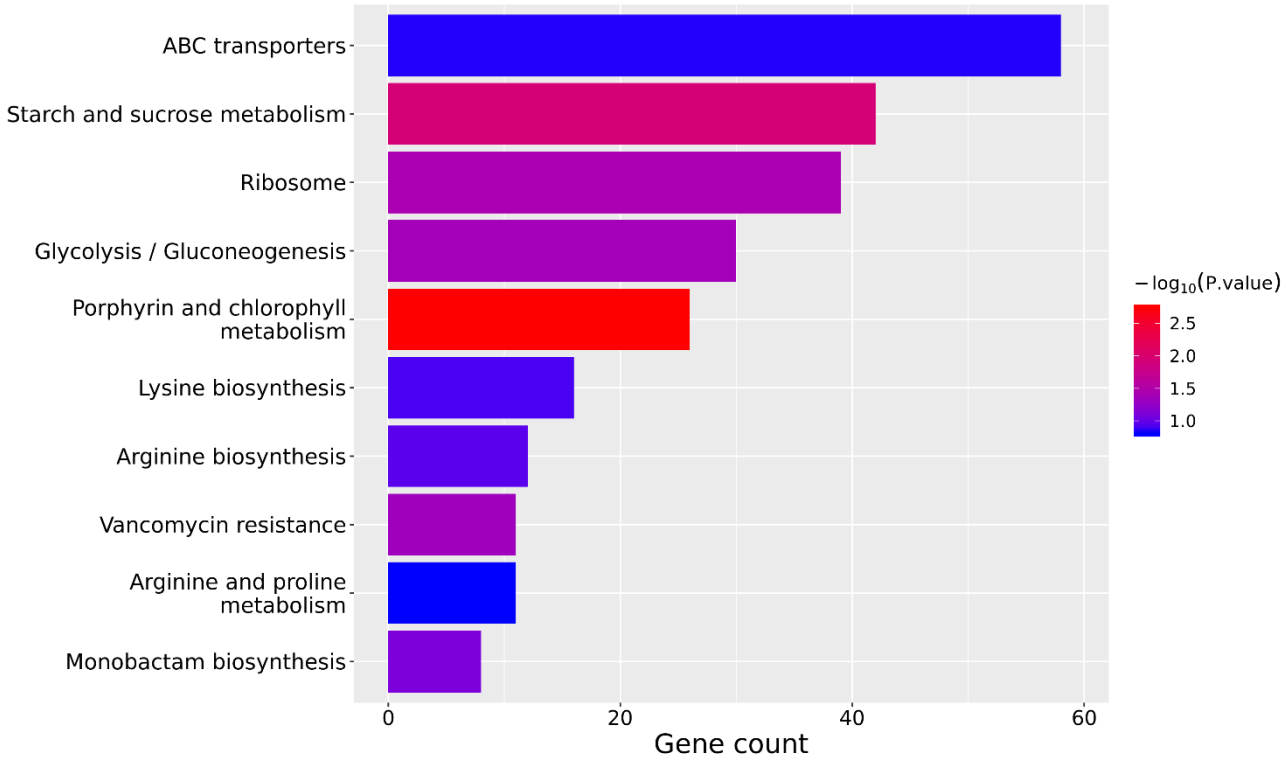
对 KEGG 中每个 Pathway 应用超几何检验进行富集分析，找出差异表达基因显著性富集的 Pathway。结果文件示例为:

表 5.2 KEGG 富集分析结果示例表

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
ko00860	Porphyrin and chlorophyll metabolism	26/654	33/1233	0.00182037	0.14042424	0.13821548	gene107/gene106/gene2508/gene1445/gene897/gene1435/gene1434/gene1443/gene1430/gene1046/gene103/gene632/gene108/gene1078/gene1436/gene104/gene1438/gene1440/gene1432/gene1919/gene1439/gene1441/gene1429/gene1446/gene2961/gene105	26
ko00500	Starch and sucrose metabolism	42/654	62/1233	0.01157030	0.44555088	0.43855241	gene964/gene965/gene609/gene969/gene1110/gene966/gene580/gene1467/gene236/gene2299/gene618/gene3049/gene2300/gene2764/gene2763/gene2298/gene3949/gene1132/gene1723/gene738/gene2396/gene2769/gene795/gene3526/gene584/gene755/gene3525/gene430/gene2892/gene469/gene1470/gene3524/gene970/gene468/gene470/gene579/gene1622/gene2675/gene2311/gene3065/gene1141/gene429	42
ko03010	Ribosome	39/654	60/1233	0.03749046	0.71050055	0.69933216	gene3216/gene3230/gene3214/gene1333/gene3223/gene3181/gene3228/gene1865/gene3200/gene3205/gene3190/gene3201/gene3198/gene1347/gene1792/gene3203/gene3213/gene3233/gene3832/gene3224/gene3218/gene3187/gene1818/gene1319/gene3182/gene2028/gene185604/gene1317/gene1845/gene1859/gene3229/gene3227/gene3207/gene3215/gene2420/gene3849/gene3834/gene1814/gene3199	39
ko00010	Glycolysis / Gluconeogenesis	30/654	45/1233	0.04224046	0.71050055	0.69933761	gene761/gene2524/gene2523/gene1633/gene4013/gene3767/gene762/gene1467/gene3885/gene3655/gene618/gene2290/gene293/gene2764/gene1093/gene563/gene795/gene3526/gene562/gene2563/gene3525/gene760/gene1470/gene3524/gene3915/gene579/gene2824/gene305665/gene1141/gene429	30
ko01502	Vancomycin resistance	11/654	14/1233	0.04613046	0.71050055	0.69933	gene2499/gene2188/gene1601/gene2980/gene3322/gene3509/gene1015/gene3394/gene537/gene3137/gene3428	11
ko00261	Monobactam biosynthesis	8/654	10/1233	0.07868061	0.86891096	0.85525	gene117/gene26/gene111/gene3710/gene309/gene118/gene2446/gene2445	8
ko00220	Arginine biosynthesis	12/654	17/1233	0.11125061	0.86891096	0.85525	gene20/gene2914/gene1057/gene2719/gene1111/gene1877/gene789/gene2455/gene2456/gene1029/gene2457/gene1530	12
ko00300	Lysine biosynthesis	16/654	24/1233	0.12580046	0.86891096	0.85525	gene2901/gene26/gene3509/gene1015/gene3710/gene309/gene2806/gene656/gene2446/gene2445/gene3137/gene523/gene2455/gene2190/gene2359/gene3428	16
ko02010	ABC transporters	58/654	99/1233	0.14734083	0.86891096	0.85525	gene3620/gene931/gene1042/gene932/gene3424/gene1040/gene933/gene3184/gene732/gene1041/gene3422/gene3423/gene2390/gene2508/gene3829/gene1428/gene1601/gene2722/gene2052/gene3740/gene2053/gene1942/gene535/gene574/gene2931/gene3117/gene2507/gene2962/gene2051/gene270/gene690/gene1515/gene528/gene317/gene3780/gene689/gene271/gene319/gene3750/gene3186/gene3739/gene3334/gene474/gene233/gene2459/gene318/gene3555/gene3744/gene1438/gene840/gene825/gene1429/gene1538/gene843/gene3654/gene1537/gene1536/gene3782	58
ko00330	Arginine and proline metabolism	11/654	16/1233	0.15505048	0.86891096	0.85525	gene3275/gene2914/gene2399/gene1057/gene1111/gene1877/gene656/gene2287/gene2663/gene2664/gene2658	11

ID:KEGGID Description 通路的名称 GeneRatio 在该 genelist 中该条目的占所有有 KEGG 注释的比例 BgRatio 在该物种所有有注释的基因中该条目所占的比值（即背景值） Fisher 精确检验的 p 值 p.adjust p 值的校正值 qvalue q 值 geneID 所找到的基因 ID

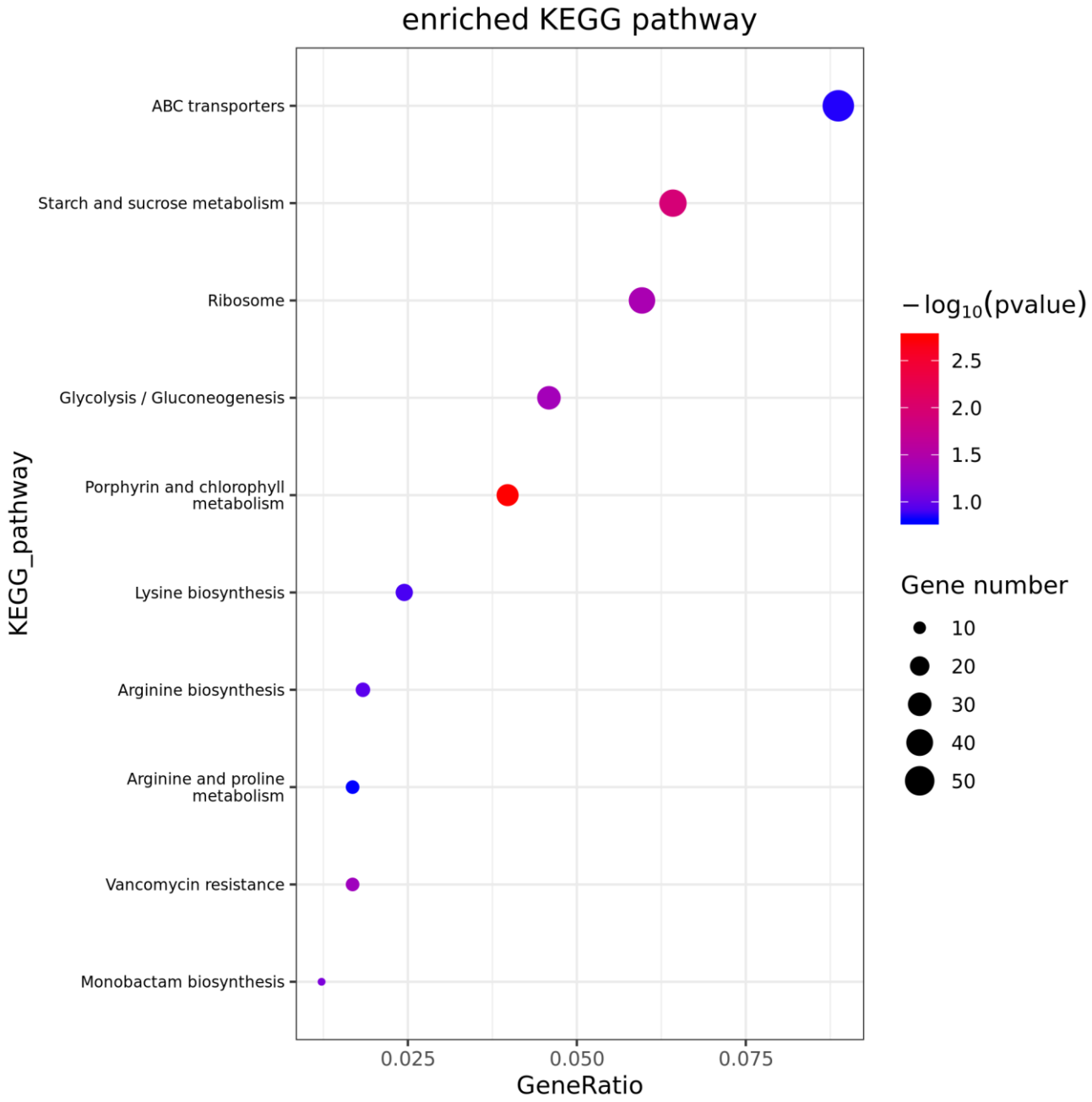
根据样品在该通路的富集程度 q 值做出分布图，结果如下：



KEGG_bar

图 5.4 富集通路 p.adjust 值分布图

取所有样品中富集的 KEGG 条目进行分析，纵坐标为 KEGG 的条目，横坐标为不同的样品名称，不同的颜色代表不同的富集程度。



KEGG_dot

图 5.5 KEGG 富集 p.adjust 值结果图

每个点表示该 KEGG 条目的富集程度，颜色越趋近于红色表示富集程度越高。每个点的大小表示富集到该 KEGG 条目的基因的个数，点越大表示富集到该 KEGG 条目的基因越多，反之则越少。

6. 附录

附 1 参考信息整理

参考数据库-GO

基因本体论联合会建立的数据库（Gene Ontology，http://geneontology.org/）。GO 的产生主要是为了解决同一基因在不同数据库定义的混乱性以及不同物种的同一基因在功能定义上的混乱性。它是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（Controlled Vocabulary）来全面描述生物体中基因和基因产物的属性。GO 涵盖三个方面，分别描述基因的分子功能（Molecular Function）、细胞的组件作用（Cellular Component）、参与的生物学过程（Biological Process）。基因或蛋白质可以通过 ID 对应或者序列注释的方法找到与之对应的 GO 编号，而 GO 编号可用于对应到 GO Term，即功能类别或者细胞定位。

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。GOSlim 是缩减版的 GO 术语，它提供了 GO 注释的概述性结果。

参考数据库-KEGG

京都基因与基因组百科全书（Kyoto Encyclopedia of Genes and Genomes，http://www.kegg.jp/）是一个整合了基因组、化学和系统功能信息的数据库。把从已经完整测序的基因组中得到的基因目录与更高级别的细胞、物种和生态系统水平的系统功能关联起来是 KEGG 数据库的特色之一。KEGG 注释主要包括：（1）KO （KEGG Ortholog）注释，即将分子网络的相关信息进行跨物种注释； （2）KEGG Pathway 注释，即代谢通路注释，获得物种内分子间相互作用和反应的网络。

附 2 所用软件介绍

表 1 所用软件介绍

软件	功能	参数
Cutadapt	数据过滤	至少 10 bp Overlap（AGATCGGAAG），允许 20% 的碱基错误率
FastQC	质量控制	默认参数
RSeQC	FPKM 饱和度分析	默认参数
DEGseq	差异分析	|log2foldchang|≥$log_2$2 和 qvalue≤0.001
ggplot2	绘制火山图、MA 图	默认参数
Pheatmap	聚类分析	默认参数
clusterprofiler	GO 和 KEGG 富集分析	默认参数
RSEM	转录组表达定量	默认参数

附 3 参考文献

[1] 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, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes[J]. BMC Bioinformatics. 2003 Sep 11,4:41.

[2] Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges[J]. Nucleic Acids Res. Epub 2011 Nov 16; PubMed 22096231.

[3] The Gene Ontology Consortium, Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology[J]. Nat Genet. 2000 May, 25（1）: 25－29.

[4] Minoru Kanehisa,* Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome[J]. Nucleic Acids Res. 2004 January 1; 32（Database issue）: D277－D280.

[5] Michael I Love,Wolfgang Huber,Simon Anders.(2014).Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.Genome Biology.

[6] 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.

[7] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids researpartA_ch16:D480－484.

[8] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology 2013, 31(1):46-53.

[9] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. Journal of molecular biology 1990, 215(3):403-410.

[10] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: BLAST/: architecture and applications. BMC bioinformatics 2009, 10:421.

[11] Conesa A, Gotz S: Blast2GO: A comprehensive suite for functional analysis in plant genomics. International journal of plant genomics 2008, 2008:619832.

[12] Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research 2008, 36(10):3420-3435.

[13] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research 2007, 35(Web Server issue):W182-185.

附 4 常用术语

名词解释

- **转录本 / Transcript：**是由一条基因通过转录形成的一种或多种可供编码蛋白质的成熟的 mRNA。
- **Read / Reads：**测序中每一条序列称为一个 Read。
- **Raw Data / Raw Reads：**测序下机的原始数据。
- **Clean Data / Clean Reads：**去除接头和低质量 Reads 后的数据，后续分析均基于 Clean Data。
- **接头/ Adapter：**接头是测序时在序列两端分别加上的一段人工序列，接头上含有与测序引物互补结合的序列，通过和测序引物结合来对目的片段进行测序。当加上接头后的序列片段比实际测序读长短时， 3'端会测到接头序列，接头序列在分析之前需要去除掉。
- **模糊碱基/ N：**测序中不能确定的碱基，以 N 表示。一条序列中 N 越多说明该序列质量越低，一般该种序列需要剔除掉。
- **log2FoldChange：**表达倍数差异，同一基因在两个样品中的表达量之商对 2 取对数，即 log2（sampleA/sampleB）。
- **P-value：**显著性，统计学根据显著性检验方法所得到的 P 值，一般以 P < 0.05 为显著， P <0.01 为非常显著，其含义是样本间的差异由抽样误差所致的概率小于 0.05 或 0.01。
- **cSNP：**SNP（Single Nucleotide Polymorphisms）是指在基因组上由单个核苷酸变异形成的遗传标记，其数量很多，多态性丰富。cSNP 是指在编码区出现的 SNP，这些 SNP 直接影响到氨基酸密码子。
- **SNP 转换：**嘧啶变成嘧啶或嘌呤变成嘌呤，即 A、G 互换，T、C 互换。
- **SNP 颠换：**嘧啶突变成嘌呤或者相反，即 A、T 互换，A、C 互换，G、T 互换，G、C 互换。
- **InDel：**Insertion-Deletion，指相对于参考基因组，样本中发生的小片段的插入或者缺失，该插入缺失可能含有一个或多个碱基。InDel 可作为一种基因标记用于研究系统进化或物种鉴定。

FASTQ 格式

FASTQ 格式（http://en.wikipedia.org/wiki/FASTQ_format）是一种文本格式，常用于存储生物学序列及其对应的质量分值。FASTQ 格式文件可以采用文本编辑软件（如写字板、UltraEdit 、EditPlus 等工具）打开，由于文件较大，对电脑的内存要求较高。FASTQ 格式中，每个 Read 由四行信息表示。

第一行为序列名称，以 @ 开头，其后是序列描述；第二行为碱基序列；第三行为 “+” 号，不代表任何意义；第四行碱基质量，与第二行的碱基序列一一对应。示例如下：

```
@M00200:111:000000000-A6VNV:1:1101:15594:1337 1:N:06
        ACGCGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGCGCCTCAGTGTCAGTTAC
        +
        ABABADBBDFFFGGGFGGGFGGHGBGHGGHGGGGGGGHGGGGGGGHHGGFBGEGGEG
```

质量值

我们使用 Sanger 质量值来评估下机数据的测序质量。质量值，简称 Q 值，是碱基读取错误率 p 的取整映射结果，等于 Phred quality score，计算公式为：

Phred quality score 计算公式

测序错误率与 Q 值的简明对应方式如**表 1** 所示。

表 1 错误率与 Q 值对应关系

测序错误率	**Q 值**
5%	13
1%	20
0.1%	30
0.01%	40

不同的测序平台，采用不同的方案对 FASTQ 文件中的碱基进行质量编码， Q 值与碱基质量的对应关系为：Q 值加上一个偏移数值，得到的结果按照 ASCII 码值对照表（见**表 2**）转换成对应的字符，参考信息如下所示：

phred_range

我们的 FASTQ 文件采用 Illumina 1.8+ 版本编码，将所有字符的 ASCII 值减去偏移值 33，即可得到碱基的 Q 值。例如，字符 I 的 ASCII 值为 73，减去 33 后得到 40，那么该字符对应位置的碱基质量为 40，测序错误率则为 0.01%。

表 2 ASCII 码表

十进制	字符	Q 值	十进制	字符	Q 值	十进制	字符	Q 值	十进制	字符	Q 值
32			48	0	15	64	@	31	80	P	47
33	!	0	49	1	16	65	A	32	81	Q	48
34	“	1	50	2	17	66	B	33	82	R	49
35	#	2	51	3	18	67	C	34	83	S	50
36		3	52	4	19	68	D	35	84	T	51
37	%	4	53	5	20	69	E	36	85	U	52
38	&	5	54	6	21	70	F	37	86	V	53
39	’	6	55	7	22	71	G	38	87	W	54
40	(	7	56	8	23	72	H	39	88	X	55
41	）	8	57	9	24	73	I	40	89	Y	56
42	*	9	58	:	25	74	J	41	90	Z	57
43	+	10	59	:	26	75	K	42	91	[	58
44	,	11	60	<	27	76	L	43	92		59
45	-	12	61	=	28	77	M	44	93	]	60
46	.	13	62	>	29	78	N	45	94	^	61
47	/	14	63	?	30	79	O	46	95	_	62

Sam / Bam 格式

Sam（sequence alignment/map format）是一种由 Sanger 制定的序列比对格式标准，以 Tab 为分割符的文本格式，可用文本编辑软件打开（如写字板、UltraEdit 、EditPlus 等工具），主要应用于测序序列比对到基因组上的结果表示，当然也可以表示任意的多重比对结果。当把 fastq 文件比对到基因组上之后，我们通常会得到一个 Sam 或者 Bam 为扩展名的文件。而

Bam 就是 Sam 的二进制文件(B 取自 binary)，占用空间更小，不可打开，只能用 samtools 等软件转换为 Sam 格式后打开。

Sam 分为两部分，注释信息（header section）和比对结果（alignment section）。注释信息可有可无，每一行都是以@开头，用不同的 tag 表示不同的信息，tag 包括 @HD（符合标准的版本、对比序列的排列顺序说明）、@SQ（参考序列说明）、@RG（比对上的序列说明）、@PG（使用的程序说明）、@CO（任意的说明信息）。比对结果部分的每一行表示一个片段（segment）的比对信息，包括 11 个必须的字段（mandatory fields）和一个可选的字段，字段之间用 Tab 分割。示例及介绍如下：

Sam 格式示例：

```
HWI-ST1001:137:C12FPACXX:7:1115:14131:66670 0 chr1 12805 1 42M4I5M * 0
OTTGGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCACCAATATG
CCCFFFFFHHGHHJJJJJHJJJJJJJJJJJJJJJJIJJJJJJJJJJJJIJJAS:i:-28 XN:i:0 XM:i:2 XO:i:1
XG:i:4 NM:i:6 MD:Z:2C41C2 YT:Z:UU NH:i:3 CC:Z:chr15 CP:i:102518319 XS:A:+ HI:i:0
```

1. QNAME，比对片段的编号
2. FLAG，位标识，1 表示该 read 是 pair 中的一条（read 表示本条 read，mate 表示 pair 中的另一条 read ），2 表示 pair 一正一负地比对上参考序列，4 表示这条 read 没有比对上，8 表示 mate 没有比对上，16 表示这条 read 比对上负链，32 表示 mate 比对上负链，64 表示这条 read 是 read1，128 表示这条 read 是 read2 等，FLAG 的值是符合情况的数字相加总和，即 83=（64+16+2+1）表示该 read 为 read1，比对到负链上，其 mate 比对到正链上
3. RNAME，参考序列的编号
4. POS，比对上的位置，注意是从 1 开始计数，如果没有比对上，此处为 0
5. MAPQ，比对的质量，越高则位点越独特，计算方法：Q = -10 log 10p，p 是该序列不来自这个位点的估计值
6. CIGAR（Compact Idiosyncratic Gapped Alignment Report），使用数字加字母表示比对结果，如 M 表示 match/mismatch，I 表示 insertion，D 表示 deletion 等，数字表示碱基个数，即 42M4I5M 为该序列 42 个碱基匹配，4 个 insertion，5 个碱基匹配
7. RNEXT，mate 的名称，如果没有 mate，用 * 表示
8. PNEXT，mate 的位置，如果没有 mate，用 0 表示
9. TLEN， paired reads 间的距离，当 mate 序列位于本序列上游时该值为负值，如果比对区域仅有一个区段，或者不可用时，此处为 0 10.SEQ，read 序列 11.QUAL，read 质量 12.Optional Fields，可选字段，格式如：TAG:TYPE:VALUE，其中 TAG 由两个大写字母组成，每个 TAG 代表一类信息，如 AS 表示匹配的得分，XS 表示第二好的匹配得分，YS 表示 mate 序列匹配的得分等，TYPE 表示 TAG 对应值的类型，可以是字符串（Z）、整数（i）等

RPKM / FPKM

不同样品过滤后获得的数据量不可能完全一致，不同基因长度也有很大差异。为了能够在样品内（不同基因）以及样品间（不同分组）比较基因的表达量， 需要采用 RPKM 对表达量进行标准化（Normalization）。

RPKM（Reads Per Kilo bases per Million reads），为每百万 Reads 中来自某一基因每千碱基长度的 Reads 数目，是一种普遍采用的基因表达量标准化方法，这种方法同时考虑了测序深度和基因长度对基因表达量计数的影响。其计算公式如下所示：

RPKM

RPKM 计算公式

目前以 FPKM 为基准的表达量标准化方法逐渐被其他统计学方法所取代，但是作为一种绝对化的标准化方法，其生物学意义明确，利于不同项目之间比较。在有参转录组当中， 我们一般认为 FPKM>1 的基因是表达的。这个阈值是主流杂志推荐的， 也能够很好的反应基因的表达水平。

此外，也可通过 FPKM 值描述基因表达量，FPKM 与 RPKM 计算方法基本一致，不同点在于：对于 Pair-End 测序，每个 Fragments 会有两个 Reads，FPKM 只计算两个 Reads 能比对到同一个转录本的 Fragments 数量， RPKM 计算是比对到转录本的 Reads 数量。

GFF /GTF 格式

gff 格式是一种 Sanger 研究所定义的，可以简单方便地描述 DNA、RNA 以及蛋白质序列的特征的数据格式，已经成为序列注释的通用格式，许多软件都支持输入或者输出 gff 格式。每一行代表一个特征条目（如基因、转录本、CDS、exon 等），每行有 9 列，以 Tab 为分割符，每列分别列出该特征条目的一些信息。 gff 可用文本编辑软件打开（如写字板、UltraEdit 、EditPlus 等工具）。目前 gff 的最新版本是版本 3。示例如下：

```
##gff-version 3
        ##sequence-region ctg123 1 1497228
        ctg123 PFAM gene 1000 5000 . + . ID=gene001;Name=EDEN
        ctg123 PFAM TF_binding_site 1000 1012 . + . Parent=gene001
        ctg123 PFAM mRNA 1050 5000 . + . ID=mRNA001;Parent=gene001
        ctg123 PFAM mRNA 1050 5000 . + . ID=mRNA002;Parent=gene001
        ctg123 PFAM exon 1300 1500 . + . Parent=mRNA001
        ctg123 PFAM exon 1050 1500 . + . Parent=mRNA001,mRNA002
        ctg123 PFAM CDS 1201 3902 . + 0 ID=cds001;Parent=mRNA001
        ctg123 PFAM CDS 3000 4600 . + 2 ID=cds001;Parent=mRNA001
        ctg123 PFAM CDS 1201 1500 . + 1 ID=cds002;Parent=mRNA002
```

1. 序列编号，可能是染色体或者 scaffold 的名称
2. 来源，产生这一特征条目的程序、数据库或者项目
3. 类型，如 gene，transcript，CDS，mRNA，exon，five/three_prime_utr，start/stop_codon 等
4. 起始位点，这一特征条目在序列上的起始位置，从 1 开始计数
5. 终止位点，这一特征条目在序列上的终止位置，不能大于序列的长度
6. 得分，是注释信息可能性的说明，可以是序列相似性比对时的 E-values 值或者基因预测是的 P-values 值。“.”表示为空
7. 序列的方向， + 表示正义链,- 反义链 ,? 表示未知

8. 相位，仅对类型为“CDS”的条目有效，有效值为 0、1、2，0 表示这一特征条目的第一个碱基是一个密码子的第一个碱基，1 表示这一特征条目的第二个碱基是一个密码子的第一个碱基，以此类推
9. 属性，以多个键值对组成的注释信息描述，键与值之间用“=”，不同的键值对用“;”隔开，一个键可以有多个值，不同值用“,”分割。键可以为 ID（该特征条目的编号，在一个 gff 文件中必须唯一），Name（该特征条目的名称，可以重复），Parent（该特征条目的父级特征条目，值为父级特征条目的编号，比如外显子所属的转录本编号，转录本所属的基因的编号。值可以为多个）等

gtf 格式与 gff 格式前 8 列基本相同，不同之处在于第 9 列，虽然同样是标签与值配对的情况，但 gtf 格式的标签与值之间以空格分开，且每个属性之后都要有分号;（包括最后一个属性），而且第 9 列必须以 gene_id 以及 transcript_id 开头。

FASTA 格式

在生物信息学中，FASTA 格式（又称为 Pearson 格式），是一种基于文本用于表示核苷酸序列或氨基酸序列的格式，可用文本编辑软件打开（如写字板、UltraEdit 、EditPlus 等工具）。序列文件的第一行是由大于号“>”或分号“;”打头的任意文字说明（习惯常用“>”作为起始），用于序列标记。从第二行开始为序列本身，只允许使用既定的核苷酸或氨基酸编码符号。通常核苷酸符号大小写均可，而氨基酸常用大写字母。文件每行的字母一般不应超过 80 个字符。示例如下：

```
>Seq1
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDAD*
```