Long noncoding RNA 生物信息分析

结题报告

2013 年 9 月

RNA 研究部

rl@novogene.cn

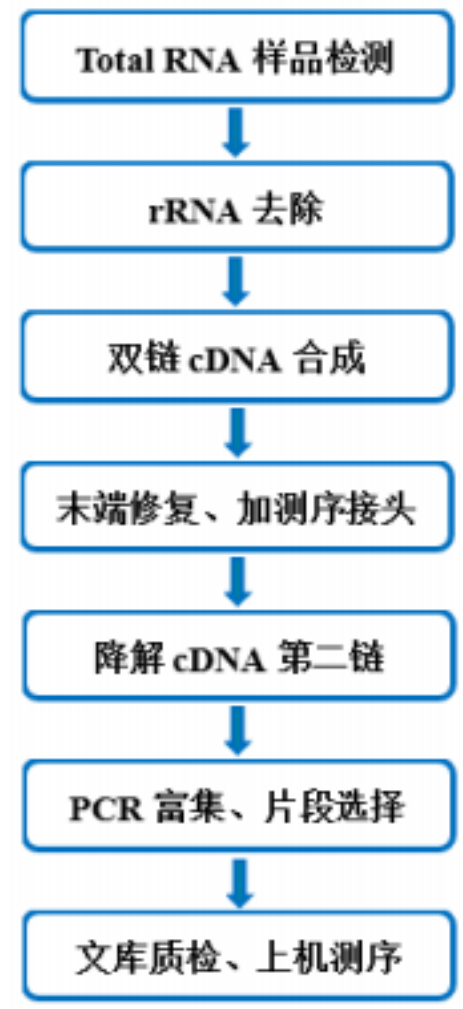
北京诺禾致源生物信息科技有限公司

地址： 北京市海淀区学清路 38 号金码大厦 B 座 21 层 网址： www.novogene.cn 电话： 010-8283 7567

lncRNA 生物信息分析结题报告

一、建库测序流程

从 RNA样品到最终数据获得，样品检测、建库、测序每一个环节都会对数据质量和数量产生影响，而数据质量又会直接影响后续信息分析的结果。因此，获得高质量数据是保证生物信息分析正确、全面、可信的前提。为了从源头上保证测序数据的准确性、可靠性，诺禾致源对样品检测、建库、测序每一个生产步骤都严格把控，从根本上确保了高质量数据的产出。实验流程图如下：



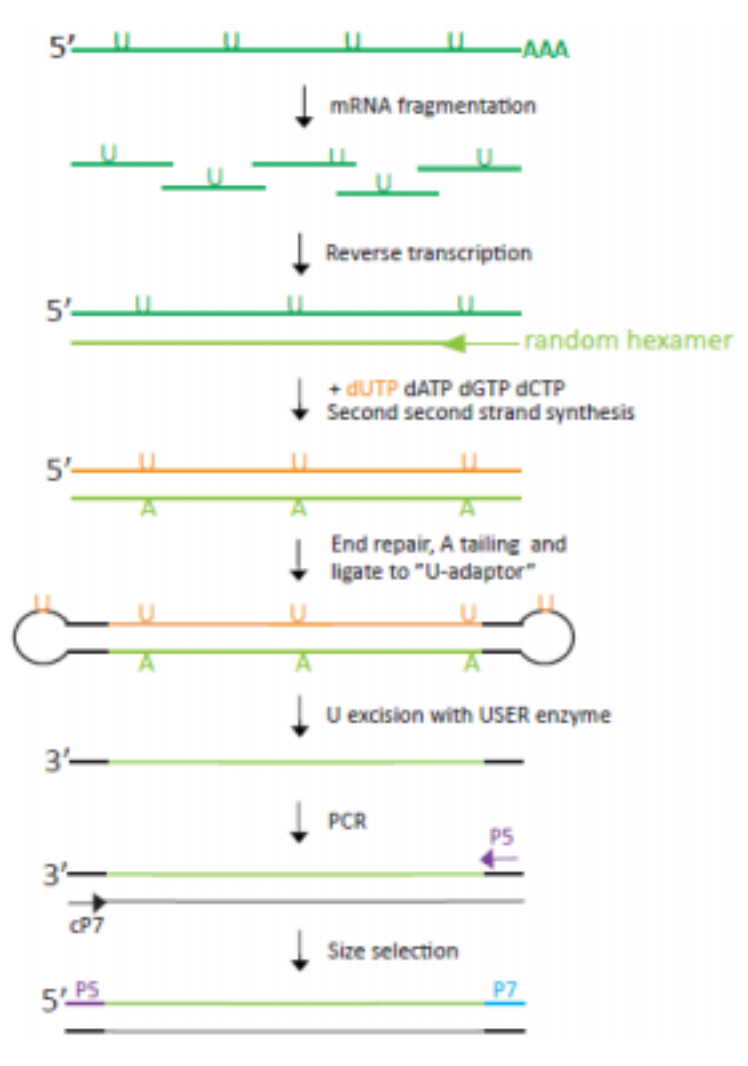
1 Total RNA 样品检测

诺禾致源对 RNA 样品的检测主要包括 4 种方法：

(1) 琼脂糖凝胶电泳分析 RNA 降解程度以及是否有污染
(2) Nanodrop 检测 RNA 的纯度（ OD260/280 比值）
(3) Qubit 对 RNA 浓度进行精确定量
(4) Agilent 2100 精确检测 RNA 的完整性

2 文库构建

RNA 检测合格后，通过 epicentre Ribo-ZeroTM 试剂盒去除 rRNA，随后加入 fragmentation buffer 将 RNA 打断成短片段，以短片段 RNA 为模板，用六碱基随机引物（ random hexamers）合成一链 cDNA，然后加入缓冲液、 dNTPs（dUTP，dATP，dGTP 和 dCTP）和 DNA polymerase I 合成二链 cDNA，随后利用 AMPure XP beads 纯化双链 cDNA。纯化的双链 cDNA 再进行末端修复、加 A 尾并连接测序接头，然后用 AMPure XP beads 进行片段大小选择。之后用 USER 酶降解含有 U 的 cDNA 第一链，最后进行 PCR 富集得到链特异性 cDNA 文库。文库构建原理图如下：



3 库检

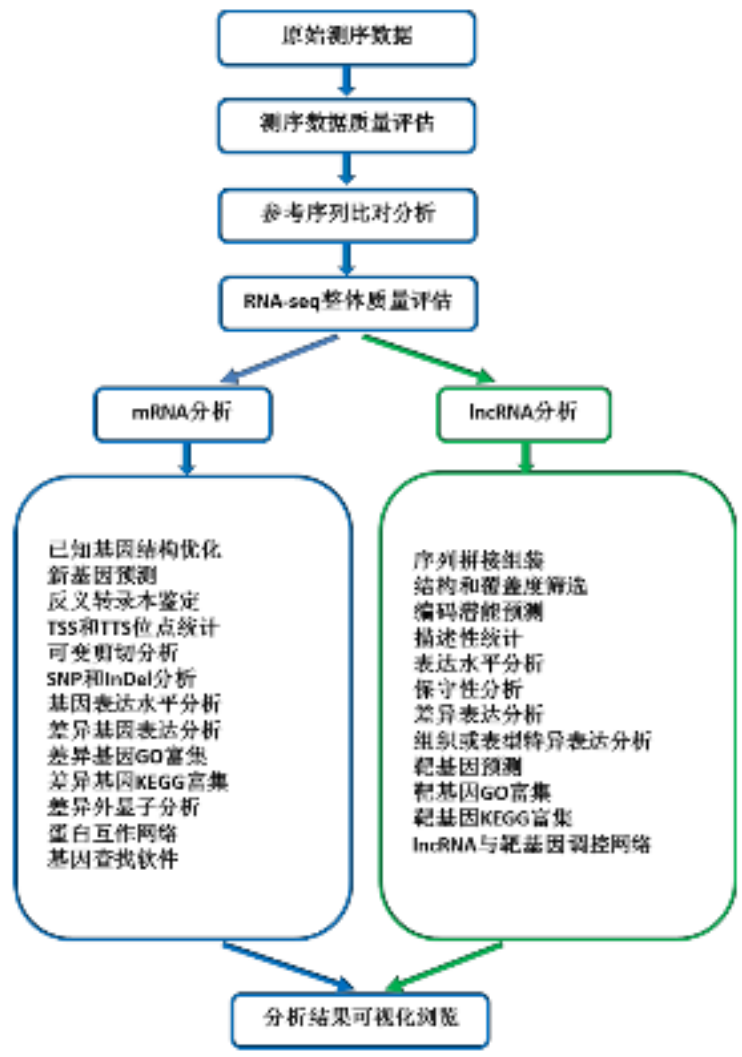
文库构建完成后，先使用 Qubit2.0 进行初步定量，稀释文库至 1ng/ul ，随后使用 Agilent 2100 对文库的 insert size 进行检测， insert size 符合预期后，使用 Q-PCR 方法对文库的有效浓度进行准确定量（文库有效浓度 ＞2nM），以保证文库质量。

4 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 HiSeq/MiSeq 测序。

二、生物信息分析流程

获得原始测序序列（Sequenced Reads）后，在有相关物种参考序列或参考基因组的情况下，通过如下流程进行生物信息分析：



三、项目结果说明

1 原始序列数据

高通量测序 (如illumina HiSeq ™2000/MiSeq等测序平台) 测序得到的原始图像数据文件经碱基识别 (Base Calling) 分析转化为原始测序序列 (Sequenced Reads) ，我们称之为 Raw Data或Raw Reads，结果以 FASTQ(简称为 fq) 文件格式存储，其中包含测序序列 (reads) 的序列信息以及其对应的测序质量信息。

FASTQ格式文件中每个 read由四行描述，如下：

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT
+
@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF
```

其中第一行以“ @”开头，随后为 illumina 测序标识符 (Sequence Identifiers) 和描述文字 (选择性部分) ；第二行是碱基序列；第三行以“ +”开头，随后为 illumina 测序标识符 (选择性部分) ；第四行是对应序列的测序质量 (Cock et al.) 。

illumina 测序标识符详细信息如下：

EAS139	Unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane
2104	Tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	Index sequence

第四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。如果测序错误率用 e表示， illumina HiSeq ™2000/MiSeq的碱基质量值用Q_{phred} 表示，则有下列关系：

公式一： $Q_{phred} = -10\log_{10}(e)$

illumina Casava 1.8 版本测序错误率与测序质量值简明对应关系如下：

测序错误率	测序质量值	对应字符
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

2 测序数据质量评估

2.1 测序错误率分布检查

每个碱基测序错误率是通过测序 Phred数值(Phred score, Q_{phred})通过公式 1转化得到，而 Phred 数值是在碱基识别 (Base Calling) 过程中通过一种预测碱基判别发生错误概率模型计算得到的，对应关系如下表所显示：

illumina Casava 1.8 版本碱基识别与 Phred分值之间的简明对应关系

Phred 分值	不正确的碱基识别	碱基正确识别率	Q-sorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

测序错误率与碱基质量有关，受测序仪本身、测序试剂、样品等多个因素共同影响。对于 RNA-seq技术，测序错误率分布具有两个特点： (1) 测序错误率会随着测序序列 (Sequenced Reads) 长度的增加而升高，这是由于测序过程中化学试剂的消耗而导致的，并且为 illumina 高通量测序平台都具有的特征。 (2) 前6个碱基的位置也会发生较高的测序错误率，而这个长度也正好等于在 RNA-seq建库过程中反转录所需要的随机引物的长度。所以推测前 6个碱基测序错误率较高的原因为随机引物和 RNA模版的不完全结合 (Jiang et al.) 。

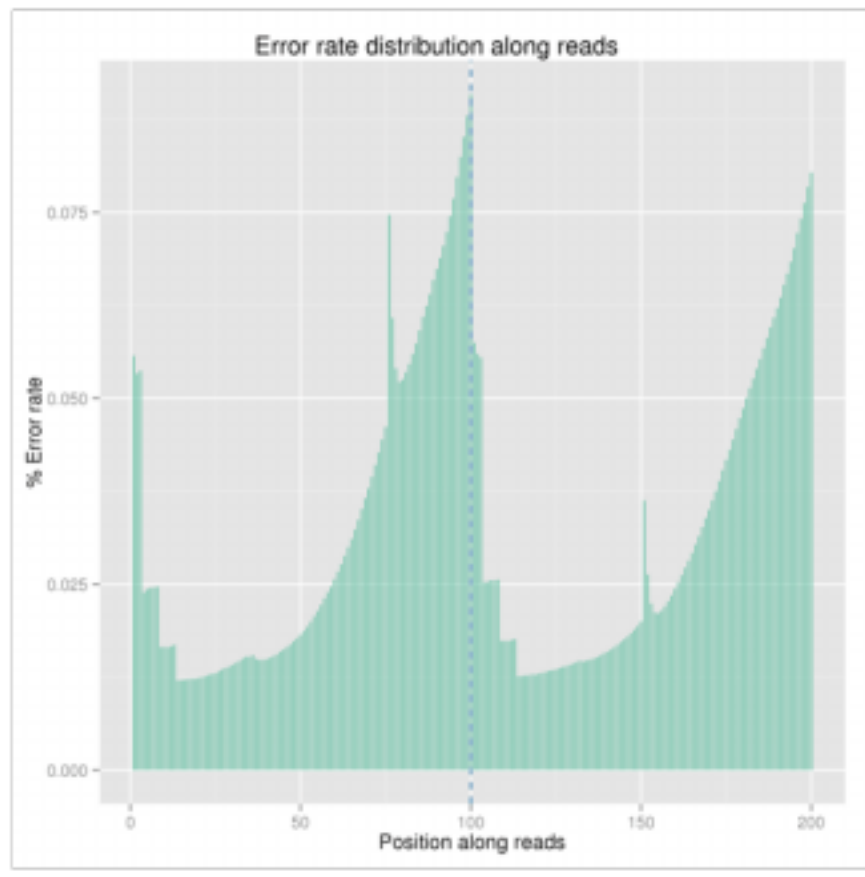


图2.1 测序错误率分布图

横坐标为 reads 的碱基位置，纵坐标为单碱基错误率

2.2 GC含量分布检查

GC含量分布检查用于检测有无 AT、GC分离现象，而这种现象可能是测序或者建库所带来的，并且会影响后续的定量分析。

在illumina 测序平台的转录组测序中，反转录成 cDNA时所用的 6bp 的随机引物会引起前几个位置的核苷酸组成存在一定的偏好性。而这种偏好性与测序的物种和实验室环境无关，但会影响转录组测序的均一化程度 (Hansen et al.) 。除此之外，理论上 G和C碱基及A和T碱基含量每个测序循环上应分别相等，且整个测序过程稳定不变，呈水平线。对于 DGE测序来说，由于随机引物扩增偏差等原因，常常会导致在测序得到的每个read前6-7个碱基有较大的波动，这种波动属于正常情况。

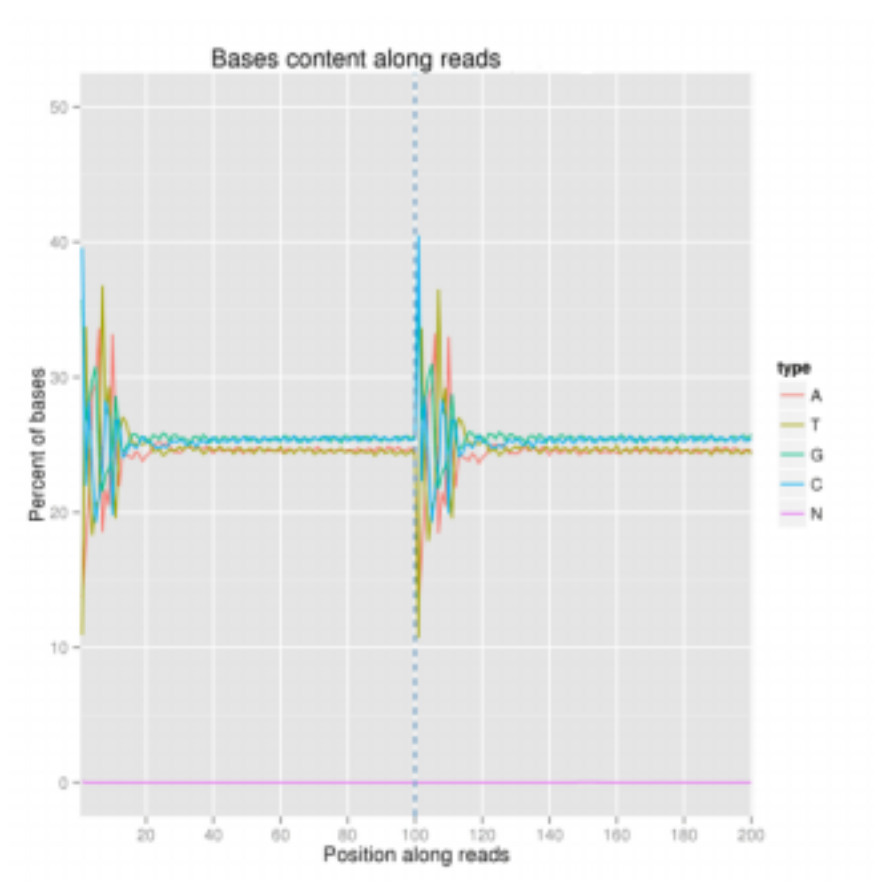


图2.2 GC含量分布图

横坐标为 reads 的碱基位置，纵坐标为单碱基所占的比例；不同颜色代表不同的碱基类型

2.3 测序数据过滤

测序得到的原始测序序列，里面含有带接头的、低质量的 reads，为了保证信息分析质量，必须对 raw reads 进行过滤，得到 clean reads ，后续分析都基于 clean reads 。

数据处理的步骤如下：

(1) 去除带接头 (adapter) 的reads ；
(2) 去除N(N表示无法确定碱基信息) 的比例大于 10%的reads ；
(3) 去除低质量 reads。

RNA-seq 的接头 (Adapter, Oligonucleotide sequences for TruSeq ™RNA and DNA Sample Prep Kits) 信息：

RNA 5’ Adapter (RA5), part # 15013205 :
5 ’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3
RNA 3’ Adapter (RA3), part # 15013207 :
5 ’-GATCGGAAGAGCACACGTCTGAACTCCAGTCAXXXXXXATCTCGTATGCCGTCTTCTGCTTG-3
(Index)

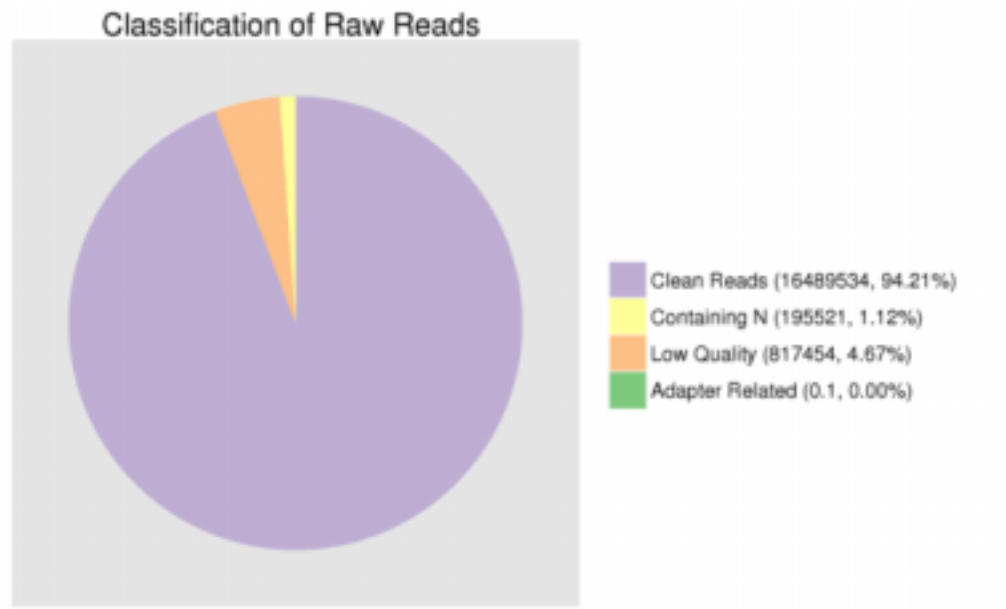


图2.3 原始数据过滤结果

2.4 数据产出情况汇总

表2.4 数据产出质量情况一览表

Sample name	Raw reads	Clean reads	clean bases	Error rate(%)	Q20(%)	Q30(%)	GC content(%)
sample1_A_1	39176275	37815985	3.78G	0.03	97.85	92.56	51.62
sample1_A_2	39176275	37815985	3.78G	0.04	96.58	90.47	52.34
sample1_B_1	35876315	34617593	3.46G	0.03	97.86	92.61	51.52
sample1_B_2	35876315	34617593	3.46G	0.04	96.53	90.51	52.00
sample1_C_1	37973817	36491666	3.65G	0.03	97.78	92.37	51.93
sample1_C_2	37973817	36491666	3.65G	0.04	96.24	89.92	52.60
sample2_A_1	40470350	38887996	3.89G	0.03	97.89	92.54	52.58
sample2_A_2	40470350	38887996	3.89G	0.04	96.54	90.29	53.39
sample2_B_1	35590714	34300840	3.43G	0.03	97.84	92.54	51.54
sample2_B_2	35590714	34300840	3.43G	0.04	96.56	90.57	52.09
sample2_C_1	43366207	41733549	4.17G	0.03	97.80	92.39	52.19
sample2_C_2	43366207	41733549	4.17G	0.04	96.33	90.06	52.84

数据质量情况详细内容如下：

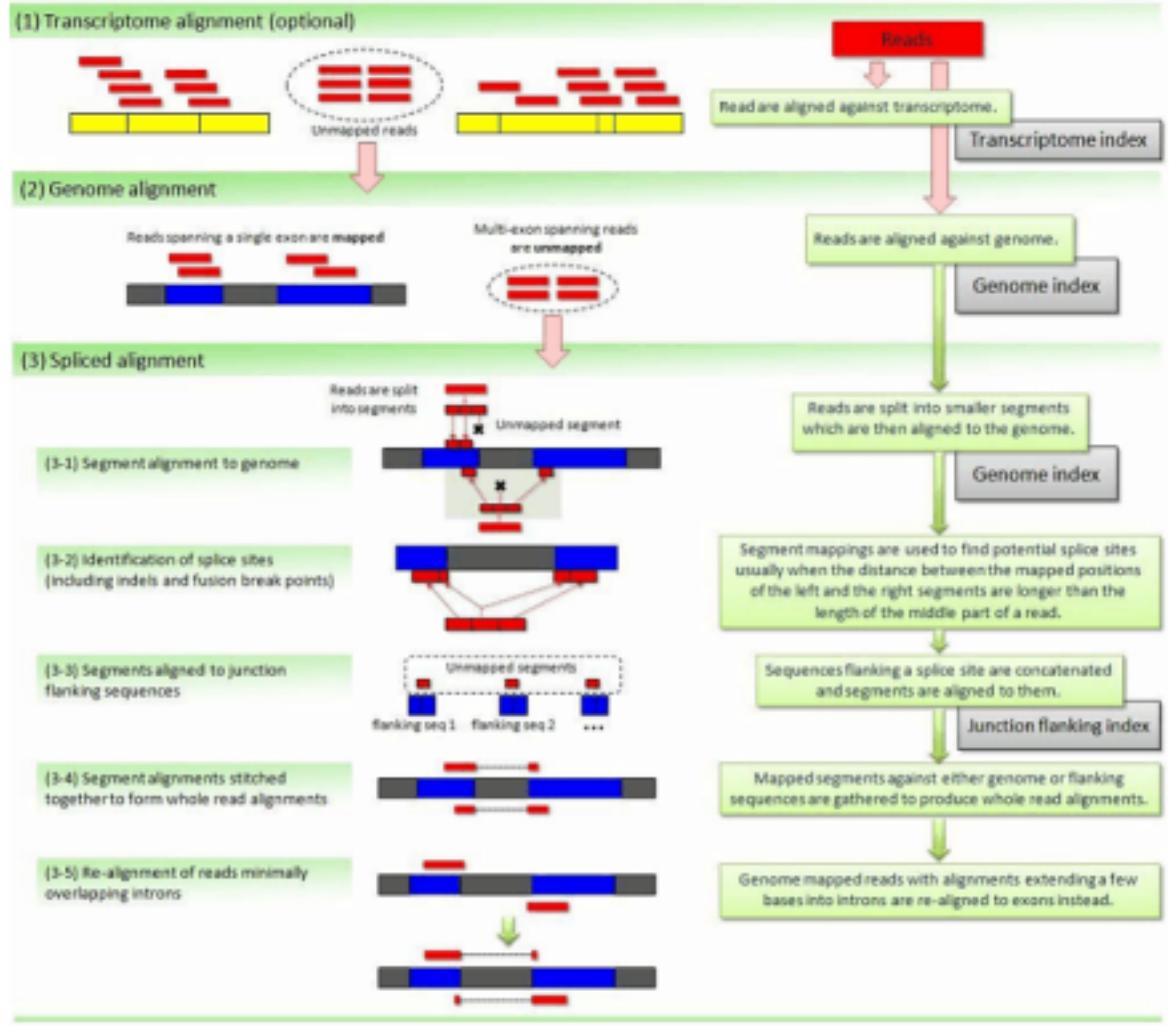
(1) Raw reads ：统计原始序列数据，以四行为一个单位，统计每个文件的测序序列的个数。
(2) Clean reads ：计算方法同 Raw Reads，只是统计的文件为过滤后的测序数据。后续的生物信息分析都是基于 Clean reads 。
(3) Clean bases ：测序序列的个数乘以测序序列的长度，并转化为以 G为单位。
(4) Error rate ：通过公式 1计算得到。
(5) Q20 、Q30：分别计算 Phred 数值大于 20、30的碱基占总体碱基的百分比。
(6) GC content ：计算碱基 G和C的数量总和占总的碱基数量的百分比。

3 参考序列比对分析

我们采用 Tophat2(Kim et al, 2013) 对过滤后的测序序列进行参考基因组的比对分析 .TopHat2 的算法主要分为三个部分：

(1) 将测序序列和转录组进行比对 (可选)
(2) 将测序序列整段比对到基因组外显子上
(3) 将测序序列分段比对到基因组的两个外显子上

下图为 TopHat2的算法示意图 (Kim et al, 2013) ：



TopHat2的算法主要分为三个部分：

(1) 将测序序列和转录组进行比对 (可选)
(2) 将测序序列整段比对到基因组外显子上
(3) 将测序序列分段比对到基因组的两个外显子上

如果参考基因组选择合适，而且相关实验不存在污染，实验所产生的测序序列的定位的百分比正常情况下会高于 70% (Total Mapped Reads or Fragments) ，其中具有多个定位的测序序列 (Multiple Mapped Reads or Fragments) 占总体的百分比通常不会超过 10%。

3.1 Reads与参考基因组比对情况统计

表3.1 Reads与参考基因组比对情况一览表

Sample name	sample1_A	sample1_B	sample1_C	sample2_A	sample2_B	sample2_C
Total reads	75631970	69235186	72983332	77775992	68601680	83467098
Total mapped	61696177 (81.57%)	55550679 (80.23%)	60199063 (82.48%)	64037897 (82.34%)	54922912 (80.06%)	67922044 (81.38%)
Multiple mapped	13545289 (17.91%)	10352982 (14.95%)	12735274 (17.45%)	17911638 (23.03%)	9822139 (14.32%)	13648290 (16.35%)
Uniquely mapped	48150888 (63.66%)	45197697 (65.28%)	47463789 (65.03%)	46126259 (59.31%)	45100773 (65.74%)	54273754 (65.02%)
Read-1	24392816 (32.25%)	22926103 (33.11%)	24117085 (33.04%)	23409920 (30.1%)	22875927 (33.35%)	27536787 (32.99%)
Read-2	23758072 (31.41%)	22271594 (32.17%)	23346704 (31.99%)	22716339 (29.21%)	22224846 (32.4%)	26736967 (32.03%)
Reads map to '+'	24031121 (31.77%)	22560794 (32.59%)	23673696 (32.44%)	23005913 (29.58%)	22485859 (32.78%)	27060491 (32.42%)
Reads map to '-'	24119767 (31.89%)	22636903 (32.7%)	23790093 (32.6%)	23120346 (29.73%)	22614914 (32.97%)	27213263 (32.6%)
Non-splice reads	34011691 (44.97%)	32712405 (47.25%)	34860803 (47.77%)	32180604 (41.38%)	32703491 (47.67%)	38273075 (45.85%)
Splice reads	14139197 (18.69%)	12485292 (18.03%)	12602986 (17.27%)	13945655 (17.93%)	12397282 (18.07%)	16000679 (19.17%)
Reads mapped in proper pairs	42068616 (55.62%)	39542002 (57.11%)	41329630 (56.63%)	41295514 (53.1%)	39636200 (57.78%)	47555248 (56.97%)
Proper-paired reads map to different chrom	526 (0%)	520 (0%)	582 (0%)	450 (0%)	504 (0%)	632 (0%)

比对结果统计详细内容如下：

(1) Total reads ：测序序列经过测序数据过滤后的数量统计 (Clean data) 。

(2) Total mapped ：能定位到基因组上的测序序列的数量的统计；一般情况下，如果不存在污染并且参考基因组选择合适的情况下，这部分数据的百分比大于 70%。

(3) Multiple mapped ：在参考序列上有多个比对位置的测序序列的数量统计；这部分数据的百分比一般会小于 10%。

(4) Uniquely mapped ：在参考序列上有唯一比对位置的测序序列的数量统计。

(5) Reads map to '+' ， Reads map to '-' ：测序序列比对到基因组上正链和负链的统计。

(6) Splice reads ：(2) 中，分段比对到两个外显子上的测序序列 (也称为 Junction reads) 的统计， Non-splice reads 为整段比对到外显子的将测序序列的统计， Splice reads 的百分比取决于测序片段的长度。

3.2 Reads在参考基因组不同区域的分布情况

对Total mapped reads 的比对到基因组上的各个部分的情况进行统计，定位区域分为 Exon(外显子)、Intron(内含子)和Intergenic(基因间隔区域)。

正常情况下， Exon (外显子) 区域的测序序列定位的百分比含量应该最高，定位到 Intron (内含子) 区域的测序序列可能是由于非成熟的mRNA的污染或者基因组注释不完全导致的，而定位到 Intergenic(基因间隔区域) 的测序序列可能是因为基因组注释不完全以及背景噪音。

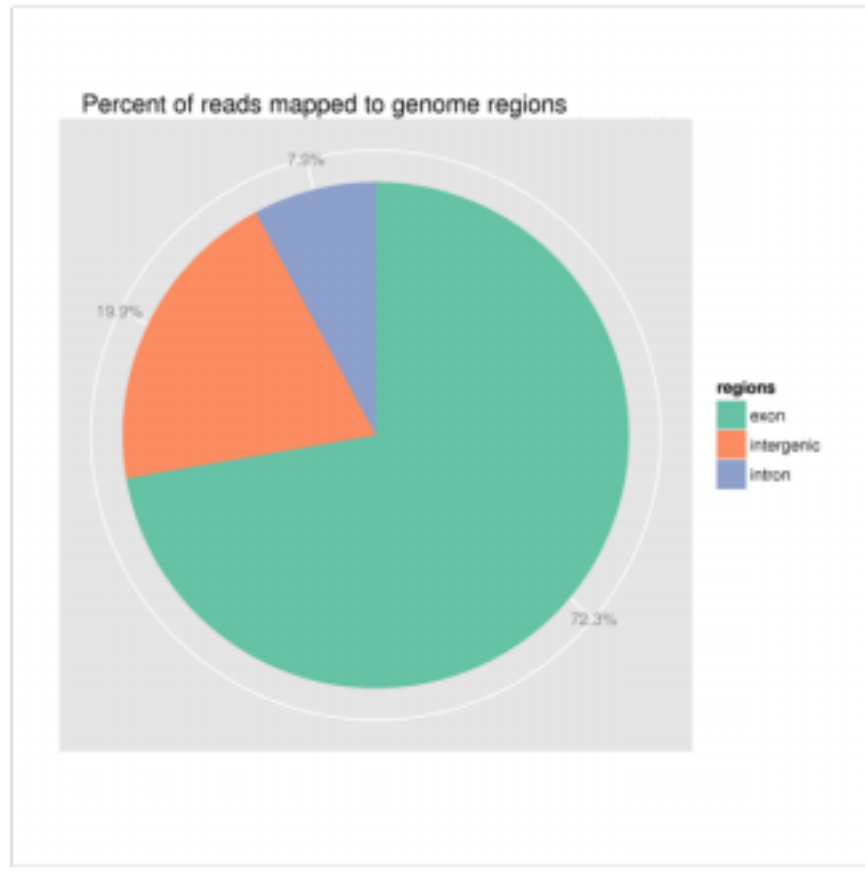
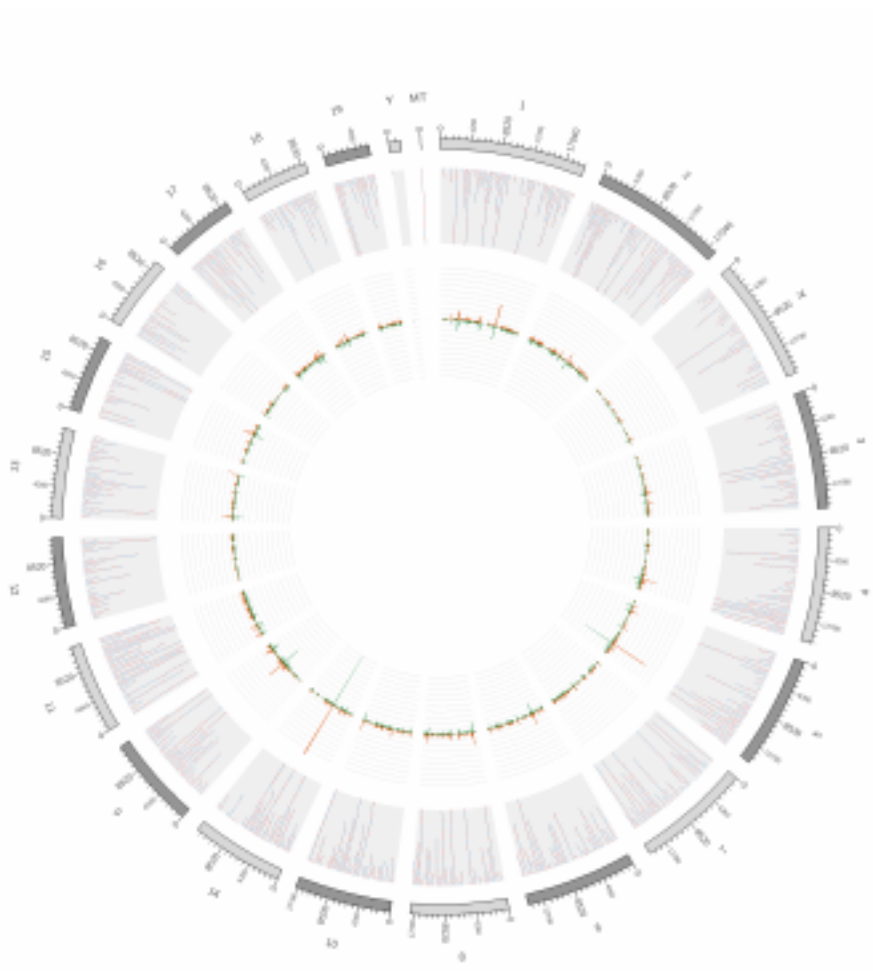


图3.2 Reads在参考基因组不同区域的分布情况

3.3 Reads在染色体上的密度分布情况

对Total mapped reads 的比对到基因组上的各个染色体（分正负链）的密度进行统计，如下图所示，具体作图的方法为用滑动窗口 (window size) 为1K，计算窗口内部比对到碱基位置上的 reads 的中位数，并转化成 $\log_2$。正常情况下，整个染色体长度越长，该染色体内部定位的 reads 总数会越多 (Marquez et al.) 。从定位到染色体上的 reads 数与染色体长度的关系图中，可以更加直观看出染色体长度和 reads 总数的关系。



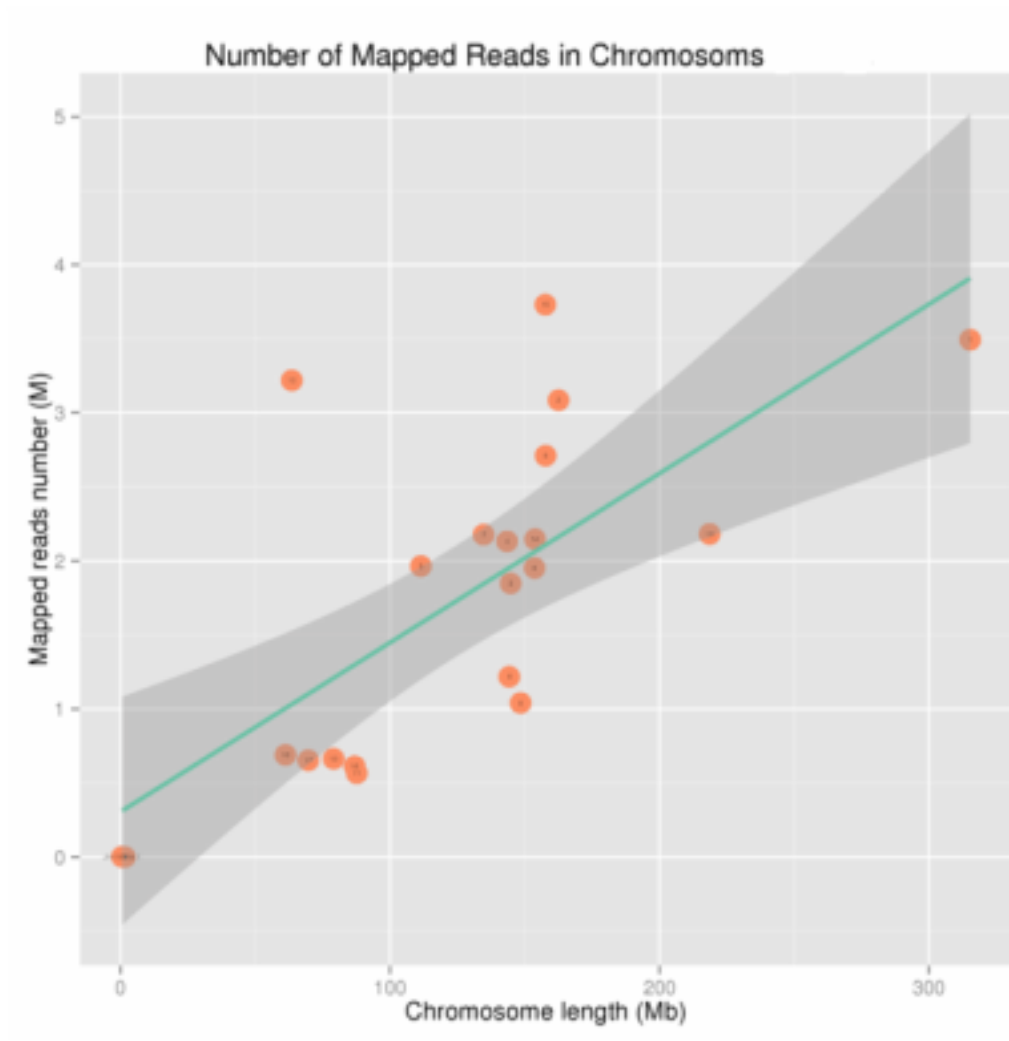


图3.3 Reads在染色体上的密度分布图

左图：图中最外圈是选择展示的各条染色体；中间的灰色背景区是抽取其中了 10000条reads 的分布情况，红色 mapping到正链，蓝色到负链；最里面的圆圈区是比对到该染色体上的所有 reads，橘黄色为正链 coverage 分布，绿色为负链 coverage 分布，超过所有 coverage 集均值+3倍标准差的奇异点将被舍弃。右图：横坐标为染色体的长度信息 (单位为Mb)，纵坐标为 mapped到染色体上的 reads 数(单位为M)，图中灰色区域表示 95%的置信区间

3.4 Reads比对结果 IGV可视化浏览

我们提供 RNA-seq Reads在基因组上比对结果的 bam格式文件，部分物种还提供相应的参考基因组和注释文件，并推荐使用 IGV (Integrative Genomics Viewer) 浏览器对 bam文件进行可视化浏览。 IGV浏览器具有以下特点： (1) 能在不同尺度下显示单个或多个读段在基因组上的位置，包括读段在各个染色体上的分布情况和在注释的外显子、内含子、剪接接合区、基因间区的分布情况等； (2) 能在不同尺度下显示不同区域的读段丰度，以反映不同区域的转录水平； (3) 能显示基因及其剪接异构体的注释信息； (4) 能显示其他注释信息； (5) 既可以从远程服务器端下载各种注释信息，又可以从本地加载注释信息。 IGV浏览器使用方法可参考我们提供的使用说明文档（ IGVQuickStart.pdf ）。

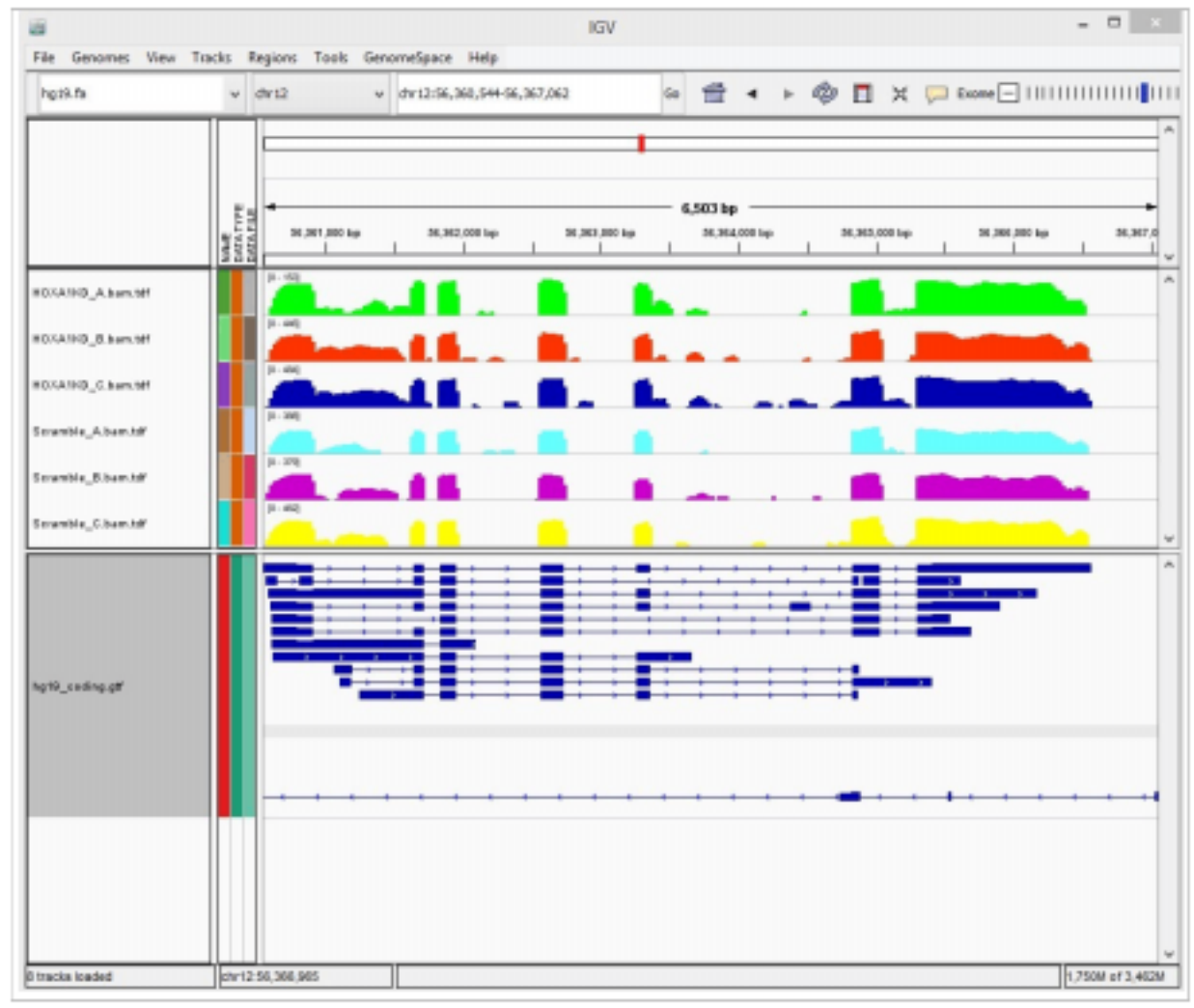


图3.4 IGV浏览器界面

4 基因表达分析

一个基因表达水平的直接体现就是其转录本的丰度情况，转录本丰度程度越高，则基因表达水平越高。在 RNA-seq 分析中，我们可以通过定位到基因组区域或基因外显子区的测序序列 (reads) 的计数来估计基因的表达水平。通过不同 Reads 计数除了与基因的真实表达水平成正比外，还与基因的长度和测序深度成正相关。为了使不同基因、不同实验间估计的基因表达水平具有可比性，人们引入了 RPKM 的概念， RPKM(Reads Per Kilo bases per Million reads) 是每百万 reads 中来自某一基因每千碱基长度的 reads 数目。RPKM 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) 。

4.1 已知注释类型基因含量分布

采用 HTSeq 软件对各样品不同基因进行表达水平分析，使用的模型为 union 。根据表达量统计样品中各类型基因的的表达分布，结果展示如下：

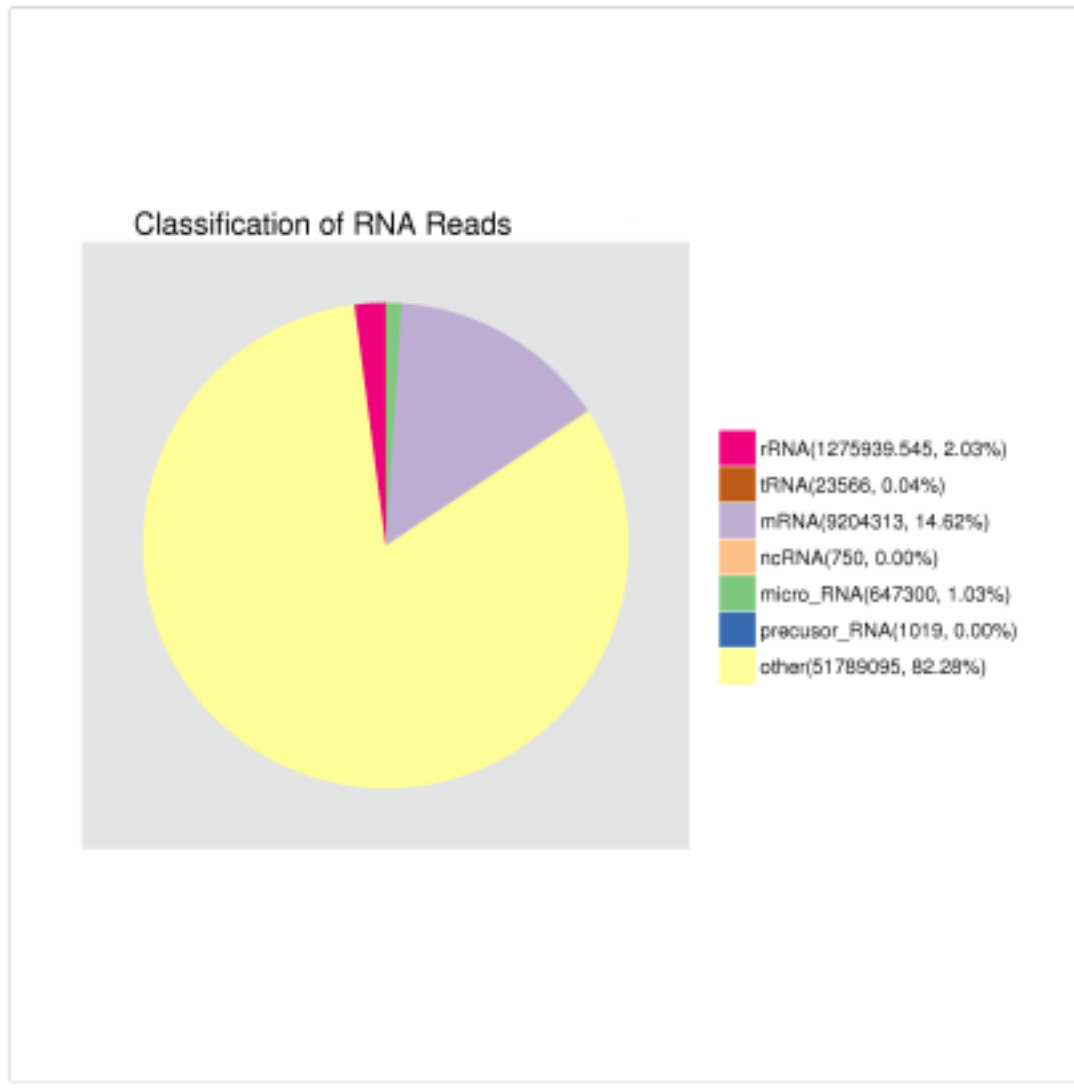


图4.1 各类已知基因表达分布图

4.2 已知基因表达水平分析

分别统计不同表达水平下基因的数量以及单个基因的表达水平。一般情况下，RPKM数值0.1或者1作为判断基因是否表达的阈值，不同的文献所采用的阈值不同。

表4.2.1 不同表达水平区间的基因数量统计表

RPKM Interval	sample1_A	sample1_B	sample1_C	sample2_A	sample2_B	sample2_C
0-1	36250(69.51%)	36137(69.29%)	35782(68.61%)	37473(71.86%)	36140(69.30%)	36837(70.64%)
1-3	6275(12.03%)	6772(12.99%)	6786(13.01%)	5768(11.06%)	6712(12.87%)	6025(11.55%)
3-15	6540(12.54%)	6421(12.31%)	6441(12.35%)	6026(11.56%)	6453(12.37%)	6247(11.98%)
15-60	2121(4.07%)	1968(3.77%)	2222(4.26%)	1940(3.72%)	2007(3.85%)	2135(4.09%)
>60	964(1.85%)	852(1.63%)	919(1.76%)	943(1.81%)	838(1.61%)	906(1.74%)

表4.2.2 基因表达水平统计表

geneID	sample1_A	sample1_B	sample1_C	sample2_A	sample2_B	sample2_C
ENSSSCG00000000001	0.125771932862869	0.133671497323016	0.192770243221167	0.0317267050358883	0.168501918079879	0.0800376669084703
ENSSSCG00000000002	1.26990067319925	0.712742543165399	0.335330135060644	0.892632434789404	0.764646758439737	0.496378572092531
ENSSSCG00000000003	2.10526950417426	1.77135307842206	3.38359531428205	2.2127772289762	3.36111680005984	3.53540758519473
ENSSSCG00000000004	0.061509184557666	0.0980587394793265	0.0471374661962334	0.03103210248962	0.0824064267883713	0.0652378146350261

5 RNA-seq整体质量评估

5.1 样品间相关性检查

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 (Hansen et al.) 。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，另一个为后续的差异基因分析所需要的。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。 Encode计划建议皮尔逊相关系数的平方 (R^2) 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 R^2 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。

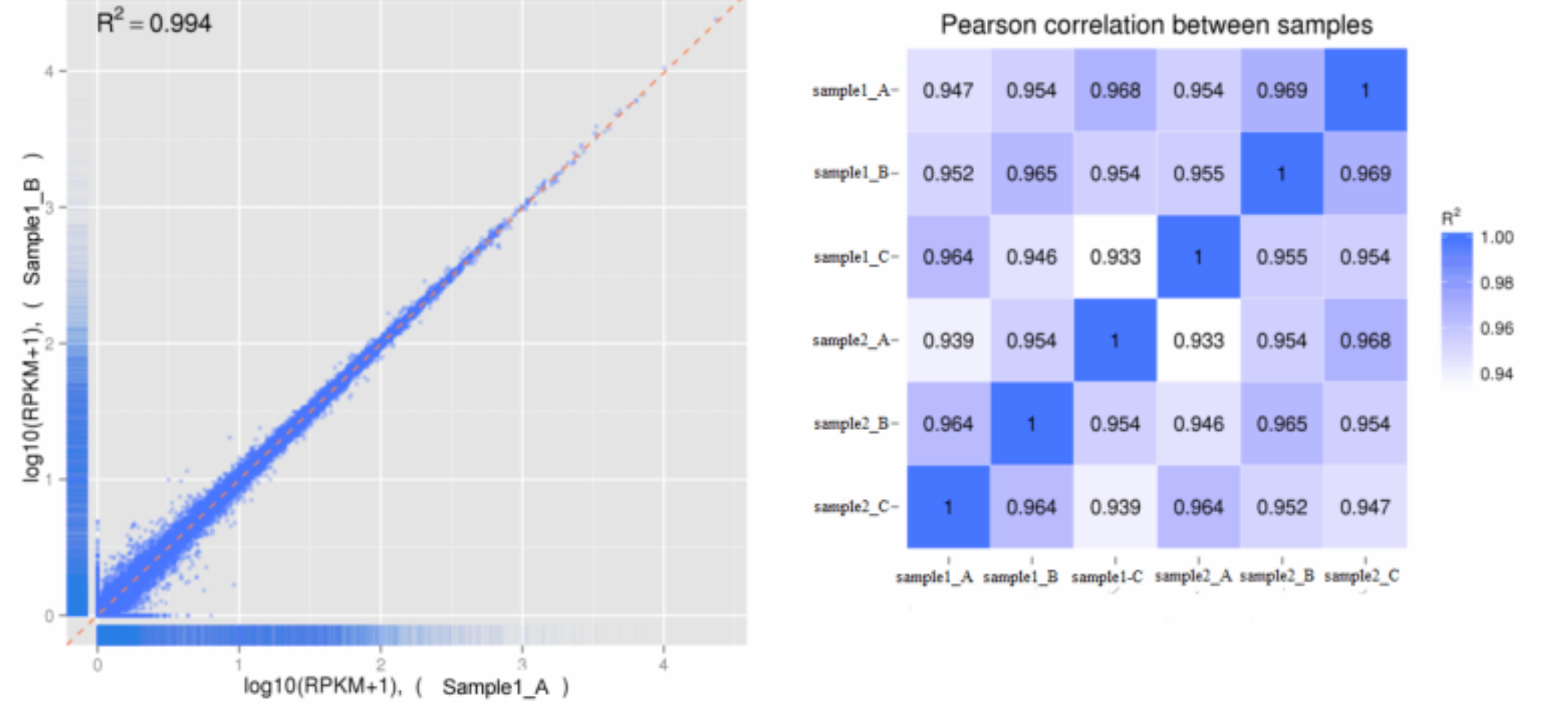


图5.1 样品间相关性检查

左图：样品间的相关系数散点图， R^2:pearson 相关系数的平方；右图：样品间相关系数热图

5.2 样品间聚类及 PCA分析

当样本数目较多时（ >=4），可以利用基因的表达量进行样本间聚类分析及 PCA分析，对样本间关系进行探究或者对实验设计进行验证。 PCA为主成分分析，可以从不同维度展现样品间的关系。样本聚类距离或者 PCA距离越近，说明样本越相似。

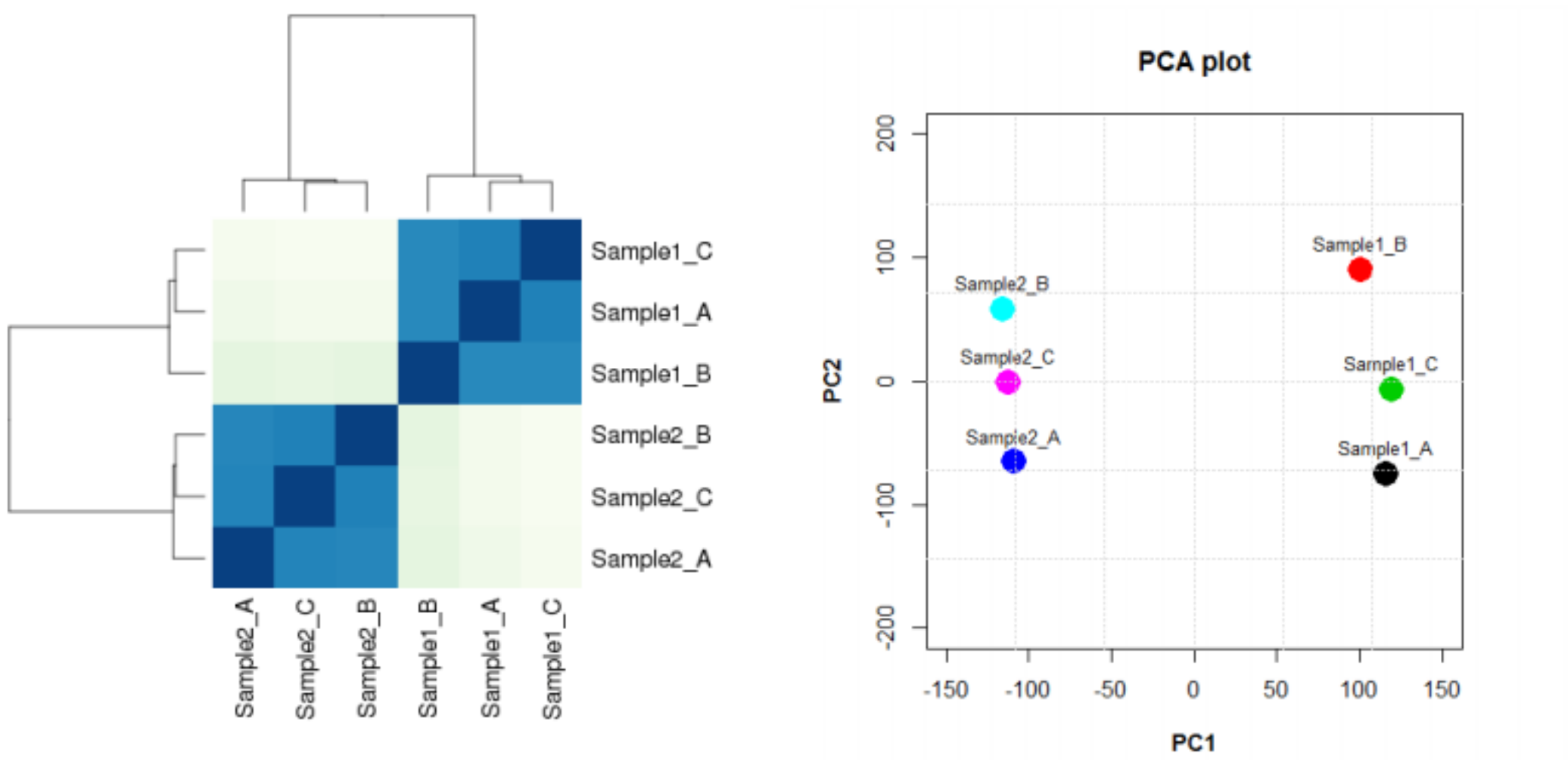


图5.2 样本间聚类及 PCA分析

左图：样品的基因表达水平的层次聚类图；右图：样品的基因表达水平的 PCA聚类图

5.3 均一性分布检查

理想条件下，对于 RNA-seq 技术来说，测序序列 (reads) 之间为独立抽样并且 reads 在所有表达的转录本上的分布应该呈现均一化分布。然而很多研究表明，很多偏好型的因素都会影响这种均一化的分布 (Dohm et al., 2008) 。例如，在 RNA-seq 建库过程中，片段破碎和 RNA 反转录的顺序不一样会导致 RNA-seq 最终的数据呈现严重的 3’偏好性。其他因素还包括转录区域的 GC 含量不同、随机引物等等，并且生物体内从 5’或者 3’的降解过程同样会导致不均一性分布。

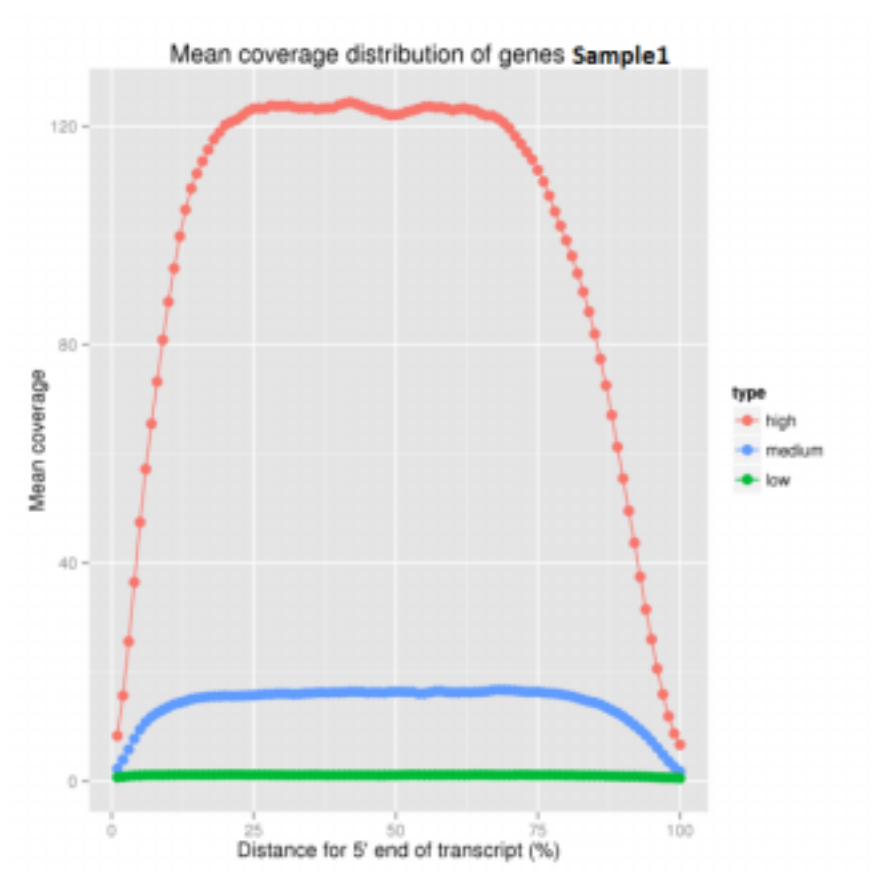


图5.3 不同表达水平的转录本的 reads密度分布图

High：高表达量转录本； Medium: 中度表达量转录本； Low: 低表达量转录本；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

6 转录本拼接

采用Cufflinks(Trapnell et al,2013) 和Scripture(Guttman et al,2010) 两种软件同时对比对结果进行组装，在此基础上进行 lncRNA的筛选。

6.1 cufflinks 拼接

cufflinks 使用概率模型，可同时组装和量化尽可能小的 isoform 集的表达水平，提供表达数据在给定位点的最大似然说明，针对链特异性文库有特定的参数可准确提供链的信息。 cufflinks 拼接结果展示如下：

表6.1 cuffliks 拼接结果展示（部分）

Chr Num	Source	Type	Start Site	End Site	.	Strand	.	Descripture
1	Cufflinks	exon	753520	753582	.	+	.	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; oId "CUFF.5.1"; tss_id "TSS1";
1	Cufflinks	exon	754103	754327	.	+	.	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "2"; oId "CUFF.5.1"; tss_id "TSS1";
1	Cufflinks	exon	898560	898633	.	+	.	gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "1"; oId "CUFF.8.1"; tss_id "TSS2";
1	Cufflinks	exon	898717	898773	.	+	.	gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "2"; oId "CUFF.8.1"; tss_id "TSS2";
1	Cufflinks	exon	898858	898884	.	+	.	gene_id "XLOC_000003"; transcript_id "TCONS_00000003"; exon_number "1"; oId "CUFF.9.1"; tss_id "TSS3";

表格说明如下：

第1列：染色体序号；第 2列：来源描述；第 3列：类型；第 4列：起始坐标；第 5列：终止坐标；第 7列：链的信息；第 9列：id 等描述信息；

6.2 scripture 拼接

scripture 基于统计分割模型来区分表达位点和实验背景噪音，报告给定位点所有统计学表达显著的 isoform ，较适用于长转录本的拼接。scripture 拼接结果展示如下：

表6.2 scripture 拼接结果展示（部分）

Chr Num	Start Site	End Site	Transcript Id	Score	Strand	thickStart	thickEnd	itemRgb	blockCount	blockSizes	blockStarts
1	14656	15045	chr1:14656-15045	0.0	-	14656	15045	0,0,0	2	173,76,	0,313,
1	16727	18365	chr1:16727-18365	0.0	-	16727	18365	0,0,0	6	38,198,136,137,147,98,	0,130,505,878,1187,1540,
1	24848	29351	chr1:24848-29351	0.0	-	24848	29351	0,0,0	2	43,31,	0,4472,
1	135989	569483	chr1:135989-569483	0.0	+	135989	569483	0,0,0	3	24,1,3,	0,433490,433491,
1	135989	569519	chr1:135989-569519	0.0	+	135989	569519	0,0,0	6	24,1,4,3,6,26,	0,433490,433491,433495,433498,433504,

表格说明如下：

第1列：染色体序号；第 2列：起始坐标；第 3列：终止坐标；第 4列：转录本 id ；第6列：链的信息；第 10列：exon个数；第 11列：exon长度；第 12列：exon起始位置；

7 候选 lncRNA筛选

lncRNA为一类长度 >200bp的长链非编码 RNA，根据与编码序列的位置关系可分为 intergenic lncRNA(简称lincRNA)，intronic lncRNA, anti-sense lncRNA ，sense lncRNA，bidirectional lncRNA 等类型。其中 lincRNA所占比例最高，这里主要进行前 3种类型的筛选。我们根据 lncRNA的特点设置一系列严格的筛选条件，基于 cufflinks 和scripture 的拼接结果同时进行以下步骤的筛选，最终选择在 >=2个样本中出现或者被两个拼接软件同时拼接出来的 lncRNA作为最终的候选 lncRNA集进行后续分析。

7.1 基本筛选

基本筛选主要由三个部分组成：

step1: 选择长度 >=200bp, Exon个数>=2的转录本；

step2: 通过cufflinks 计算每条转录本的 reads 覆盖度，选择 Reads最小覆盖度 >=3的转录本；

step3: 通过与已知非 lncRNA基因比较过滤掉非 lncRNA基因，并利用 cuffcompare 的结果进行位置筛选（对不同种类的 lncRNA选择不同的 class_code）；

下图展示的是 lncRNA的筛选过程统计．

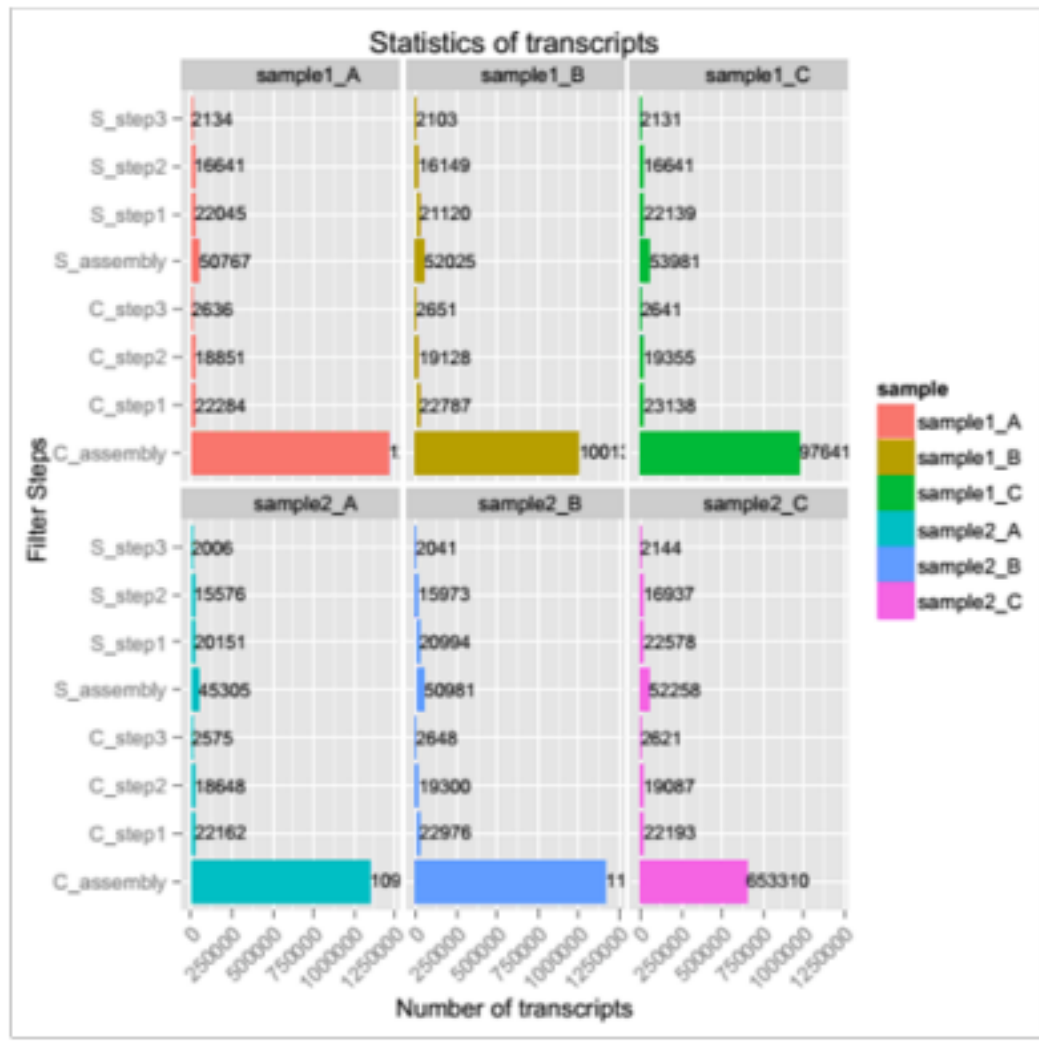


图7.1.1 lncRNA 的筛选统计图

纵坐标为筛选步骤（ C代表cufflinks ，S代表scripture ，assembly为原始拼接出来的转录本条数），横坐标为对应步骤筛选过后的转录本条数

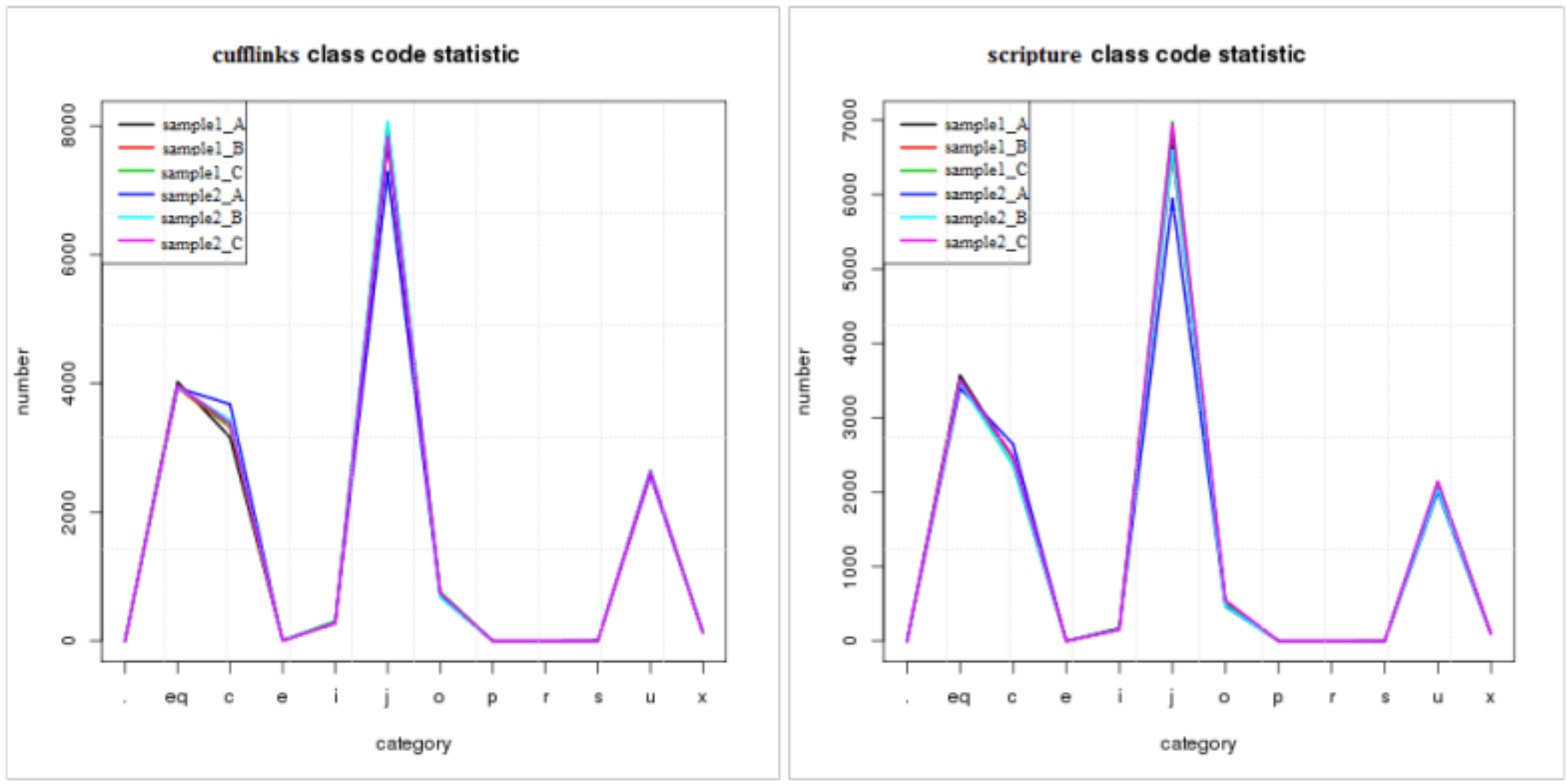


图7.1.2 lncRNA 的筛选统计图

横坐标为各 class_code 类型，纵坐标为对应类型的转录本条数；左图为 cufflinks 的结果，右图为 scripture 的结果

7.2 编码潜能筛选

具有编码潜能与否是判断转录本是否为 lncRNA的关键条件，我们综合了目前主流的编码潜能分析方法进行该项筛选，主要包括： CPC分析、CNCI 分析、 pfam蛋白结构域分析、 PhyloCSF分析四种方法。

(1) CPC 分析

CPC(CodingPotential Calculator) 结果展示如下：

TCONS_00001692	179	noncoding	-0.959246
TCONS_00002087	175	noncoding	-0.0656937
TCONS_00003397	302	coding	3.04452
TCONS_00003398	602	coding	1.6424
TCONS_00003399	2215	coding	8.34615
TCONS_00003407	3849	coding	5.56963

(2) CNCI 分析

CNCI(Coding-Non-Coding Index) 结果展示如下：

>TCONS_00001692	noncoding	score: -0.0981794389676737	start: 6	stop: 9
>TCONS_00002087	noncoding	score: -0.144992537603425	start: 0	stop: 21
>TCONS_00003397	noncoding	score: -0.0289064254521316	start: 24	stop: 147
>TCONS_00003398	noncoding	score: -0.166017172756653	start: 381	stop: 447
>TCONS_00003399	coding	score: 0.174295904605794	start: 0	stop: 1824
>TCONS_00003407	coding	score: 0.199483408774546	start: 0	stop: 1953

(3) pfam 蛋白结构域分析

pfam蛋白结构域搜索结果展示如下：

seq id	alignment start	alignment end	envelope start	envelope end	hmm acc	hmm name	type	hmm start	hmm end	hmm length	bit score	E-value	significance	clan
TCONS_00003399-0	7	104	7	108	PF00752.12	XPG_N	Family	1	97	101	23.7	3.5e-05	1	CL0280
TCONS_00003399-0	122	244	121	318	PF12813.2	XPG_I_2	Domain	2	118	246	68.6	4.4e-19	1	CL0464
TCONS_00003407-0	2	184	2	185	PF06466.6	PCAF_N	Domain	70	251	252	359.1	8.8e-108	1	No_clan
TCONS_00003407-0	411	483	409	484	PF00583.19	Acetyltransf_1	Family	3	82	83	33.2	3.2e-08	1	CL0257
TCONS_00003407-0	594	675	593	676	PF00439.20	Bromodomain	Domain	2	83	84	89.2	9.8e-26	1	No_clan

(3) phyloCSF 分析

phyloCSF(phylogenetic codon substitution frequency) 进化密码子置换频率分析，利用多物种间的全基因组序列比对文件定义一段基因组区域是否有编码潜能。通过文献查询，我们发现不同的物种间 phyloCSF阈值不尽相同，故首先随机选择本项目研究物种一定数目的已知 coding 和lncRNA基因进行阈值分析，再筛选候选转录本分析结果。 phyloCSF结果展示如下：

TCONS_00001692	max_score(decibans)	18.2411	39	140	MKGQDEASQQLLCENEDWCSLGAGHVEWASLWGC
TCONS_00005370	max_score(decibans)	12.7010	69	203	MSRSLAELQRVSLLLKGRGLPQGSTQFPCLTALGCANRLWLEKLA
TCONS_00003569	max_score(decibans)	40.1837	630	746	MRLRVLSQPLLSGLRIRRCRDLWWRLQTRLGSRVAVALA
TCONS_00005065	max_score(decibans)	28.4149	48	140	MVLGTPVKIEVGFEGSSPKSFVLATETSTVA

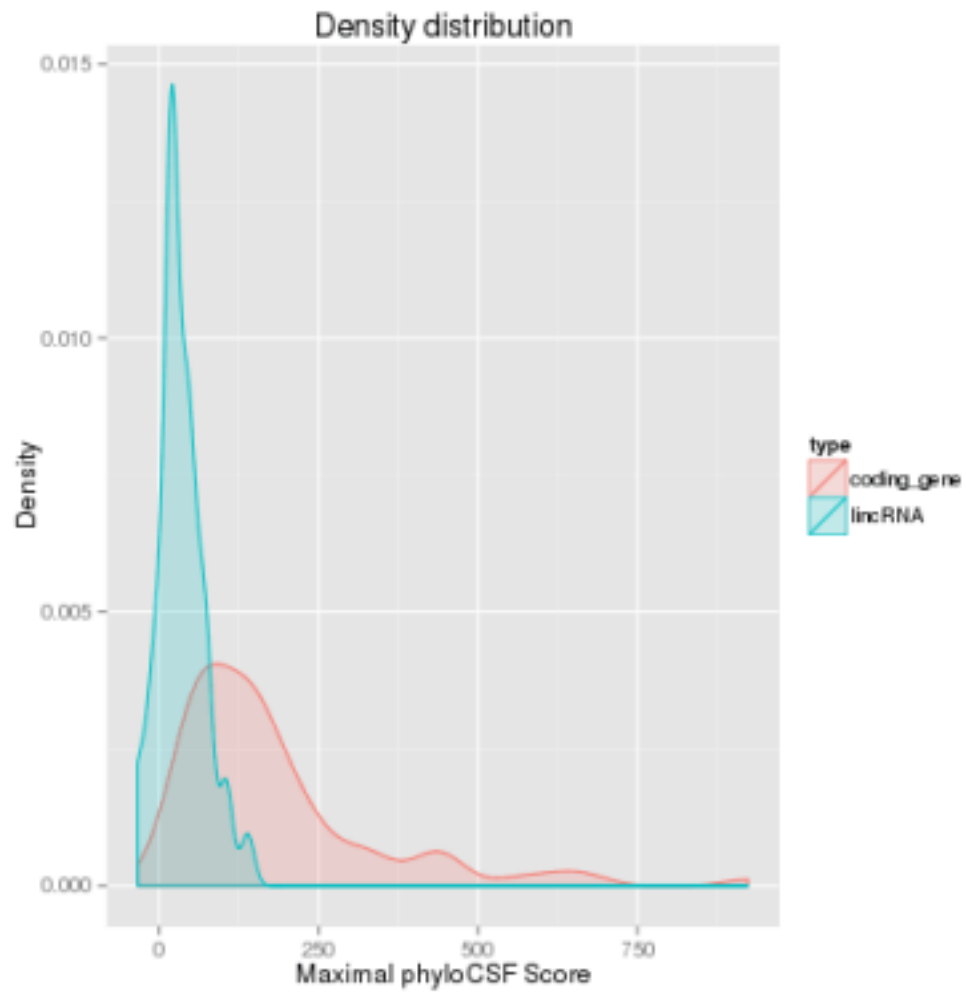


图7.2.1 lncRNA 的筛选统计图

横坐标为 phyloCSF的分值，纵坐标为对应分值的转录本占所有转录本条数的比例

将4种软件的结果取交集：

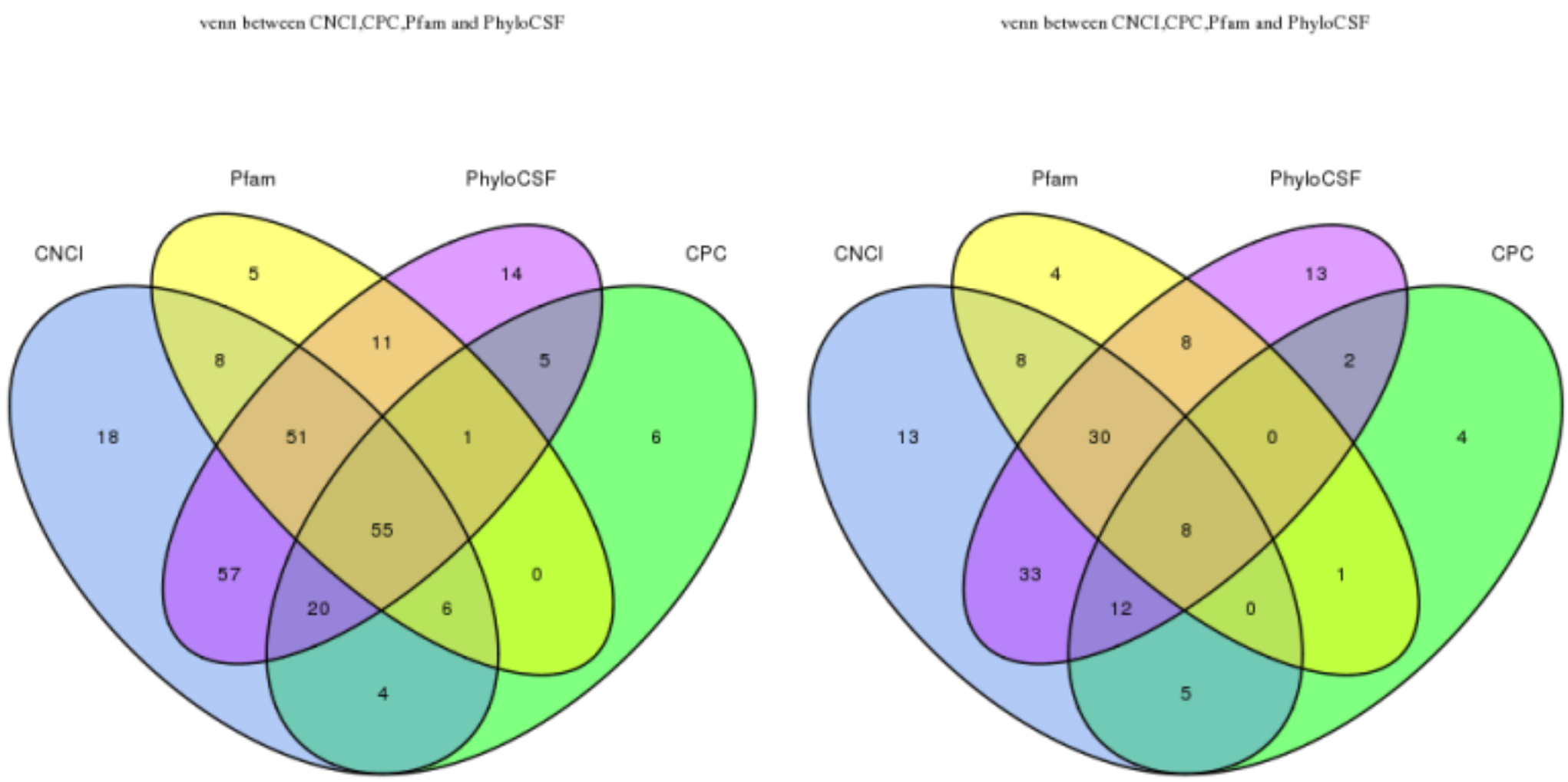


图7.2.2 4 种方法结果维恩图展示

左边为cuffliks 的结果，右边为 scripture 的结果

7.3 重现性筛选

首先选择在 >=2个生物学重复样本中出现的转录本为候选 lncRNA集，然后对剩下的转录本进行一项“挽救”措施，即将同时被两个拼接软件拼接得到的 lncRNA也筛选出来，得到最终的候选 lncRNA集进行后续分析。

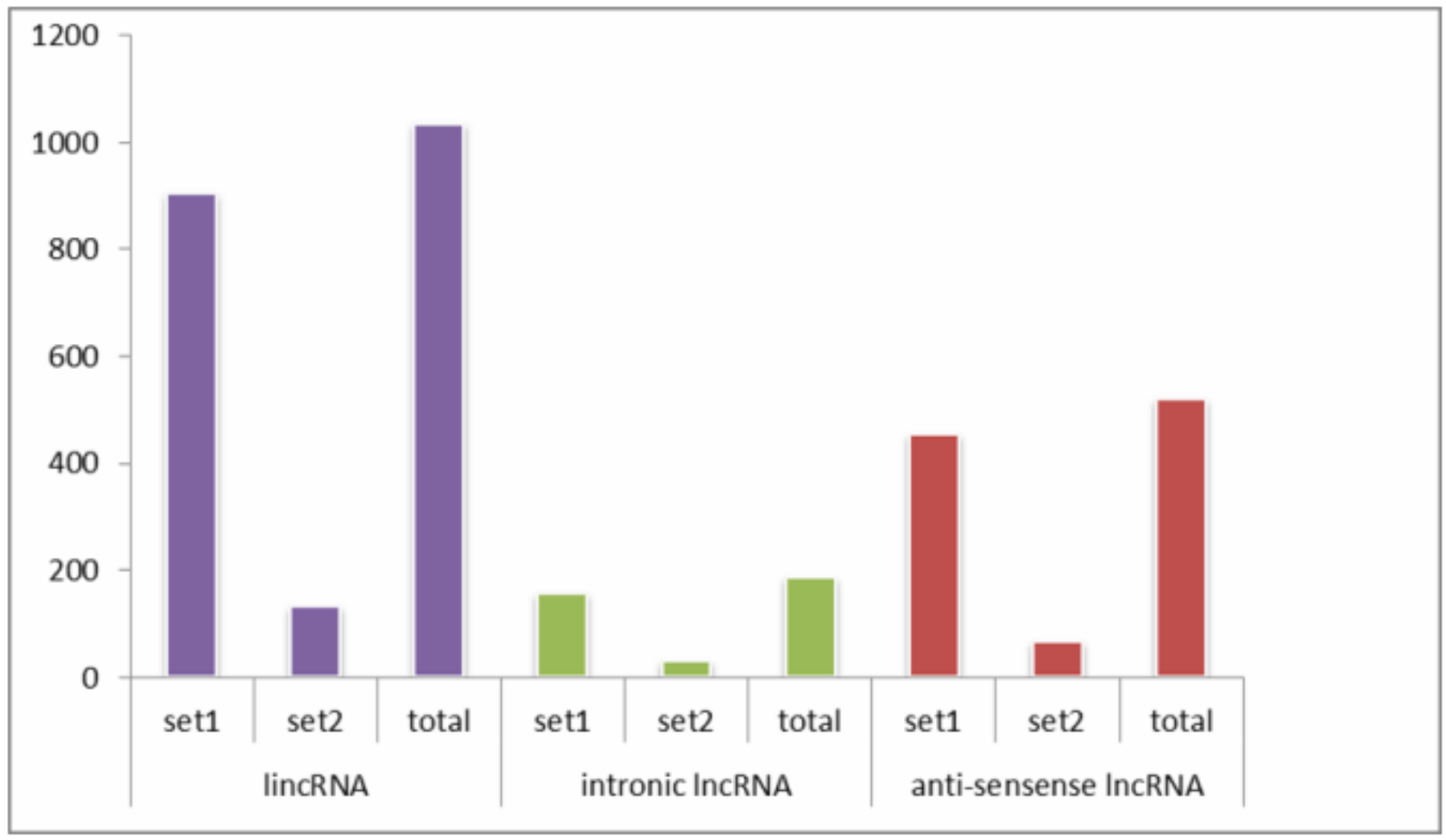


图7.3 lncRNA 的筛选统计图

横坐标为 lncRNA类型，纵坐标为对应类型的转录本条数

8 候选 lncRNA描述性统计

对筛选得到的 lncRNA进行长度， exon个数等方面的统计，有助于进一步观察筛选得到的候选 lncRNA的特点，并通过与本物种已知 lncRNA比较得到已知 lncRNA和新预测的 lncRNA。

8.1 长度分布统计

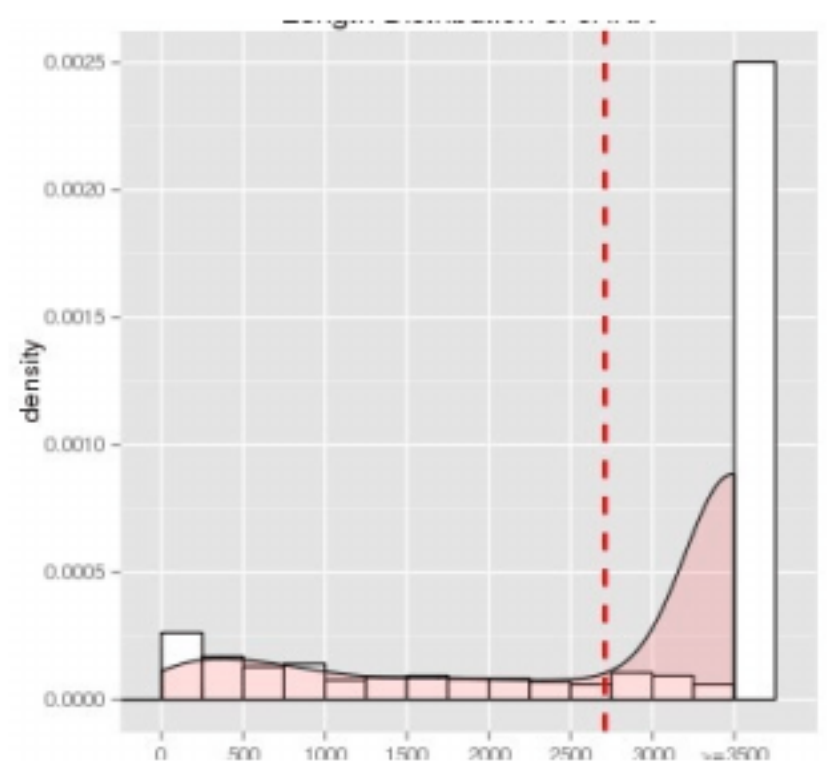


图8.1 lncRNA 长度分布图

横坐标为 lncRNA长度(bp)， 纵坐标为对应长度的转录本密度

8.2　外显子数目统计

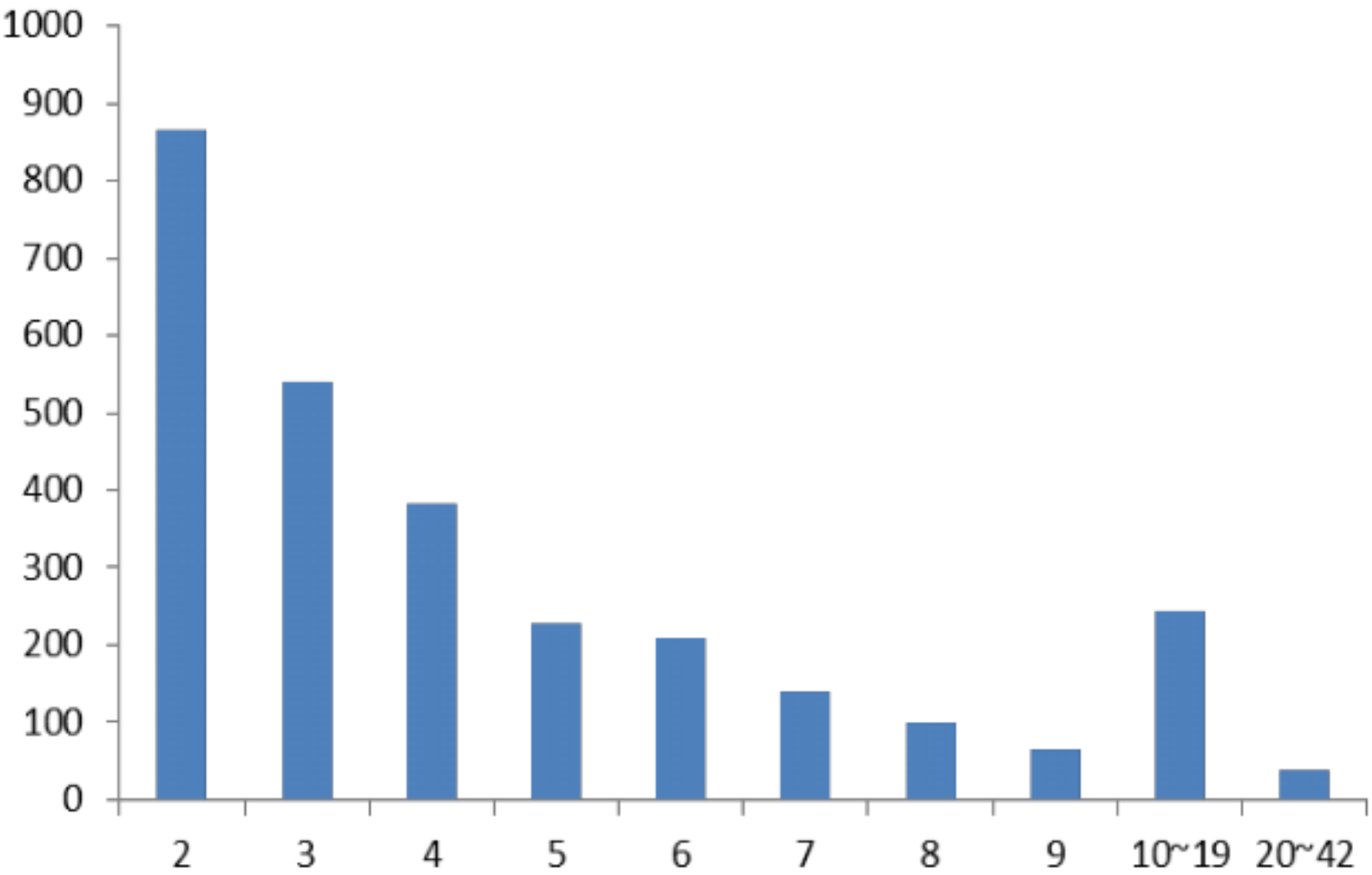


图8.2　外显子数目统计

横坐标为外显子个数，纵坐标为对应转录本的数目

8.3 已知和预测 lncRNA统计

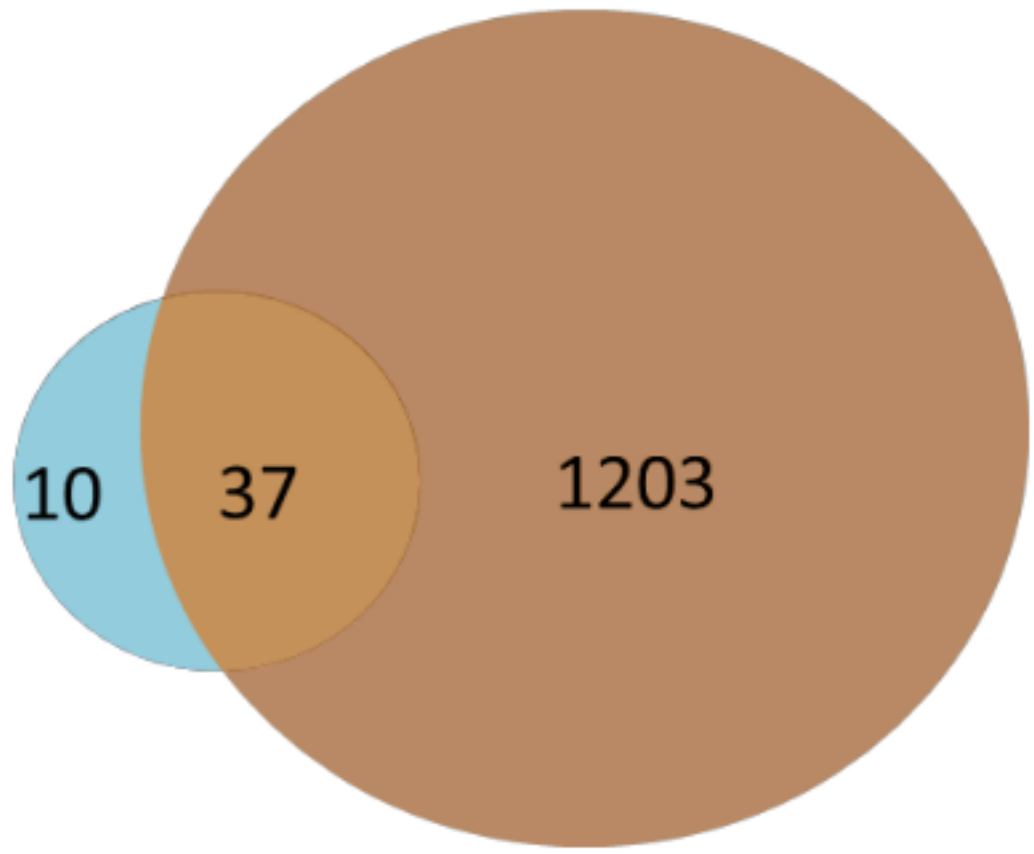


图8.3 已知和预测 lncRNA维恩图

9 lncRNA保守性分析

9.1 序列保守性分析

lncRNA的序列保守性相对蛋白编码基因要低，采用 phastCons(http://compgen.bscb.cornell.edu/phast/) 分别对蛋白编码基因和 lncRNA基因进行保守性打分，得到如下保守性分值累积分布图。

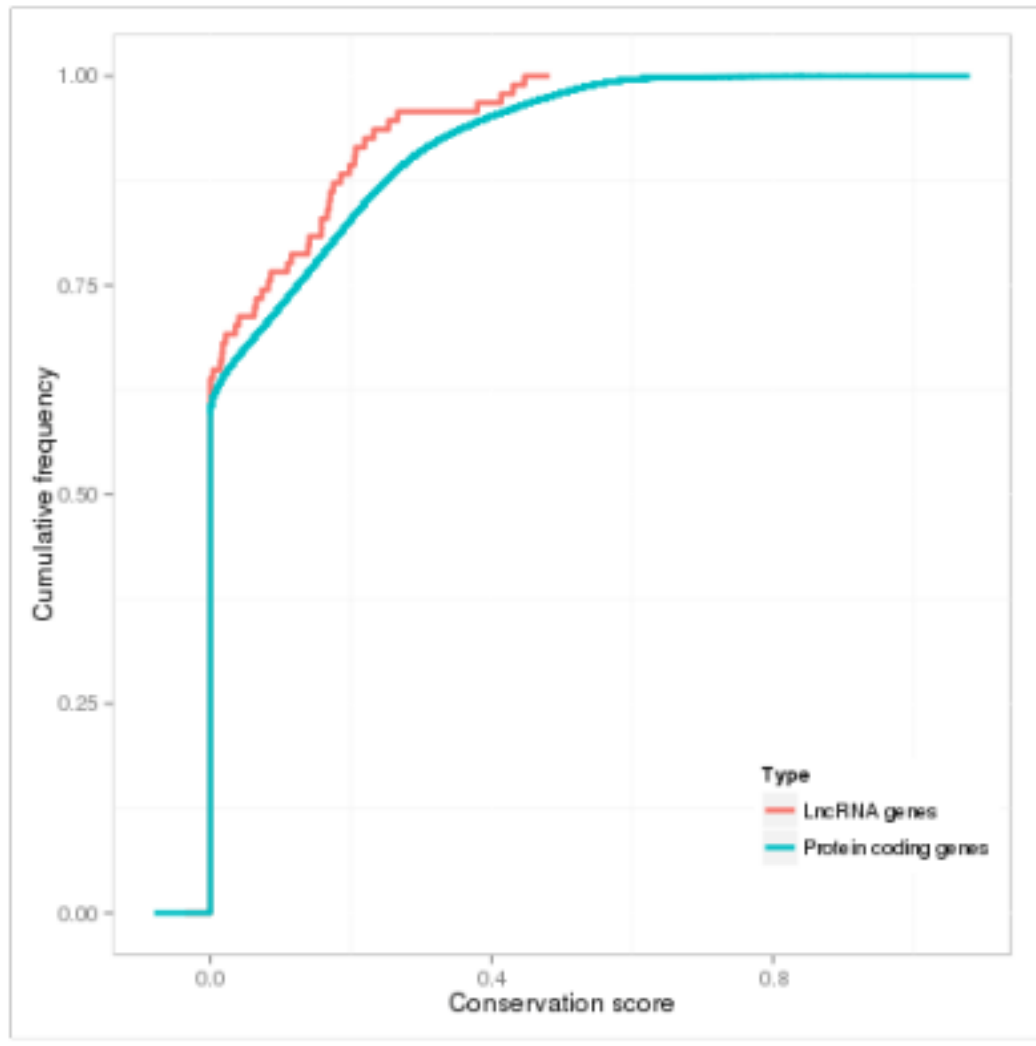


图9.3 lncRNA 和蛋白编码基因的保守性分值累积分布图

9.2 位点保守性分析

lncRNA的序列在物种间有一定的位点保守性，通过 UCSC浏览器可视化 lncRNA在不同物种中的位置。

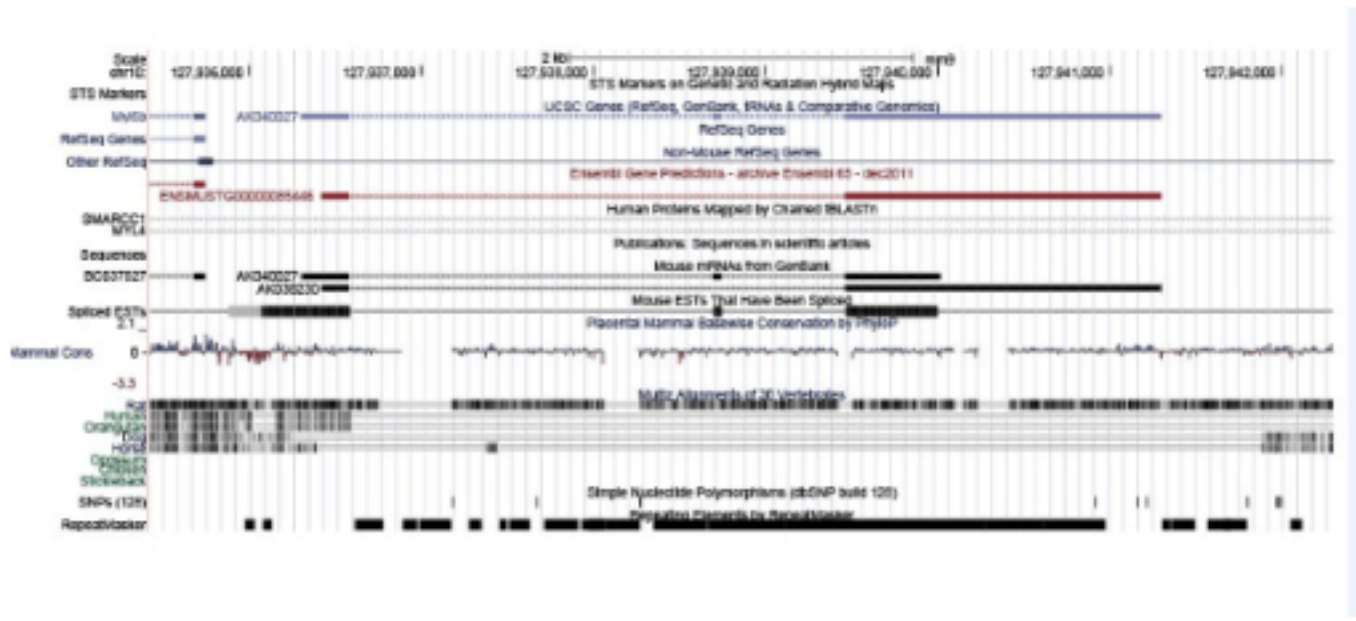


图9.3 lncRNA 和蛋白编码基因的保守性分值累积分布图

10 lncRNA差异表达分析

10.1 lncRNA表达水平对比

通过所有 lncRNA的RPKM的分布图以及盒形图对不同实验条件下的 lncRNA表达水平进行比较。对于同一实验条件下的重复样品，最终的 RPKM为所有重复数据的平均值。

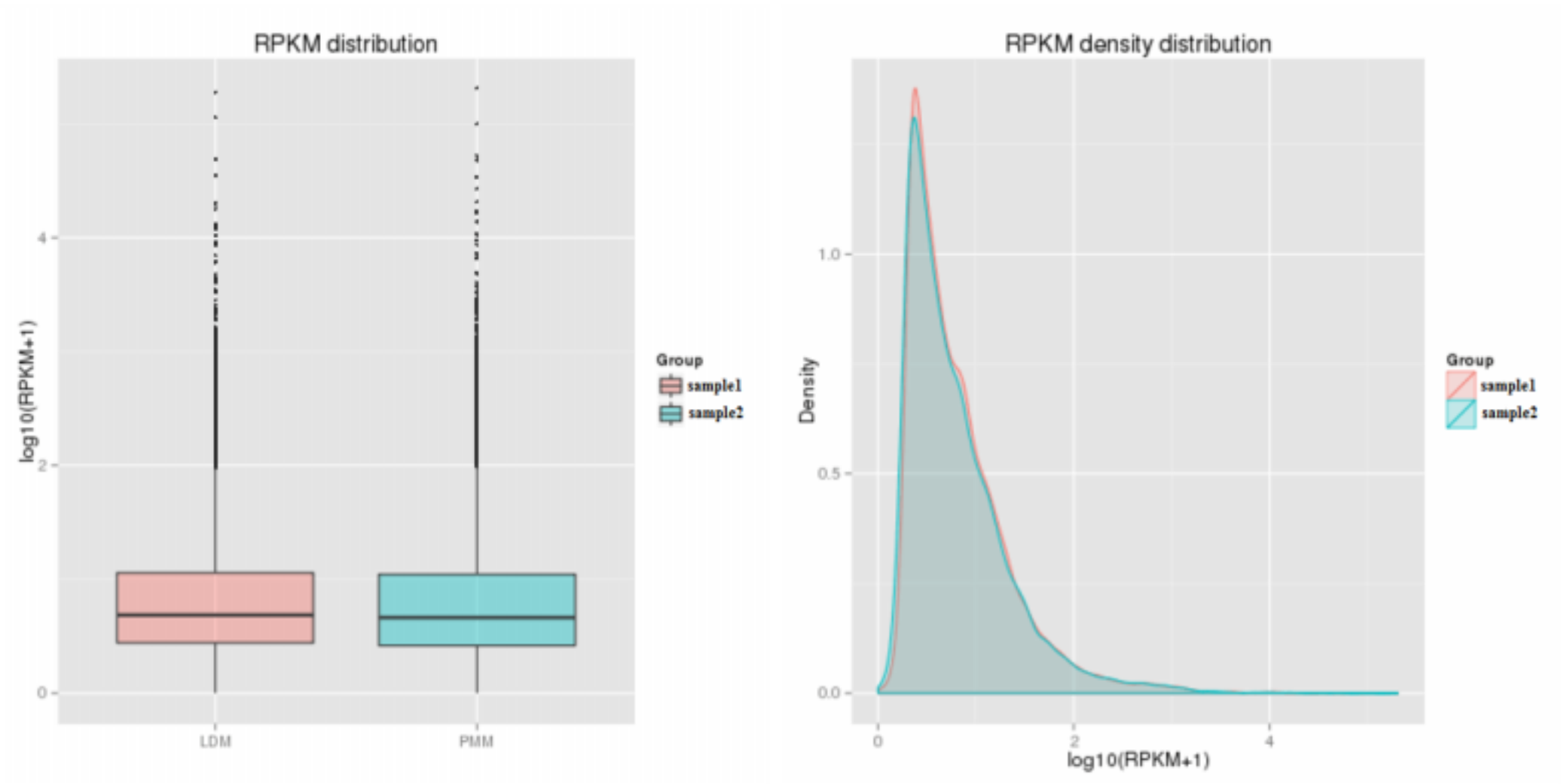


图10.1 不同实验条件下 lncRNA表达水平比对图

图一：RPKM盒形图，横坐标为样品名称，纵坐标为 $\log_{10}$(RPKM)；每个区域的盒形图对五个统计量（至上而下分别为最大值，上四分位数，中值，下四分位数和最小值）

图二：RPKM分布图，横坐标为 $\log_{10}$(RPKM)，纵坐标为基因的密度

10.2 lncRNA差异表达分析

lncRNA差异表达的输入数据为 lncRNA表达水平分析中得到的 readcount 数据。对于有生物学重复的样品，分析我们采用 DESeq（Anders et al, 2010 ）进行分析：

该分析方法基于的模型是负二项分布，第 i 个基因在第 j 个样本中的 read count 值为 K_{ij} ，则有

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$$

对于无生物学重复的样品，先采用 TMM对read count 数据进行标准化处理，之后用 DEGseq进行差异分析。差异表达基因列表如下：

表9.2 差异基因列表

gene_id	readcount_sample1	readcount_sample2	log2FoldChange	pval	padj
lincRNA_001	11.6103255151009	182.601521401153	-3.9752	2.7547e-22	1.3326e-17
lincRNA_002	143.999286139504	14.8247488638477	3.28	1.0235e-15	2.0042e-11
lincRNA_003	1.63236134587914	53.9757276864048	-5.0473	1.2429e-15	2.0042e-11
lincRNA_004	0.325322610674719	39.165115536752	-6.9116	9.0982e-15	9.7672e-11

差异lncRNA列表主要包括的内容：

(1) Gene_id: 基因编号
(2) readcount_Sample1 ：校正后样品组 1的readcount 值
(3) readcount_Sample2 ：校正后样品组 2的readcount 值
(4) log2FoldChange: $\log_2$(Sample1/Sample2)
(5) pvalue(pval): 统计学差异显著性检验指标
(6) qvalue(padj): 校正后的 pvalue 。qvalue 越小，表示基因表达差异越显著

10.3 差异表达 lncRNA筛选

用火山图可以推断差异 lncRNA的整体分布情况，对于无生物学重复的实验，为消除生物学变异，我们从差异倍数和显著水平两个水平进行评估，对差异 lncRNA进行筛选，阈值设定一般为 ：$|\log_2(\text{FoldChange})| > 1$ 且 qvalue < 0.005 。对于有生物学重复的实验，由于 DESeq已经进行了生物学变异的消除，我们对差异 lncRNA筛选的标准一般为 ：padj < 0.05 。

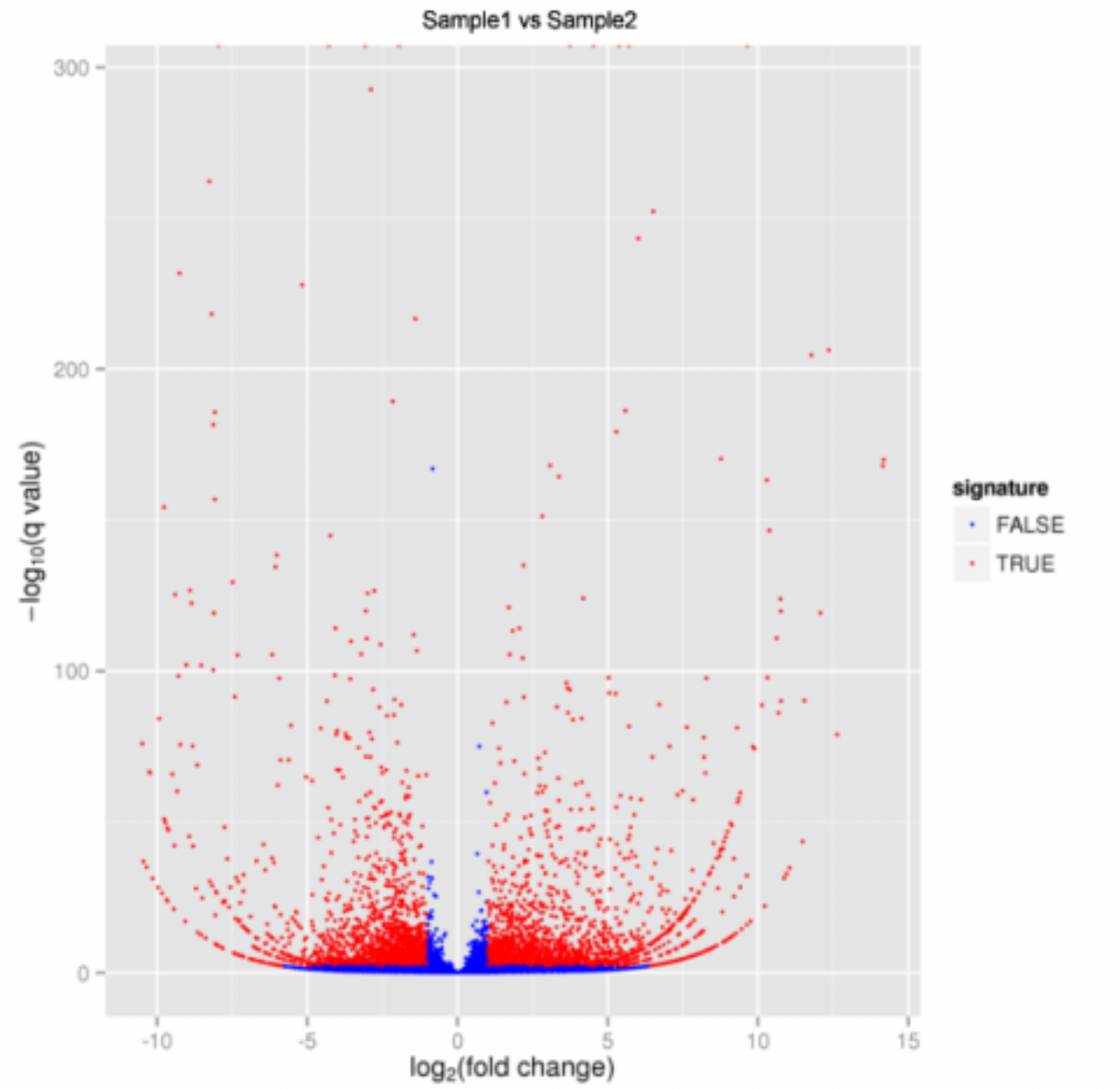


图10.3 差异 lncRNA火山图

有显著性差异表达的 lncRNA用红色点表示；横坐标代表 lncRNA在不同样本中表达倍数变化；纵坐标代表 lncRNA表达量变化差异的统计学显著性

11 lncRNA组织或表型特异性分析

11.1 lncRNA与mRNA表达聚类分析

通常认为 lincRNA 相对于 mRNA 有较高的组织表达特异性，随机抽取一定比例的 lncRNA和mRNA比较两种类型基因在不同组织中表达水平的聚类情况。

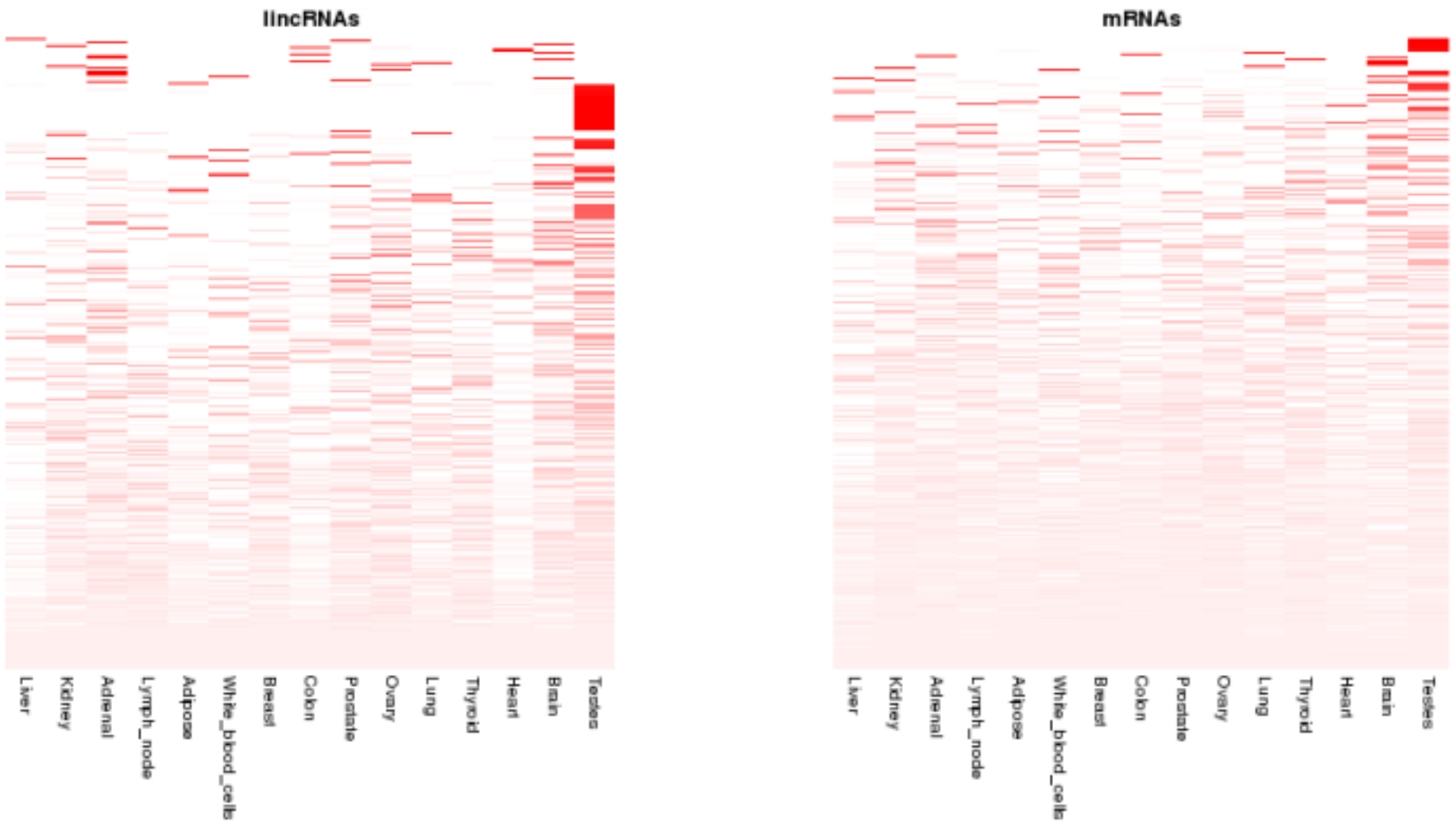


图11.l lncRNA 和mRNA的表达热图

左图为lncRNA在各样品中的表达情况，右图为 mRNA在个样品中的表达情况；横坐标为样品，纵坐标为基因，颜色越深表示表达水平越高

11.2　组织或表型特异性分析

我们基于 JS divergence 这一衡量指标对于各转录本在不同组织样本中的表达模式（ pattern ）进行分析。参考文献（ Cabili, M.N.et al. ，2011），预先设定每个转录本仅在一个组织中特异性表达有 N（N为组织个数）种模式，定义每两个转录本表达模式之间的距离为 JS的平方根，则一个转录本在 N个组织中的组织特异性定义为： $JS_{sp}(e|t) = 1 - JS_{dist}(e, e^t)$, 其中 e^t 为预先设定的转录本表达模式。选择其中最大的值作为该转录本在各组织中的特异性分值。分值范围为 0-1，分数越接近于 1，表示该转录本的组织特异性越高。

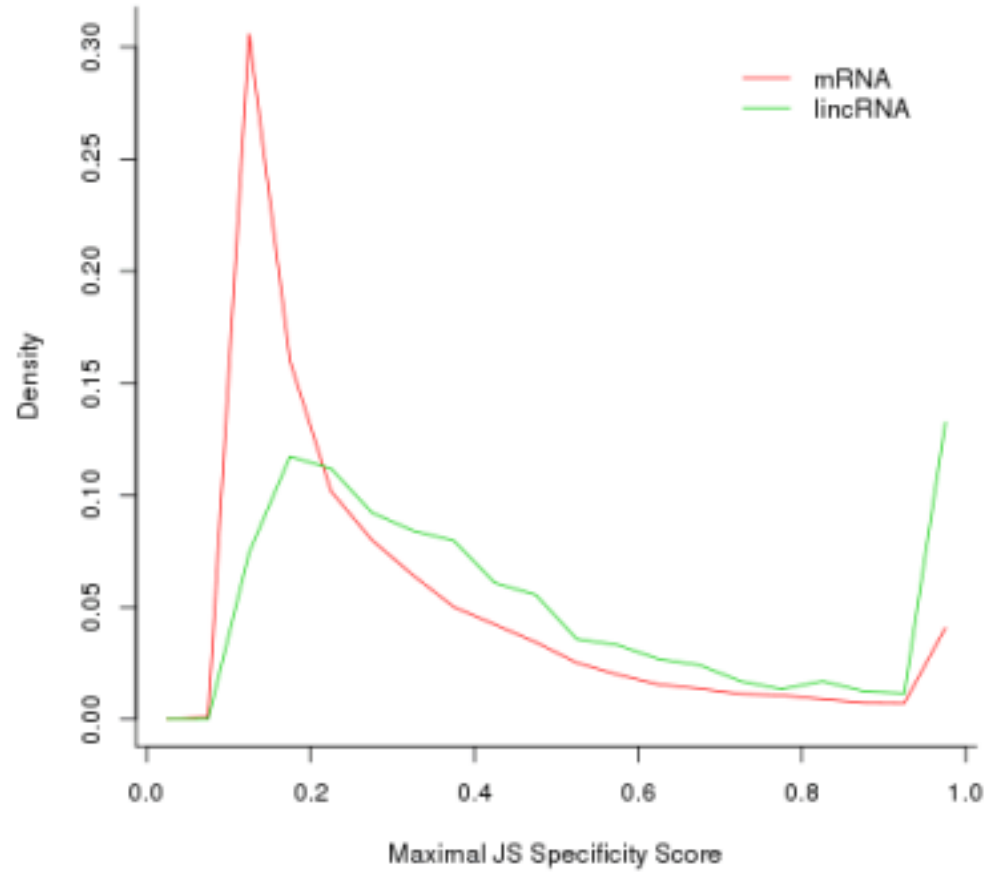


图11.2　转录本的组织特异性分值密度分布图

横坐标为转录本的组织特异性分值；纵坐标为转录本的密度

12 lncRNA靶基因预测

lncRNA功能主要通过 cis 或trans 作用于蛋白编码靶基因的方式实现，因此分成两种情况预测 lncRNA的靶基因。

12.1 cis 作用靶基因预测

cis 功能预测基本原理认为 lncRNA的功能与其坐标临近的编码蛋白基因相关，于是将 lncRNA基因临近的 (~上下游 10k-100k) 蛋白编码的基因找出进行功能富集分析，以推测 lncRNA的主要功能。 cis 作用靶基因预测结果如下表所示：

表12.1 cis 作用靶基因统计表

Expressed_Samplenum	lncRNA_num	10kb(lncRNA/mRNA)	100kb(lncRNA/mRNA)
1	47	26/29	45/160
2	18	7/10	15/78
3	10	7/12	9/36
4	11	6/7	9/59
5	9	5/5	6/32

注：

(1) Expressed_Samplenum ：lncRNA在n个样品中表达。
(2) lncRNA_num ：lncRNA在n个样品中表达的数目。
(3) 10kb(lncRNA/mRNA) ：在上下游 10kb范围内 (2) 中的lncRNA能检测到的 mRNA的数目。
(4) 100kb(lncRNA/mRNA) ：在上下游 100kb范围内 (2) 中的lncRNA能检测到的 mRNA的数目。

12.2 trans 作用靶基因预测

trans 功能预测基本原理认为 lncRNA的功能与样品中共表达的编码蛋白基因相关，可以通过样本间 lncRNA与蛋白编码基因的相关性分析或共表达分析来预测。当样本量 >=5时采用 Pearson相关系数法分析样本间 lncRNA与蛋白编码基因的相关性；当样本数 >=15时可采用 WGCNA(Langfelder et al, 2008) 将不同的组织、处理或者时间点间表达模式相似的基因聚类，以得到不同的共表达模块，根据模块内已知的蛋白编码的基因功能进一步探索研究 lncRNA的功能。

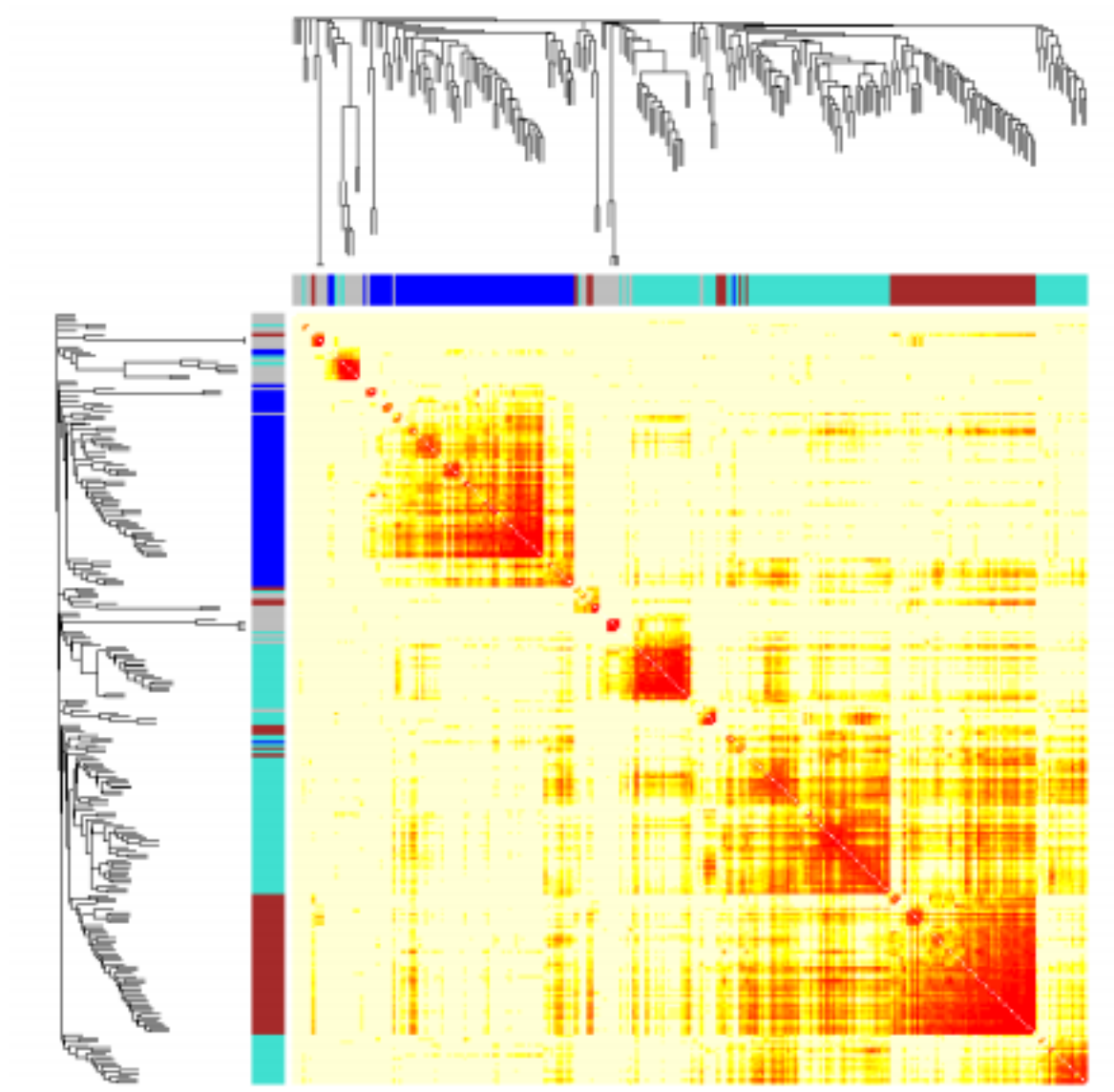


图9.6 共表达聚类热图

热图上方及左侧的层次聚类对应着共表达模块，这些模块在聚类图的下方或右侧用色彩条带展示。热图中的颜色表示基因间共表达连通性高低，饱和度高的黄色和红色表示基因间共表达连通性高。共表达连通性高的基因 "block" 对应于其上方及左侧的共表达模块。模块间的连通性高的基因位于层次聚类各分支的最顶端，因为它们与其它处于同一模块的基因具有最高的连通性。

13 特异 lncRNA靶基因功能富集分析

特异lncRNA一般指差异表达的或者组织或表型特异性表达的 lncRNA，对这些 lncRNA对应的靶基因分别进行 GO和KEGG功能富集分析。

13.1 GO富集分析

Gene Ontology（简称 GO，http://www.geneontology.org/ ）是基因功能国际标准分类体系。根据实验目的筛选特定 lncRNA后，研究该lncRNA对应的靶基因在 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 。相对于普通的超几何分布 (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 被靶基因富集的概率。

表13.1.1 靶基因 GO富集列表

GO_accession	Description	Term_type	Over_represented_pValue	Corrected_pValue	DEG_item	DEG_list
GO:0003700	sequence-specific DNA binding transcription factor activity	molecular_function	3.7832e-07	0.0030012	8	20
GO:0001071	nucleic acid binding transcription factor activity	molecular_function	3.846e-07	0.0030012	8	20
GO:0010468	regulation of gene expression	biological_process	3.4857e-06	0.013579	11	20
GO:0034654	nucleobase-containing compound biosynthetic process	biological_process	5.7706e-06	0.013579	11	20

结果表格详细内容如下：

(1) GO_accession ：Gene Ontology 数据库中唯一的标号信息

(2) Description ：Gene Ontology 功能的描述信息

(3) Term_type ：该GO的类别 (cellular_component ：细胞组分； biological_process ：生物学过程； molecular_function ：分子功能）

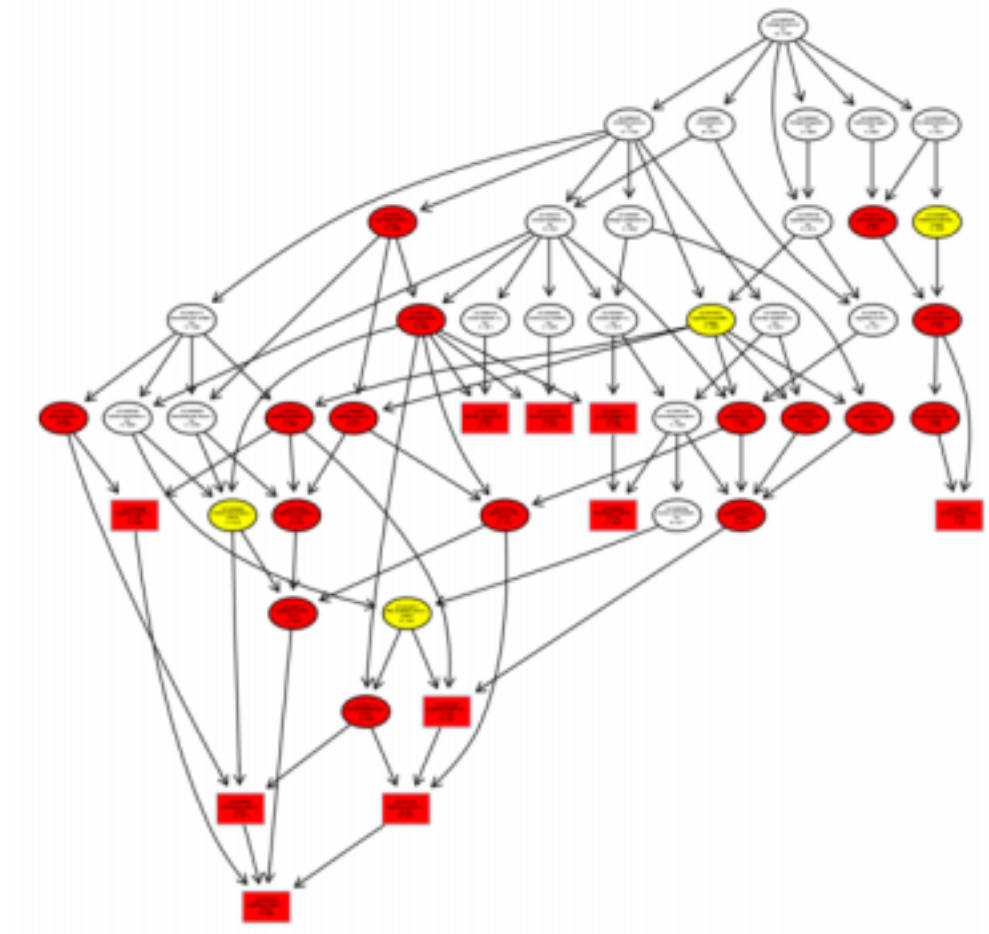
(4) Over_represented_pValue ：富集分析统计学显著水平

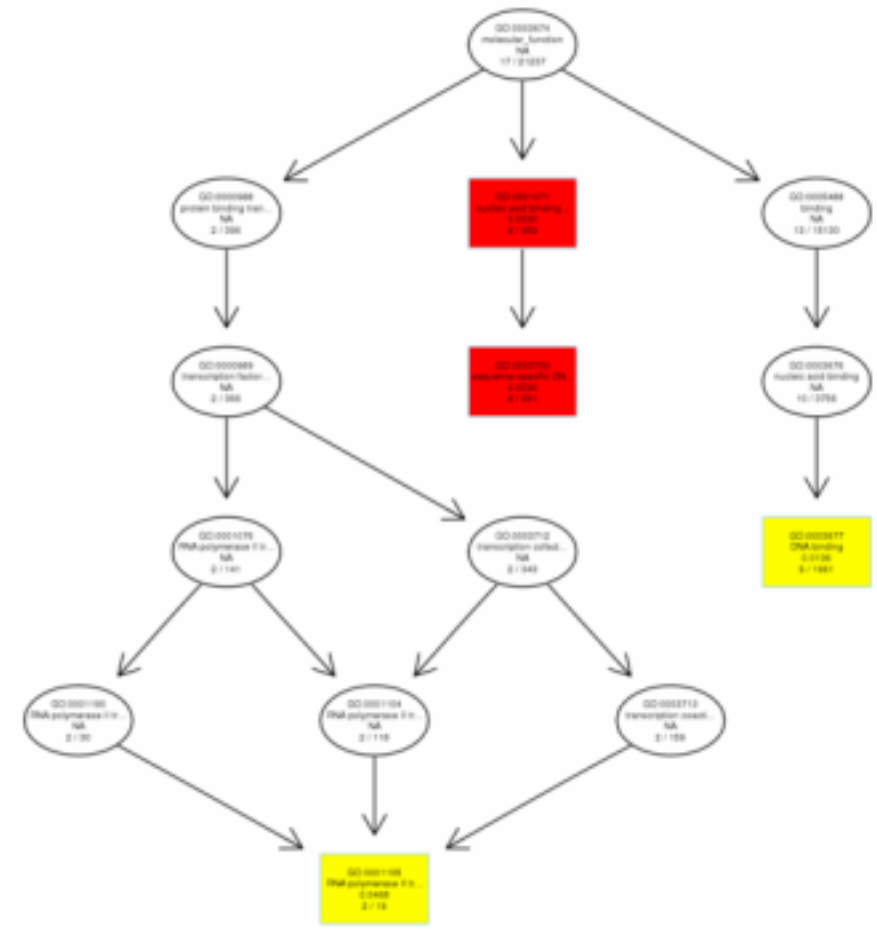
(5) Corrected_pValue ：矫正后的 P-Value，一般情况下， P-value < 0.05 该功能为富集项

(6) DEG_item ：与该GO相关的靶基因的数目

(7) DEG_list ：GO注释的靶基因数目

有向无环图 (Directed Acyclic Graph ，DAG)为差异基因 GO富集分析结果的图形化展示方式，分支代表包含关系，从上至下所定义的功能范围越来越小，一般选取 GO富集分析的结果前 10位作为有向无环图的主节点，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。我们的项目中分别绘制生物过程 (biological process) 、分子功能 (molecular function) 和细胞组分 (cellular component) 的DAG图。





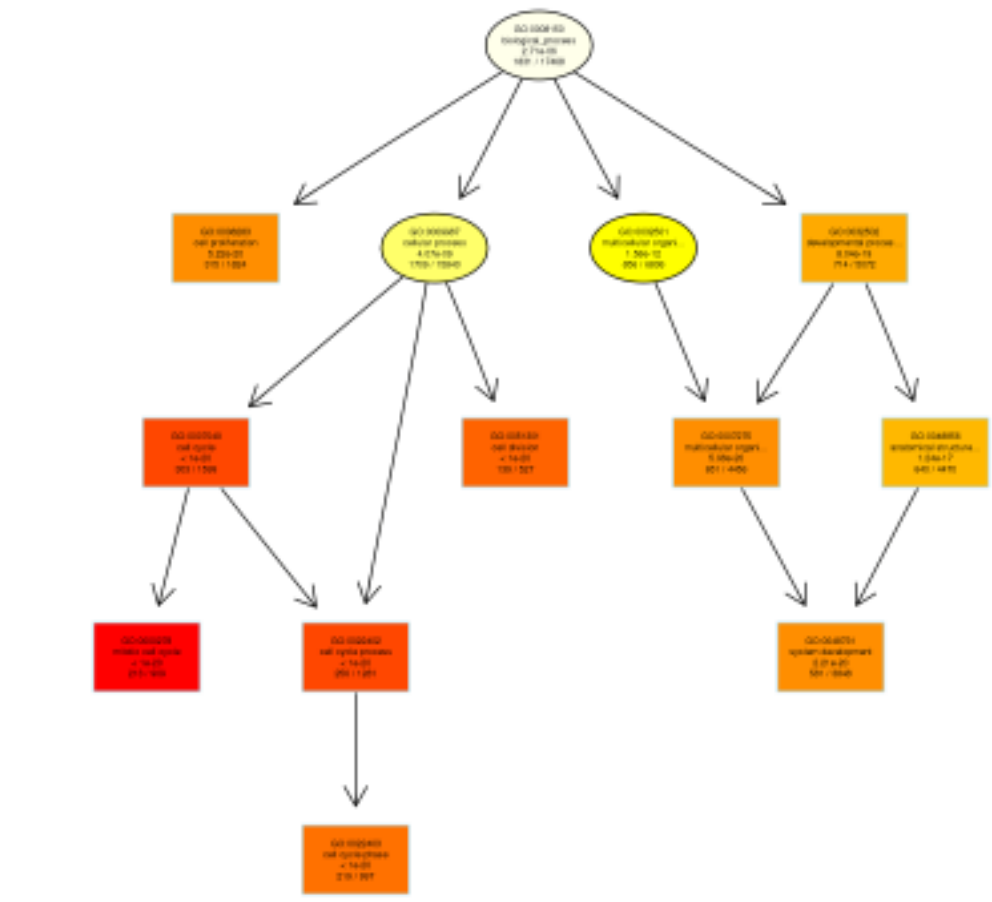


图13.1.2 GO富集有向无环图

每个节点代表一个 GO术语，方框代表的是富集程度为 TOP10的GO，颜色的深浅代表富集程度，颜色越深就表示富集程度越高，每个节点上展示了该 TERM的名称及富集分析的p-value

靶基因基因 GO富集柱状图，直观的反映出在生物过程 (biological process) 、细胞组分 (cellular component) 和分子功能 (molecular function) 富集的GO term上靶基因的个数分布情况。我们挑选了富集最显著的 30个GO term在图中展示，如果不足 30条，则全部展示。

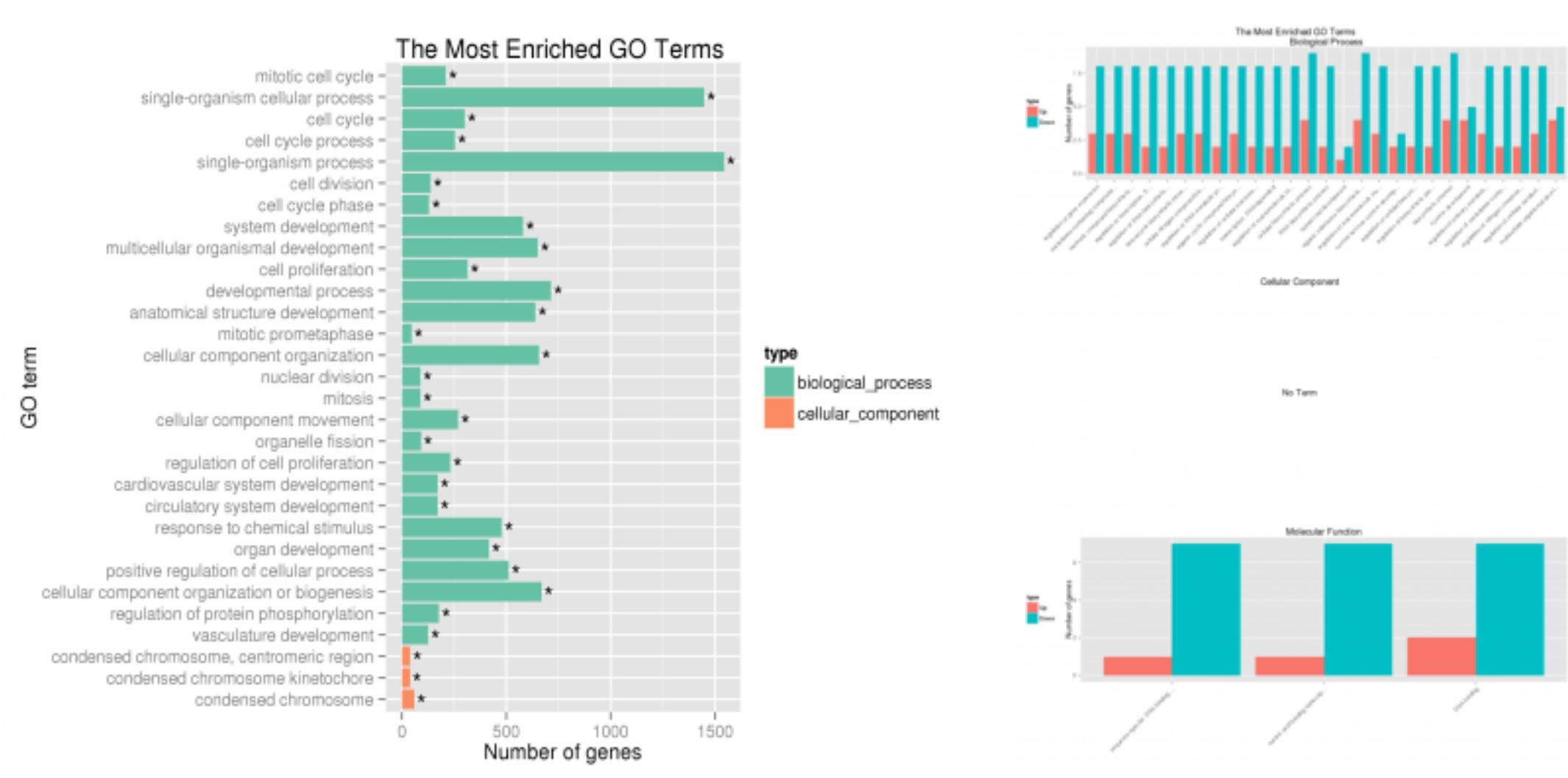


图13.1.3 GO富集柱状图

每组两张图；左图：纵坐标为富集的 GO term，横坐标为该 term中靶基因个数。不同颜色用来区分生物过程、细胞组分和分子功能，带“ * ”为富集的 GOterm右图：对图一中的 GO 按生物过程、细胞组分和分子功能三大类别及差异基因上下调分类画的三个子图

13.2 KEGG富集分析

在生物体内，不同基因相互协调行使其生物学功能，通过 Pathway显著性富集能确定靶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，在靶基因中显著性富集的 Pathway。该分析的计算公式：

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

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。 FDR≤0.05 的 Pathway 定义为在差异表达基因中显著富集的 Pathway，我们使用 KOBAS（2.0 ）进行 Pathway富集分析。

表13.2.1 差异基因 KEGG富集列表

#Term	Database	Id	Sample number	Background number	P-Value	Corrected P-Value
Glycosphingolipid biosynthesis	KEGG PATHWAY	ssc00601	1	24	0.00514620071337	1.0
TGF-beta signaling pathway	KEGG PATHWAY	ssc04350	1	75	0.0160159062787	1.0
Axon guidance	KEGG PATHWAY	ssc04360	1	111	0.0236348359067	1.0
Metabolic pathways	KEGG PATHWAY	ssc01100	1	1085	0.213505456529	1.0

结果表格详细内容如下：

(1) #Term ：KEGG通路的描述信息。
(2) Id ：KEGG数据库中通路唯一的编号信息。
(3) Sample number ：该通路下靶基因的个数。
(4) Background number ：该通路下基因的个数。
(5) P-value ：富集分析统计学显著水平。
(6) Corrected P-value ：矫正后的统计学显著水平，一般情况下， P-value < 0.05 该功能为富集项。

散点图是 KEGG富集分析结果的图形化展示方式。在此图中， KEGG富集程度通过 Rich factor 、Qvalue和富集到此通路上的基因个数来衡量。其中Rich factor 指差异表达的基因中位于该 pathway条目的基因数目与所有有注释基因中位于该 pathway条目的基因总数的比值。 Rich factor 越大，表示富集的程度越大。 Qvalue是做过多重假设检验校正之后的 Pvalue，Qvalue的取值范围为 [0,1] ，越接近于零，表示富集越显著。我们挑选了富集最显著的 20条pathway条目在该图中进行展示，若富集的 pathway条目不足 20条，则全部展示。

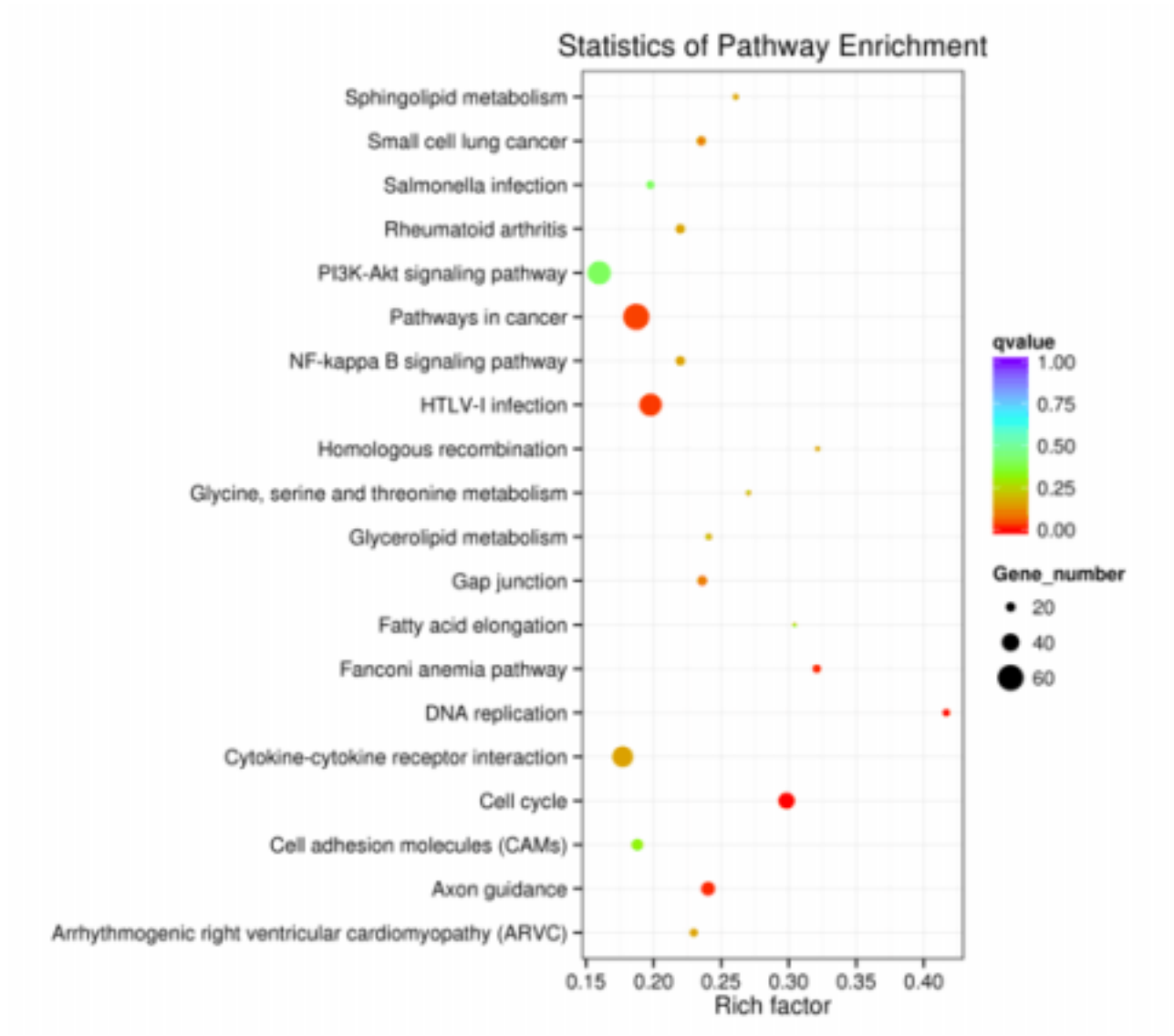


图13.2.2 富集的 KEGG代谢通路的散点图

纵轴表示 pathway名称，横轴表示 Rich factor ，点的大小表示此 pathway中差异表达基因个数多少，而点的颜色对应于不同的 Qvalue范围；

将差异基因富集出的通路图展示出来，该通路图中，包含上调基因的 KO节点标红色，包含下调基因的 KO节点标绿色，包含上下调的标黄色。鼠标悬停于标记的 KO节点，弹出差异基因细节框，标色同上，括号中数字为 log2(Fold change) 。以上步骤可脱机实现，如连接互联网，点击各个节点，可以连接到 KEGG官方数据库中各个 KO的具体信息页。

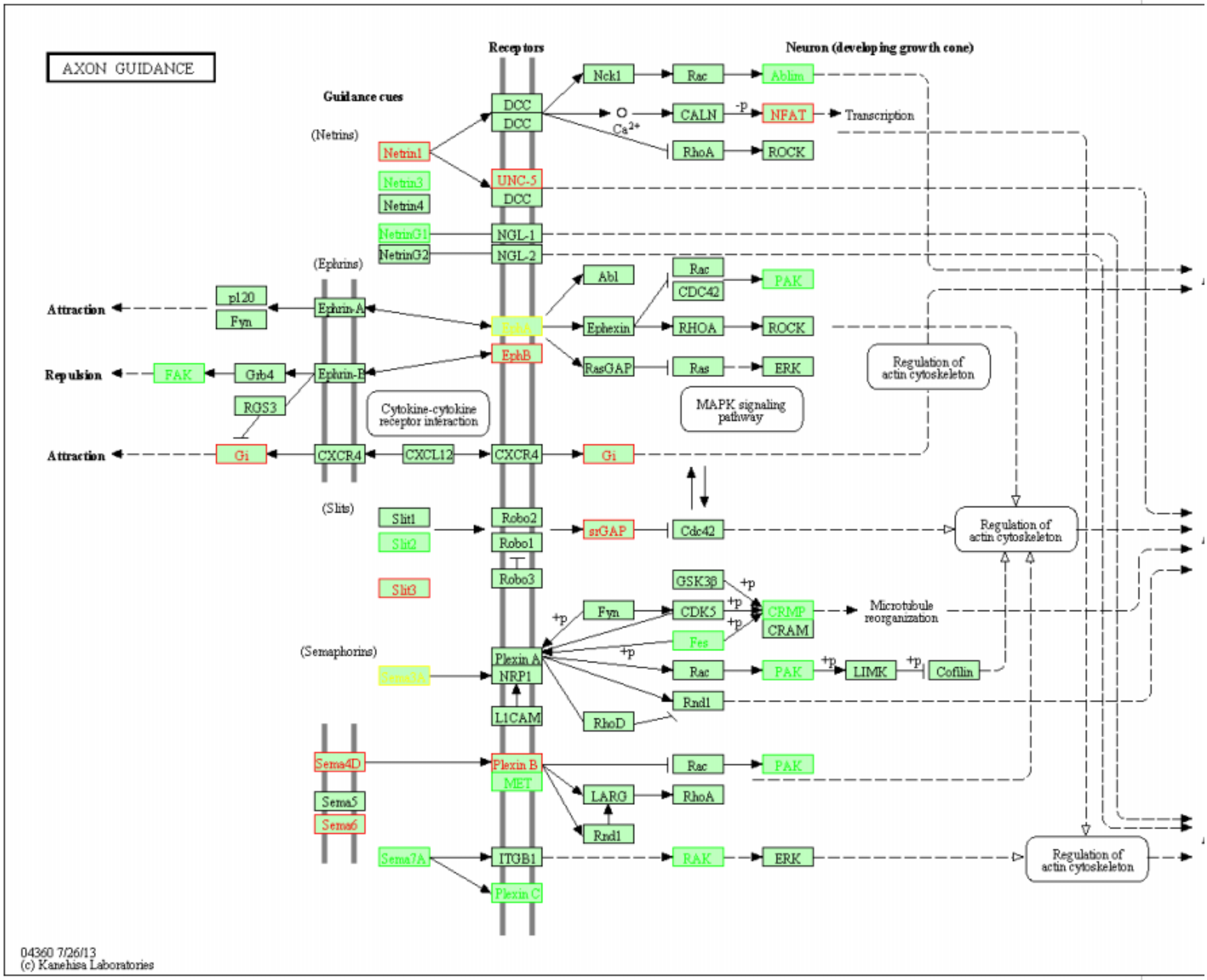


图13.2.3 显著富集的 KEGG pathway代谢通路图

KEGG代谢通路图中，包含上调基因的 KO节点标红色，包含下调基因的 KO节点标绿色，包含上下调的标黄色。在网页界面上，鼠标悬停于标记的 KO节点，弹出差异基因细节框，标色同上，括号中数字为 log2(Fold change) 。

14 特异 lncRNA与mRNA网络互作分析

lncRNA与mRNA可以通过靶向关系进行关联，mRNA和mRNA之间可以通过蛋白质互作关系进行关联，从而可以形成 lncRNA-mRNA-protein的互作网络关系。

mRNA和mRNA之间主要应用 STRING蛋白质互作数据库（ http://string-db.org/ ）中的互作关系，针对数据库中包含的物种，直接从数据库中提取出目标基因集（比如差异基因 list） 的互作关系构建网络。

我们提供特异 lncRNA与靶基因，靶基因蛋白互作网络数据文件，此文件可以直接导入 Cytoscape软件进行可视化编辑。 Cytoscape软件使用方法可参考我们提供的使用说明文档（ CytoscapeQuickStart.pdf ）。客户可以针对一些网络的拓扑属性进行统计和标示作图，比如：互作网络图中节点(node) 的大小与此节点的度 (degree) 成正比，即与此节点相连的边越多，它的度越大，节点也就越大，这些节点在网络中可能处于较为核心的位置。节点的颜色与此节点的聚集系数 (clustering coefficient) 相关，颜色梯度由绿到红对应聚集系数的值由低到高；聚集系数表示此节点的邻接点之间的连通性好坏，聚集系数值越高表示此节点的邻接点之间的连通性越好等等。根据不同的研究目的和需求，客户还可以在网络图中进行调整节点位置和颜色、标注表达量水平等操作。需要注意的是，通过 blast 比对得到的结果不能保证较好的准确性，这部分的工作只是给客户提供参考，辅助客户发现一些可能的重要的基因。按我们提供的使用说明将文件导入 Cytoscape软件后的效果图如下：

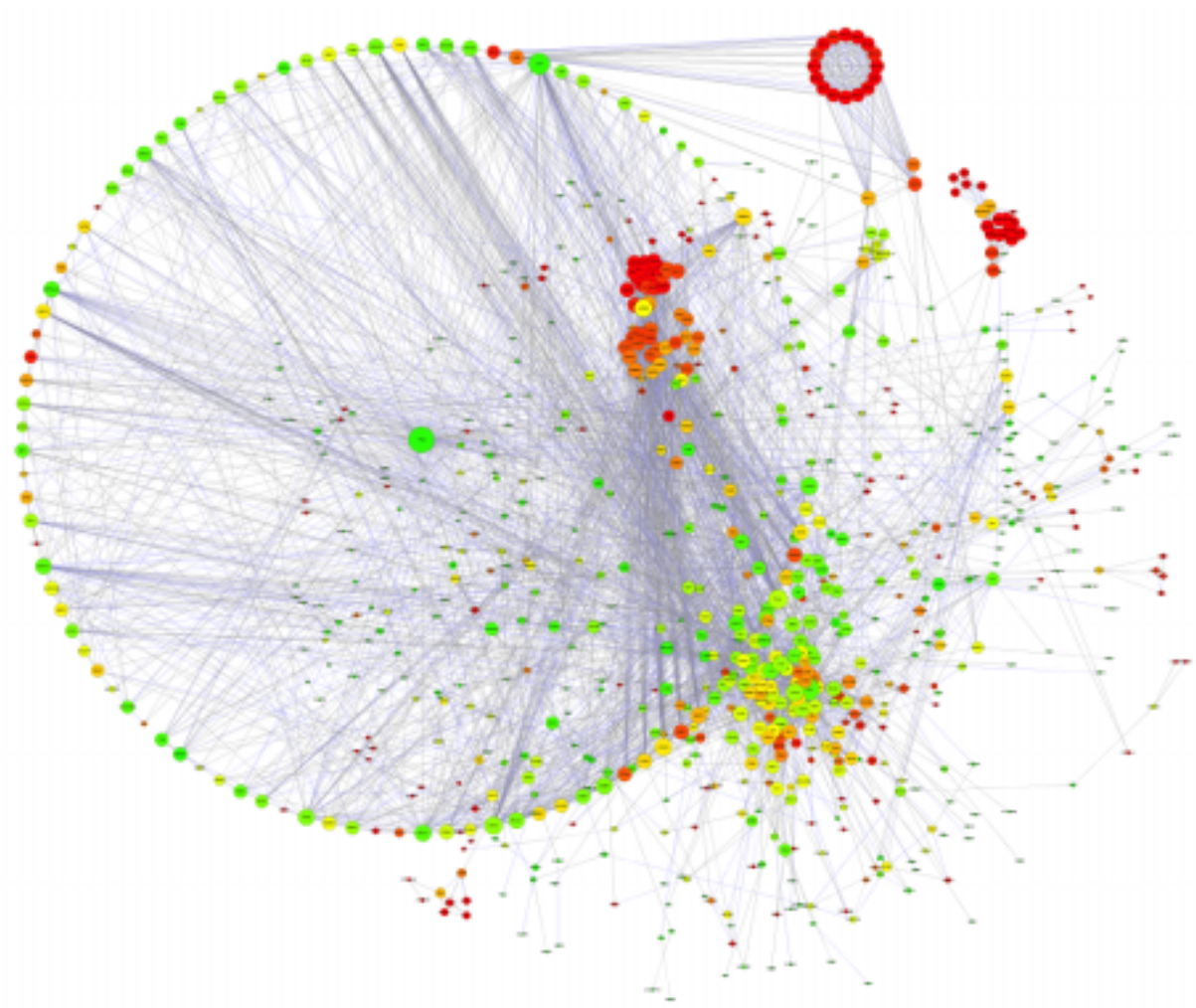


图14 Cytoscape软件界面

四、参考文献

Anders, S.(2010). HTSeq: Analysing high-throughput sequencing data with Python.(HTSeq)

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol.(DESeq)

Anders, S. and Huber, W. (2012). Differential expression of RNA-Seq data at the gene level-the DESeq package.(DEseq)

Anders S, Reyes A, Huber W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Research. (DEXSeq)

Kanehisa, M., M. Araki, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic acids research.(KEGG)

Kim, D., G. Pertea, et al. (2012). TopHat2: Parallel mapping of transcriptomes to detect InDels, gene fusions, and more.(TopHat2)

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol.(Bowtie)

Langmead, B. and S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. Nature methods.(Bowtie 2)

Mao, X., Cai, T., Olyarchuk, J.G., Wei, L. (1995). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics.(KOBAS)

Guttman, M., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature biotechnology.(scripture)

Mortazavi, A., B. A. Williams, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods.

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics.(edgeR)

Langfelder, P.,Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. (coexpression)

Robinson, M. D. & Oshlack, A. (2010)A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol.(DEGSeq)

Trapnell, C. et al. (2010).Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol.(Cufflinks)

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics.(TopHat)

Trapnell, C., A. Roberts, et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. nature protocols.(Tophat & Cufflinks)

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

Wang, Z., M. Gerstein, et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010).Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology.(GOseq)