单位 +HhocRRNA分析网页版结题报告

GENOME.cn

安诺基因

ANNOROAD 安诺优达基因科技（北京）有限公司

2015/12/29

目录

1 项目信息

1.1 基本思想

安诺优达 lncRNA测序，基于 Illumina 测序平台，鉴定某个物种在特定组织或者特定时期下表达的 Novel lncRNA，检测 mRNA 已知 lncRNA和novel lncRNA的表达量，并针对实际样品信息采用灵活的差异分析策略可以找到生物体不同时期、不同组织或不同个体间差异表达的 mRNA和lncRNA。对于 mRNA 进行功能注释，进而得到 mRNA在生物体中参与生命活动的一个清晰的生物信息图谱,mRNA的深层分析包括差异表达分析、可变剪接分析、新转录本预测和变异分析等其他个性化分析；对于 lncRNA，深层分析包括保守性分析、靶标预测、功能预测等功能分析。

1.2 实验流程

1.2.1 样本检测

安诺优达对总 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

每个样品取 3μg总RNA作为起始量构建 lncRNA文库。使用 Ribo-Zero? Gold Kits 去除样品中的 rRNA，根据 NEB Next Ultra Directional RNA Library Prep Kit for Illumina （NEB，Ispawich，USA）的操作说明分别选取不同的index 标签建库。

文库构建的具体步骤为：首先使用试剂盒去除核糖体 rRNA，向反应体系中加入 Fragmentation Buffer 使 RNA片断化成为短片段，再以片断后的 RNA 为模板，用六碱基随机引物 （Random Hexamers） 合成 cDNA 第一链，并加入缓冲液、dNTPs、RNase H 和 DNA Polymerase I 合成 cDNA 第二链，经过 QiaQuick PCR试剂盒纯化并加 EB缓冲液洗脱经末端修复、加碱基 A，加测序接

头，经琼脂糖凝胶电泳回收目的大小片段，加 UNG酶消化 cDNA二链，并进行 PCR扩增，最后琼脂糖凝胶电泳回收目的大小片段，从而完成整个文库制备工作。

构建好的文库用 Illumina HiSeq / NextSeq 500 进行测序。测序策略为PE100。

其实验流程如下：

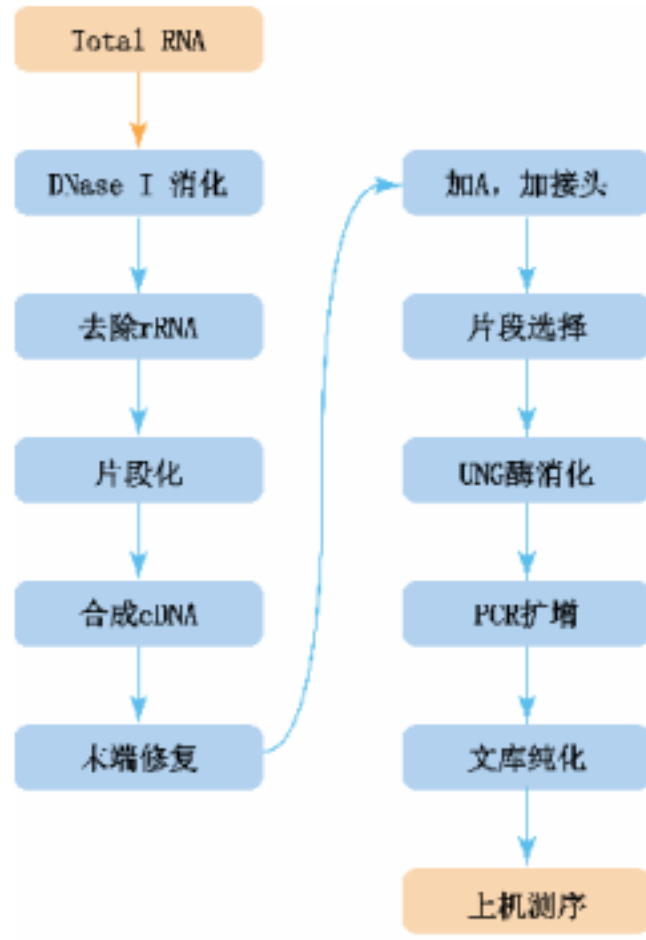


图1 实验流程图

1.3 信息分析流程

Illumina HiSeq / NextSeq 500 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。

安诺优达 lncRNA测序信息分析流程主要分为三部分：测序数据质控、数据比对分析和 lncRNA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；lncRNA深层分析包括差异表达分析、保守性分析、靶标预测、功能预测等其他个性化分析。具体的信息分析流程图如下：

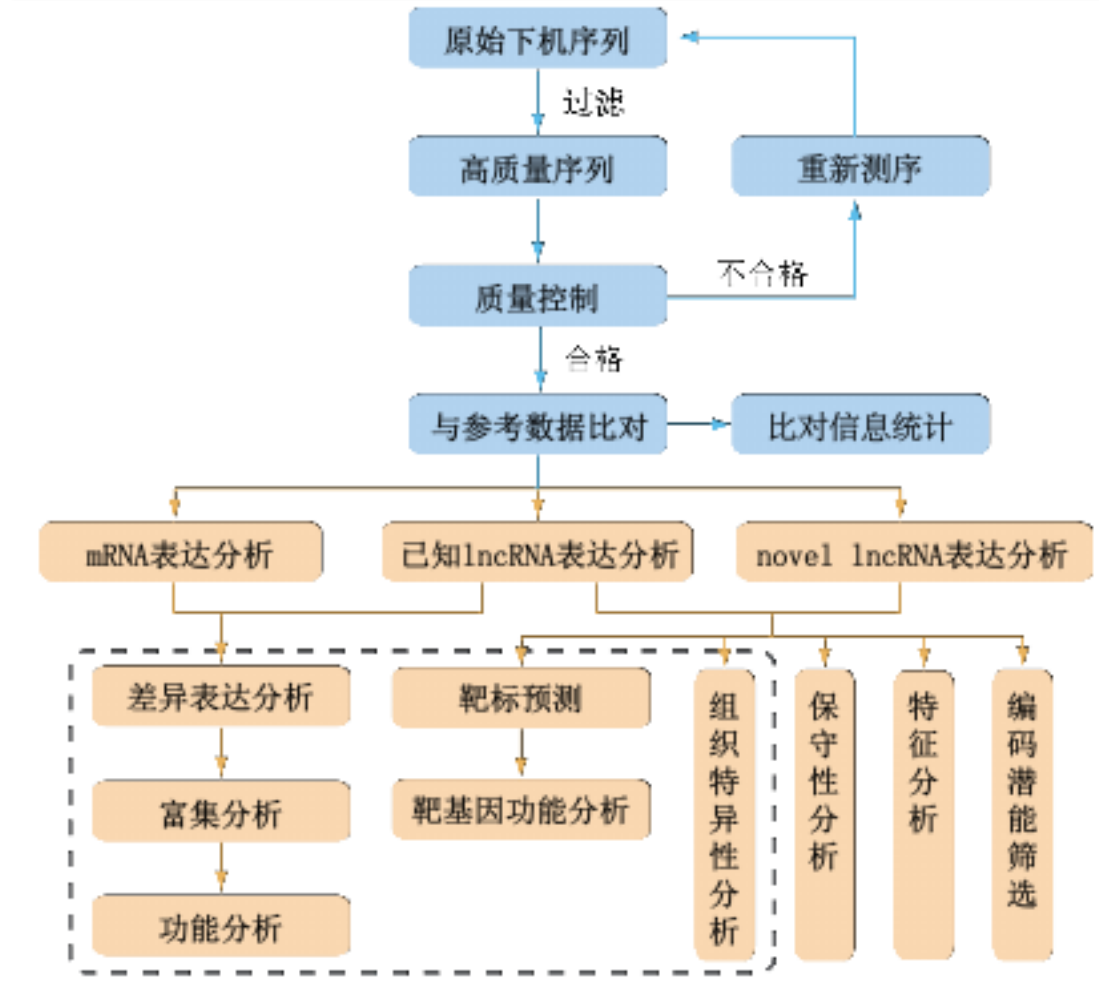


图2　信息分析流程图

如项目仅有一个样品，无法进行虚线所示的分析内容；如果项目样品数小于　　3，无法进行靶标预测中的反式靶标预测和组织特异性分析。

1.4　样品信息

本项目共 3个样本，样品信息示例如下：

表1　样品信息

Sampl e	sam pl e 1
Group	group1
Description	description

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。

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：

```
@HWI-ST1268:544:H8YO2ADXX:1:1101:1480:2221 1:N:0
TATGGTTGGCTGTTCACGGCCTGGAATTCTCGGGTGCCAAGGAACTCCA
+
1:BDFFDFFHHHHJJJJIJJJJJJIJJJIHIJJJHIDIJJJJJJJIIJJ
```

图3 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，我们对原始下机数据序列进行过滤，得到高质量的Clean Reads，再进行后续分析，后续分析都基于 Clean Reads。

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）；

（2）去除低质量的 Reads（Reads中质量值 Q≤5的碱基占总碱基的 15%以上）；

（3）去除含 N比例大于 5% 的Reads；

（4）去除与核糖体 RNA（rRNA）匹配的 Reads。

Q值是Phred Quality Score的简称；N碱基是指未知碱基。对于双端测序，有一端Reads不满足以上任意一个条件，就去除该 Reads。

数据过滤统计结果见下表：

表2　数据过滤统计分析表

#Samples	Sample1
Raw Reads Number	67,301,560
Raw Bases Number	8,412,695,000
Clean Reads Number	58,039,376
Clean Reads Rate (%)	86.2400
Clean Bases Number	7,254,922,000
Low-quality Reads Number	8,668,816
Low-quality Reads Rate (%)	12.8800
Ns Reads Number	17,872
Ns Reads Rate (%)	0.0300
Adapter Polluted Reads Number	575,496
Adapter Polluted Reads Rate (%)	0.8600
Raw Q30 Bases Rate (%)	91.9500
Clean Q30 Bases Rate (%)	96.3700
rRNA Mapping Reads Number	2,499,397
rRNA Mapping Rate(%)	4.3100
Total Clean Reads number	55,478,360
Total Clean Bases number	6,934,795,000
Total Q30(%)	96.3900

（1）Raw Reads Number: 原始下机的 Reads数；

（2）Raw Bases Number: 原始下机序列的碱基数；

（3）Clean Reads Number：过滤后得到的高质量的 Reads数；

（4）Clean Reads Rate (%) ：过滤后得到的高质量序列占 Raw Reads的比例。这个值越大，说明测序质量或者文库质量越好；

（5）Clean Bases Number ：过滤后的高质量序列的碱基数；

（6）Low-Quality Reads Number ：去除低质量的 Reads数；

（7）Low-Quality Reads Rate (%) ：去掉低质量的 Reads占Raw Reads的比例；

（8）Ns Reads Number：去掉含 N比例大于 5%的Reads数；

（9）Ns Reads Rate (%) ：去掉含 N比例大于 5%的Reads占Raw Reads的比例；

（10）Adapter Polluted Reads Number ：去掉接头污染的 Reads数；

（11）Adapter Polluted Reads Rate (%) ：去掉接头污染的 Reads占Raw Reads的比例；

（12）Raw Q30 Bases Rate (%) ：Raw Reads中测序质量值大于 30（错误率小于 0.1%）的碱基占总碱基（ Raw Reads） 的比例；

（13）Clean Q30 Bases Rate (%) ：Clean Reads中测序质量值大于 30（错误率小于 0.1%）的碱基占总碱基（ Clean Reads）的比例；

（14）rRNA Mapping Reads Number：核糖体 RNA污染的 Reads数；

（15）rRNA Mapping Rate (%) ：核糖体 RNA污染的 Reads占Raw Reads的比例；

（16）Total Clean Reads Number ：去掉核糖体 RNA污染的 Reads数；

（17）Total Clean Bases Number ：去掉核糖体 RNA污染的碱基数；

（18）Total Q30(%) ：去掉核糖体之后 Clean Reads中测序质量值大于 30（错误率小雨 0.1%）的碱基占总碱基（ Total Clean Reads ）的比例。

Q30可反应测序的碱基质量水平，本项目全部样本 Q30比例如下图：

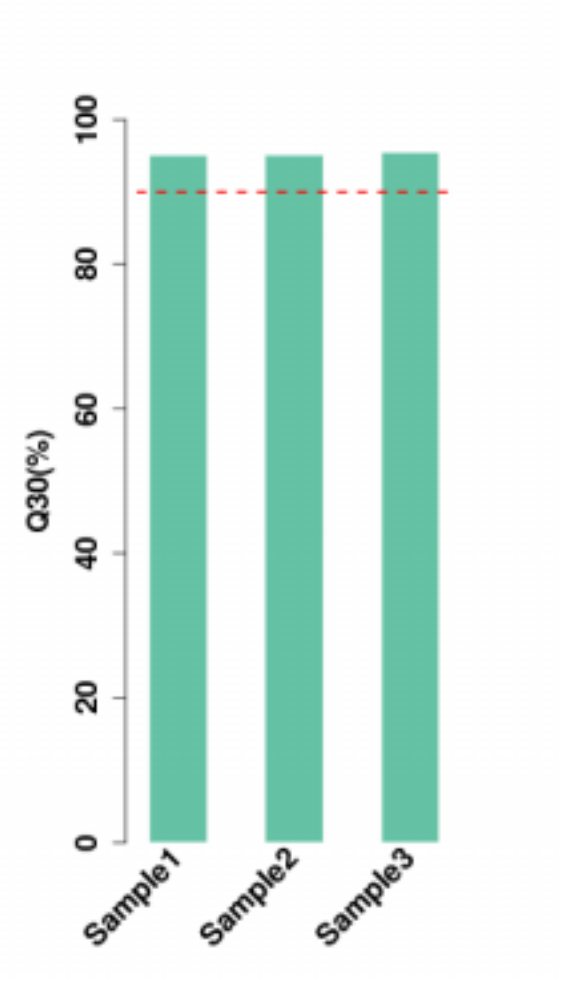


图4 Q30 质控图

所有样品过滤前各种 Reads比例分布如下图：

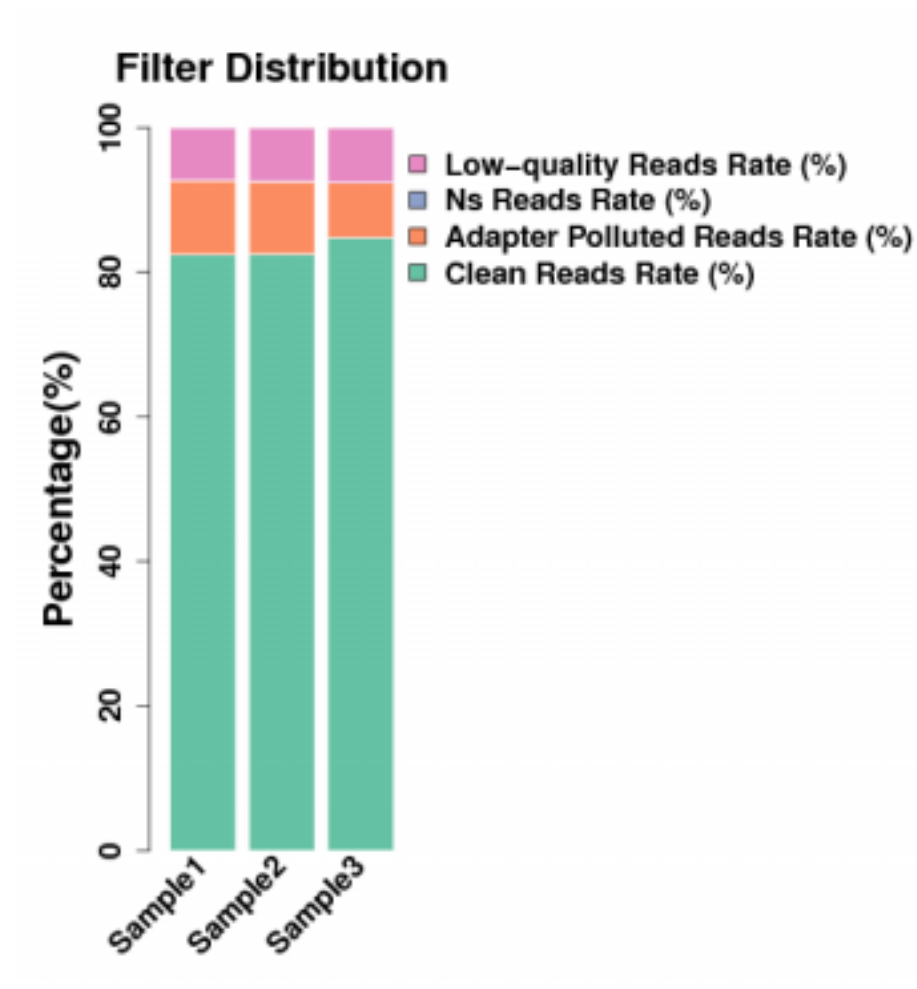


图5 过滤分布图

单个样本过滤前各种 Reads分布饼图如下：

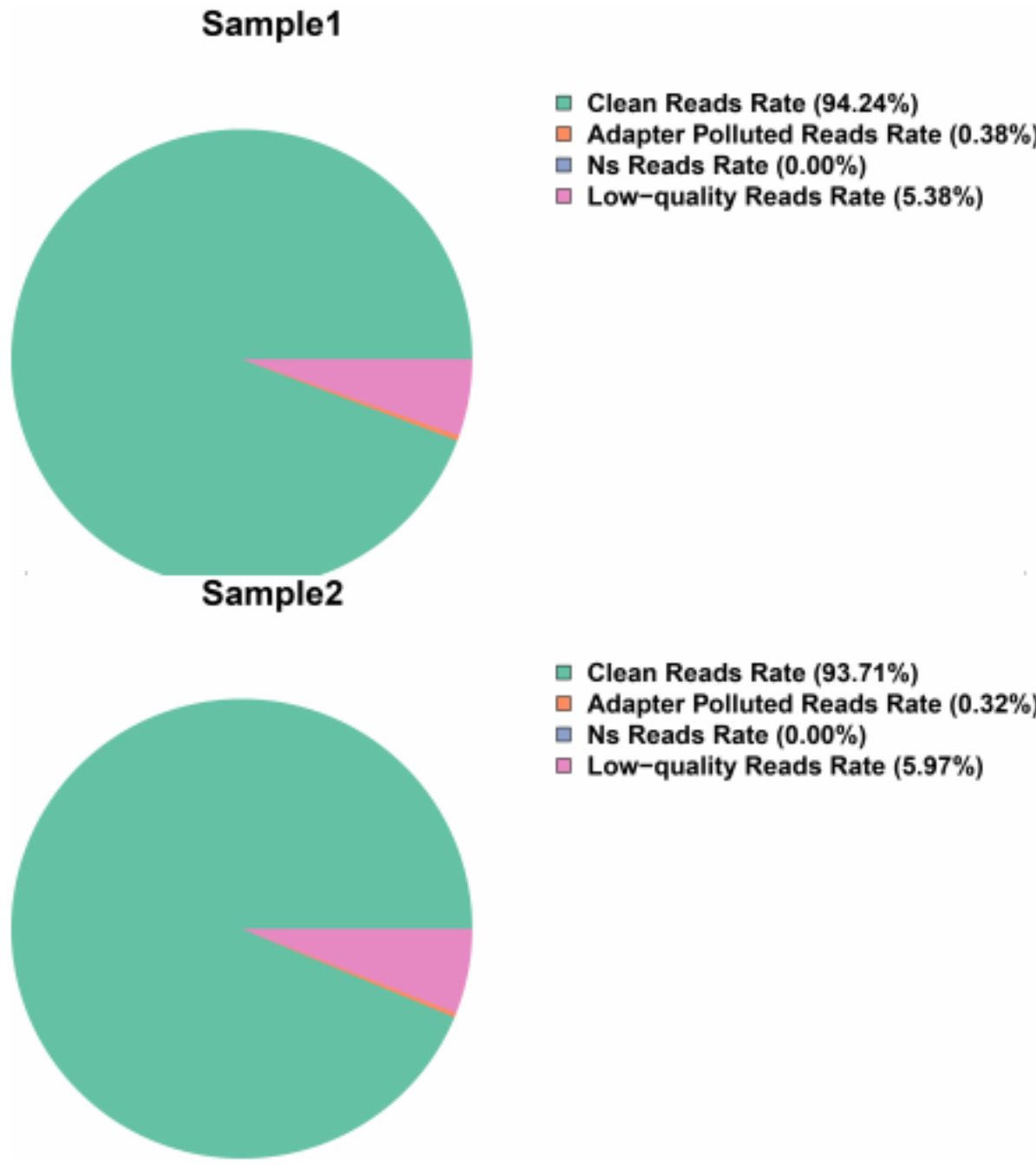


图6　单样品 Reads分布

本项目所有样本 Clean Data 数据量水平如下图：

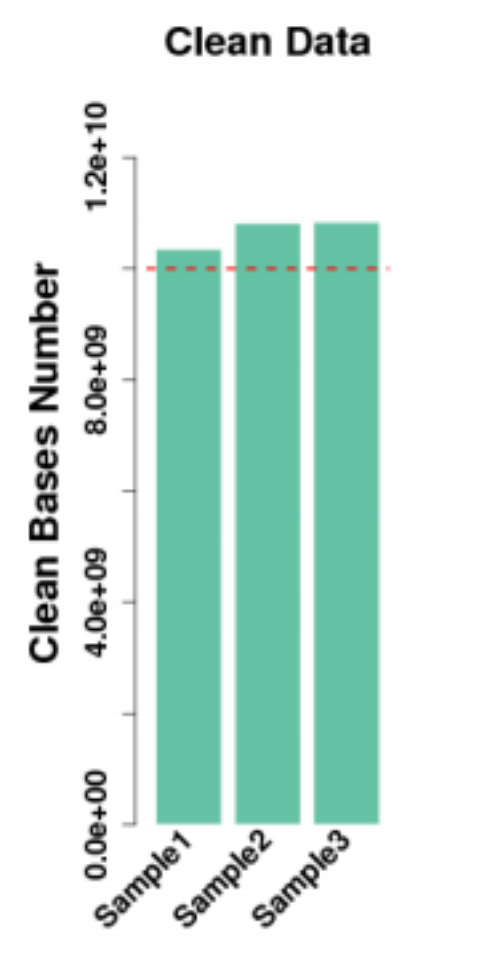


图7　Clean Data　　数据量分布

2.3　测序质量分布

测序错误率与碱基测序质量有关，受测序仪本身、测序试剂、样品等多个因素共同影响。每个碱基测序错误率是通过　Phred数值（Phred Score，Qphred）通过公式转化得到，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，对应关系如下表所示：

表3　Illumina Casava　　碱基识别与　Phred　分值之间的简明对应关系表

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

对于RNA-Seq技术，碱基质量值分布具有两个特点：

（1）碱基质量值会随着测序序列（　Sequenced Reads）长度的增加而降低，这个特点是 Illumina　高通量测序平台都具有的特征；

（2）前6个碱基的测序质量值较其他位置会低一些，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。

以过滤后高质量序列的碱基位置作为横坐标，每个位置的平均测序质量值作为纵坐标，得到下面的测序质量分布图：

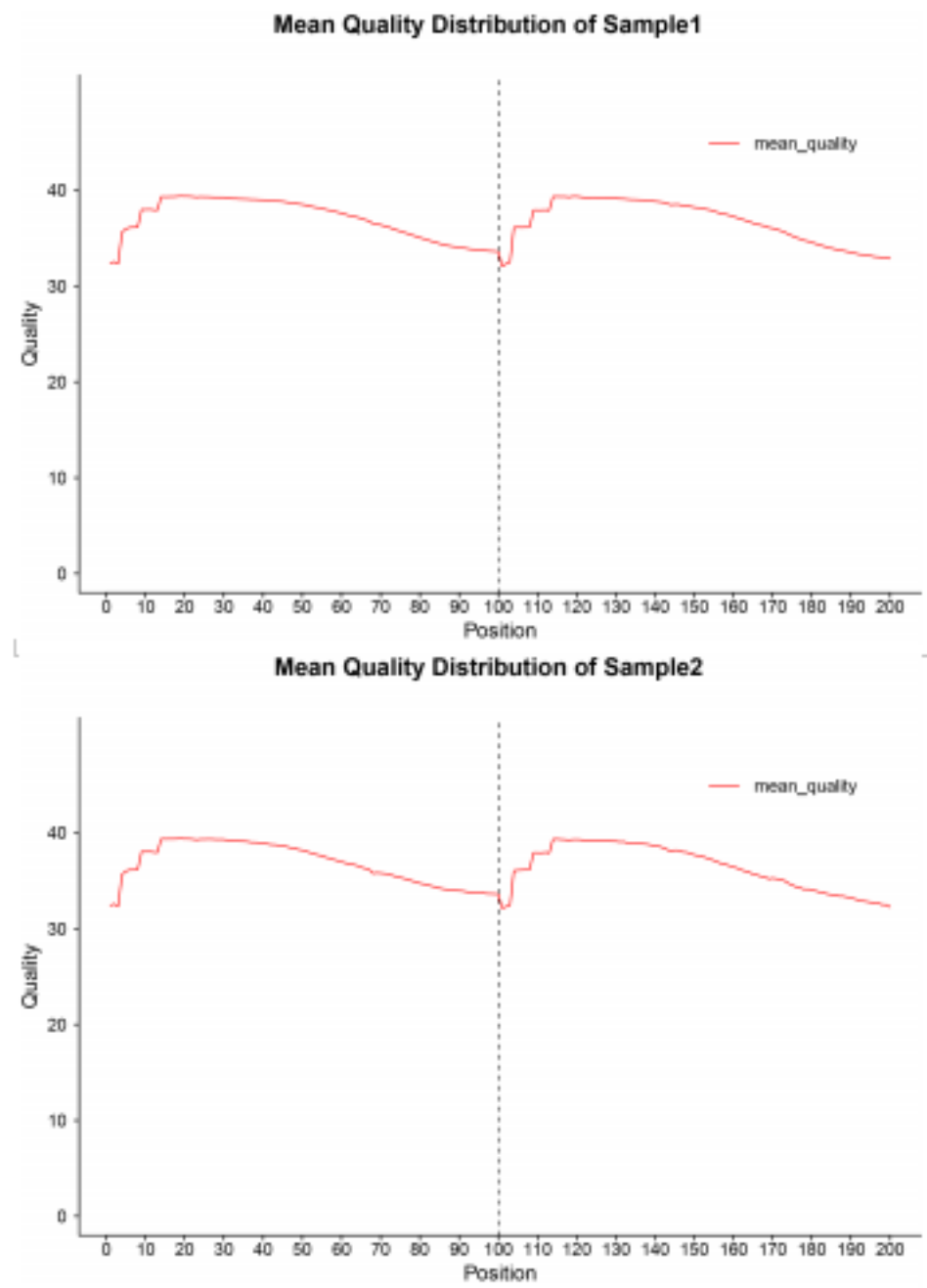


图8　样品测序质量分布图

2.4　测序碱基分布

碱基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 lncRNA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。这种偏好性与测序的物种和实验室环境无关，但会影响 lncRNA测序的均一化程度。除此之外，理论上 G碱基和C碱基及 A碱基和 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。

以过滤后序列的碱基位置作为横坐标，以每个位置的 ATCG碱基含量的比例作为纵坐标，得到碱基含量分布图。

图9 样品碱基含量分布图

3 比对分析

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，使其定位到基因组。对动植物样本，一般采用 TopHat软件比对 (Trapnell ,et al., 2009)。TopHat是专门用于数据比对的软件，其优点在于可将前期比对上的序列切分后进行二次比对，从而达到鉴定 Exon-Exon剪接位点的目的。 Tophat软件进行序列比对时调用软件Bowtie2(Langmead ,et al., 2009)对序列进行比对，使比对更加准确快速。 TopHat比对原理示意图如下：

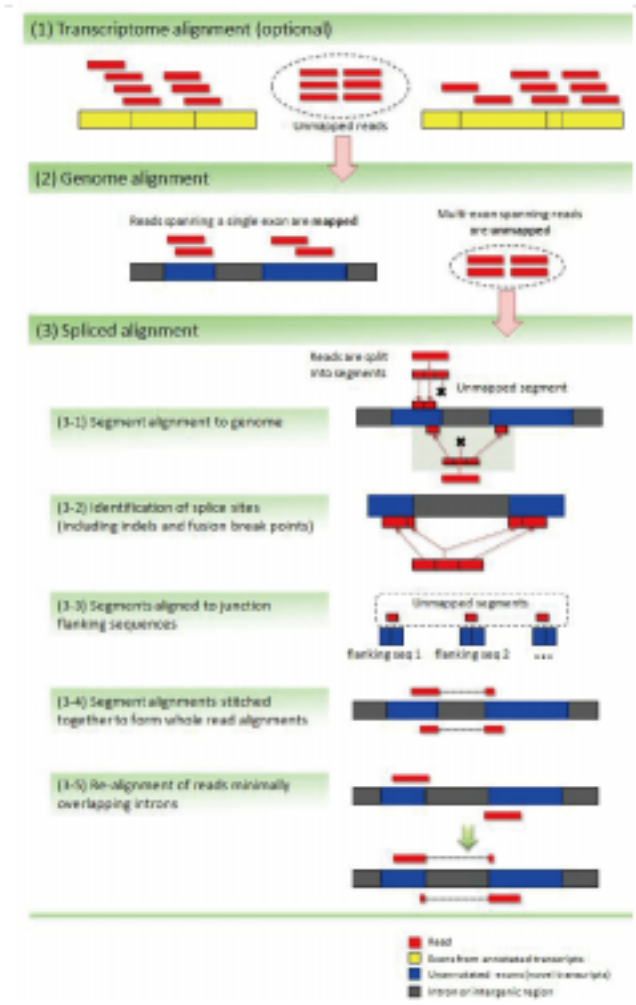


图10　Tophat比对分析原理

本项目分析中，我们使用　TopHat版本号为 v2.0.12 ，选用默认参数。

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，比对率通常都会大于 80%。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示意图表：

表4　样品的比对率统计表

Li br ar y	Sampl e1
Total Reads	102,287,578
Mapped Reads	98,284,753
Mapping Rate	0.9600
Unmapped Reads	4,002,825
MultiMap Reads	6,924,023
MultiMap Rate	0.0700

（1）Total Reads ：过滤后总的序列数；

（2）Mapped Reads: 比对上基因组的序列数；

（3）Mapping Rate ：比对上基因组的序列数的比率；

（4）Unmapped Reads: 未比对上基因组的序列数；

（5）MultiMap Reads ：比对到基因组多个位置的序列数；

（6）MultiMap Rate ：比对到基因组多个位置的序列数的比率。

比对上序列在染色体的分布指能唯一比对到基因组的 Reads在各染色体（分正负链）的密度分布。通常以 1K的滑动窗口（ Window Size ）为单位，计算该窗口中比对上的 Reads的中位数，取对数值（ $\log_2$）。一般情况下，染色体越长，定位到该染色体的 Reads数也越多。根据染色体长度与定位的 Reads数作图，可更直观展现比对上序列在染色体的分布，橙色表示正链上的分布，青色表是负链上的分布如下图：

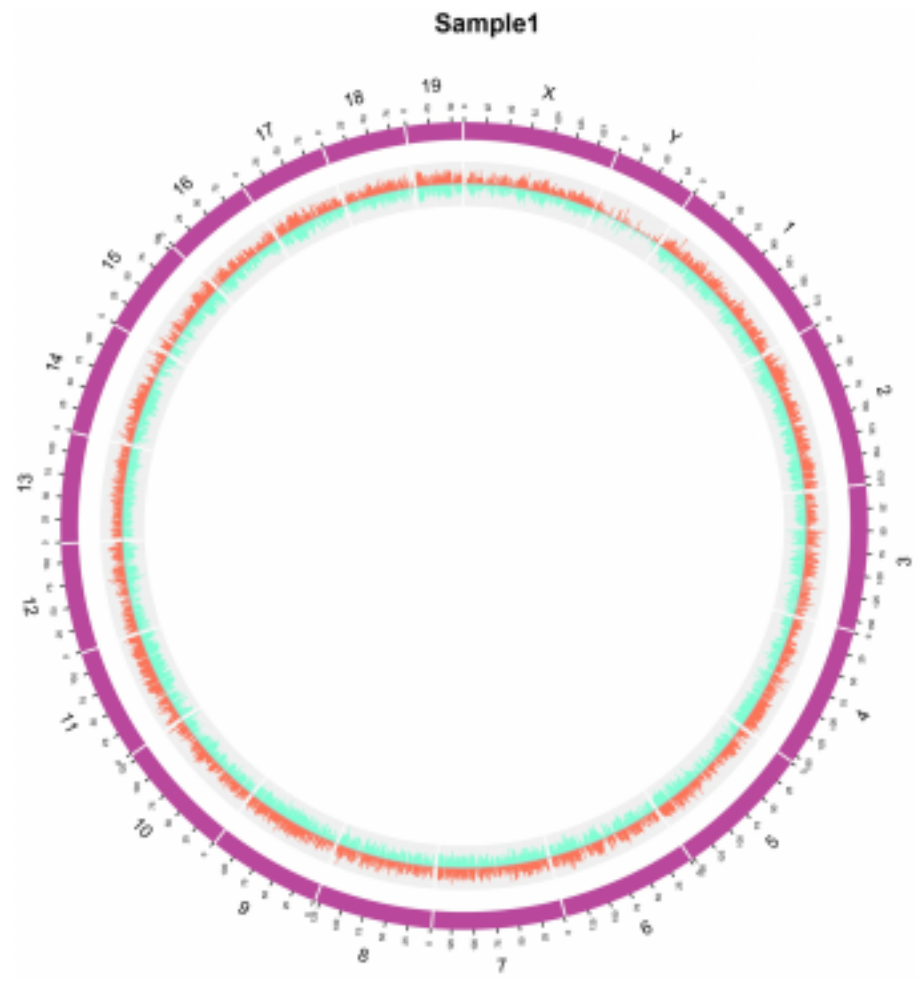


图11 唯一比对序列在参考基因组染色体的分布

3.2 基因区域分布

比对上序列在基因区域的分布是指根据相关数据库中该物种的基因注释文件，统计基因三种功能元件（ Exon, Intron 和Intergenic ）上唯一比对序列（即只比对上基因组一个位置的序列）的数目和比例。一般情况下，如果该物种的注释信息比较全面，大部分序列应该比对到 Exon区域，可变剪接、噪音表达等会导致一些序列来源于 Intron 区域，新转录本、表达噪音等会导致源于基因间区序列产生。

下表为一个样本唯一比对序列在参考基因组区域分布统计示意表：

表5　唯一比对序列在参考基因组区域分布统计

Sampl e	Sampl e1
Exon	49,992,490(68.08%)
Intron	17,426,621(23.73%)
Intergenic	6,015,880(8.19%)

下图为对应示意图：

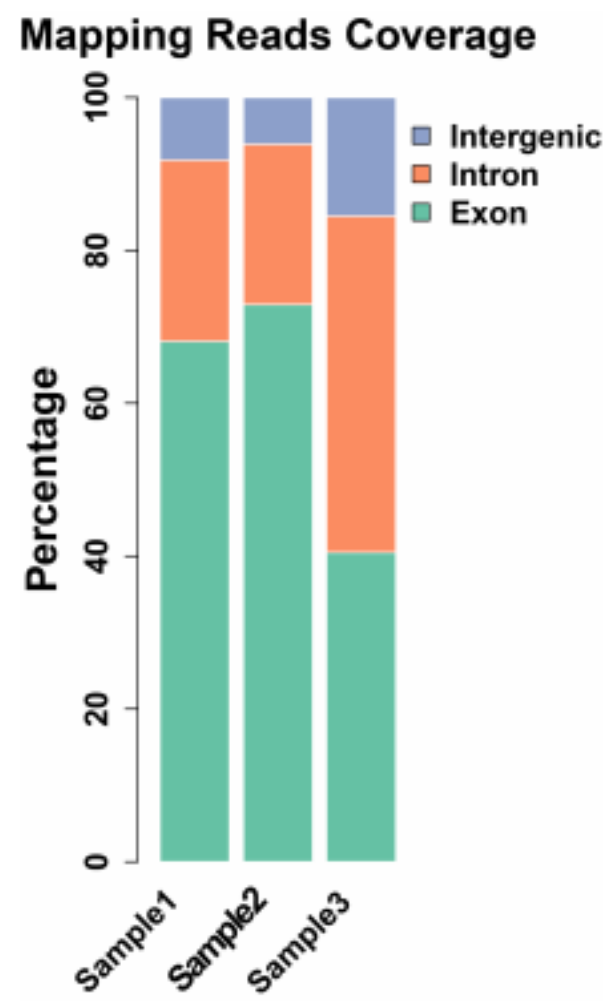


图12　唯一比对序列在参考基因组基因各区域的分布

根据序列的比对信息，分别统计比对到外显子、内含子和基因间区的序列数，并根据比例做出柱状图

单样本唯一比对序列在参考基因组各区域分布图如下：

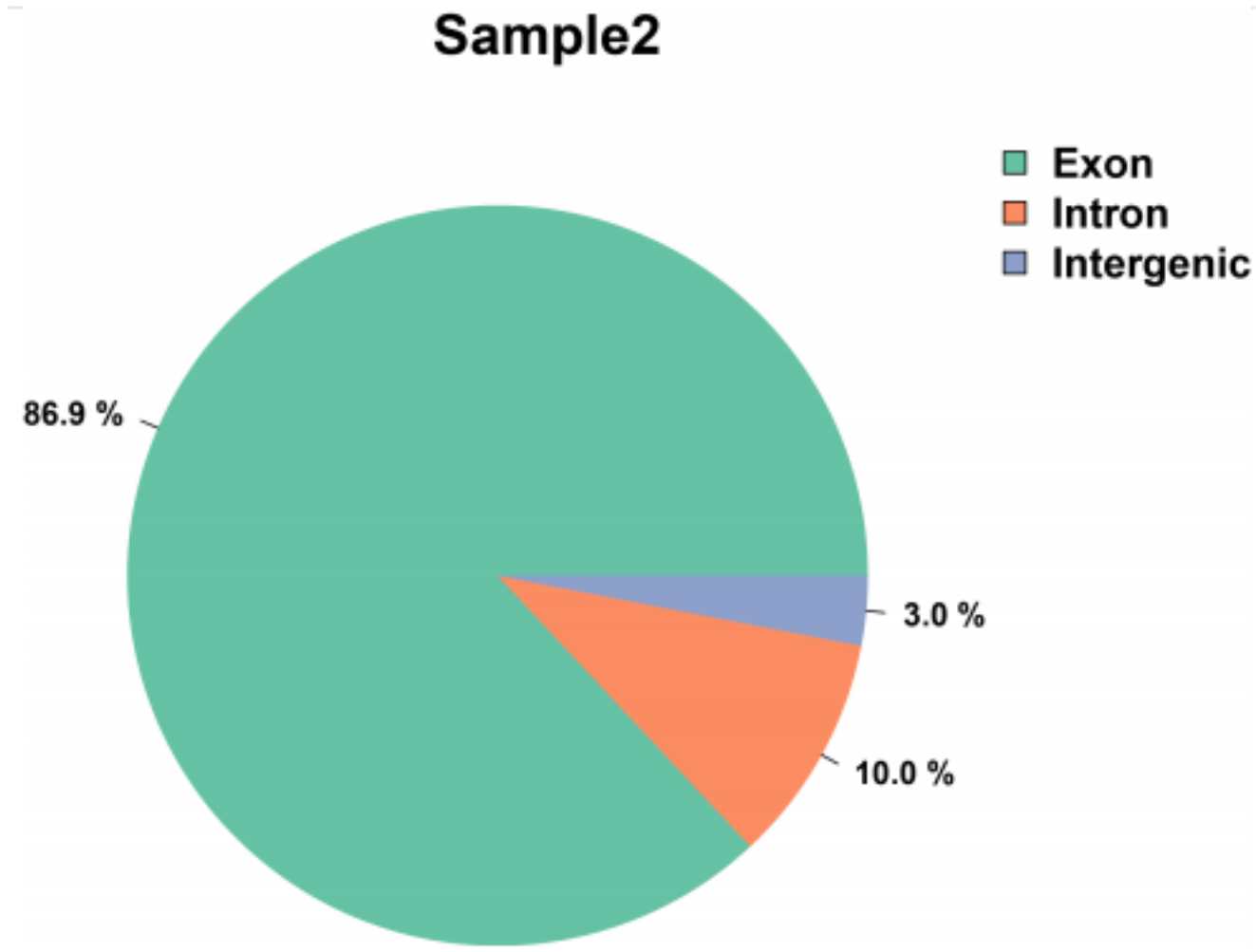


图13　单样品唯一比对序列在参考基因组基因各区域的分布

3.3 均一性分析

均一性是指测序的核酸序列的随机程度。若测序序列不偏向于基因的特定区域，则称其均一性好。若测序结果均一性很差，将直接影响 lncRNA各项分析结果。一般情况下，由于随机引物反转录的随机性和测序的随机性，整个实验对基因的扩增没有明显的偏好性，整体上应该为一条平滑的曲线。然而很多研究表明，这种均一化分布受多种因素影响。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。通过统计每个基因不同位置的深度来衡量测序结果均一性。由于不同基因长度不同，将基因外显子区平均划分为100个窗口，计算每个窗口的平均深度，并统计所有基因在该对应窗口的深度平均值，依此作出序列均一性分布图。

下图为均一性分析结果示意图：

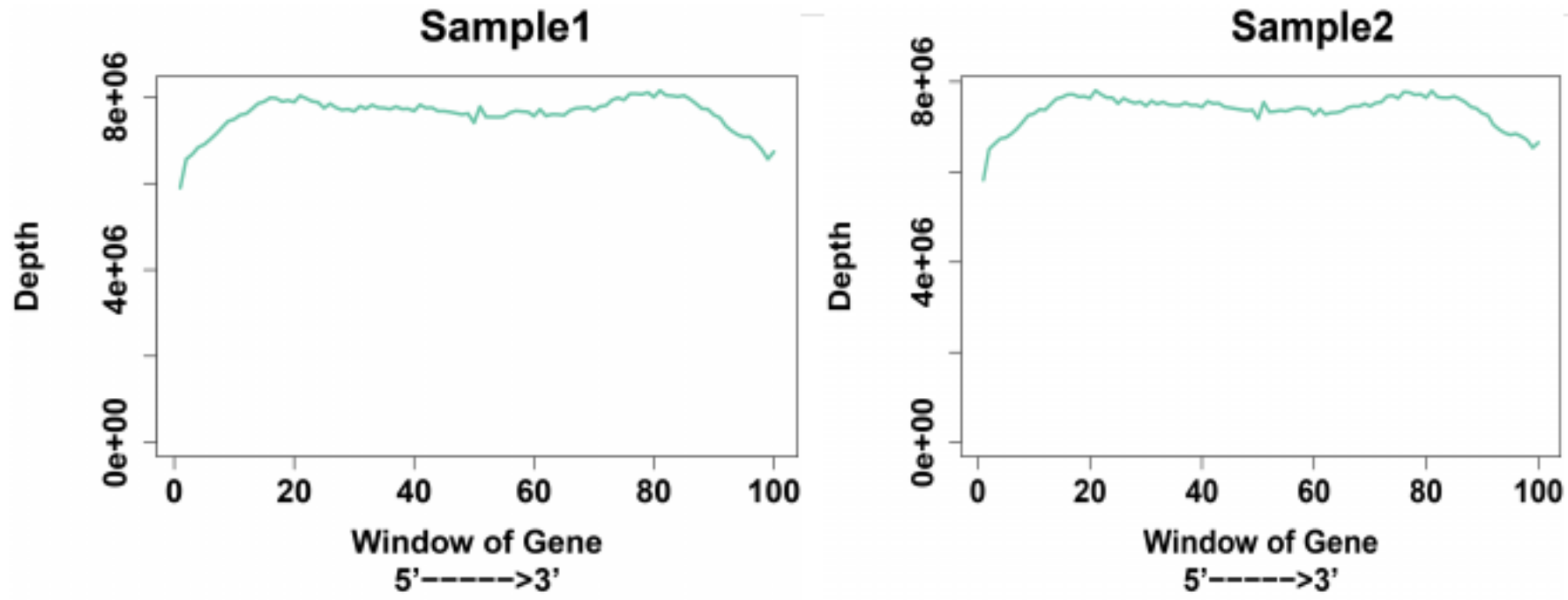


图14　均一性分析图

将基因的外显子区域平均划分为　100个窗口，统计每个窗口里面的碱基平均深度。以基因的划分窗口为横坐标，以落在该窗口内的碱基深度为纵坐标，做折线图。

3.4　比对文件可视化

Integrative Genomics Viewer　（IGV）(Thorvalds ,et al., 2013)　是一种探索大型综合基因组数据的高性能交互式可视化工具。它支持各种各样的数据类型，包括基于芯片测序、二代测序数据和基因组注释数据等。

IGV可以同时查看关于基因表达信息的其他层次，以及遗传密码中的序列变化或突变。其他的基因组细节，如拷贝数变化、染色质沉淀数据和表观遗传学修饰，也可以在 IGV中查看。此外，所有这些数据类型都可以被覆盖或叠加，以此判定某个水平的变化是如何影响其他水平。　IGV可以选择多种显示模式，以Heatmap、柱状图、散点图或其他形式查看数据。

使用IGV进行比对结果文件的可视化时，可以将文件夹下的　.tdf　文件导入IGV 中，得到下图示意的可视化结果，在选择感兴趣的区域是，可以选择要查看的染色体，并在染色体区域选择要查看的感兴趣的区域即可。

图15　IGV 可视化操作示意图

4 表达量分析

4.1 表达量估计

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，基因转录出 mRNA 的量都是不一样的。

RPKM(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。 RPKM 是 Reads per Kilobase Million Mapped Reads 的缩写，由 Mortazavi 于2008年第一次提出。其计算公式为：

$$RPKM = \frac{10^6 * R}{NL/10^3}$$

图16　RPKM计算公式

设RPKM(A)为基因 A 的表达量，则 R 为唯一比对到基因 A 的 Reads 数， N 为唯一比对到参考基因的总 Reads 数， L 为基因 A 的外显子区域的长度。 RPKM 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。

4.1.1 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：

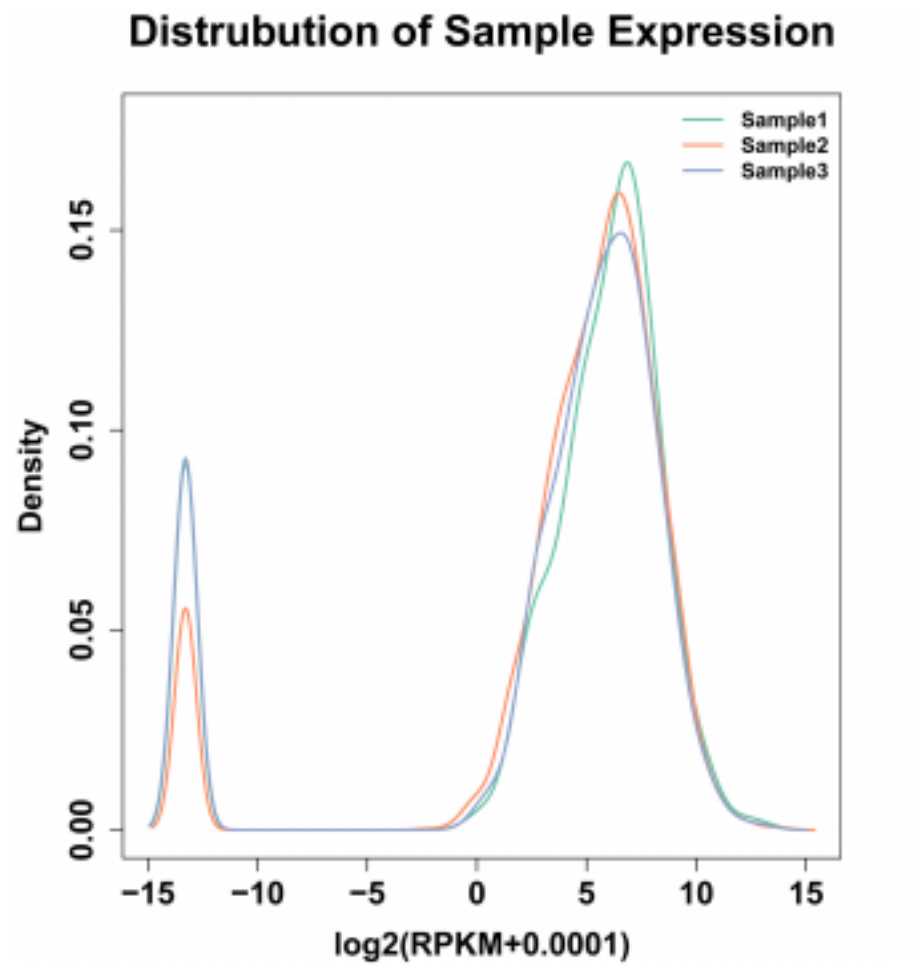


图17　表达量分布图

对每组样品的基因表达量，取以　2为底的对数后，做出密度分布图。横坐标为$\log_2(RPKM+0.0001)$，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：

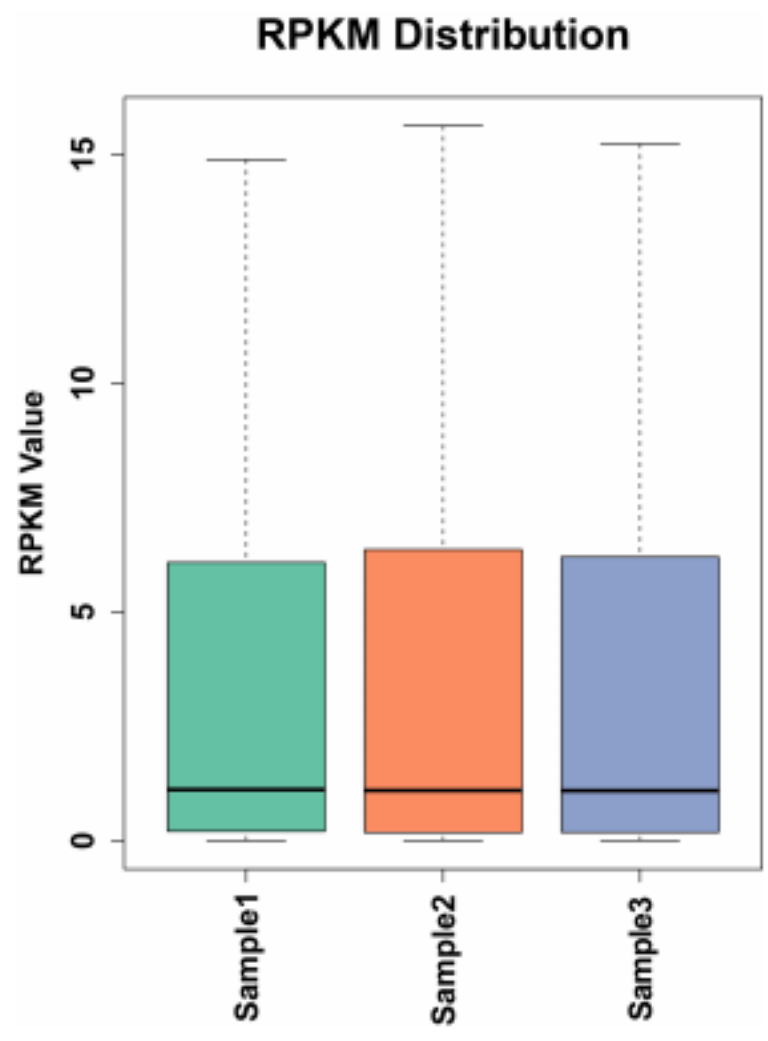


图18　表达量箱式分布图

4.1.2　饱和度分析

定量饱和曲线检查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。

根据每条饱和曲线达到平台期的数据量深度，可以推断本次实验的数据量是否满足数据分析需求。

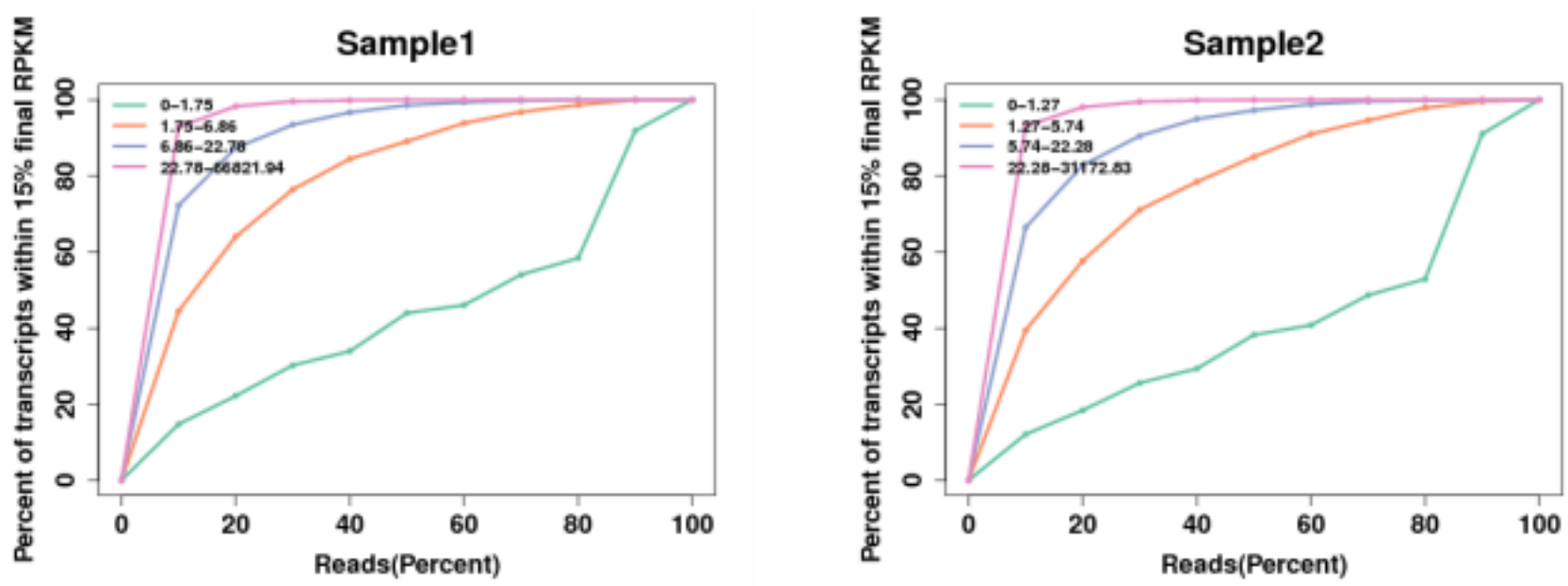


图19　饱和度分析图

4.1.3　样品实验的聚类

一般情况下，源于同一实验条件下的样品会聚类到一起，表明实验条件为影响聚类的主要因素。根据样品全部基因的表达量信息对样品进行系统聚类，得到下图：

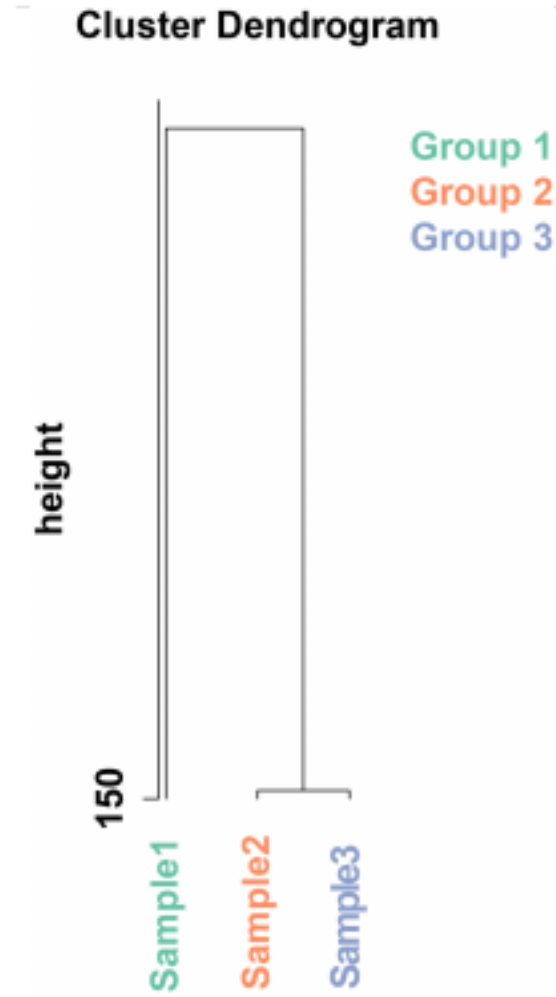


图20　样品聚类图

根据每个样品的基因表达量，计算两两之间的皮尔逊相关系数（Pearson Correlation Efficiency），来表示样品两两间的相似度。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。

4.2 差异表达分析

4.2.1 差异表达分析统计结果

对于设置生物学重复的实验，我们采用 DEseq进行基因差异表达分析，比较处理组与参考组，并选取 $|\log_2 Ratio| \geq 1$和$q<0.05$的基因作为差异表达基因，得到上下调基因个数。

对于无生物学重复样品，则采用 DEGseq(Wang ,et al., 2010)进行基因差异表达分析，比较处理组与参考组，并选取 $|\log_2 Ratio| \geq 1$和$q<0.05$的基因作为差异表达基因，得到上下调基因个数。

本项目所有组别差异表达基因结果见：

表6　组间比较得到的差异表达基因数目

name	Sample1_VS_Sample2
Up	24,947
Down	2,500
Total	27,447

（1）Up：表示在第一组中表达上调的基因；

（2）Down: 表示在第一组中表达下调的基因；

（3）Total ：表示在两组中有差异基因数目总和。

根据上表，统计的结果如下图：

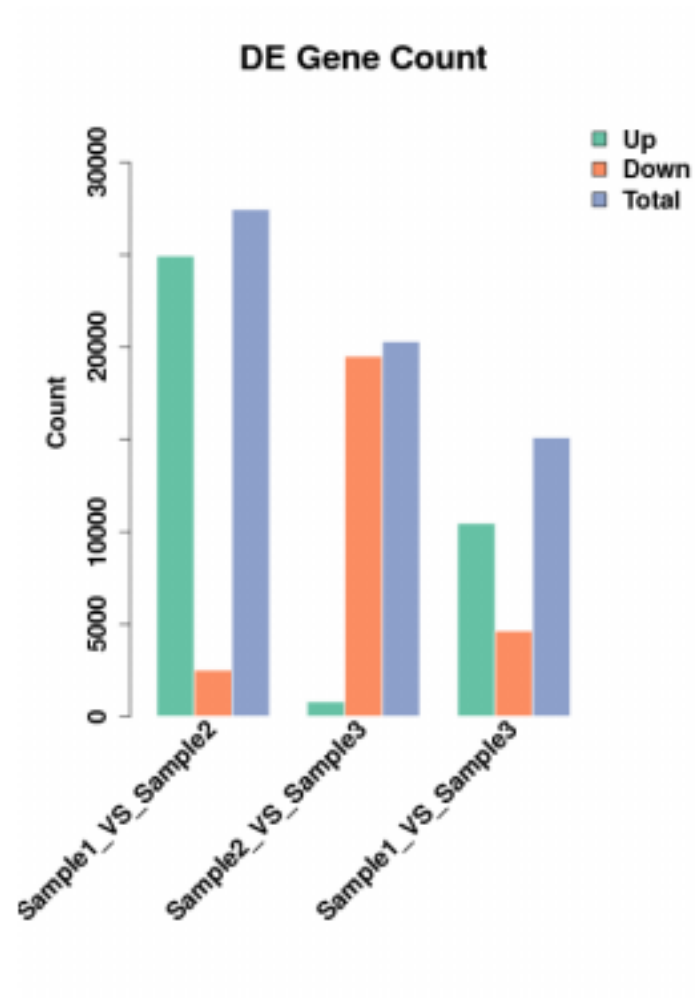


图21　差异表达基因统计图

对于比较组之间差异表达基因的分布，　venn图结果如下：

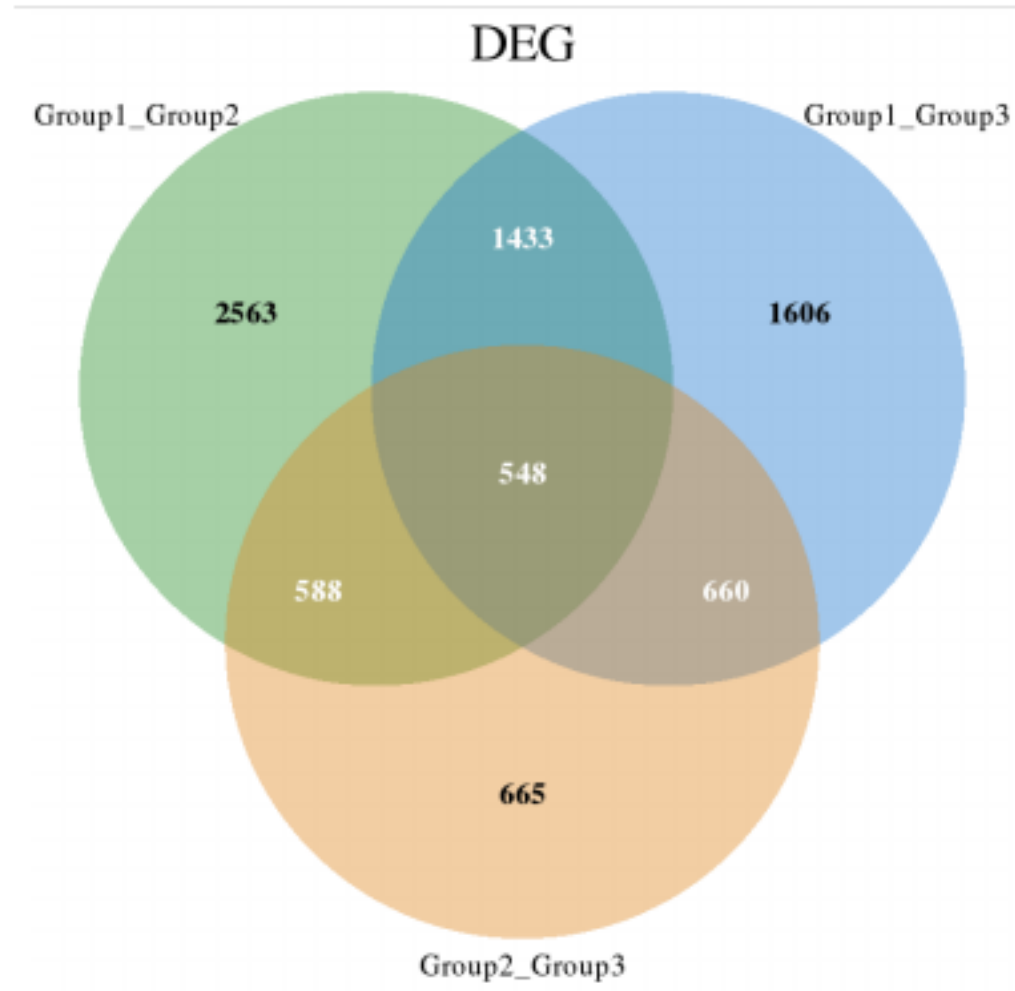


图22　组间差异基因韦恩图

根据各比较组上下调基因，绘制差异表达基因火山图：

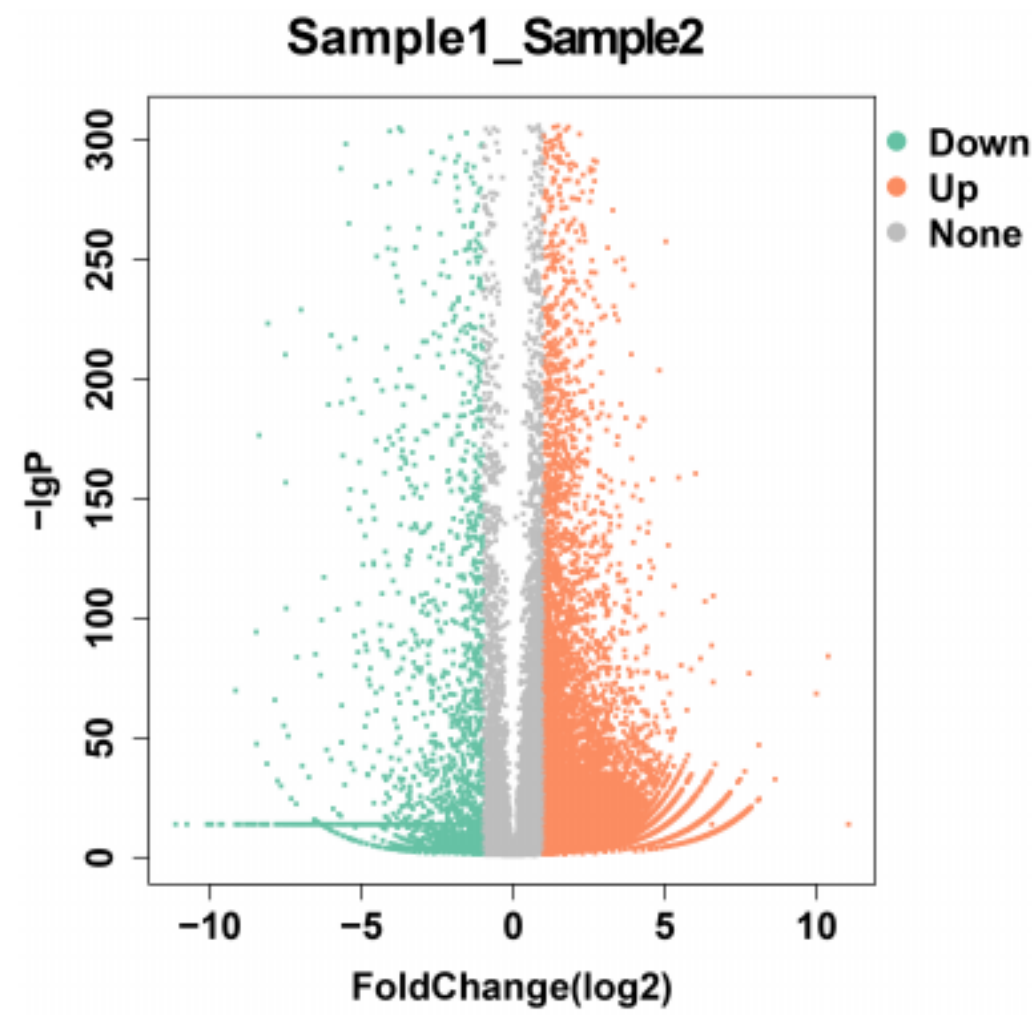


图23　差异表达基因火山图

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。

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中的差异表达基因的变化情况。我们利用 R软件 （ 版本号 ：v3.1.1 ） ，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

号　：v3.1.1　）　，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图：

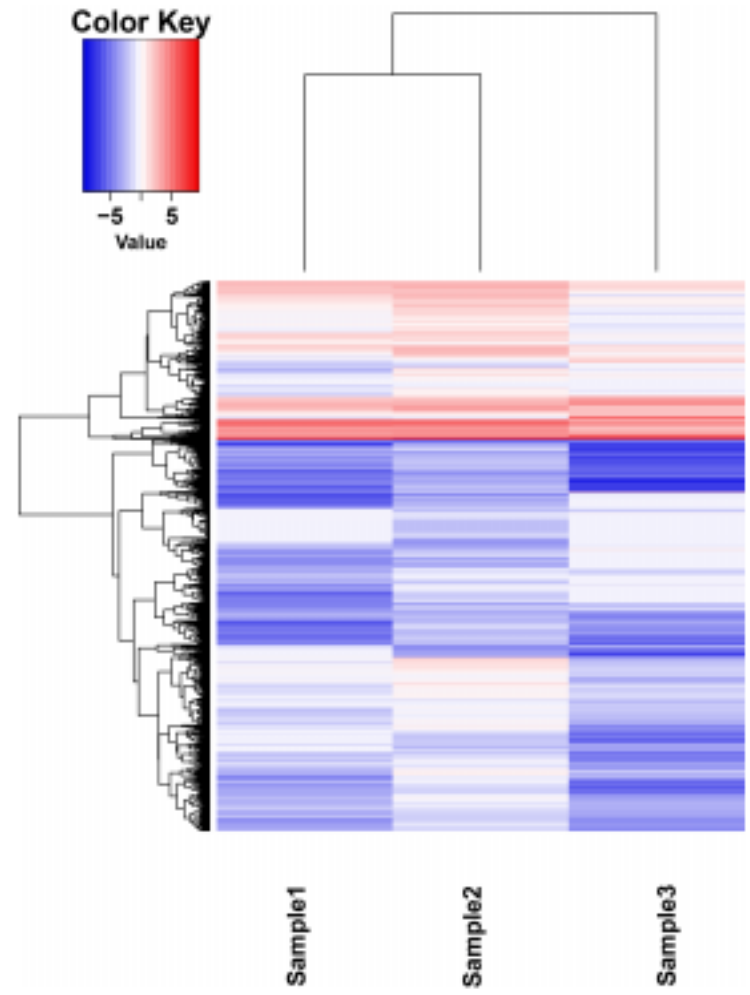


图24　差异基因聚类图

根据差异表达基因在每个样品里的表达量，取以　2为底的对数后，计算欧氏距离，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。

4.2.3　差异表达基因统计结果注释

统计差异表达基因，并利用　NCBI、Ensemble、GO和KEGG等数据库对差异表达基因进行注释，获得差异表达基因详细描述信息。

表7　差异表达基因的结果注释

gene	ENSG00000266876
S1_count	232
S1_normalize	0.2664
S2_count	21
S2_normalize	6.44E-05
FoldChange	4137.5317
Log_2FoldChange	5.4059
pval	1.76E-05
padj	2.81E-06
Up/Down	up
Significant	yes
Position	--
NR:Seq-id	gi|281364746|ref|NP_723538.3|
NR:Score	3,613
NR:Evalue	0
NR:Description	"methuselah-like 15, isoform C [Drosophila melanogaster]"
NT:Seq-id	gi|281364747|ref|NM_164896.2|
NT:Score	2,836
NT:Evalue	0
NT:Description	Drosophila melanogaster methuselah-like 15 (mthl15)
Uniprot:UniProtKB-AC	Q9V818
Uniprot:Score	36.6000
Uniprot:Evalue	3.00E-24
Uniprot:Description	Probable G-protein coupled receptor Mth-like 3
COG:gene	.
COG:Score	.
COG:Eval	.
COG:num	.
Pfam:pfam_ID	pfam00002
Pfam:pfam_Name	7tm_2
Pfam:pfam_Description	7 transmembrane receptor (Secretin family).
GO:biological_process	GO:0008340|determination of adult lifespan;
GO:cellular_component	GO:0016021|integral component of membrane;GO:0005886|plasma membrane;
GO:molecular_function	GO:0004930|G-protein coupled receptor activity;
KEGG:KO	K04599
KEGG:Description	MTH; G protein-coupled receptor Mth (Methuselah protein)

（1）*_count：样品*的Reads Count数；

（2）*_normalize ：组*标准化后的结果；

（3）FoldChange：两组的标准化后的数值倍数的比例；

（4）Log 2FoldChange：两组的标准化后的数值倍数的比例的 log 2值；

（5）pval ：计算的 p值；

（6）padj ：校正之后的 p值；

（7）Up/Down：上调还是下调表达， Up上调， Down为下调；

（8）Significant ：是否为显著性差异；

（9）NR:Seq-id ：基因同 NR数据库的最优比对结果；

（10）NR:Score：基因同 NR数据库的比对得分；

（11）NR:Evalue：基因同 NR数据库的比对 Evalue 值；

（12）NR:Description ：NR数据库中该基因的功能描述；

（13）NT:Seq-id ：基因同 NT数据库的最优比对结果；

（14）NT:Score：基因同 NT数据库的比对得分；

（15）NT:Evalue：基因同 NT数据库的比对 Evalue 值；

（16）NT:Description ：NT数据库中该基因的功能描述；

（17）Uniprot:UniProtKB-AC ：基因同 Uniprot 数据库的最优比对结果；

（18）Uniprot:Score ：基因同 Uniprot 数据库的比对得分；

（19）Uniprot:Evalue ：基因同 Uniprot 数据库的比对 Evalue 值；

（20）Uniprot:Description ：niprot 数据库中该基因的功能描述；

（21）COG:gene：比对上的 COG数据库中的基因名；

（22）COG:Score：与 COG数据库的比对得分；

（23）COG:Eval：与 COG数据库的比对 Evalue 值；

（24）COG:num：比对上的 COG数据库中的基因 ID；

（25）Pfam:pfam_ID：比对上的蛋白家族 Pfam的基因 ID；

（26）Pfam:pfam_Name：比对上的蛋白家族 Pfam的基因名；

（27）Pfam:pfam_Description ：比对上的蛋白家族 Pfam的功能描述；

（28）GO:biological_process ：注释到的描述生物进程的 GO Term；

（29）GO:cellular_component ：注释到的描述细胞组分的 GO Term；

（30）GO:molecular_function ：注释到的描述分子功能的 GO Term；

（31）KEGG:KO：注释到的 KEGG中的 ID；

（32）KEGG:Description ：KEGG中的功能描述。

4.3 蛋白互作网络

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，直接将差异表达基因集（比如差异基因 List ）映射到该物种的蛋白互作网络。

我们提供参考物种蛋白互作网络数据文件（蛋白互作网络参考数据集）和目

标基因集，构建蛋白互作网络的网络数据文件可以直接导入 Cyto sca pe 软件（Shannon ,et al. , 2003)，并根据目标基因集中的基因属性对网络进行可视化编辑。 Cytoscape软件使用方法可参考我们提供的使用说明文档。您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）的颜色与此基因的上下调特性相关。若基因上调，则其节点为红色；若基因下调，则其节点为青色。根据不同的研究目的和需求，您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。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：

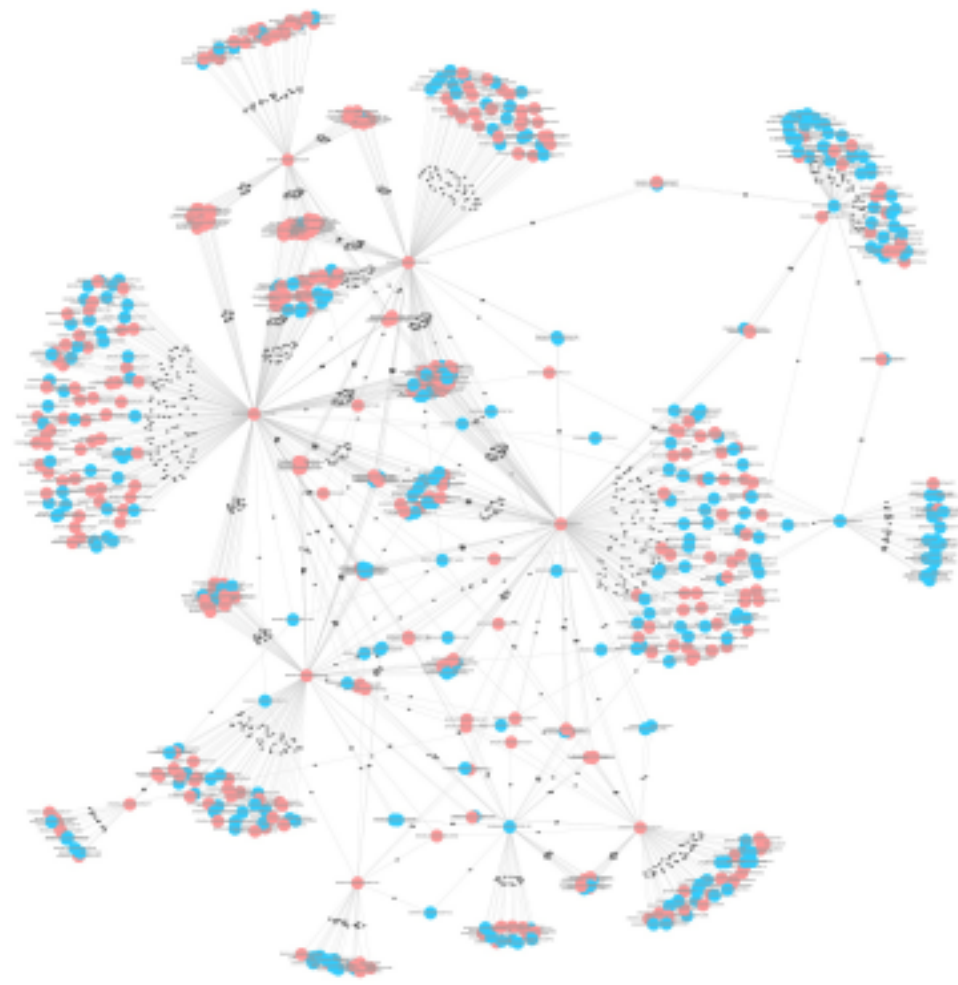


图25 差异基因蛋白互作网络图

5 功能分析

5.1 GO功能分析

5.1.1 差异表达基因的 GO统计

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数据库中总共有三个 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 Blast2GO，得到每个基因对应的 GO条目。针对 GO数据库中第三

层的条目，统计差异表达基因（区分上调表达和下调表达）在该条目里的个数，并计算百分比，得到的结果如下：

表8　GO统计示意图

GO Term	biological_process
GO Subterm	single-organism process
Up_Count	697
Up_Percent	0.5212
Down_Count	1,691
Down_Percent	0.4973

（1）GO Subterm：进行统计的 GO 的子类名称；

（2）*_Count：位于该子类的差异表达基因（上调或者下调）数目；

（3）*_Percent ：位于该子类的差异表达基因（上调或者下调）占总差异表达基因的比例。

为了直观的展示差异表达基因集合的 GO 统计结果，可以根据上表统计结果，绘制柱状图，结果如下：

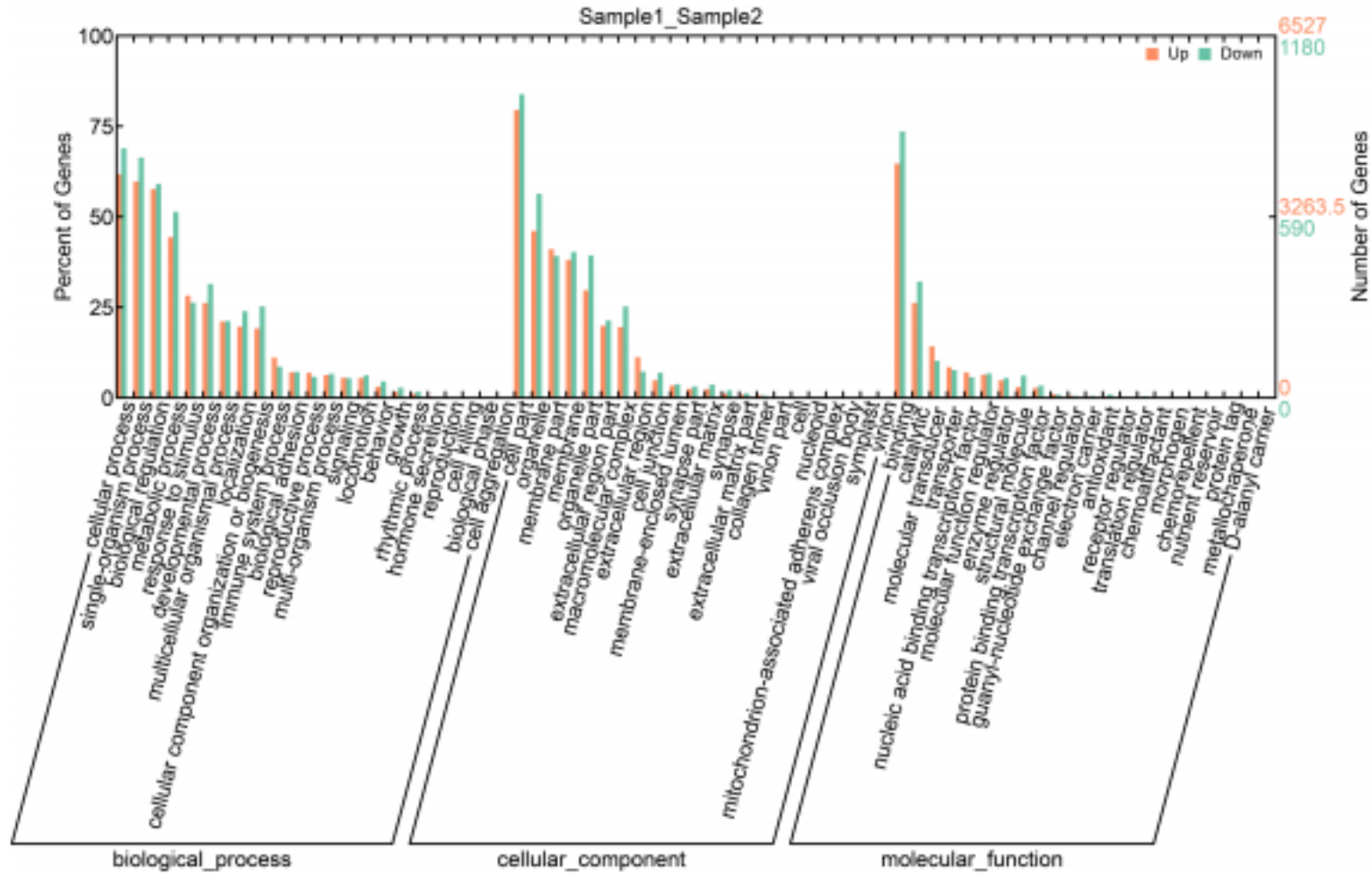


图26　差异表达基因的 GO统计柱状图

5.1.2 GO富集分析

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，在差异表达基因中显著富集的 GO 条目，其计算公式为：

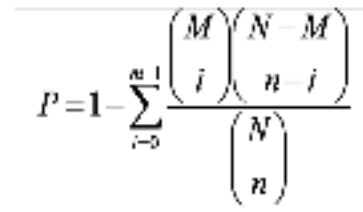


图27　GO富集计算公式

其中，N为所有基因中具有GO注释的基因数目；n为N中差异表达基因的数目；M为所有基因中注释到某特定GO条目的基因数目；m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。

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9　GO结果示例表

Description	response to stimulus
GO	GO:0050896
p	2.0e-17
FDR	5.126e-14
Significant	255
Gene_in_de	AT1G01060|AT1G02205
Annotated	1,736
Gene_in_background	AT1G01060|AT1G01300
Up_Gene	AT1G01060|AT1G0388
Up_Count	95
Down_Gene	AT1G02205|AT1G02450
Down_Count	160
Links	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO
Result	yes

（1）Description：该GO的功能描述；

（2）GO: 进行富集的GO条目；

（3）p：检验后的p值；

（4）FDR: 错误发现率；

（5）Significant：富集到该GO的差异基因数目；

（6）Gene_in_de；富集到该GO的差异基因；

（7）Annotated：富集到该GO的所有背景基因数目；

（8）Gene_in_background：富集到该GO的所有背景基因；

（9）Up_Gene: 富集到该GO的上调的基因；

（10）Up_Count：富集到该 GO 的上调的基因个数；

（11）Down_Gene：富集到该 GO 的下调的基因；

（12）Down_Count：富集到该 GO 的下调的基因个数；

（13）Links ：该 GO 条目的 GO 数据库链接；

（14）Result ：该 GO 是否显著。

对所有样品富集的 GO 取并集，并且根据样品在该 GO 的富集显著性 q值做出分布图，结果如下：

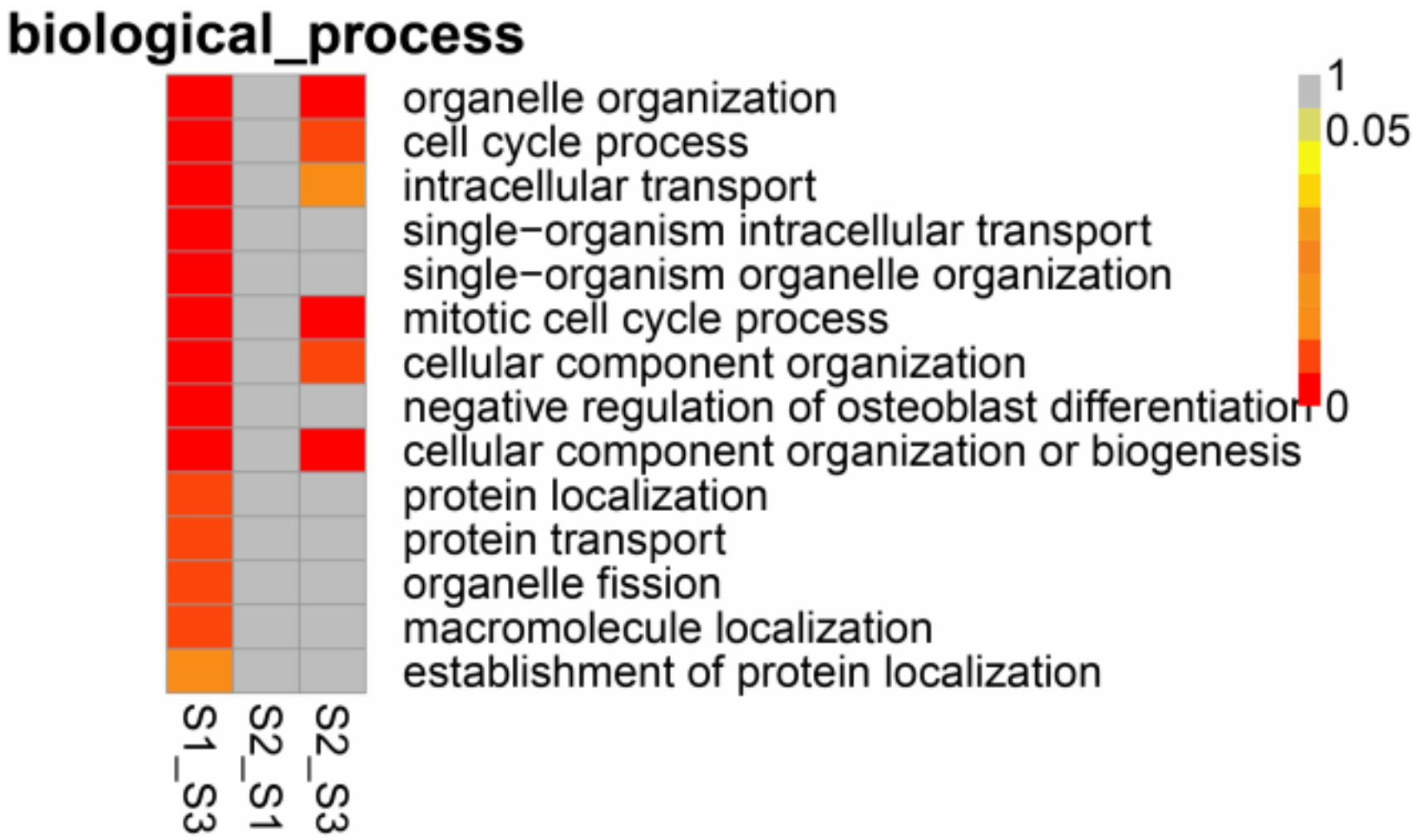


图28　富集 GO条目 q值分布图

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

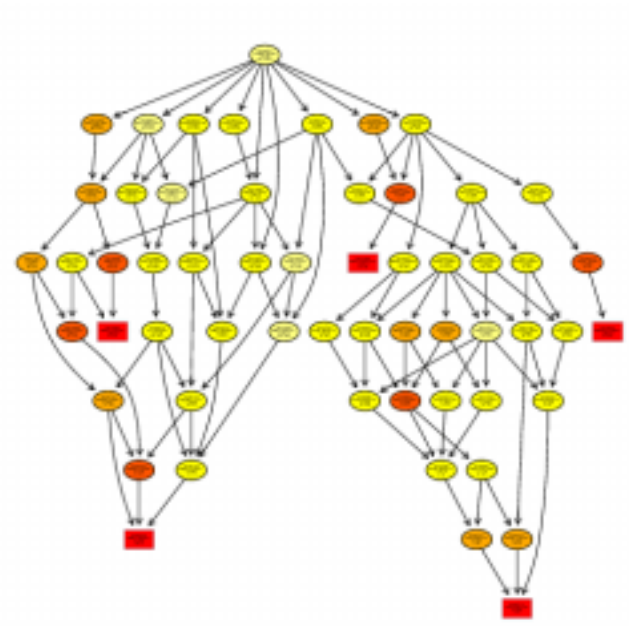
同样的，对于每个比较组而言，展示 q值富集的 GO 条目以及富集的程度，结果展示如下图：

图29　单个组　GO条目 q值富集图

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　GO富集 DAG图

DAG图也称有向无环图（ Directed　Acyclic　Graph, DAG），它将差异基因GO富集分析结果以图形方式展示。其分支呈现包含关系，下层的节点功能属于上层节点功能的下属分支，一般选取　GO富集分析结果的前 5位作为有向无环图的主节点，将互相关联的 GO　Term以此进行展示。 DAG图中节点颜色的深浅指示该节点的富集程度。颜色越红，节点富集程度越高，颜色越偏向黄色，节点富集程度越低。 GO功能的三大主分支为生物过程（　Biological　Process ）、分子功能（Molecular　Function ）和细胞组分（ Cellular　Component），具体结果中将分别绘制该三大功能的　DAG图。



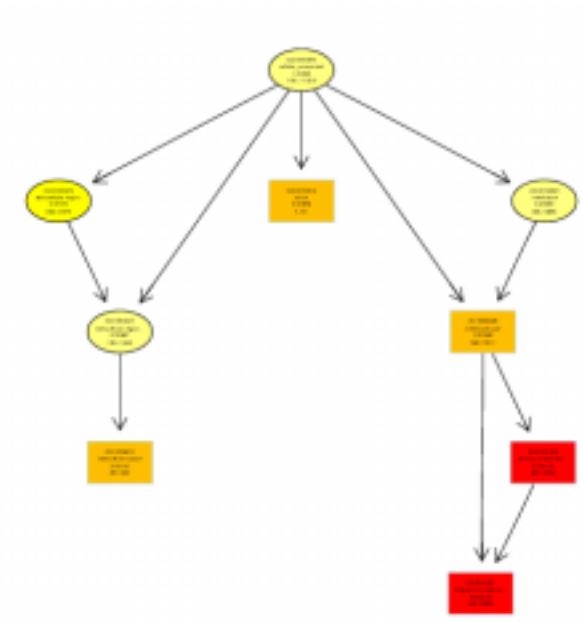


图30　GO富集有向无环图

每个节点代表一个 GO术语，方框代表的是富集程度为 TOP 5的GO，颜色代表富集程度，颜色越偏向红色表示富集程度越高，每个节点上展示了该 TERM的名称及富集分析的 p-value 。

5.3　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，找出差异表达基因中显著性富集的 Pathway。结果文件示例为：

表10　KEGG结果示例表

Name	Ri bo so me
Map	ko03010
Count1	0
Count2	307
Count3	467
Count4	4,081
p	1.99008166681922e-14
q	6.1294515338032e-12
Gene_in_background	AT3G22450|K02881;
Gene_in_DE	
Up_Gene	.
Up_Count	0
Down_Gene	.
Down_Count	0
Links	http://www.kegg.jp/pathway/map03010+
Result	yes

（1）Name: KO的名称；

（2）Map: 进行富集的 map条目；

（3）Count1，Count2，Count3，Count4：进行 Fisher 检验的四个数据，分别为上面公式里的m，M-m，n-m，N-n-M+n；

（4）p：检验后的 p值；

（5）q：多重检验校正的 p值；

（6）Gene_in_background：在背景中的基因的名称；

（7）Gene_in_DE: 差异表达基因中具有该 KO条目的基因名称；

（8）Up_Gene: 差异表达基因中具有该 KO条目的基因名称的上调基因；

（9）Up_Count：差异表达基因中具有该 KO条目的基因名称的上调基因的个数；

（10）Down_Gene: 差异表达基因中具有该 KO条目的基因名称的下调基因；

（11）Down_Count:差异表达基因中具有该 KO条目的基因名称的下调基因的个数；

（12）Links ：该KO条目的 KO数据库链接；

（13）Result ：该KO是否显著。

对所有样品的富集通路的提取并集，并且根据样品在该通路的富集程度 q值做出分布图，结果如下：

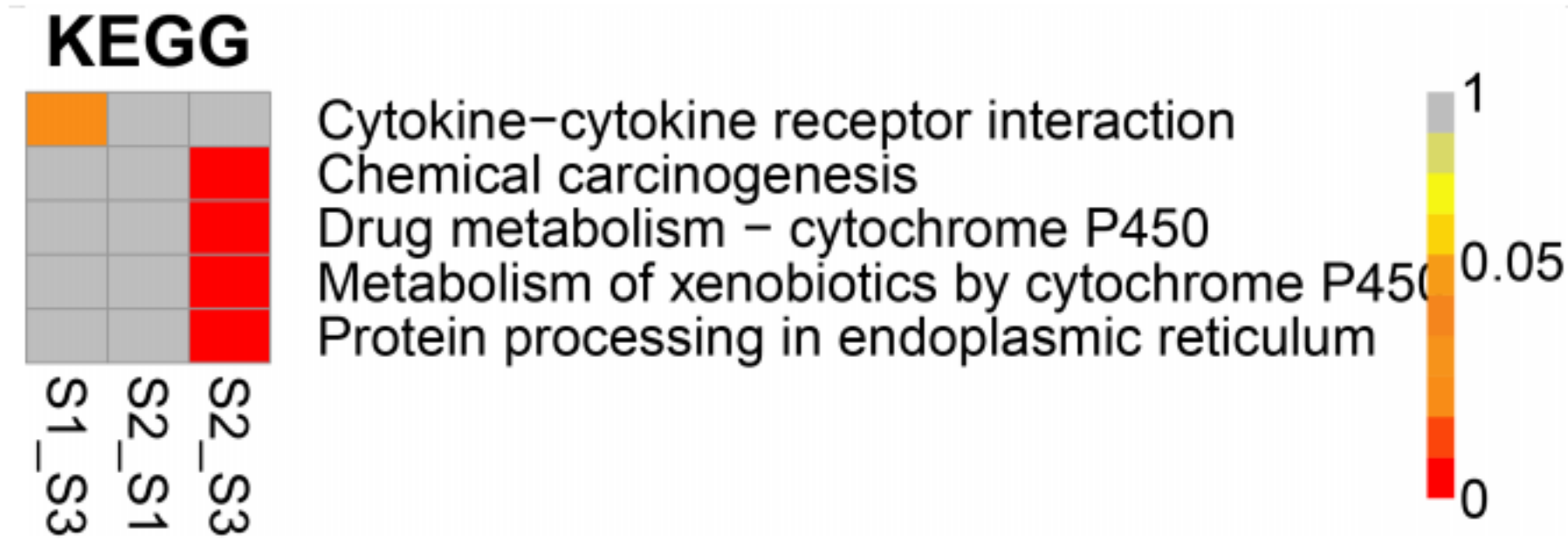


图31　富集通路　q 值分布图

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。

对于每个组进行 q值的KEGG条目的分析，结果展示如下图：

图32　单个组的　KEGG富集 q 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。

对每个比较的通路图进行注释，得到的结果如下：

6 可变剪接

6.1 可变剪切分析

mRNA前体经不同的剪接方式或选择不同的剪接位点将产生多种mRNA剪切异构体，该过程即称为可变剪切（AS）。可变剪切(Florea,et al.,2013)广泛存在于真核生物中，是调节基因表达和蛋白质多样性的重要机制。

可变剪切事件的类型整体有以下5类，如下图所示：

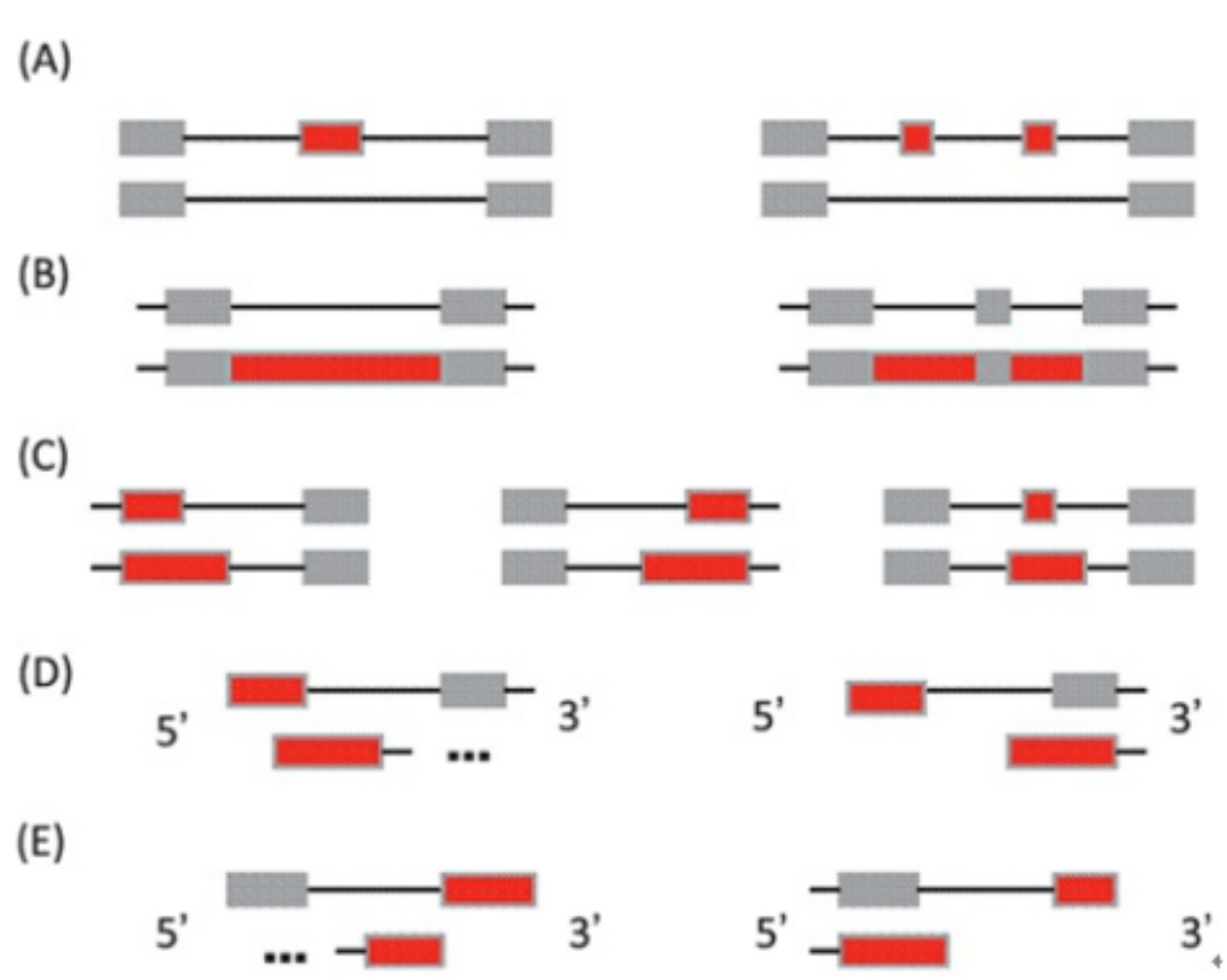


图33　选择性剪切事件类型分类

（1）外显子跳跃（SKIP，左）和盒式外显子跳跃（MSKIP，右）；

（2）内含子滞留（IR，左）和多重内含子滞留（MIR，右）；

（3）可变5’端或3’端剪切（AE）；

（4）转录起始区域可变剪切（TSS）；

（5）转录结束区域可变剪切（TTS）。

利用ASprofile软件在已知基因模型的基础上，分析并统计各样本的可变剪切事件及表达量。

6.1.1　可变剪切事件分类和数量统计

可变剪切事件统计结果由ASprofile软件分析获得。

表11　可变剪切事件统计结果

AS_type	Sample1
AE	21,401
IR	8,838
MIR	2,176
MSKIP	30,166
SKIP	70,910
TSS	77,495
TTS	43,604
XAE	4,386
XIR	8,728
XMIR	1,586
XMSKIP	9,074
XSKIP	19,934

（1）AE：Alternative Exon Ends　（5’，3’或both），可变5’或3’端剪切；

（2）IR：Intron Retention　，单内含子保留；

（3）MIR：Multi-IR　，多内含子保留；

（4）MSKIP: Multi-exon SKIP　，多外显子跳跃；

（5）SKIP：Skipped Exon，单外显子跳跃；

（6）TSS: Transcription Start Site　　，转录起始区域可变剪切；

（7）TTS: Transcription Terminal Site　　，转录结束区域可变剪切；

（8）XAE: 边界模糊型 5’或3’端可变剪切；

（9）XIR：边界模糊型单内含子保留；

（10）XMIR: 边界模糊型多内含子保留；

（11）XMSKIP: 边界模糊型多外显子跳跃；

（12）XSKIP: 边界模糊型单外显子跳跃。

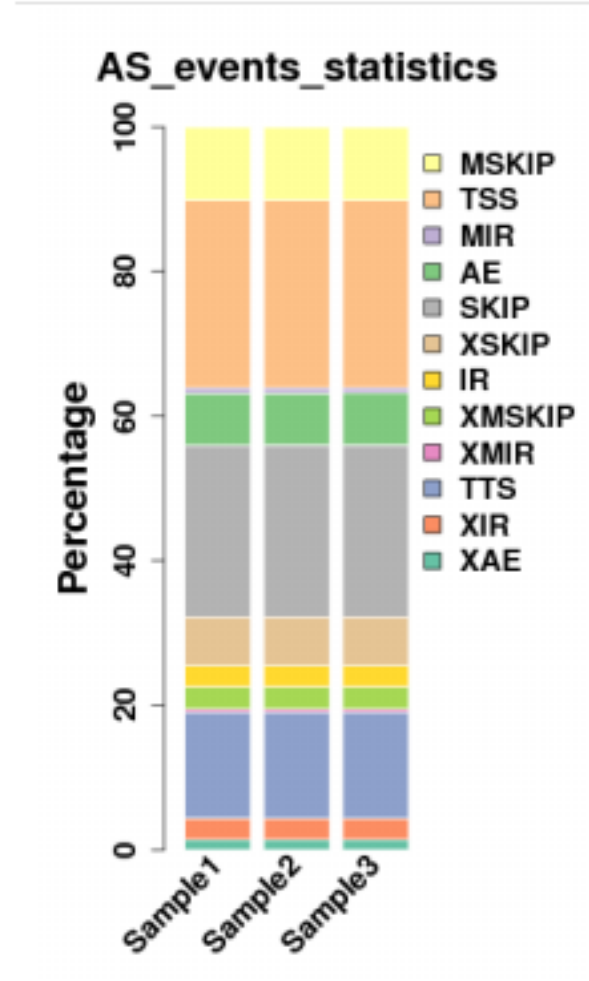


图34　可变剪切结果统计

6.1.2　可变剪切事件结构和表达量

用ASprofile 软件的RPKM工具分析获得可变剪切事件结构详情和表达量，同时利用已有参考基因信息对可变剪切事件基因信息作注释。

表12　可变剪切结构和表达量统计

Eve nt I　D	1 0 0 0 0 0 1
Event Type	TSS
Gene ID	CUFF.1
Chrom	chr1
Event Start	120775
Event End	120932
Event Pattern	120775
Strand	-
Fpkm	0.1531
Ref Gene ID	RP11-34P13.7,RP11-34P13.8

（1）Event ID ：可变剪切事件 ID；

（2）Event Type ：可变剪切类型；

（3）Gene ID：该基因 Cufflink 编号；

（4）Chrom: 该可变剪切事件所处染色体；

（5）Event Start ：该可变剪切事件在染色体的起始位置；

（6）Event End ：该可变剪切事件在染色体的终止位置；

（7）Event Pattern ：该可变剪切事件模式；

（8）Strand ：该事件所处正负链；

（9）Fpkm: 基因的表达量；

（10）Ref Gene ID ：该事件基因的 NM 或 NR 号。

6.2 新转录本预测

利用 Cuffcompare （版本号： v2.2.1 ；参数：默认参数）将 Cufflinks （版本号： v2.2.1 ；参数：默认参数）构建的转录本与参考基因组的已知转录本进行比较(Trapnell ,et al., 2010)，可以评估转录本的构建情况，发现新的未知基因（相对于原有基因注释文件），即新转录本本分析 , 在新转录本分析中，我们还提取出新转录本的 FASTA 序列信息，并与 NT 库作 Blastn 比对，根据比对结果为新转录本添加基因信息注释。

新转录本预测结果为 GTF 格式文件。

GTF 文件格式详细说明见： http://mblab.wustl.edu/GTF2.html

表13　新转录本预测注释结果

Se qnam e	chr　1 0
Source	Cufflinks
Feature	exon
Start	1223422
End	1225059
Score	.
Strand	-
Frame	.
Attributes	gene_id "Novel_000002"; transcript_id "Novel_000002";
Novel_transp_anno:GeneID	105
NM	AF034837
NR	EAW86508
Ensemble_gene	--
Ensemble_rna	--
Symbol	ADARB2
Description	adenosine deaminase, RNA-specific, B2 (non-functional)
GO_Process	GO:0006397|mRNA processing
GO_Component	GO:0005634|nucleus;GO:0005730|nucleolus;
GO_Function	GO:0003725|double-stranded RNA binding;
KO	--
Map	--

（1）Seqname：新转录本编号；

（2）Source：来源标签，这里的 Novel Gene指新基因；

（3）Feature ：区域类型，预测的外显子区域；

（4）Start ：起始坐标；

（5）End：终止坐标；

（6）Score：意义不显著；

（7）Strand ：正负链信息；

（8）Frame：意义不显著；

（9）Attributes ：属性，包括基因编号、转录本编号等信息；

（10）Novel Transp Anno:GeneID ：新转录本注释的基因；

（11）NM：该基因在 NCBI数据库中转录本编号；

（12）NR：该基因在 NCBI数据库中蛋白编号；

（13）Ensemble_gene：该基因在 Ensembl数据库中的基因编号；

（14）Ensemble_rna：该基因在 Ensembl数据库中的转录本编号；

（15）Symbol：该基因的名称；

（16）Description ：该基因的相关功能描述；

（17）GO_Process：该基因的 GO数据库里生物过程的注释；

（18）GO_Component：该基因在 GO 数据库里细胞组分的注释；

（19）GO_Function：该基因在 GO 数据库中分子功能的注释；

（20）KO：该基因在 KEGG 数据库中的注释；

（21）Map：该基因在 KEGG 中通路的编号。

7 变异分析

SNP（Single Nucleotide Polymorphysm），即单核苷酸多态性，是指等位基因发生突变，致使出现不同等位型。SNP 的位点极其丰富，几乎遍及整个基因组。利用 Samtools-0.1.19(Li ,et al., 2009)将经过排序、去 PCR 重复后的比对文件同参考序列进行比对，得到变异检测结果。

统计结果见下表：

表14 变异统计结果

#Sample	Sample1
SNP	221,509
InDel	10,065
Total	231,574

（1）SNP：检测到的 SNP 的个数；

（2）InDel ：检测到的 InDel 的个数；

（3）Total ：总的变异的个数。

根据检测到的 SNP 位点，统计每个突变类型的频率。

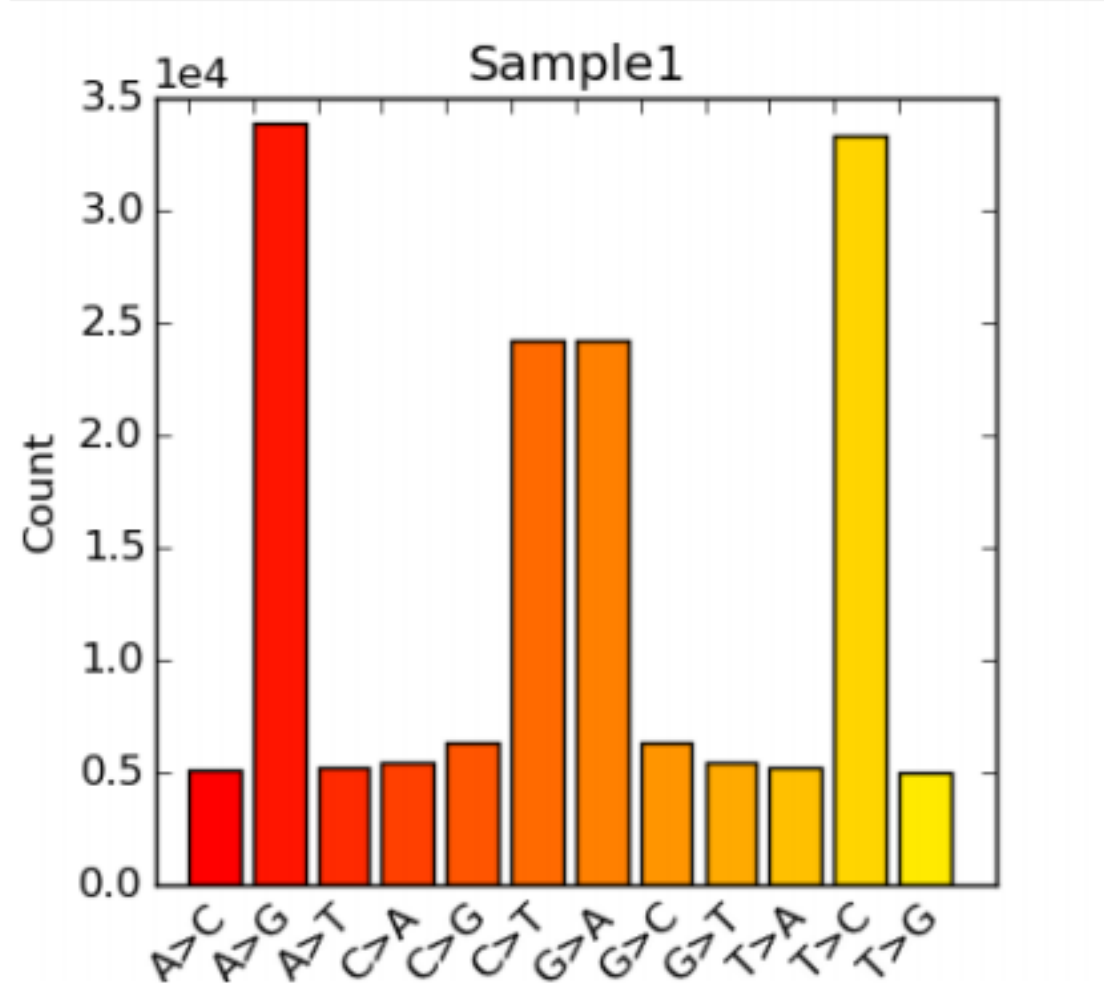


图35　SNP突变频率分布

统计每个突变类型的个数，并作出上述的柱状图。其中　　A>C表示的是由 A突变为 C的SNP位点的个数。

根据检测到的 Indel ，统计 Indel 长度的分布。

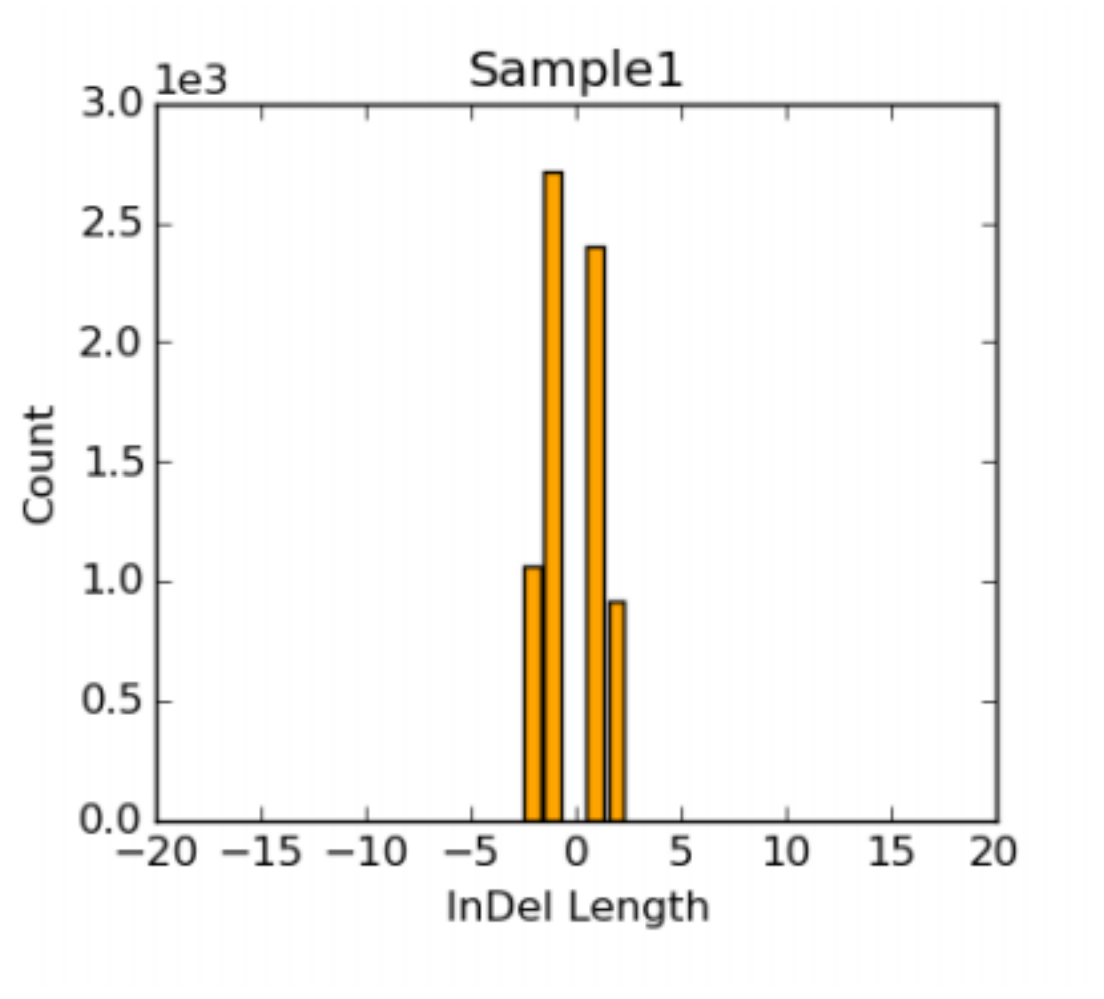


图36　Indel　突变频率分布

8 已知 lncRNA

8.1 表达量分析

8.1.1 表达量统计

利用计算 mRNA表达量的相同的方法，用 RPKM定量已知 lncRNA（已有基因组注释的 lncRNA）的表达水平。

表15 已知 lncRNA 表达量注释结果

Gene	ENSG00000248939
Sample1	0.1014
Sample2	0.0000
Sample3	0.0000

（1）name: 表示该基因在各个样品中的表达量，列表示样品。

8.1.2 表达量分布统计

根据所有样品的已知 lncRNA表达量，得到该样品 Novel lncRNA的表达量密度图如下：

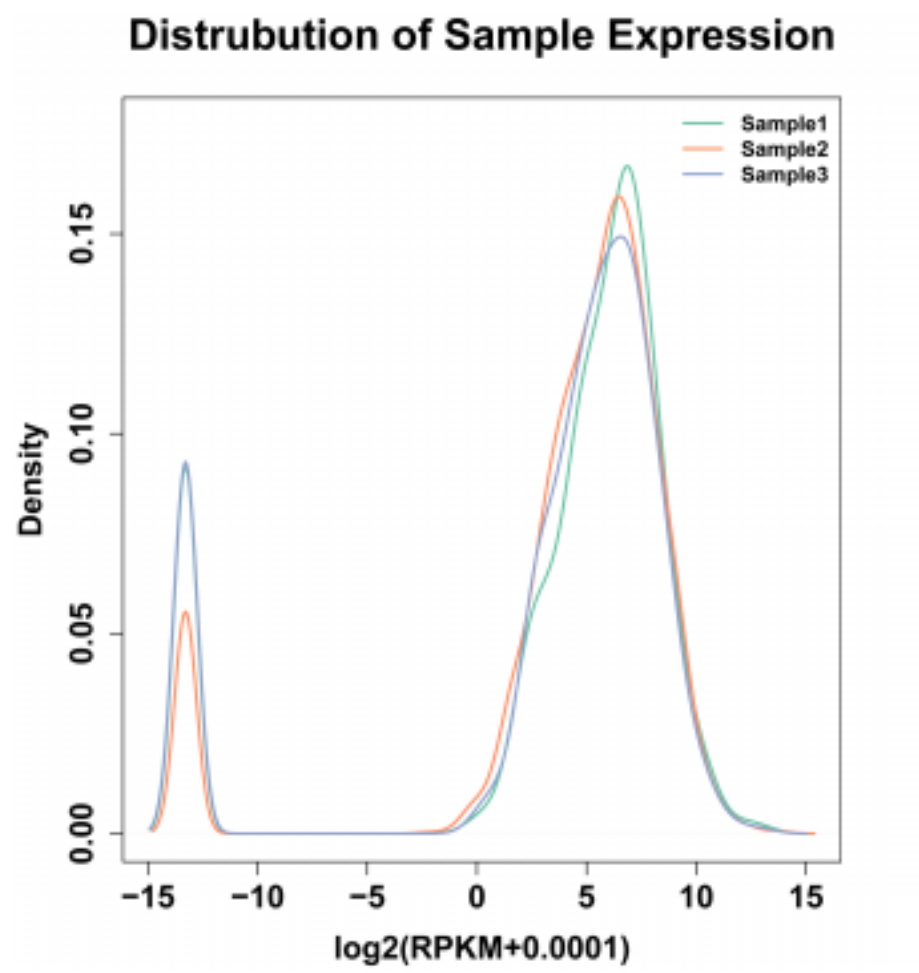


图37 已知 lncRNA 表达量分布；

对每组样品的 lncRNA表达量，取以 2为底的对数后，做出密度分布图。横坐标为 $\log_2(RPKM+0.0001)$，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：

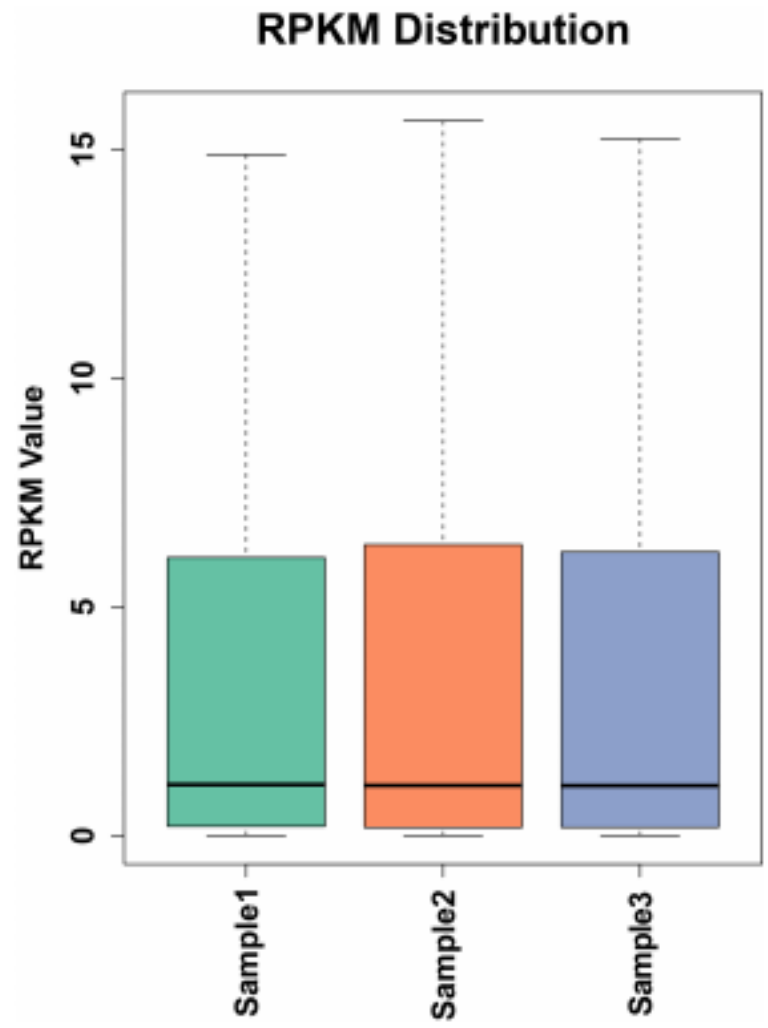


图38　表达量箱式分布图

8.1.3　样品实验的聚类

根据样品全部已知 lncRNA的表达量信息对样品进行系统聚类，得到下图：

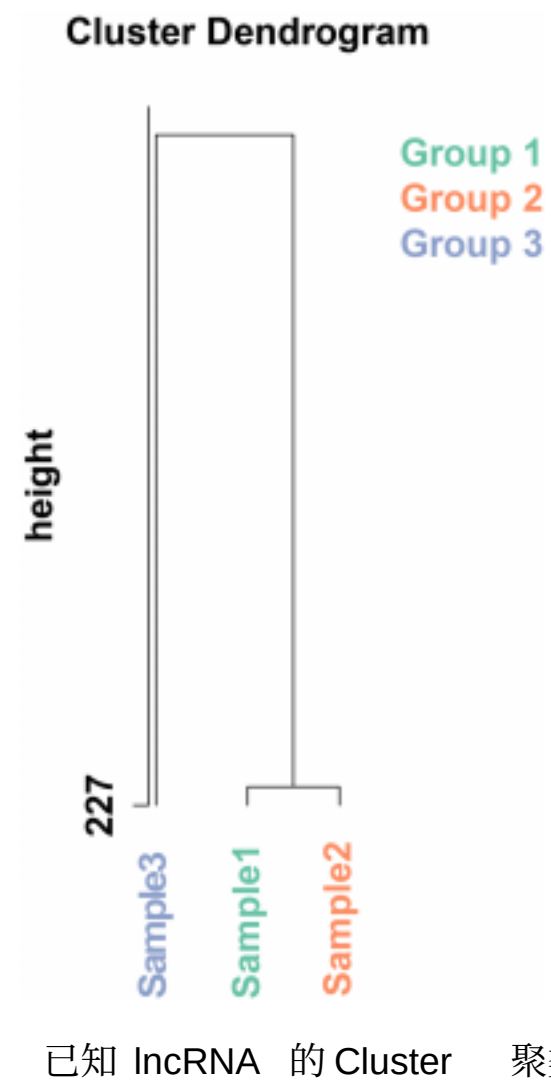


图39　已知 lncRNA 的 Cluster　聚类图

根据每个样品的已知 lncRNA表达量，计算两两样品之间的欧氏距离，然后根据类平均距离法度量两两类之间的距离，再利用系统聚类法（ Hierarchical Cluster ）进行聚类，最终得到样品的整体聚类结果。

8.2　已知差异表达分析

8.2.1 已知差异表达分析统计结果

对于设置生物学重复的实验，我们采用 DEseq进行lncRNA差异表达分析，对处理组与参考组进行比较，并选取 $|\log_2 Ratio| \geq 1$和q<0.05的lncRNA作为差异表达lncRNA，得到上下调 lncRNA个数。

对于无生物学重复样品，则采用 DEGseq进行lncRNA差异表达分析，对处理组与参考组进行比较，并选取 $|\log_2 Ratio| \geq 1$和q<0.05的lncRNA作为差异表达lncRNA，得到上下调 lncRNA个数。

本项目根据上下调 lncRNA，绘制差异表达 lncRNA火山图：

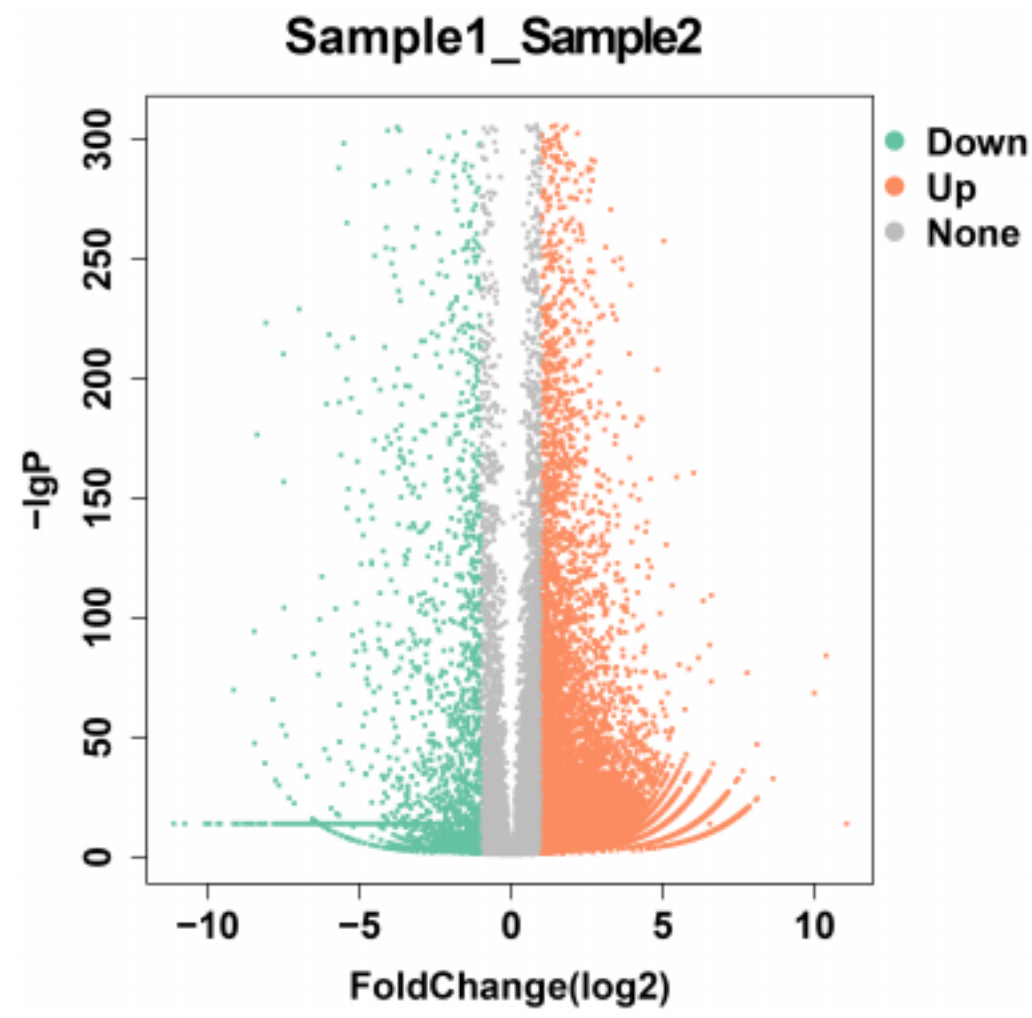


图40 差异表达 lncRNA 火山图

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。

本项目所有组别差异表达 lncRNA结果如下：

表16　组间比较得到的差异表达　　lncRNA　数目

name	Sample1_VS_Sample2
Up	3,477
Down	669
Total	4,146

（1）Up：表示在第一组（例如　Sample1）中表达上调的 lncRNA；

（2）Down: 表示在第一组（例如　Sample2）中表达下调的 lncRNA；

（3）Total　：表示在两组中有差异　lncRNA数目总和。

根据上表，统计的结果如下图：

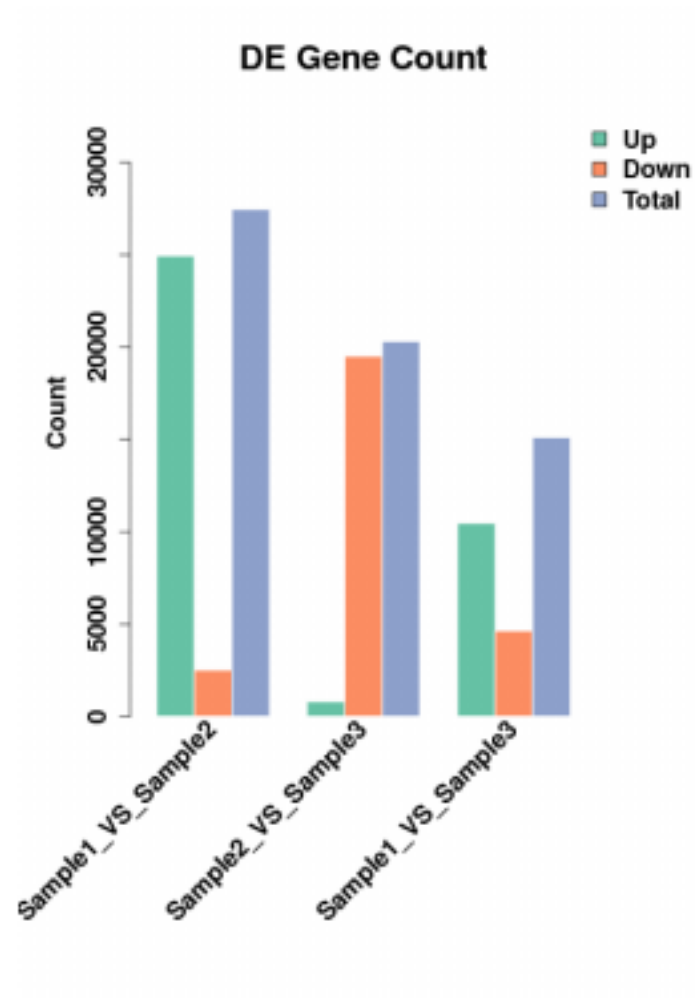


图41　差异表达　lncRNA　统计图

通过比较处理组和参考组，对差异表达　lncRNA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　lncRNA的变化情况。我们利用　R软件（版本号：v3.3.1 ），对差异表达 lncRNA和不同样本 / 实验条件同时进行分层聚类分析。下图为两组样本的差异表达　lncRNA聚类示例。

本项目所有差异表达 lncRNA聚类分析结果如下：

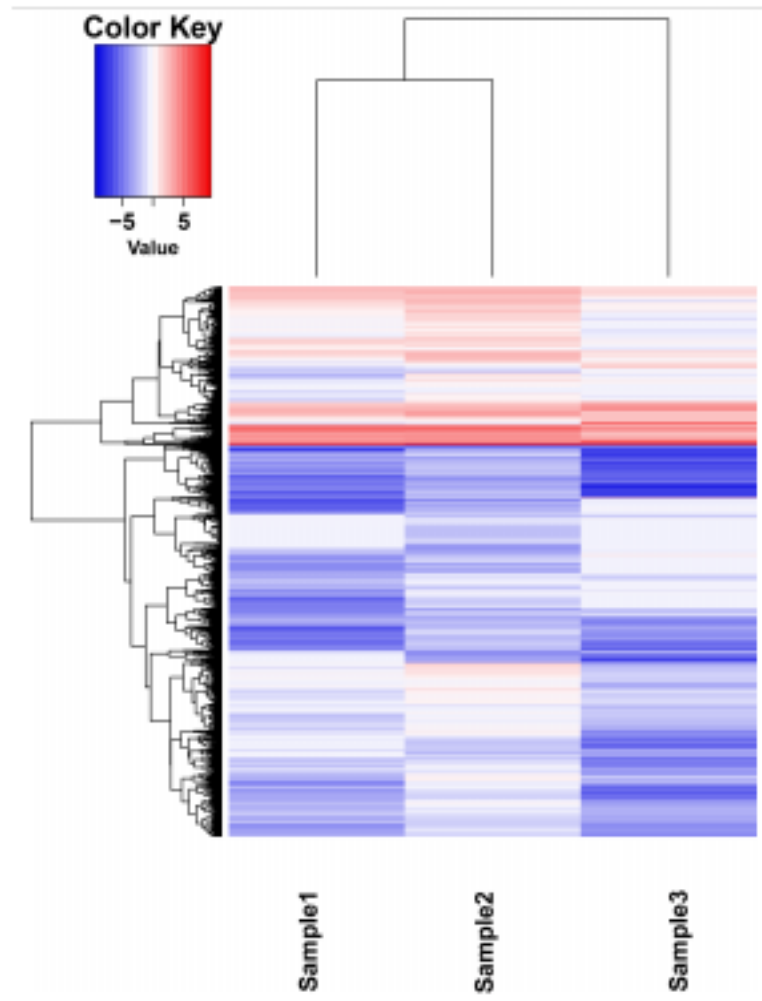


图42　差异 lncRNA 聚类图

根据差异表达 lncRNA在每个样品里的表达量，取以 2为底的对数后，计算欧氏距离，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。

8.3 已知 GO功能分析

8.3.1 已知差异表达 lncRNA的GO统计

针对GO数据库中第三层的条目，统计差异表达 lncRNA（区分上调表达和下调表达）在该条目里的个数及其百分比，得到的结果如下：

表17 GO统计示意图

GO Term	biological_process
GO Subterm	cellular process
Up_Count	4,033
Up_Percent	0.6178
Down_Count	813
Down_Percent	0.6889

（1）GO Term：GO大类的名称；

（2）GO Subterm：GO子类的名称；

（3）*_Count：位于该子类的上调（下调）差异表达 lncRNA数目；

（4）*_Percent ：位于该子类的差异表达 lncRNA占总差异表达 lncRNA的比例。

为了直观的展示差异表达 lncRNA集合的GO统计结果，根据上表统计结果绘制柱状图，结果如下：

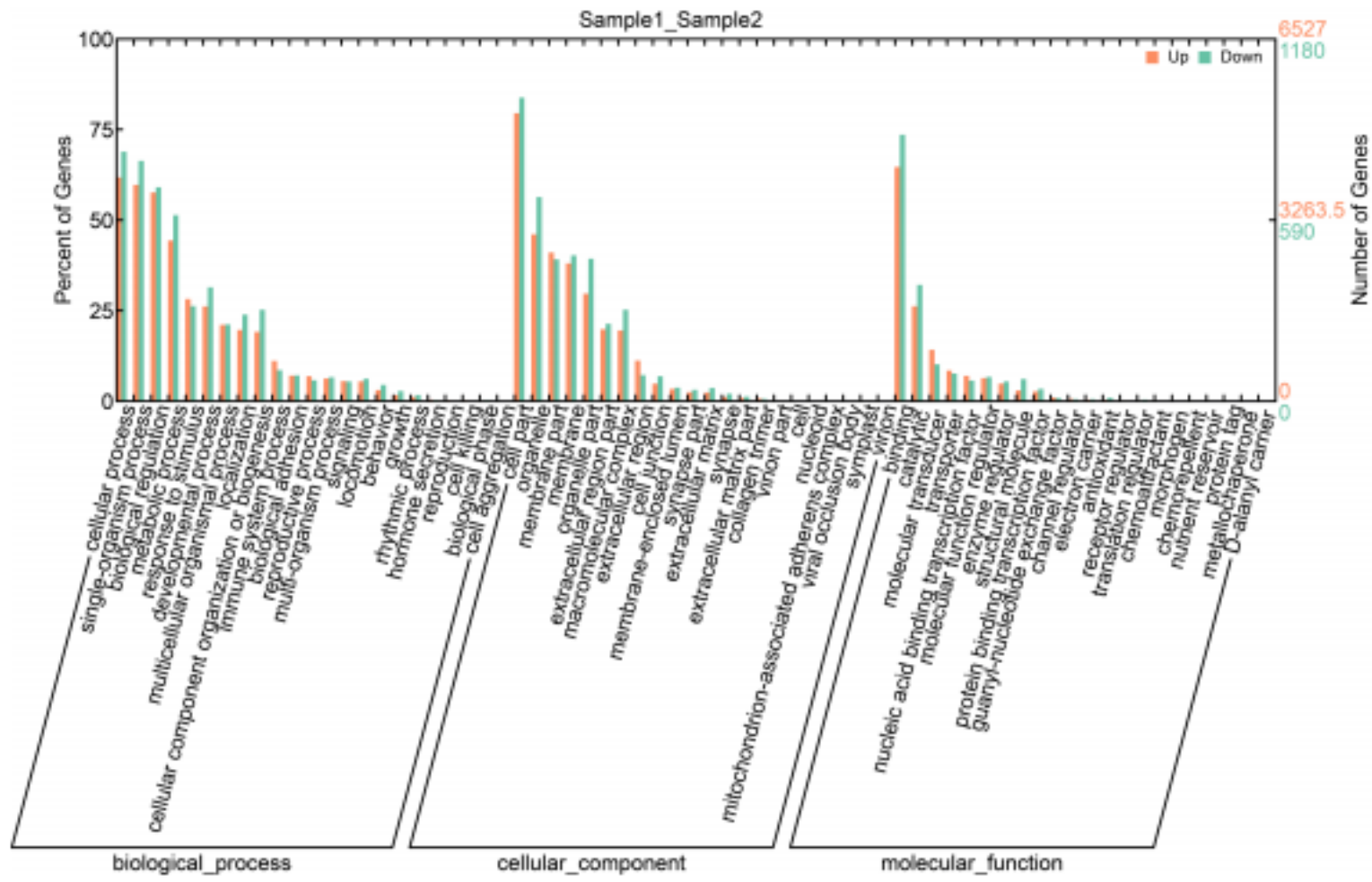


图43 差异表达 lncRNA 的GO统计柱状图

横坐标为GO下的三个大类及其具体子类，左侧纵坐标为差异基因在该子类中所占的比例，右侧纵坐标为该子类中的差异基因数。不同的颜色代表不同的组别。

8.3.2 已知差异表达 lncRNA的GO富集分析

应用超几何检验，找出与整个基因组背景相比，在差异表达 lncRNA中显著富集的GO条目，其计算公式为：

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

图44 GO富集计算公式

其中，N为所有基因中具有GO注释的基因数目；n为N中差异表达lncRNA的数目；M为所有基因中注释到某特定GO Term的基因数目；m为注释到某特定GO Term的差异表达lncRNA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 Term定义为在差异表达lncRNA中显著富集的GO Term。通过GO功能显著性富集分析能确定差异表达lncRNA行使的主要生物学功能。

每两组比较得到的差异表达 lncRNA的GO富集统计结果示例见下表：

表18 GO结果示例表

Description	response to stimulus
GO	GO:0050896
p	2.0e-17
FDR	5.126e-14
Significant	255
Gene_in_de	AT1G01060|AT1G02205
Annotated	1,736
Gene_in_background	AT1G01060|AT1G01300
Up_Gene	AT1G01060|AT1G0388
Up_Count	95
Down_Gene	AT1G02205|AT1G02450
Down_Count	160
Links	http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO
Result	yes

（1）Description ：该GO的功能描述；

（2）GO: 进行富集的 GO条目；

（3）p：检验后的 p值；

（4）FDR: 错误发现率；

（5）Significant ：富集到该 GO的差异基因数目；

（6）Gene_in_de；富集到该 GO的差异基因；

（7）Annotated ：富集到该 GO的所有背景基因数目；

（8）Gene_in_background：富集到该 GO的所有背景基因；

（9）Up_Gene: 富集到该 GO的上调的基因；

（10）Up_Count: 富集到该 GO的上调的基因个数；

（11）Down_Gene: 富集到该 GO的下调的基因；

（12）Down_Count: 富集到该 GO的下调的基因个数；

（13）Links ：该GO条目的 GO数据库链接；

（14）Result ：该GO是否显著。

不同的样品比较组可能得到不同或相似的差异表达 lncRNA集合，而不同的差异表达lncRNA却可能富集相同的 GO功能，即具有相同的差异功能。对所有样品比较组的差异表达 lncRNA所显著富集的 GO条目取并集，并根据样品比较组在

该GO条目的富集显著性 q值做出分布图，直观的展现了不同的样品比较组在差异功能水平上的异同。结果如下图：

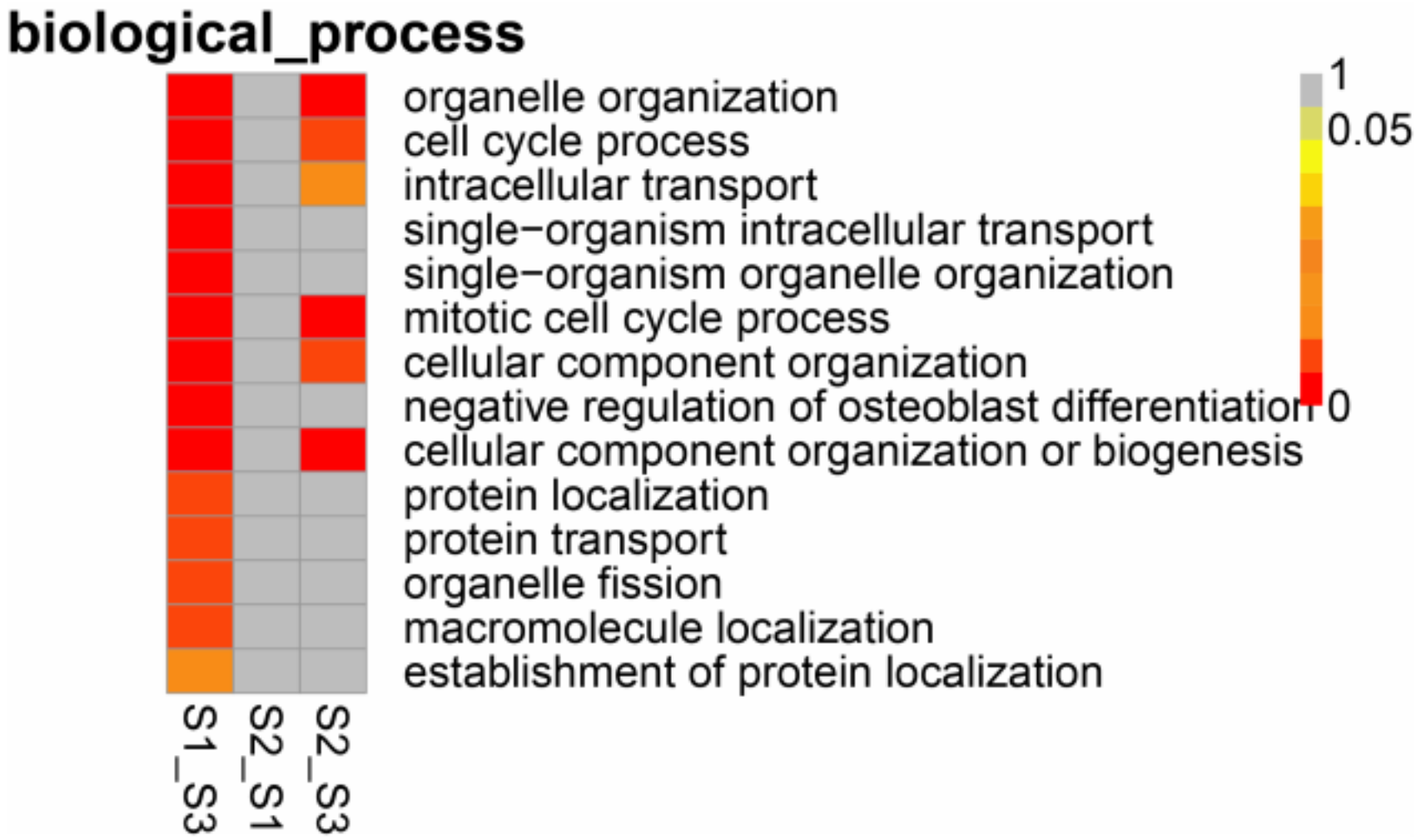


图45　富集 GO条目 q值分布图

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

8.3.3　已知　GO富集 DAG图

DAG图也称有向无环图（ Directed　Acyclic　Graph, DAG），它将差异基因GO富集分析结果以图形方式展示。其分支呈现包含关系，下层的节点功能属于上层节点功能的下属。

DAG有向无环图示例如下：

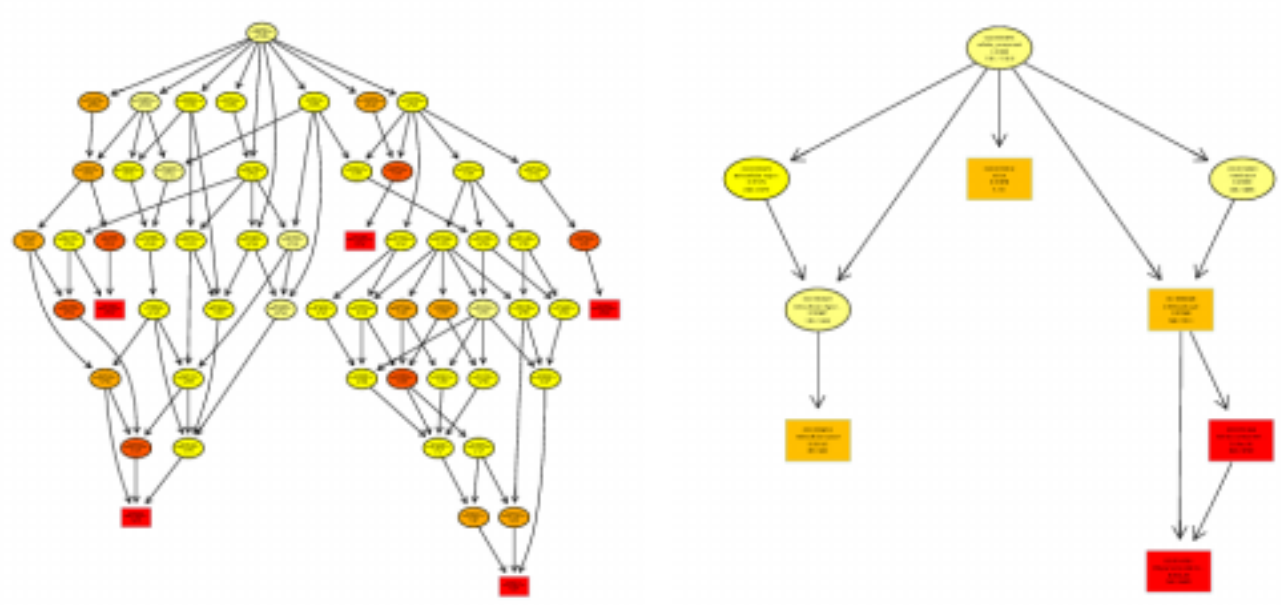


图46　GO富集有向无环图

每个节点代表一个 GO术语，方框代表的是富集程度为 TOP 5的GO，颜色代表富集程度，颜色越偏向红色表示富集程度越高，每个节点上展示了该 TERM的名称及富集分析的 p-value 。

8.4 已知 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）网络在各种细胞活动起的作用作出预测。

KEGG的PATHWAY数据库整合当前在分子互动网络（比如通道、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。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。

表19　KEGG结果示例表

Name	Ri bo so me
Map	ko03010
Count1	0
Count2	307
Count3	467
Count4	4,081
p	1.99008166681922e-14
q	6.1294515338032e-12
Gene_in_background	AT3G22450|K02881;AT4G33865|K02980;
Gene_in_DE	
Up_Gene	.
Up_Count	0
Down_Gene	.
Down_Count	0
Links	http://www.kegg.jp/pathway/map03010
Result	yes

（1）Name: KO的名称；

（2）Map: 进行富集的 map条目；

（3）Count1，Count2，Count3，Count4：进行 Fisher 检验的四个数据，分别为上面公式里的m，M-m，n-m，N-n-M+n；

（4）p：检验后的 p值；

（5）q：多重检验校正的 p值；

（6）Gene_in_background：在背景中的基因的名称；

（7）Gene_in_DE: 差异表达基因中具有该 KO条目的基因名称；

（8）Up_Gene: 差异表达基因中具有该 KO条目的基因名称的上调基因；

（9）Up_Count：差异表达基因中具有该 KO条目的基因名称的上调基因的个数；

（10）Down_Gene: 差异表达基因中具有该 KO条目的基因名称的下调基因；

（11）Down_Count:差异表达基因中具有该 KO条目的基因名称的下调基因的个数；

（12）Links ：该KO条目的 KO数据库链接；

（13）Result ：该KO是否显著。

对所有样品的富集通路的提取并集，并且根据样品在该通路的富集程度 q值做出分布图，结果如下：

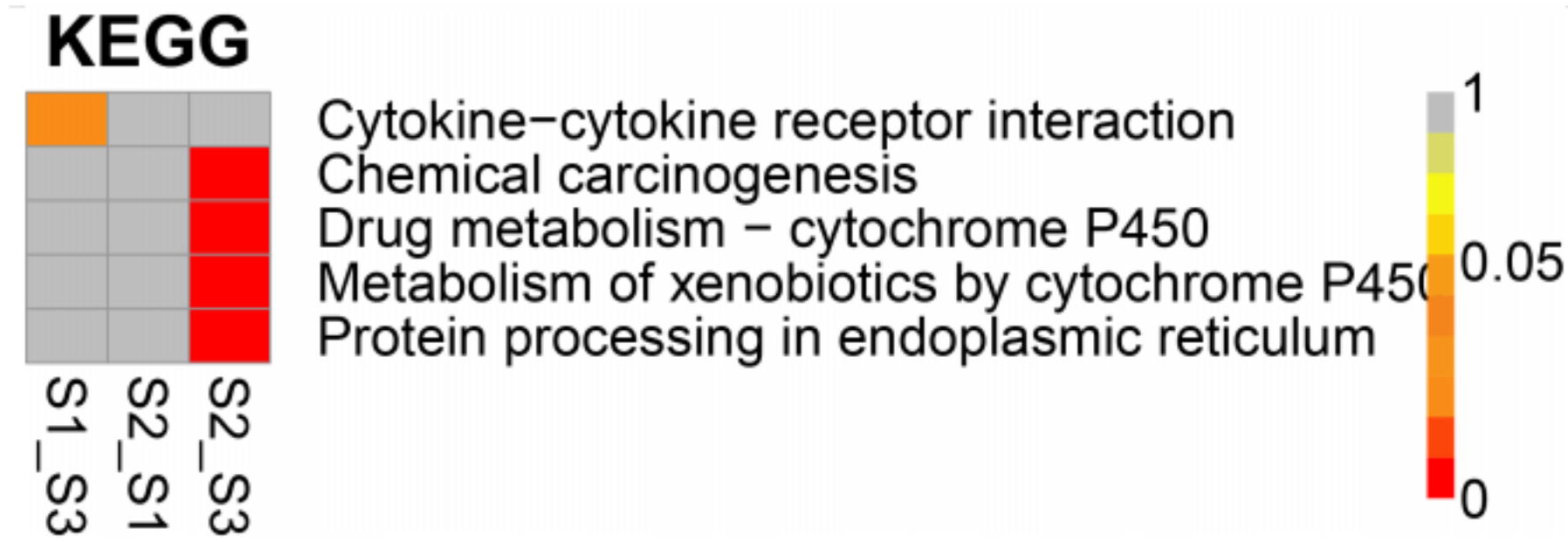


图47　富集通路　q值分布图

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。

对每个比较的通路图进行注释，得到的结果如下：

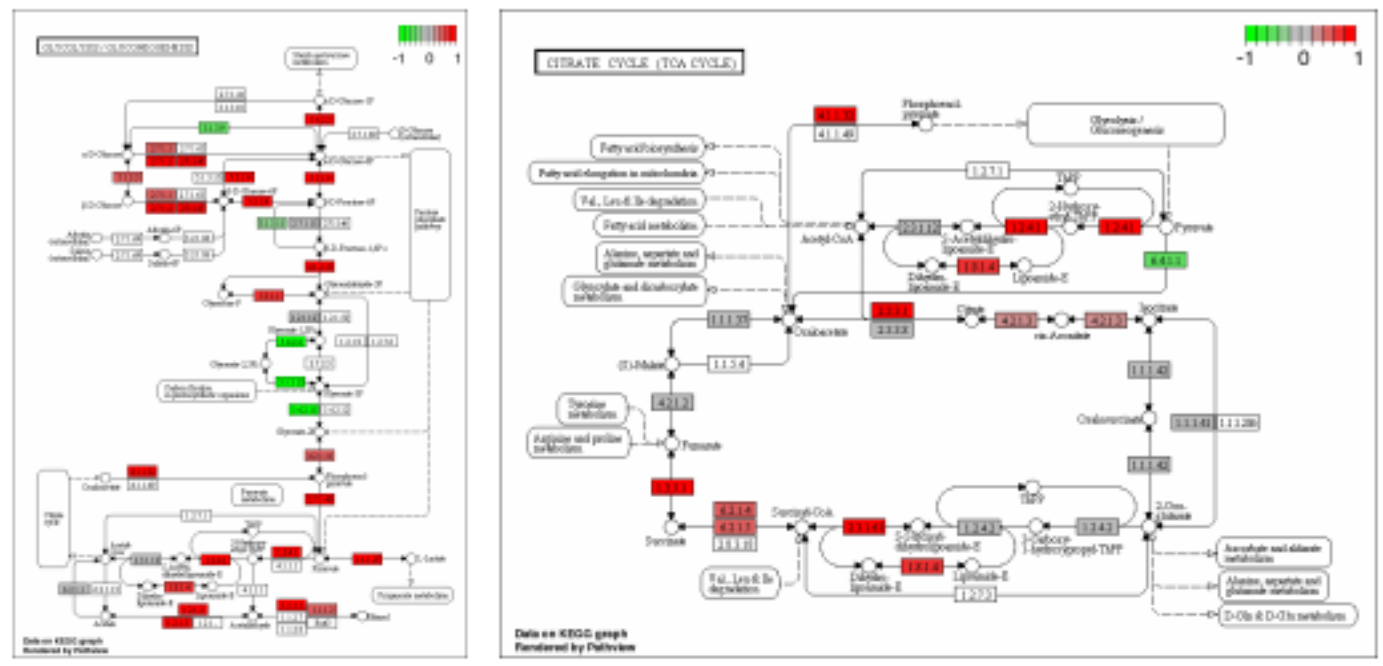


图48　显著富集的　KEGG代谢通路图

9 Novel　l　ncRNA

9.1　Novel ln　cRNA鉴定

lncRNA(Chen ,et　al.,　2013)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 l　inc RNA）, intronic　lncRNA, anti-sense　lncRNA, sense　lncRNA, bidirectional lncRNA等类型，其中 lincRNA所占比例最高。我们根据 lncRNA的特点设置一系列严格的筛选条件，筛选条件 (Cabili ,et al., 2011)　为：

（1）选择长度≥ 200bp，Exon个数≥ 2的转录本；

（2）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；

（3）去除已知的 mRNA转录本；

（4）去除已知的非编码 RNA转录本；

（5）去除有蛋白家族的转录本；

（6）去除有编码潜能的 RNA（CPAT、CNCI、CPC、PLEK）；

（7）去除只存在于一个样本中的转录本。

每个样品剩余的转录本结果统计如下：

表20　Novel lncRNA　　统计结果

Sampl e	Sampl e1
Total Assemble Transcripts	226,025
Short Transcript	59,349
Low Abundant Transcripts	30,666
Remove Known Transcripts	130,571
CPAT Remove Transcripts	1,006
Remove Protein Family Transcripts	629
CNCI Remove Transcripts	361
CPC Remove Transcripts	13
Less Than Two Sample Transcripts	1,030
Potential lncRNA Transcripts	2,400

（1）Total Assemble Transcripts　：组装出来的总转录本数；

（2）Short Transcript　：由于长度过短，被去除的转录本数目；

（3）Low Abundant Transcripts　：由于丰度过低，被去除的转录本数目；

（4）Remove Known Transcripts　：去除同已知 mRNA转录本相关的组装后转录本条数；

（5）CPAT Remove Transcripts　：软件 CPAT去除具有蛋白编码潜能的组装后转录本条数；

（6）Remove Protein Family Transcripts　：去除有蛋白家族（ PFam）注释的转录本条数；

（7）CNCI Remove Transcripts　：软件 CNCI去除具有蛋白编码潜能的组装后转录本条数；

（8）CPC Remove Transcripts　：软件 CPC去除具有蛋白编码潜能的组装后转录本条数；

（9）Less Than Two Sample Transcripts　：去除只存在一个样本中的组装后转录本条数；

（10）Potential lncRNA Transcripts　：剩余的潜在 lncRNA转录本条数。

根据各种规则去除转录本的个数的结果展示如下图：

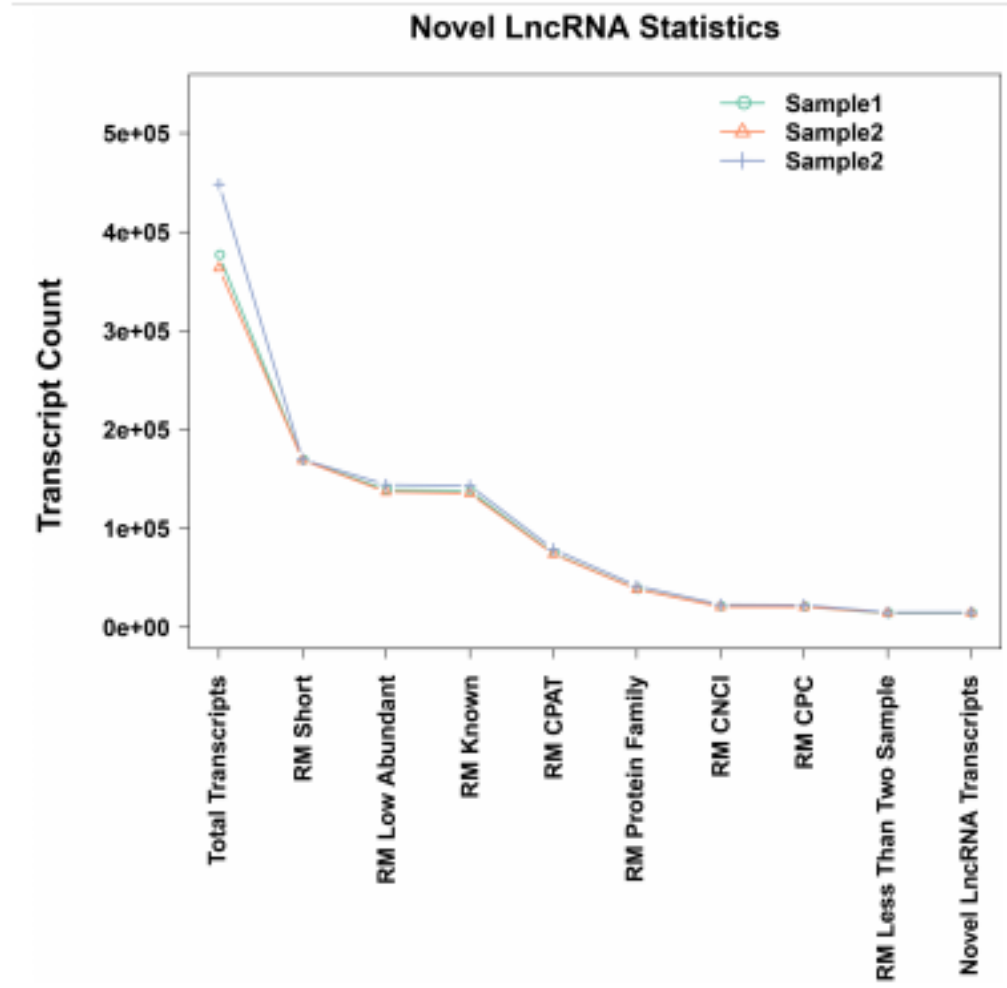


图49　根据规则识别　Novel lncRNA　的评估

9.2　编码潜能分析

编码潜能分析采用了目前主流的四个用于预测　lncRNA编码潜能的软件，包括CPAT、CNCI、CPC和PLEK，其中PLEK仅适用于人类。

9.2.1　CPAT分析

CPAT(Liguo　,et　al.2013)是一个无需比对的快速分类工具，可以从大量的候选转录本中快速的区分出编码和非编码转录本。它使用　ORF长度、ORF覆盖度、Fickett　TESTCODE统计量和核苷酸六联体使用频率偏性等建立逻辑回归模型，来评估候选 lncRNA的编码潜能。

表21　CPAT分析结果

Transcript ID	Transcript Length	ORF Size	Coding Probility
TCONS_00008693	457	102	0.0067
TCONS_00000009	676	177	0.0362
TCONS_00000025	437	87	0.0003
TCONS_00000032	498	87	0.0003
TCONS_00008723	1,620	414	0.5269
TCONS_00008724	1,166	378	0.8968

（1）Transcript ID　：转录本 ID；

（2）Transcript Length　：转录本长度；

（3）ORF Size：编码区域长度；

（4）Coding Probility ：转录本具有编码能力的可能性。

9.2.2 CNCI 分析

CNCI(Liang ,et al.2013) 是一个高效的区分编码和非编码转录本的工具，对lncRNA测序组装出来的转录本具有很高的分类准确性。该工具不依赖于已知的注释文件，而是基于相邻核苷酸三联体推断最佳编码区，利用支持向量机判断候选lncRNA是否为非编码 RNA。

表22 CNCI 分析结果

Transcript ID	Type	Score	Start	End	Transcript Length
TCONS_00008693	noncoding	-0.0041	150	297	1,187
TCONS_00000009	coding	0.0500	24	579	918
TCONS_00000025	coding	0.1110	0	588	995
TCONS_00000032	noncoding	-0.0778	0	72	402
TCONS_00008723	noncoding	-0.0270	240	393	802
TCONS_00008724	coding	0.1110	0	588	995

（1）Transcript ID ：转录本 ID；

（2）Type：鉴定结果（编码或非编码）；

（3）Score：对转录本编码潜能的打分；

（4）Start ：CDS起始位置；

（5）End：CDS终止位置；

（6）Transcript Length ：转录本长度。

9.2.3 CPC分析

CPC(Lei ,et al.2007) 是一种计算蛋白质编码潜能的工具，利用 framefinder 提取候选 lncRNA的3个ORF指标，并将候选 lncRNA与已知蛋白数据库进行 BLASTX 获取 3个比对指标，然后利用这些指标通过支持向量机对候选lncRNA进行编码潜能计算，从而判定其是否为 lncRNA。

表23　CPC分析结果

Transcript ID	Transcript Length	Type	Score
TCONS_00000051	1,944	coding	5.7102
TCONS_00001156	809	coding	0.3685
TCONS_00001157	771	coding	0.3791
TCONS_00000011	457	noncoding	-1.0467
TCONS_00000029	676	noncoding	-1.0655
TCONS_00000031	607	noncoding	-0.9138

（1）Transcript ID　：转录本 ID；

（2）Transcript Length　：转录本长度；

（3）Type：鉴定结果（编码或非编码）；

（4）Score：对转录本编码潜能的打分。

9.2.4　PLEK分析

PLEK(Li ,et al. 2014)　是一个快速的区分 lncRNA和mRNA的工具，它同样无需比对，而是基于优化的 k-mer策略和支持向量机算法，识别出 lncRNA。该方法在人类lncRNA的预测中准确率高达 95.6% 。

表24　PLEK分析结果

Transcript ID	Type	Score
TCONS_00000011	Non-coding	-1.4893
TCONS_00000029	Non-coding	-2.0140
TCONS_00000031	Non-coding	-2.2391
TCONS_00000032	Non-coding	-2.4465
TCONS_00000033	Non-coding	-1.6264
TCONS_00000034	Non-coding	-2.4080

（1）Transcript ID　：转录本 ID；

（2）Type：鉴定结果（结果文件只列出非编码转录本）；

（3）Score：转录本得分，负数代表不具有编码能力。

9.3　特征分析

9.3.1　Novel lncRNA长度统计

对所有样品的 Novel lncRNA 进行长度分布统计，统计结果如下图：

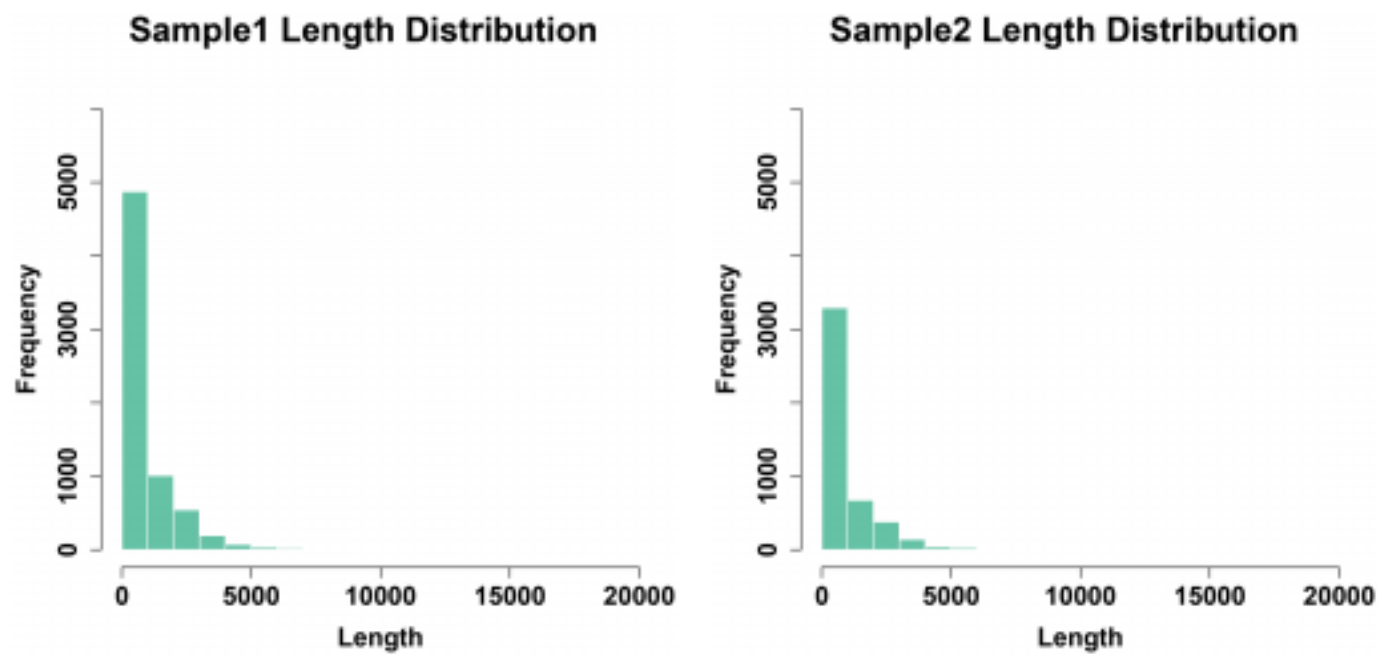


图50　Novel lncRNA　　长度分布

横坐标为 Novel lncRNA 的长度，纵坐标为该长度下的频率分布。

根据所有样本的长度分布，总体长度分布的趋势展示如下图：

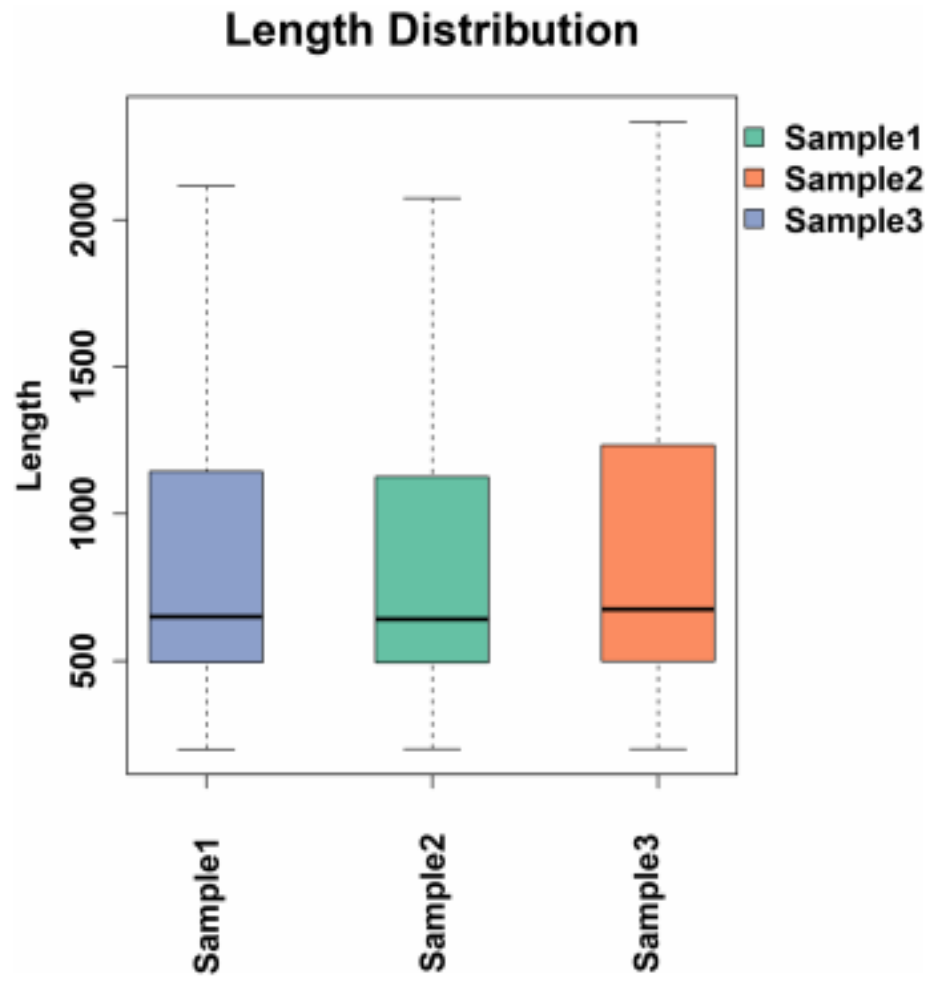


图51　样本长度分布统计趋势结果图

9.3.2　Novel ln　cRNA外显子个数统计

根据所有样品的 Novel　lncRNA进行外显子（ Exon）个数统计，统计结果如下图：

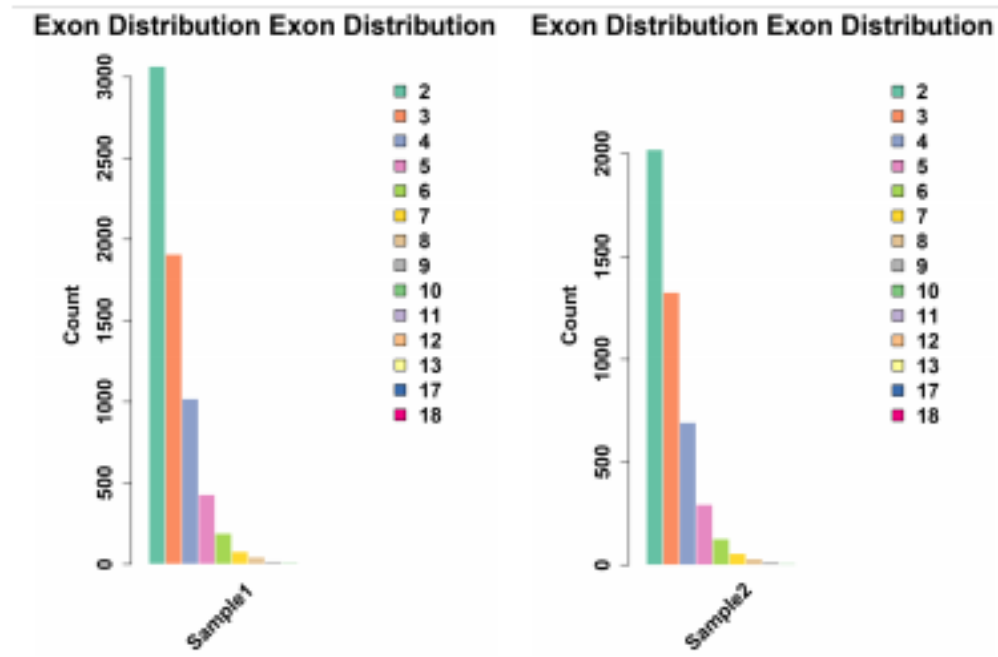


图52　Novel lncRNA　外显子个数统计结果；

横坐标为样本，纵坐标为具有该外显子个数的　Novel lncRNA 的数目。

9.3.3　编码基因与　lncRNA转录本的长度分布比较

根据编码基因和 lncRNA的转录本的长度分布状态，将两者转录本的长度分布趋势进行比较，查看两者分布的差异性与一致性，结果展示如下图：

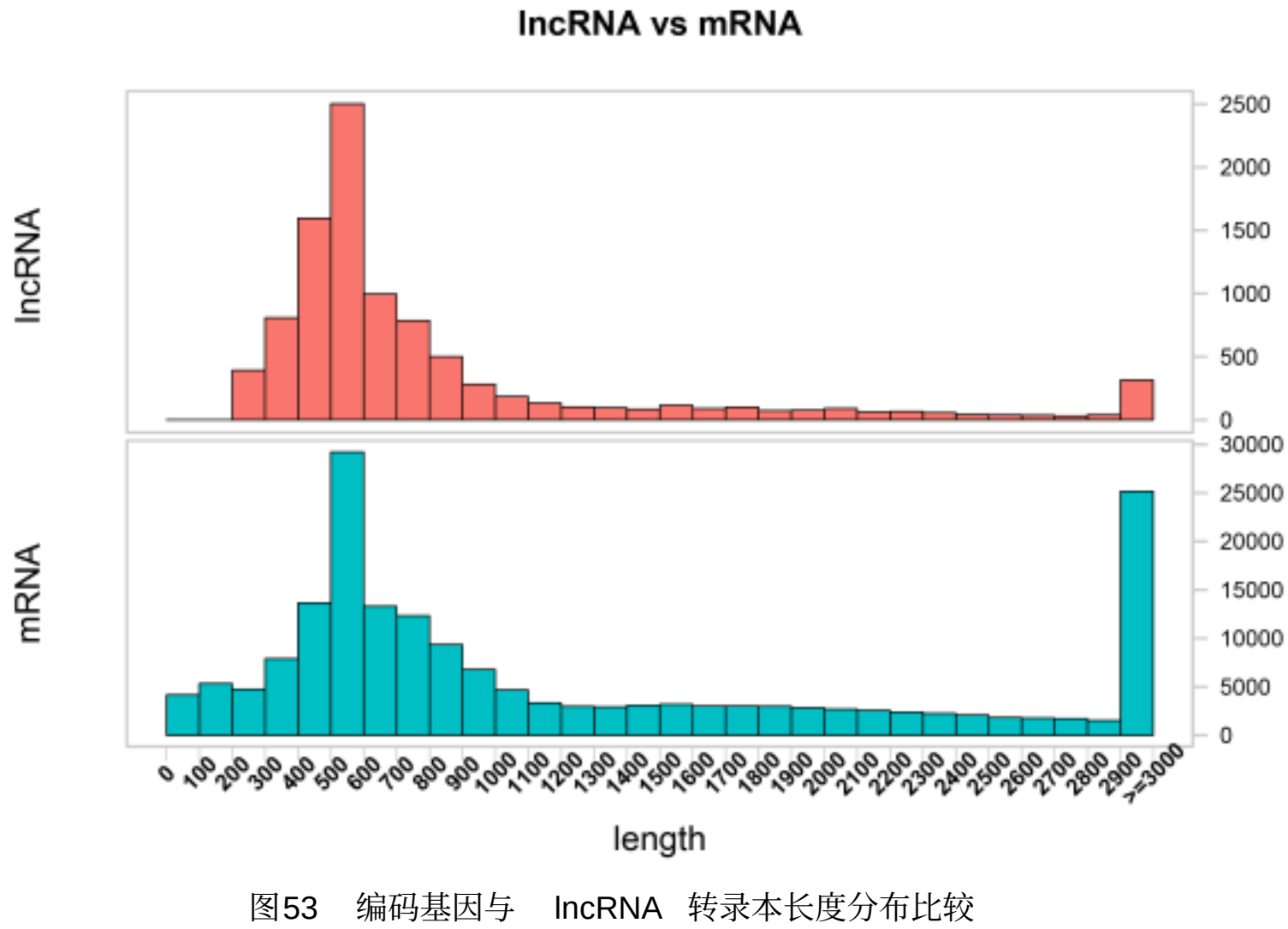


图53　编码基因与　lncRNA　转录本长度分布比较

9.3.4　编码基因与　lncRNA转录本外显子分布比较

根据编码基因与 lncRNA的转录本的外显子个数的分布，将两者进行比较，查看两者分布的差异性与一致性，结果展示如下图：

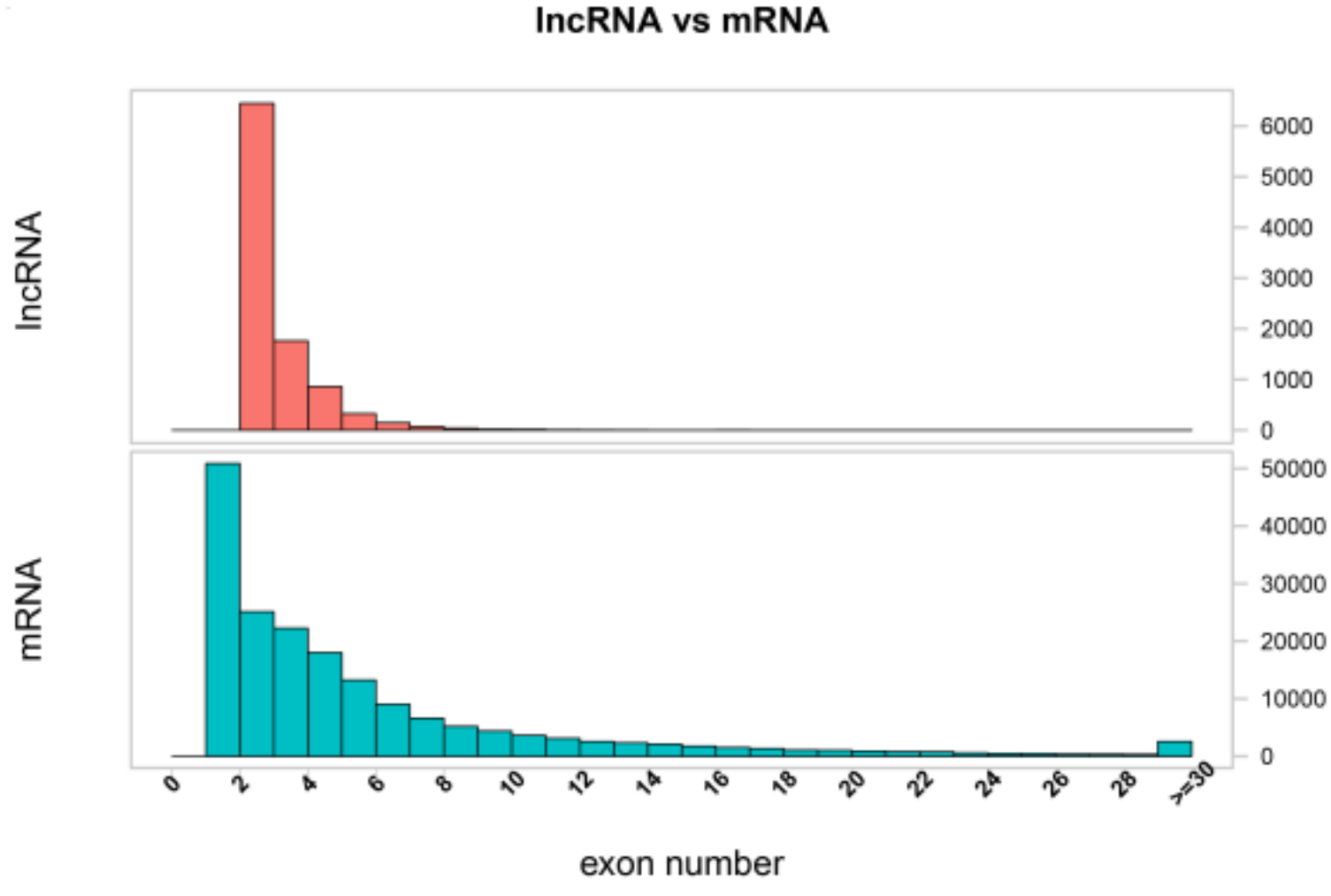


图54　编码基因与　lncRNA　转录本外显子个数的分布比较

9.4　保守性分析

对于识别到的　Novel　　lncRNA，希望能够了解其对应的序列的保守性及位点保守性分析，以便推测　Novel　　lncRNA 的变异程度。根据每个样品识别的Novel　　lncRNA的GTF文件，使用 CuffMerge合并所有样品的 GTF文件，然后利用PhastCons(Florea　　,et　　al.　　2013)根据从 UCSC数据库下载的模式生物（人类）保守性估计的文件，对　Novel lncRNA 分析相对的保守性。

PhastCons　Score是用种系发生的隐马尔科夫模型的方法衡量许多脊椎动物某段序列进化的保守性分数。该方法不同于其他保守性打分的程序，不依赖于固定大小的滑动窗口，短序列高度保守的区域以及较长的比较保守区域可以获得较高的分值，可表征每个位点处碱基保守性的概率。从　UCSC-Download下载人类基因组HG19版本下每条染色体包含的每个基因上各位点的　PhastCons分数，以 0~1之间的数值表示。

9.4.1　位点保守性

根据识别的 lncRNA的转录本中对应的位点，计算得到每个位点的保守性得分，得到所有 lncRNA覆盖到的每个位点的保守性得分。

表25　位点保守性得分

Chr omo so me	chr 10
Location	978,966
PhastCons	0.0000

（1）Chromosome：染色体信息；

（2）Location ：位点信息；

（3）PhastCons：位点保守性得分。

同时会计算编码基因的每个位点的保守性得分，并根据每条染色体比较两者的位点保守性得分。橙色表示为编码基因的位点保守性得分，蓝色表示 lncRNA的保守性得分。

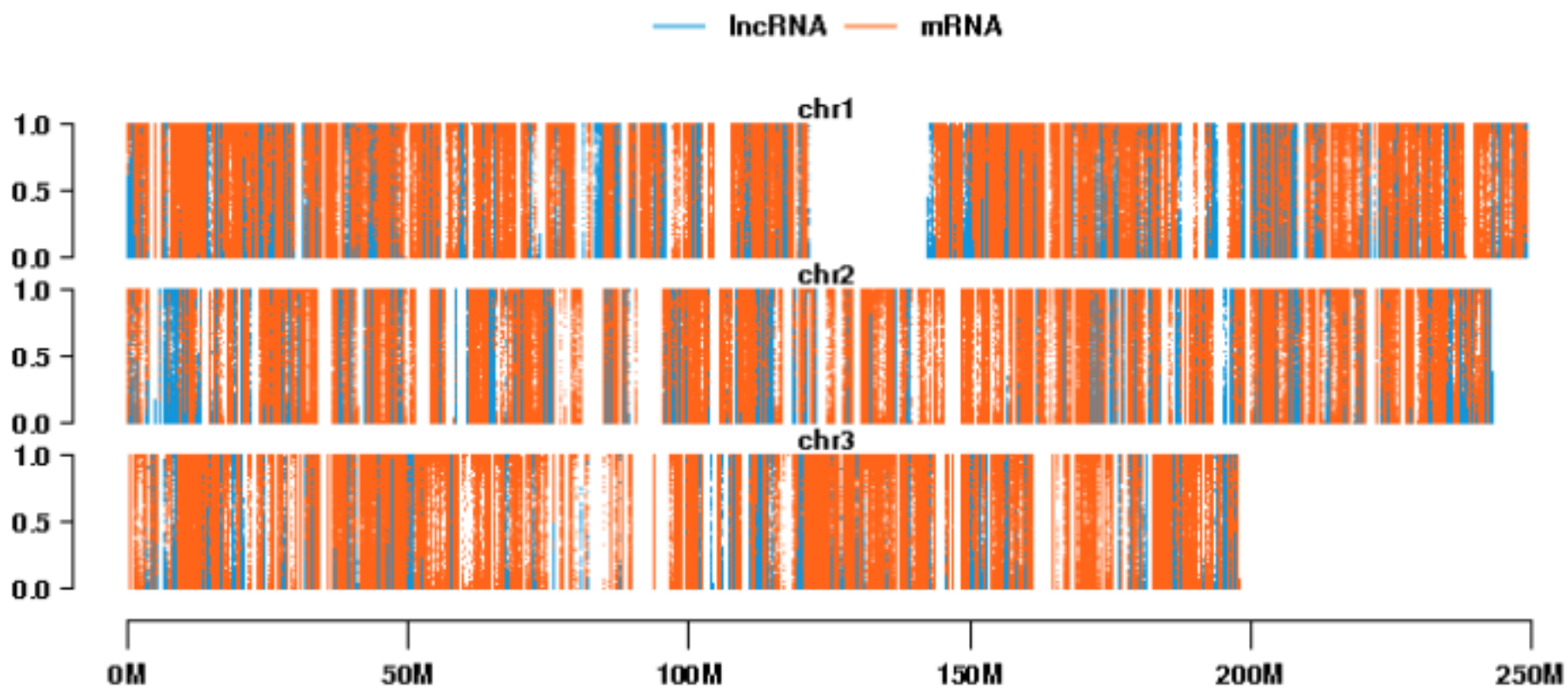


图55　位点保守性得分

lncRNA和mRNA的保守性得分 wig文件导入 UCSC（具体操作参考操作文档说明），可以展示任意精度的保守性分布，示例如下图：

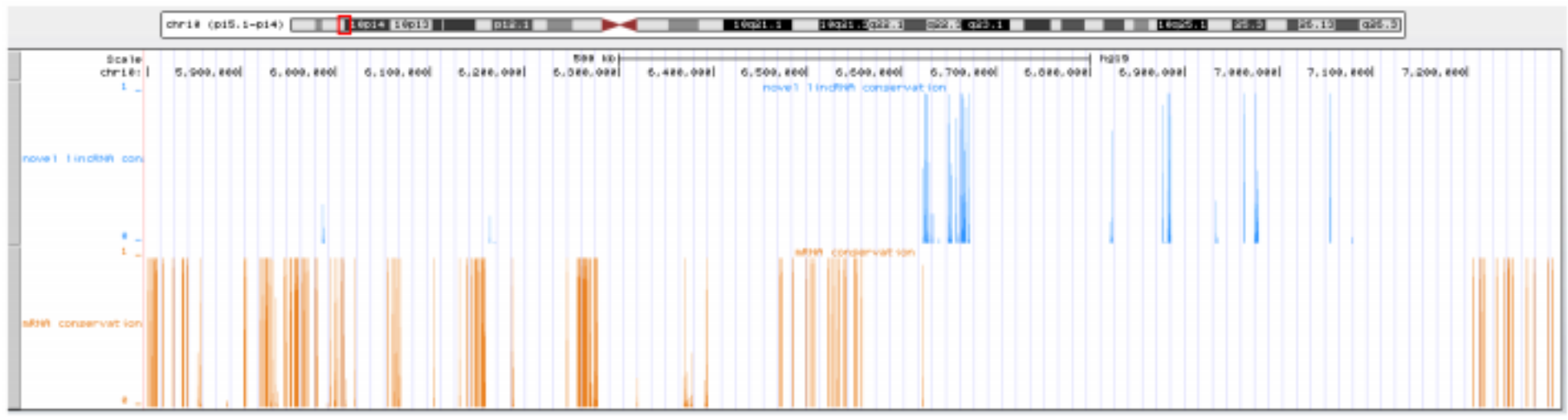


图56　位点保守性得分——　UCSC tracks

上面的 Track 为 lncRNA（蓝色），下面的 Track 是 mRNA（橙色），高度代表保守性强度，柱子越高，保守性越强。

同时根据两者的位点保守性得分，分析编码基因与 lncRNA 两者的得分分布频率，结果展示如下：

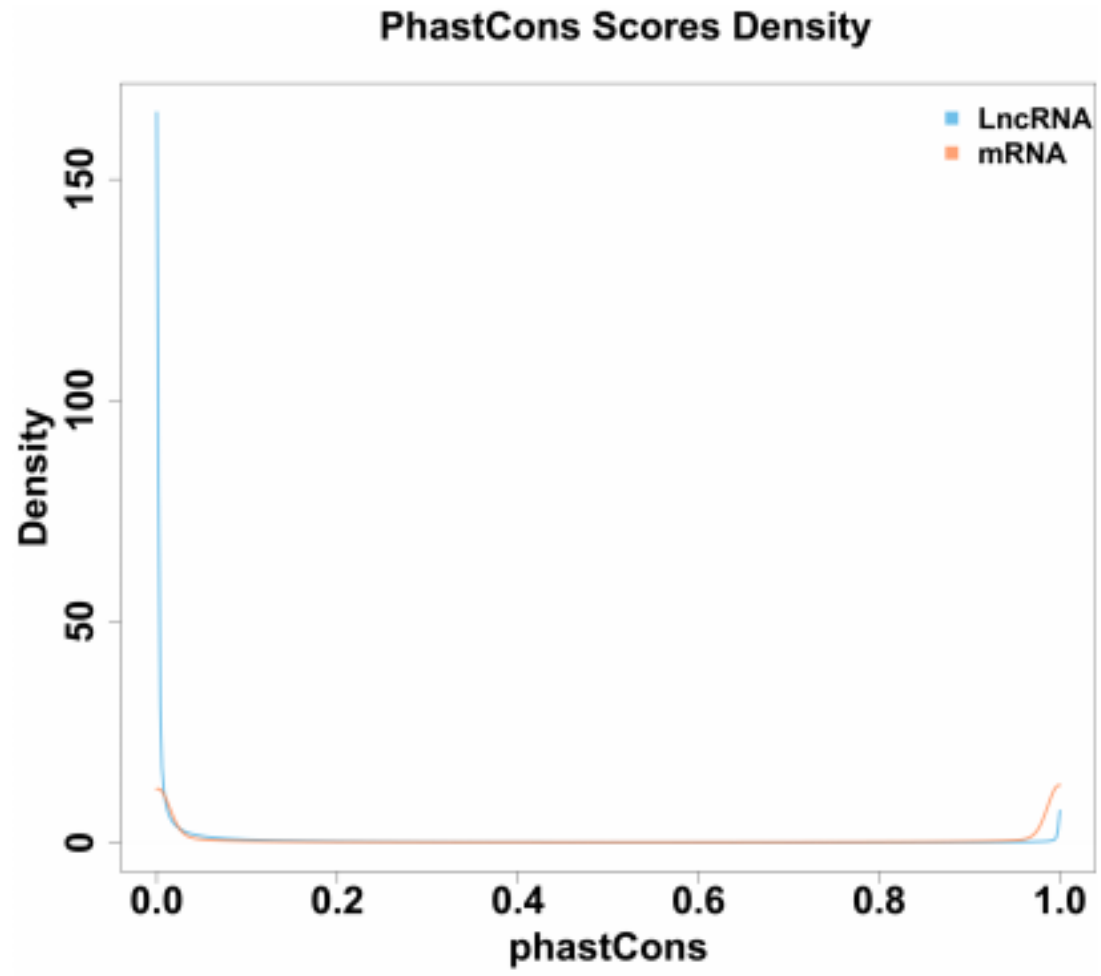


图57　位点保守得分密度分布图

9.4.2　序列保守性

根据 lncRNA 的每个位点的保守性得分，计算每个 lncRNA 的转录本的保守性得分，根据序列中的位点保守性得分的均值，作为转录本的保守性得分。

表26　序列保守性得分

Chromosome	chr10
Start	126951720
End	126974964
Strand	+
Novel_lncRNA	Novel_Lnc_00001003
Score	0.1446
Exon_Length	678
Exon_count	5

（1）Chromosome：染色体信息；

（2）Start ：Novel_lncRNA起始位置；

（3）End：Novel_lncRNA终止位置；

（4）Strand ：Novel_lncRNA正负链信息；

（5）Novel_lncRNA：Novel_lncRNA对应名字信息；

（6）Score：PhastCons Score 值；

（7）Exon_Length：Novel_lncRNA外显子长度；

（8）Exon_count：Novel_lncRNA对应外显子长度。

根据上表信息中得到的 PhastScore 值绘制频率直方图，展示结果如下：

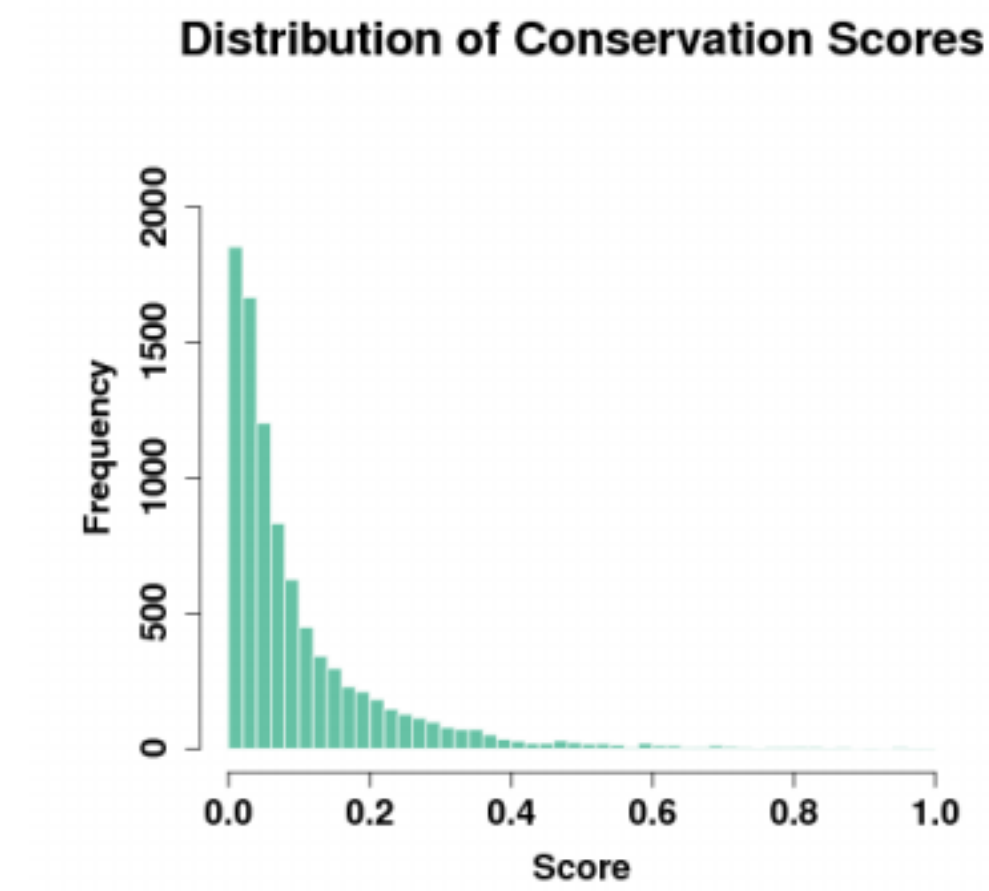


图58 保守性 PhastScores 值分布统计结果

9.5 估计表达量

9.5.1 Novel lncRNA表达量估计

同mRNA表达量估计原理一致，根据每个样本识别到的 Novel lncRNA的GTF文件，使用 CuffMerge 整合所有的 GTF文件，并根据整合之后的 GTF文件以及TopHat比对结果计算 Novel lncRNA 的表达量。

9.5.2 编码基因与 lncRNA表达量的比较

根据编码基因的表达量与 lncRNA的表达量大小进行比较，去除低表达量的基因，在每个样本中比较两者的区别，结果展示如下图：

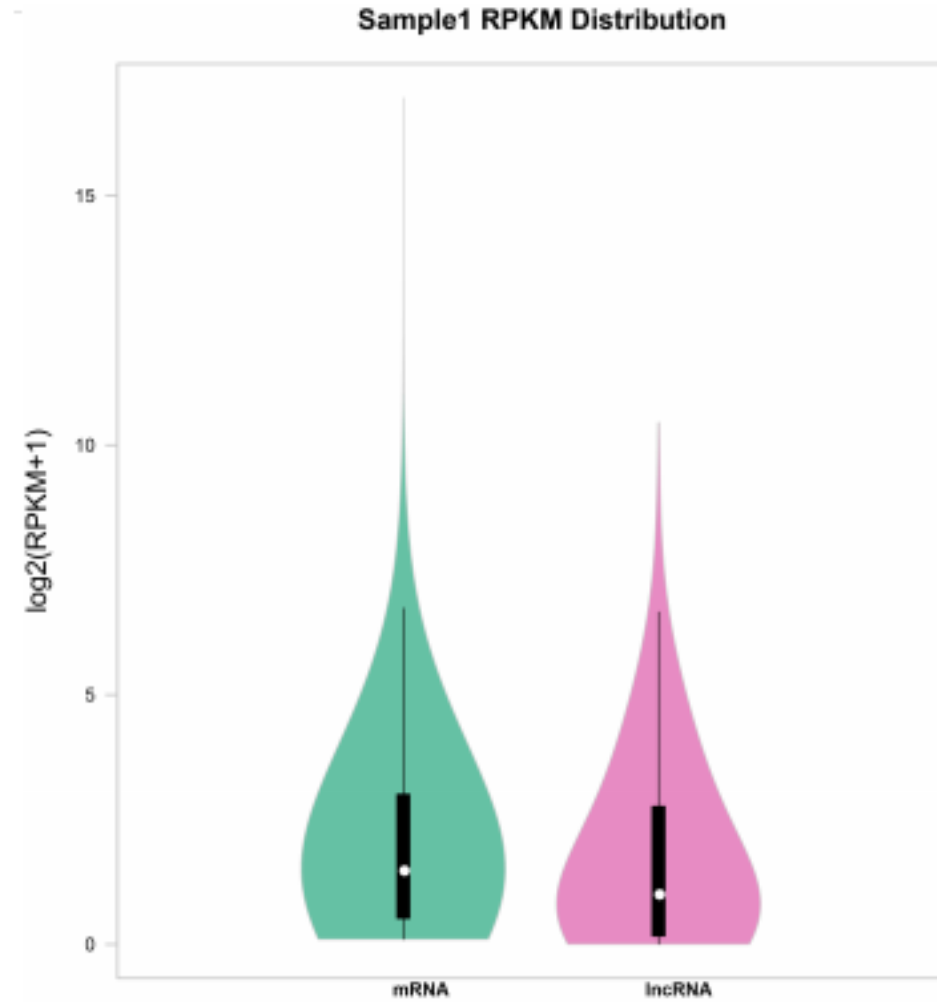


图59　编码基因与　lncRNA　表达量的比较

9.5.3　Novel ln　cRNA表达量分布统计

一般而言，差异表达 lncRNA的数量只占整体 lncRNA的小部分，因此少量的差异表达lncRNA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 Novel　lncRNA表达量，得到该样品 Novel　lncRNA的表达量密度图如下：

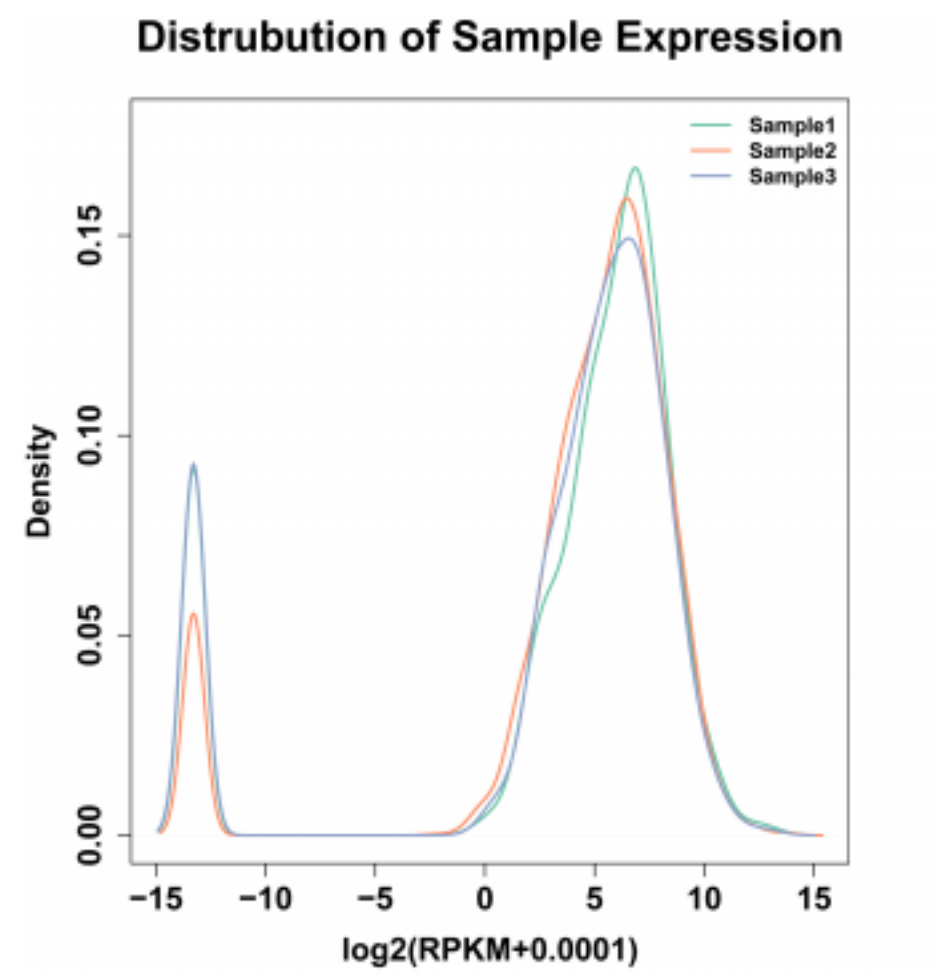


图60　Novel lncRNA　表达量分布

对每组样品的 lncRNA表达量，取以 2为底的对数后，做出密度分布图。横坐标为$\log_2$(RPKM+0.0001)，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：

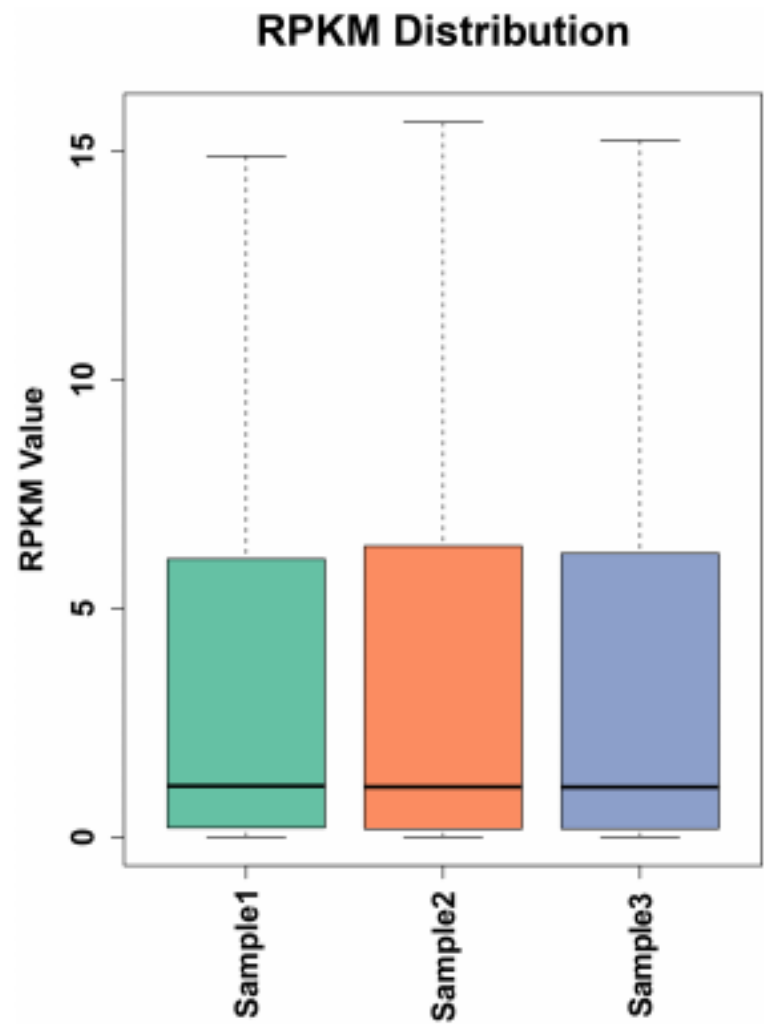


图61　表达量箱式分布图

9.5.4　Novel ln　cRNA样品实验的聚类

一般情况下，源于同一实验条件下的样品会聚类到一起，表明实验条件为影响聚类的主要因素。根据样品全部已知或　Novel　lncRNA的表达量信息对样品进行系统聚类（注：只有当样品数目≥　3时才会有样品聚类图）

根据样品全部 Novel　lncRNA 的表达量信息对样品进行系统聚类，得到下图：

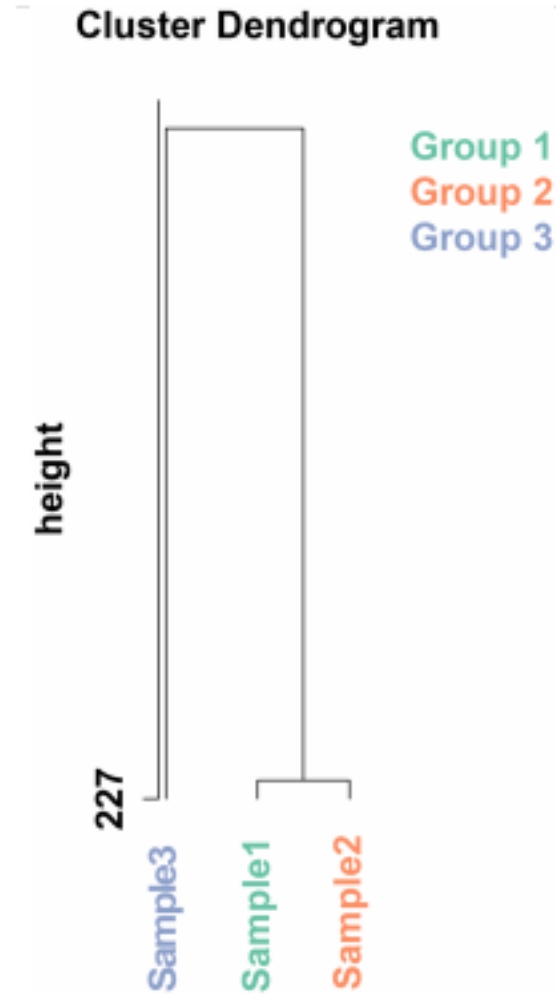


图62　Novel lncRNA　的 Cluster　聚类图

根据每个样品的 lncRNA表达量，计算两两样品之间的欧氏距离，然后根据类平均距离法度量两两类之间的距离，再利用系统聚类法（　Hierarchical Cluster　）进行聚类，最终得到样品的整体聚类结果。

9.6　Novel 差异表达分析

9.6.1　lncRNA差异表达分析统计结果

对于设置生物学重复的实验，我们采用　DESeq进行lncRNA差异表达分析，比较处理组与参考组，并选取　$|\log_2 Ratio| \geq 1$和q<0.05的lncRNA作为差异表达lncRNA，得到上下调 lncRNA个数。

对于无生物学重复样品，则采用　DEGseq进行lncRNA差异表达分析，比较处理组与参考组，并选取　$|\log_2 Ratio| \geq 1$和q<0.05的lncRNA作为差异表达 lncRNA，得到上下调 lncRNA个数。

本项目根据上下调 lncRNA，绘制差异表达 lncRNA火山图：

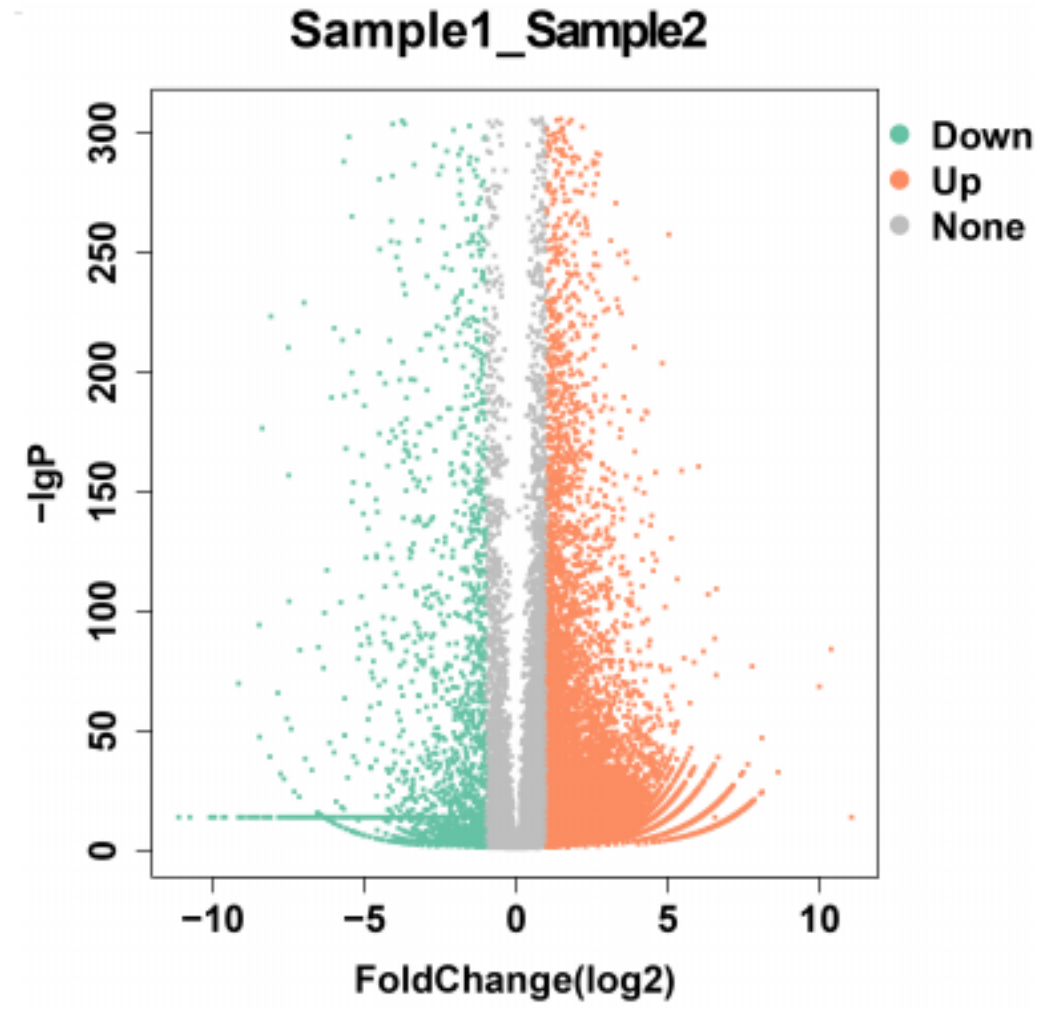


图63　差异表达　lncRNA　火山图

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。

本项目所有组别差异表达 lncRNA结果如下：

表27　组间比较得到的差异表达　　lncRNA　数目

name	Sample1_VS_Sample2
Up	3,477
Down	669
Total	4,146

（1）Up：表示在第一组（例如 Sample1）中表达上调的 lncRNA；

（2）Down: 表示在第一组（例如 Sample2）中表达下调的 lncRNA；

（3）Total ：表示在两组中有差异 lncRNA数目总和。

根据上表，统计的结果如下图：

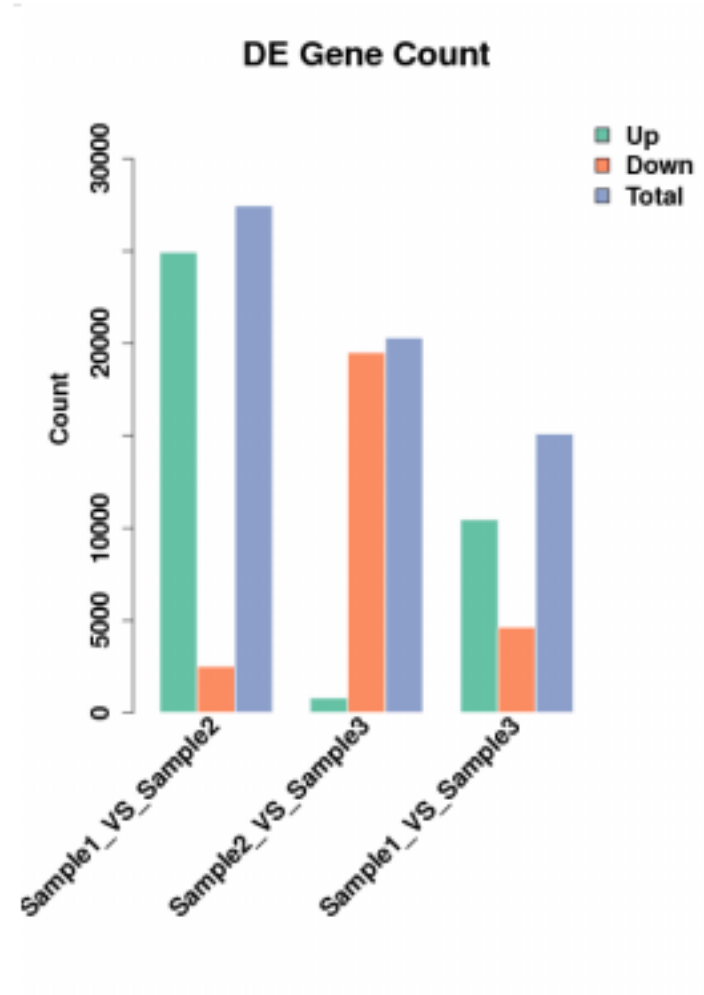


图64　差异表达　lncRNA　统计图

通过比较处理组和参考组，对差异表达　lncRNA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　lncRNA的变化情况。我们利用　R软件（版本号：v3.3.1 ），对差异表达 lncRNA和不同样本 / 实验条件同时进行分层聚类分析。下图为两组样本的差异表达　lncRNA聚类示例。

本项目所有差异表达 lncRNA聚类分析结果如下：

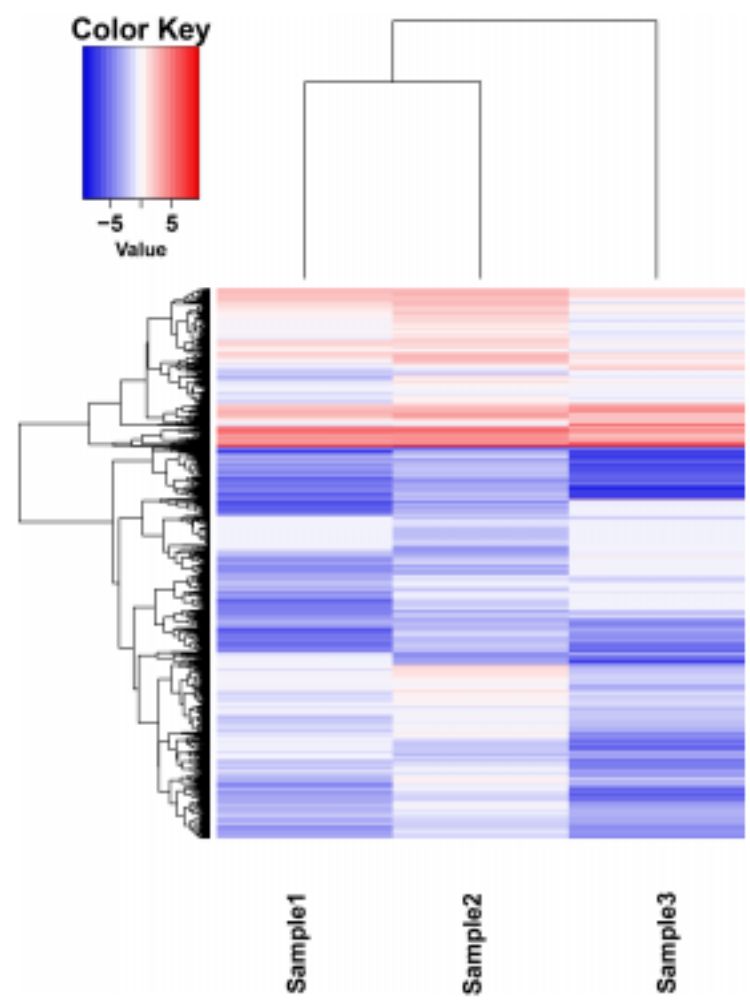


图65　差异 lncRNA　聚类图

根据差异表达 lncRNA在每个样品里的表达量，取以　2为底的对数后，计算欧氏距离，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。

9.7　靶标预测分析

部分已知 lncRNA在GO等功能数据库中存在功能注释信息，可以直接进行功能分析；另外的已知 lncRNA以及鉴定的 Novel lncRNA尚无功能信息，因此采用通过预测靶基因（ Cis作用靶标和 Trans作用靶标）间接预测其功能的策略。

9.7.1　Novel ln cRNA的 Ci s作用靶标预测及功能注释

Cis功能预测基本原理是依据 lncRNA的功能与其坐标临近的编码蛋白基因相关。根据找出同 lncRNA基因相邻的（上下游 20K）蛋白编码的基因对 lncRNA筛选，同时计算 lncRNA和 mRNA的表达量的相关性， Cis靶标预测相关性系数筛选标准是保留相关性大于 0.9的基因对关系。

Cis作用靶基因预测结果及功能注释如下表所示：

表28　**Cis作用靶标预测**

LncRN A	TCONS_0 0 1 6 4 733
LncRNA_position	chr8:105603126-105606837:-
mRNA	ENSG00000147650
Correlation	0.9730
Distance	1,709
Position	chr8:105501459-105601417:-
NR:Seq-id	gi|281364746|ref|NP_723538.3|
NR:Score	3,613
NR:Evalue	0
NR:Description	"methuselah-like 15, isoform C [Drosophila melanogaster]"
NT:Seq-id	gi|281364747|ref|NM_164896.2|
NT:Score	2,836
NT:Evalue	0
NT:Description	Drosophila melanogaster methuselah-like 15 (mthl15)
Uniprot:UniProtKB-AC	Q9V818
Uniprot:Score	36.6000
Uniprot:Evalue	3.00E-24
Uniprot:Description	Probable G-protein coupled receptor Mth-like 3
COG:gene	.
COG:Score	.
COG:Eval	.
COG:num	.
Pfam:pfam_ID	pfam00002
Pfam:pfam_Name	7tm_2
Pfam:pfam_Description	7 transmembrane receptor (Secretin family).
GO:biological_process	GO:0008340|determination of adult lifespan;
GO:cellular_component	GO:0016021|integral component of membrane;GO:0005886|plasma membrane;
GO:molecular_function	GO:0004930|G-protein coupled receptor activity;
KEGG:KO	K04599
KEGG:Description	MTH; G protein-coupled receptor Mth (Methuselah protein)

（1）LncRNA：lncRNA的名称；

（2）LncRNA_position ：lncRNA的位置；

（3）mRNA：转录本名称；

（4）Correlation　：表达量的相关性；

（5）Distance ：lncRNA与转录本之间的距离。

（6）Position　：基因的坐标；

（7）NR:Seq-id ：基因同 NR数据库的最优比对结果；

（8）NR:Score：基因同 NR 数据库的比对得分；

（9）NR:Evalue：基因同 NR 数据库的比对 Evalue 值；

（10）NR:Description ：NR 数据库中该基因的功能描述；

（11）NT:Seq-id ：基因同 NT 数据库的最优比对结果；

（12）NT:Score：基因同 NT 数据库的比对得分；

（13）NT:Evalue：基因同 NT 数据库的比对 Evalue 值；

（14）NT:Description ：NT 数据库中该基因的功能描述；

（15）Uniprot:UniProtKB-AC ：基因同 Uniprot 数据库的最优比对结果；

（16）Uniprot:Score ：基因同 Uniprot 数据库的比对得分；

（17）Uniprot:Evalue ：基因同 Uniprot 数据库的比对 Evalue 值；

（18）Uniprot:Description ：niprot 数据库中该基因的功能描述；

（19）COG:gene: 比对上的 COG 数据库中的基因名；

（20）COG:Score: 与 COG 数据库的比对得分；

（21）COG:Eval：与 COG 数据库的比对 Evalue 值；

（22）COG:num: 比对上的 COG 数据库中的基因 ID；

（23）Pfam:pfam_ID：比对上的蛋白家族 Pfam 的基因 ID；

（24）Pfam:pfam_Name: 比对上的蛋白家族 Pfam 的基因名；

（25）Pfam:pfam_Description ：比对上的蛋白家族 Pfam 的功能描述；

（26）GO:biological_process ：注释到的描述生物进程的 GO Term;

（27）GO:cellular_component ：注释到的描述细胞组分的 GO Term;

（28）GO:molecular_function ：注释到的描述分子功能的 GO Term;

（29）KEGG:KO: 注释到的 KEGG 中的 ID；

（30）KEGG:Description ：KEGG 中的功能描述。

9.7.2 Novel lncRNA 的 Trans 靶标预测及功能注释

lncRNA 的 Trans 靶标预测，样本数大于 3 个才可以进行样品之间的靶标预测分析。首先根据 lncRNA 和基因的表达量计算 lncRNA 和基因之间的相关性，并且根据相应 GTF 文件和基因组文件提取对应的 Fasta 序列，使用 RNAPlex 根据提取的 Fasta 序列，计算 lncRNA 和基因序列之间的结合自由能来预测反式作用靶标。根据自由能和相关性筛选预测靶标，筛选标准 Cor>0.75。

反式靶标预测及功能注释示例如下：

表29　**Trans**作用靶基因预测

LncRN A	XLOC_0 0 0 2 37
LncRNA_position	chr16:4296375-4303780:-
Gene_id	ENSG00000185338
Correlation	0.8769
Energy	-104.3000
Position	chr16:11348262-11350036:-
NR:Seq-id	gi|281364746|ref|NP_723538.3|
NR:Score	3,613
NR:Evalue	0
NR:Description	"methuselah-like 15, isoform C [Drosophila melanogaster]"
NT:Seq-id	gi|281364747|ref|NM_164896.2|
NT:Score	2,836
NT:Evalue	0
NT:Description	Drosophila melanogaster methuselah-like 15 (mthl15)
Uniprot:UniProtKB-AC	Q9V818
Uniprot:Score	36.6000
Uniprot:Evalue	3.00E-24
Uniprot:Description	Probable G-protein coupled receptor Mth-like 3
COG:gene	.
COG:Score	.
COG:Eval	.
COG:num	.
Pfam:pfam_ID	pfam00002
Pfam:pfam_Name	7tm_2
Pfam:pfam_Description	7 transmembrane receptor (Secretin family).
GO:biological_process	GO:0008340|determination of adult lifespan;
GO:cellular_component	GO:0016021|integral component of membrane;GO:0005886|plasma membrane;
GO:molecular_function	GO:0004930|G-protein coupled receptor activity;
KEGG:KO	K04599
KEGG:Description	MTH; G protein-coupled receptor Mth (Methuselah protein)

（1）LncRNA：lncRNA的名称；

（2）LncRNA_position ：lncRNA的位置；

（3）Gene_id：基因的名称；

（4）Correlation　：表达量的相关性；

（5）Energy：结合自由能；

（6）Position　：基因的坐标；

（7）NR:Seq-id ：基因同 NR数据库的最优比对结果；

（8）NR:Score：基因同 NR 数据库的比对得分；

（9）NR:Evalue：基因同 NR 数据库的比对 Evalue 值；

（10）NR:Description ：NR 数据库中该基因的功能描述；

（11）NT:Seq-id ：基因同 NT 数据库的最优比对结果；

（12）NT:Score：基因同 NT 数据库的比对得分；

（13）NT:Evalue：基因同 NT 数据库的比对 Evalue 值；

（14）NT:Description ：NT 数据库中该基因的功能描述；

（15）Uniprot:UniProtKB-AC ：基因同 Uniprot 数据库的最优比对结果；

（16）Uniprot:Score ：基因同 Uniprot 数据库的比对得分；

（17）Uniprot:Evalue ：基因同 Uniprot 数据库的比对 Evalue 值；

（18）Uniprot:Description ：niprot 数据库中该基因的功能描述；

（19）COG:gene: 比对上的 COG 数据库中的基因名；

（20）COG:Score: 与 COG 数据库的比对得分；

（21）COG:Eval：与 COG 数据库的比对 Evalue 值；

（22）COG:num: 比对上的 COG 数据库中的基因 ID；

（23）Pfam:pfam_ID：比对上的蛋白家族 Pfam 的基因 ID；

（24）Pfam:pfam_Name: 比对上的蛋白家族 Pfam 的基因名；

（25）Pfam:pfam_Description ：比对上的蛋白家族 Pfam 的功能描述；

（26）GO:biological_process ：注释到的描述生物进程的 GO Term;

（27）GO:cellular_component ：注释到的描述细胞组分的 GO Term;

（28）GO:molecular_function ：注释到的描述分子功能的 GO Term;

（29）KEGG:KO: 注释到的 KEGG 中的 ID；

（30）KEGG:Description ：KEGG 中的功能描述。

9.7.3 WGCNA 预测 Tra ns 靶标

当样本个数大于 25个、组内重复大于等于 3个重复时，通过如下方法预测Trans 靶标。

WGCNA（加权重共表达网络） (Langfelder ,et al. 2008) 是一个用于共表达网络分析的工具，通过计算每对基因两两间的相关系数，可以得到两两基因的相关性，再根据相关性进行聚类，最终得到具有相同表达模式分类。

通过 WGCNA 软件预测 Trans 靶标，在预测 Trans 靶标的过程中，通过计算lncRNA 与编码基因的表达情况（表达量），预测与 lncRNA 表达情况相似的编码基因的集合，该编码基因的集合即预测到的 lncRNA 的 Trans 靶标。

预测 lncRNA 和编码基因的共表达结果示例如下图：

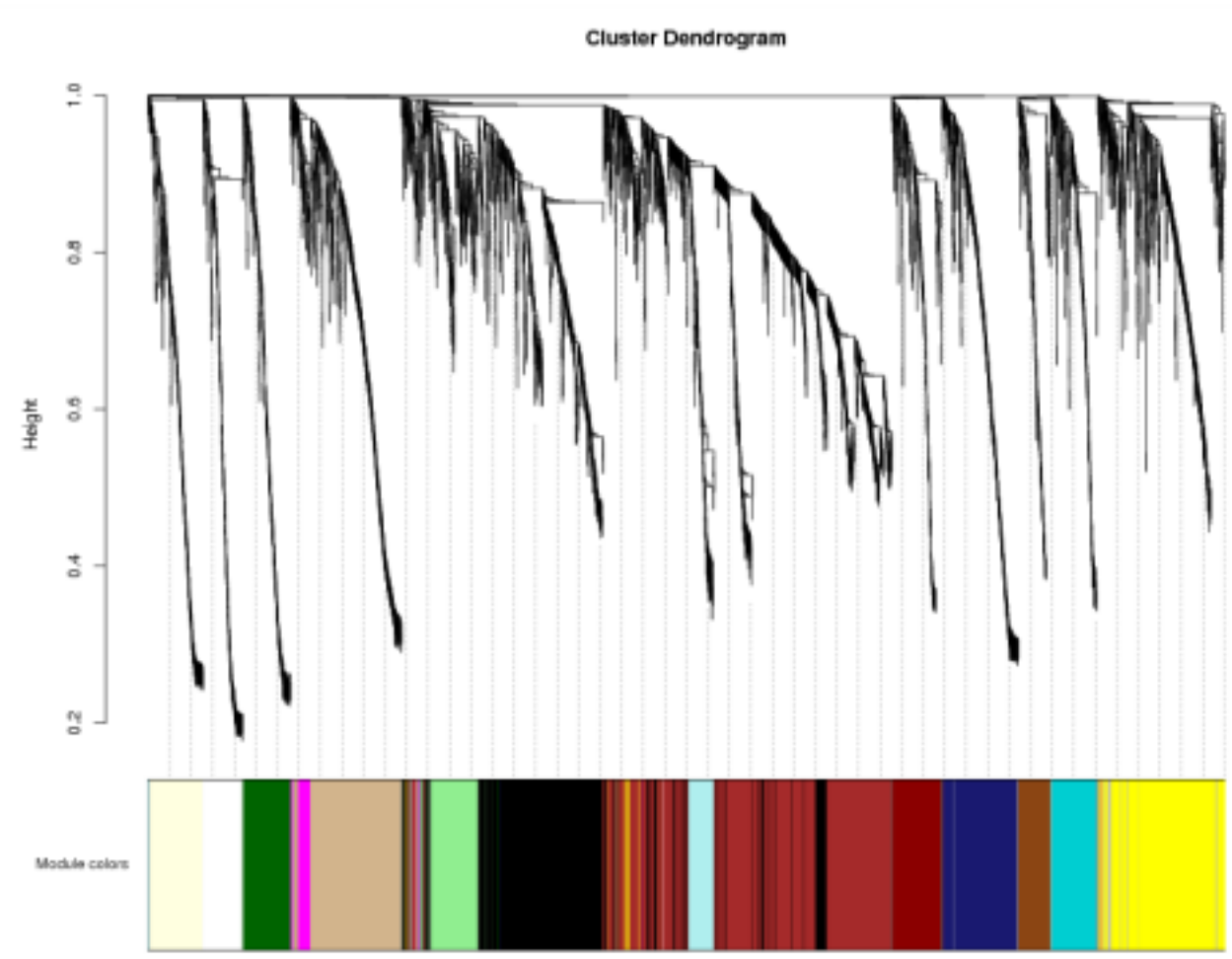


图66　lncRNA 的 Trans 靶标预测

根据lncRNA预测的 Trans 靶标的编码基因进行功能注释，结果示例如下表：

表30　WGCNA预测 **Trans**靶标lncRNA的功能注释结果

LncRNA	XLOC_000237
LncRNA_position	chr16:4296375-4303780:-
Gene_id	ENSG00000185338
Position	chr16:11348262-11350036:-
NR:Seq-id	gi|281364746|ref|NP_723538.3|
NR:Score	3,613
NR:Evalue	0
NR:Description	"methuselah-like 15, isoform C [Drosophila melanogaster]"
NT:Seq-id	gi|281364747|ref|NM_164896.2|
NT:Score	2,836
NT:Evalue	0
NT:Description	Drosophila melanogaster methuselah-like 15 (mthl15)
Uniprot:UniProtKB-AC	Q9V818
Uniprot:Score	36.6000
Uniprot:Evalue	3.00E-24
Uniprot:Description	Probable G-protein coupled receptor Mth-like 3
COG:COG_gene	.
COG:COG_Score	.
COG:COG_Eval	.
COG:COG_num	.
Pfam:pfam_ID	pfam00002
Pfam:pfam_Name	7tm_2
Pfam:pfam_Description	7 transmembrane receptor (Secretin family).
GO:biological_process	GO:0008340|determination of adult lifespan;
GO:cellular_component	GO:0016021|integral component of membrane;GO:0005886|plasma membrane;
GO:molecular_function	GO:0004930|G-protein coupled receptor activity;
KEGG:KO	K04599
KEGG:Description	MTH; G protein-coupled receptor Mth (Methuselah protein)

（1）LncRNA：lncRNA的名称；

（2）LncRNA_position ：lncRNA的位置；

（3）Gene_id：基因的名称；

（4）Position ：基因的坐标；

（5）NR:Seq-id ：基因同 NR数据库的最优比对结果；

（6）NR:Score：基因同 NR数据库的比对得分；

（7）NR:Evalue：基因同 NR数据库的比对 Evalue 值；

（8）NR:Description ：NR数据库中该基因的功能描述；

（9）NT:Seq-id ：基因同 NT数据库的最优比对结果；

（10）NT:Score：基因同 NT数据库的比对得分；

（11）NT:Evalue：基因同 NT数据库的比对 Evalue 值；

（12）NT:Description ：NT数据库中该基因的功能描述；

（13）Uniprot:UniProtKB-AC ：基因同 Uniprot 数据库的最优比对结果；

（14）Uniprot:Score ：基因同 Uniprot 数据库的比对得分；

（15）Uniprot:Evalue ：基因同 Uniprot 数据库的比对 Evalue 值；

（16）Uniprot:Description ：niprot 数据库中该基因的功能描述；

（17）COG:gene：比对上的 COG数据库中的基因名；

（18）COG:Score：与 COG数据库的比对得分；

（19）COG:Eval：与 COG数据库的比对 Evalue 值；

（20）COG:num：比对上的 COG数据库中的基因 ID；

（21）Pfam:pfam_ID：比对上的蛋白家族 Pfam的基因 ID；

（22）Pfam:pfam_Name：比对上的蛋白家族 Pfam的基因名；

（23）Pfam:pfam_Description ：比对上的蛋白家族 Pfam的功能描述；

（24）GO:biological_process ：注释到的描述生物进程的 GO Term；

（25）GO:cellular_component ：注释到的描述细胞组分的 GO Term；

（26）GO:molecular_function ：注释到的描述分子功能的 GO Term；

（27）KEGG:KO：注释到的 KEGG中的 ID；

（28）KEGG:Description ：KEGG中的功能描述。

9.7.4 Novel lncRNA靶基因调控网络分析

根据识别出的差异表达 lncRNA基因与 mRNA基因及 lncRNA的顺式、反式靶标预测的基因的关系，绘制差异 lncRNA与靶基因的调控网络分析。三角形表示mRNA基因，菱形表示 lncRNA基因；红色表示 mRNA基因，青色表示 lncRNA基因。调控关系分为 Trans、Cis调控关系，示例结果图见下图：

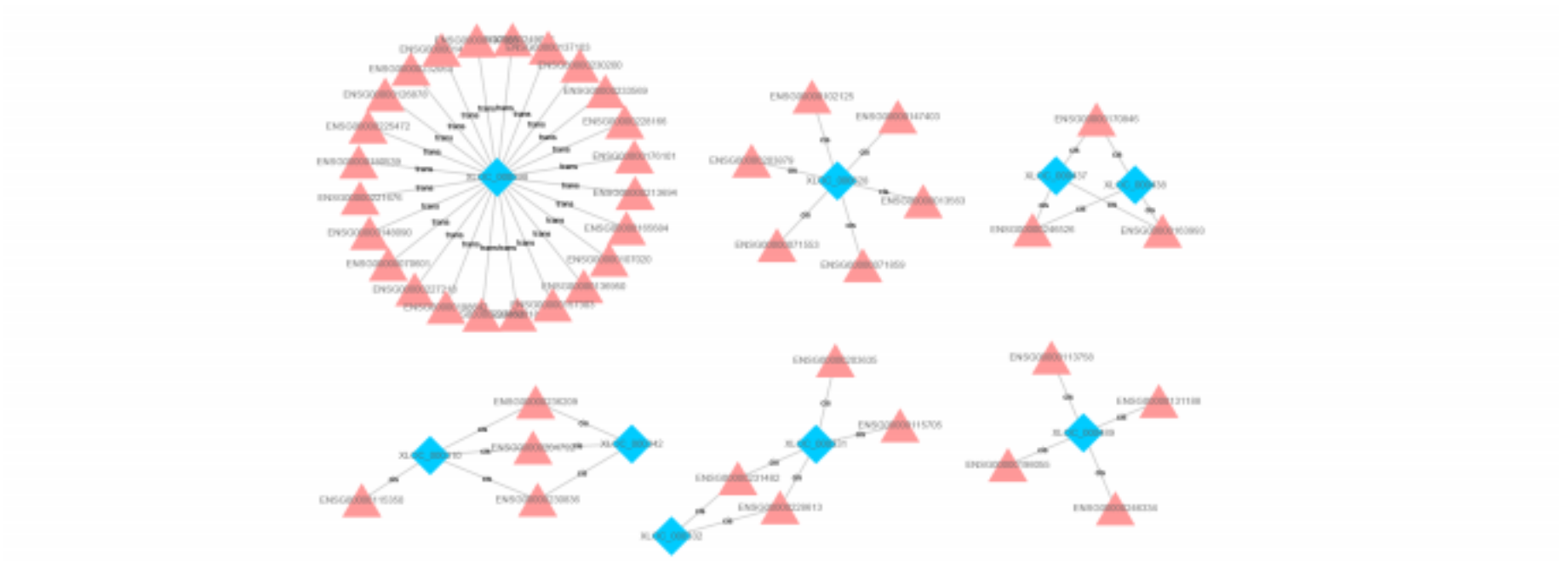


图67 lncRNA 与靶基因的调控网络展示图

9.8 组织特异性

通常认为， lncRNA表达的组织特异性比蛋白编码基因更强。当检测样品包含多种组织（或实验条件）时，可以评估基因（ lncRNA或蛋白编码基因）在各组织（实验条件）间的表达偏好性，即组织特异性。

首先根据单个基因的表达值，基于信息熵的方法，计算其在各组织（实验条件）中的 Jensen－Shannon divergence ，即JS(Cabili ,et al. 2011)分；该基因对应的最大 JS得分作为其组织特异性得分。组织特异性得分越高，该基因在各组织（实验条件）中表达越不均匀，具有越强的组织表达偏性；反之，该基因在各组织（实验条件）的表达差异越不明显。

根据编码基因和 lncRNA的表达量信息，分别计算 JS得分，根据每个基因的JS得分分布可以评估基因的表达是否有组织特异性。

所有基因的 JS得分绘制密度分布图结果如下：

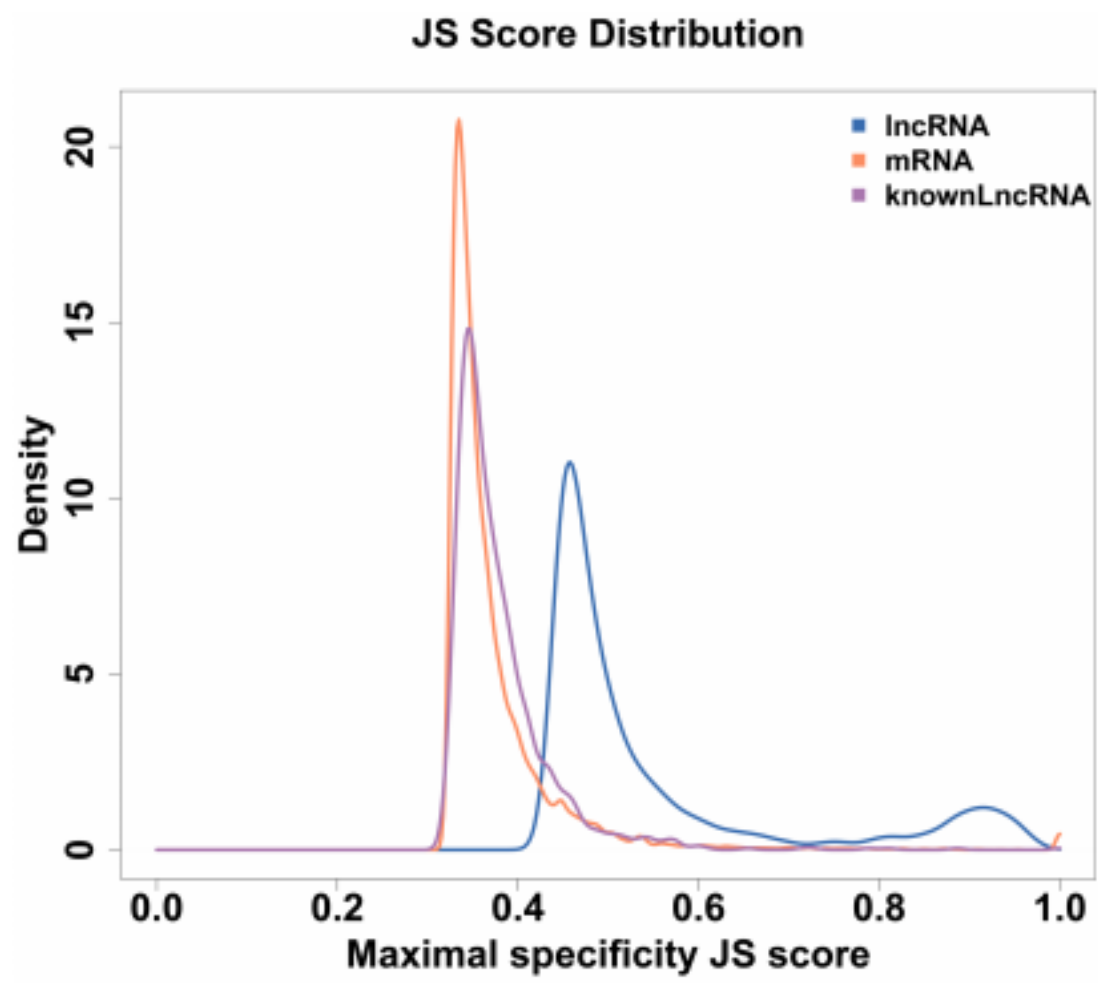


图68 JS Score 分布图

并且根据基因的 JS得分，筛选 JS得分大于 0.5 的基因，根据其表达量绘制热图，以查看其表达趋势及组织表达特异性，结果展示如下：

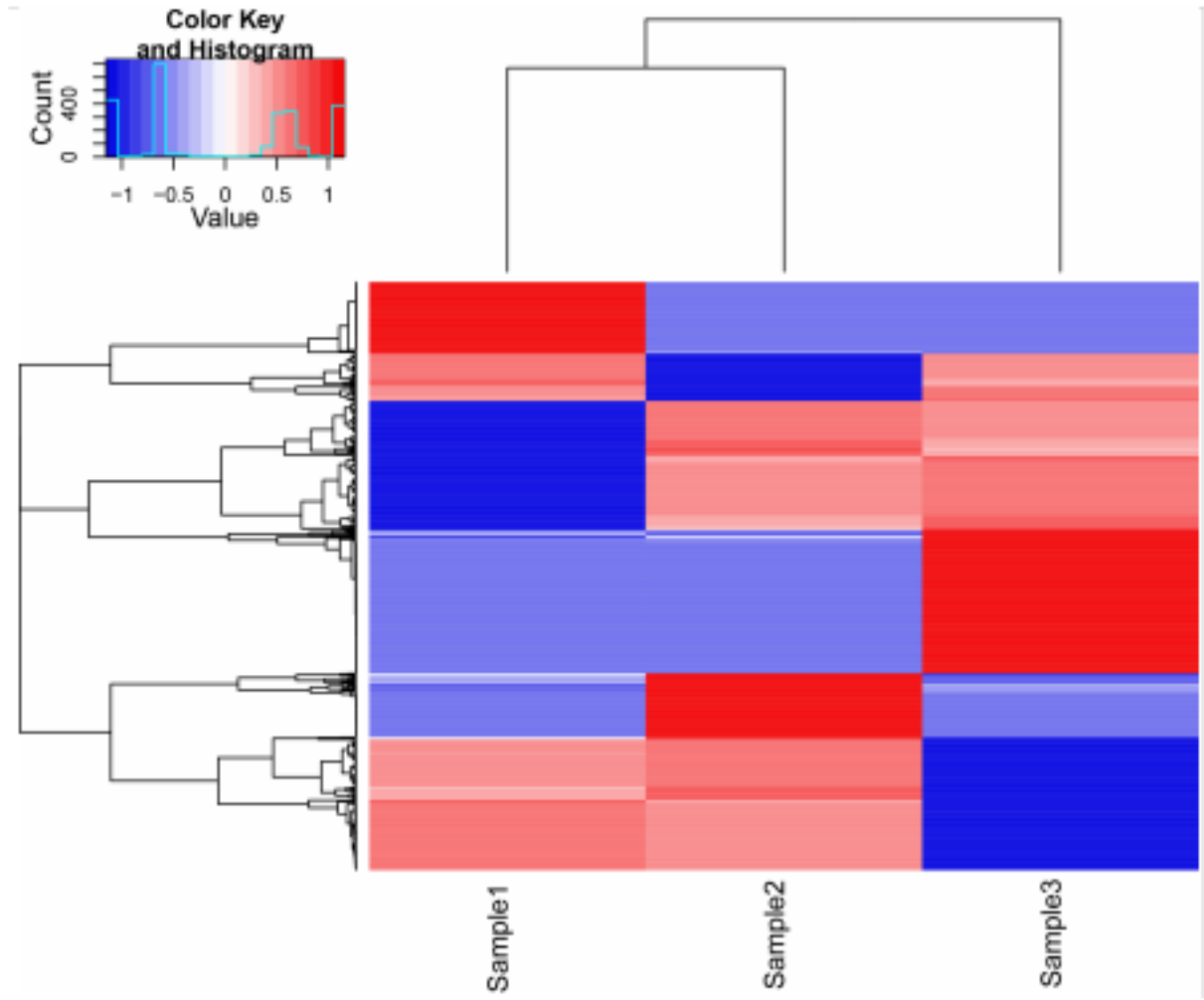


图69　筛选基因表达量分布

10 附录

10.1 参考文献

Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25, 1915-1927 (2011).

Felsenstein, J. & Churchill, G. A. A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular Biology & Evolution 13, 93-104 (1996).

Florea, L., Song, L. & Salzberg, S. L. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. F1000Research (2013).

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25-10 (2009).

Lei, K. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Research 35, W345-W349 (2007).

Li, A., Zhang, J. & Zhou, Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics 15, 311 (2014).

Li, H. et al. The sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25, 2078-2079 (2009).

Shannon, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research 13, 2498-2504 (2003).

Liang, S. et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research 41, e166-e166

(2013).

Thorvaldsd ó ttir, H., Robinson, J.T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics 14, 178-192 (2013).

Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. Bioinformatics 25, 1105-1111 (2009).

Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28, 511-515 (2010).

Wagner, G.P., Kin, K. & Lynch, V.J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131, 281-285 (2012).

Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559-559 (2008).

Liguo, W. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Research 41, e74-e74 (2013).

Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an r package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26, 136-138 (2010).

10.2 软件与方法说明

lncRNA分析的各种软件说明及分析方法的中英文版介绍请见下面的下载链接。

10.3 结果目录