伯远生物RNASeq结题报告

合同编号	Bior220801-005
合同名称	RNA-Seq测序技术服务
委托方（甲方）	XXX研究院
受托方（乙方）	武汉伯远生物科技有限公司
结题日期	2023-01-09

物种信息

物种名称	拉丁学名	参考基因组
human	*homo sapiens*	hg38_gencode_v40

1 背景及分析流程

RNA-seq即转录组测序技术，就是把mRNA，small RNA，和 Non-coding RNA（ncRNA）等全部或部分RNA，进行提取富集，然后用高通量测序技术进行测序分析，反映出它们的表达水平的一个技术。转录组是某个物种或者特定细胞类型产生的所有转录本的集合。转录组研究能够从整体水平研究基因的表达和功能，揭示特定生物学过程以及疾病发生过程中的转录层面的分子机理。

1.1 实验原理

RNA-seq 的实验原理是获得total RNA, 然后进行纯化，反转录，加接头等构建文库；然后对富集得到的片段进行高通量测序[1]。具体步骤如下：

(1). 使用RNA试剂盒提取RNA，并检测RNA样品的纯度、浓度和完整性；(2). 用带有Oligo（dT）的磁珠富集真核生物mRNA, 并将mRNA进行随机打断；(3) 以mRNA为模板，合成第一条cDNA链，然后加入缓冲液、dNTPs、RNase H和DNA polymerase I 合成第二条cDNA链，利用AMPure XP beads纯化cDNA；(4) 纯化的双链cDNA再进行末端修复、加A尾并连接测序接头，进行片段大小选择；(5) 最后通过PCR富集得到cDNA文库;(6). 对文库的浓度和插入片段大小（Insert Size）进行检测，使用Q-PCR方法对文库的有效浓度进行准确定量，以保证文库质量。

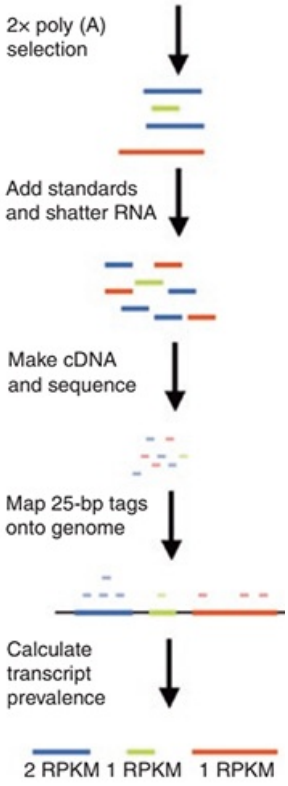


图1.1 RNA-seq实验原理

1.2 生信分析流程

信息分析主要分为3部分：1、数据预处理。去接头序列、污染序列、低质量碱基，获得clean data序列，并进行相关数据统计；2、clean data定位到参考基因组上，得到bam文件，并去除重复序列，保留唯一比对的序列；3、SNP、InDel检测及统计。

图1.2 RNA-Seq信息分析流程

1.3 样本信息

进行分析的样本的信息，由客户或实验人员提供样本详细的处理操作信息。

表1.1 样本信息

sampleID	sampleName	description
1	test1	MECP2-/-, hESCs
2	test2	MECP2-/-, hESCs
3	con1	Wild type, hESCs
4	con2	Wild type, hESCs

2 数据处理

2.1 Raw Data 说明

Illunima等高通量测序平台得到的原始图像数据文件，经碱基识别（Base Calling）转化为原始测序序列（Sequenced Reads），我们称之为 Raw Reads。Raw Reads 以 FASTQ 格式存储，包含序列以及对应的测序质量信息。Fastq 数据示例如下：

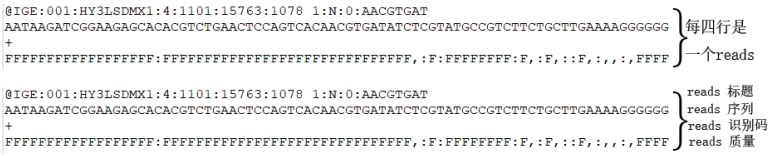


图2.1 FASTQ格式文件说明

每条序列由 4 行组成，第 1 行是序列名称；第 3 行也是序列名称，一般省略，只保留“+”；第 2 行是序列；第 4 行是序列的测序质量。第四行中每个字符对应的 ASCII 值减去 33，即为第二行对应碱基的测序质量值: $Q = -10\log_{10}(E)$。

测序错误率与对应字符换算示例如下所示：

表2.1 测序错误率与测序质量值对应关系

测序错误率	测序质量值	对应ASCII码
5%	13	.
1%	20	5
0.10%	30	?
0.01%	40	I

2.2 Raw Data 数据评估

对原始数据，使用软件 FastQC（version: 0.11.5） 质控处理，结果如下所示。

2.2.1 原始数据碱基质量分析

X轴是read中的碱基位置，Y轴是碱基质量；盒形图中间的红线表示中位数（median value）；黄色部分代表四分位距（25-75%）；上下分割线代表90%和10%的上下临界值；蓝色的线代表碱基质量的平均值。

碱基质量（Q值），-10*$\log_{10}$(p)，p为测错的概率。所以一条read某位置出错概率为0.01时，其quality就是20，通常认为Q20反映了数据的质量。Y轴将质量值被划分三部分：绿色（高质量），橘黄色（中等质量）和红色（低质量）。由于在每一个测序反应开始后，碱基的信号质量会逐渐降低，因此在每个读长最后的碱基通常都会处于橘黄色的中等质量区。

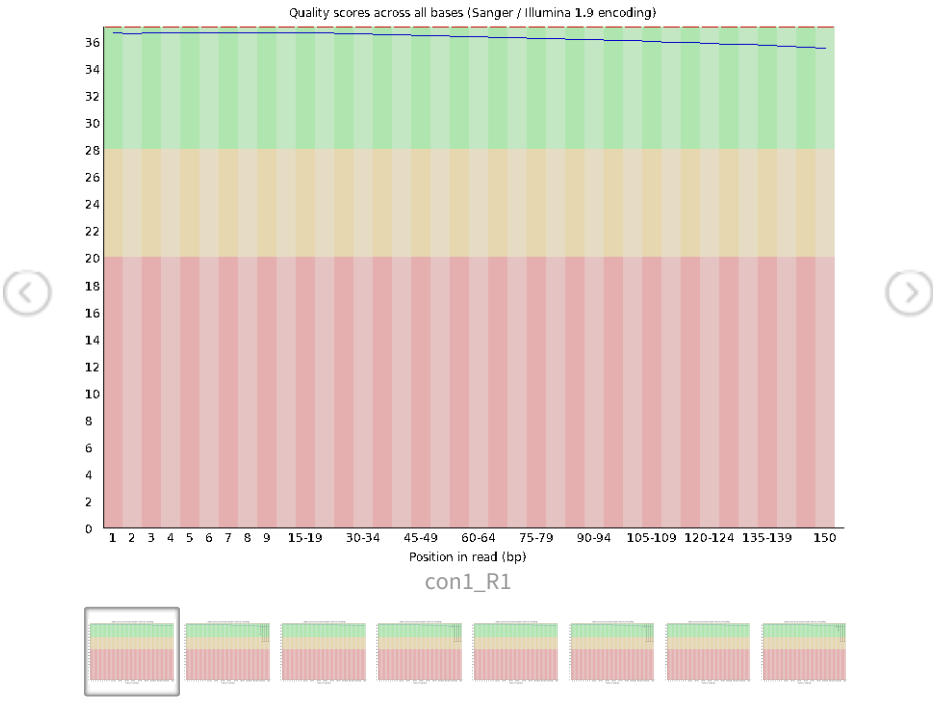


图2.2 原始数据碱基质量分析

2.2.2 原始数据碱基分布图

横坐标为read中的碱基位置，纵坐标为对应位点上单个碱基所占的比例。不同颜色代表不同的碱基类别。正常情况下四种碱基的出现频率应该是接近的，而且没有位置差异。因此好的样本中四条线应该平行且接近。当部分位置碱基的比例出现bias时，即四条线在某些位置纷乱交织，往往提示我们有overrepresented sequence的污染。当所有位置的碱基比例一致的表现出bias时，即四条线平行但分开，往往代表文库有bias (建库过程或本身特点)，或者是测序中的系统误差。

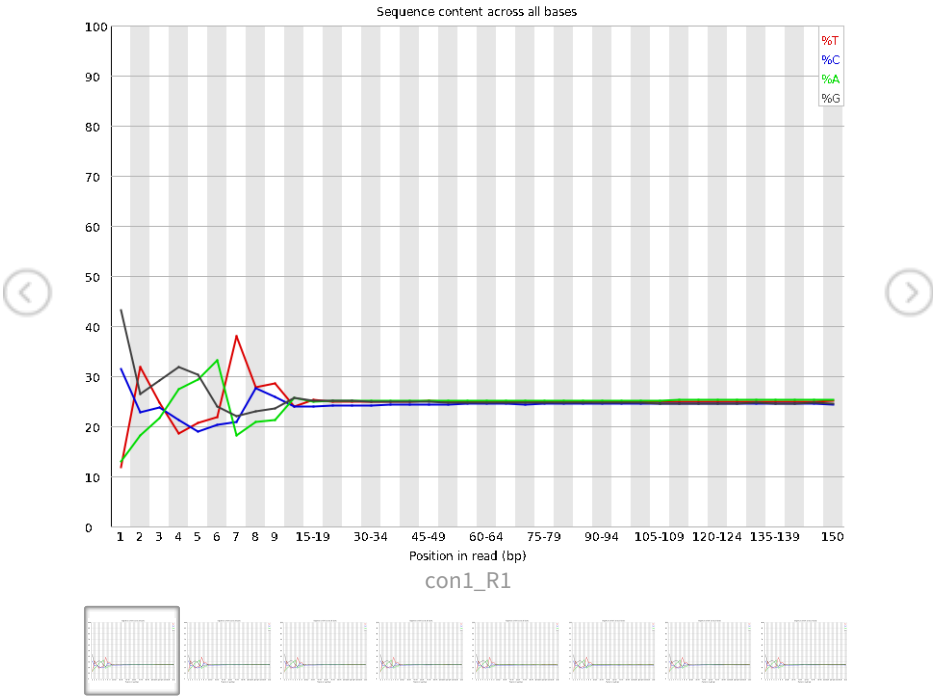


图2.3 原始数据碱基分布图

2.2.3 原始数据N碱基含量图

横坐标为序列长度，纵坐标为N碱基的比例。当测序仪无法识别具体是哪种碱基时，就会给出N, N比例越小越好。当某个位点N碱基的比例大于5%时，会给出警告信息，大于20%时，会给出错误信息。

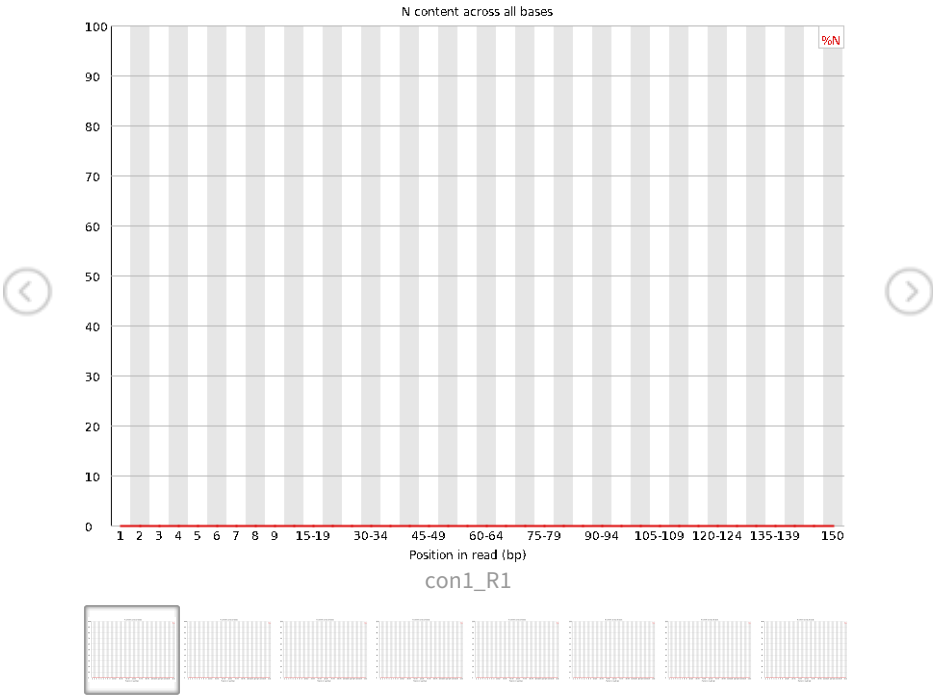


图2.4 原始数据N碱基含量图

2.3 clean Data 数据评估

使用fastp软件[2]进行数据过滤，截掉末尾的adapter序列，保留最短长度16bp。对过滤数据，使用软件 FastQC（version: 0.11.5）质控处理，结果如下所示。

2.3.1 过滤数据碱基质量分析

X轴是read中的碱基位置，Y轴是碱基质量；盒形图中间的红线表示中位数（median value）；黄色部分代表四分位距（25-75%）；上下分割线代表90%和10%的上下临界值；蓝色的线代表碱基质量的平均值。

碱基质量（Q值），$-10*\log_{10}(p)$，p为测错的概率。所以一条read某位置出错概率为0.01时，其quality就是20，通常认为Q20反映了数据的质量。Y轴将质量值被划分三部分：绿色（高质量），橘黄色（中等质量）和红色（低质量）。由于在每一个测序反应开始后，碱基的信号质量会逐渐降低，因此在每个读长最后的碱基通常都会处于橘黄色的中等质量区。

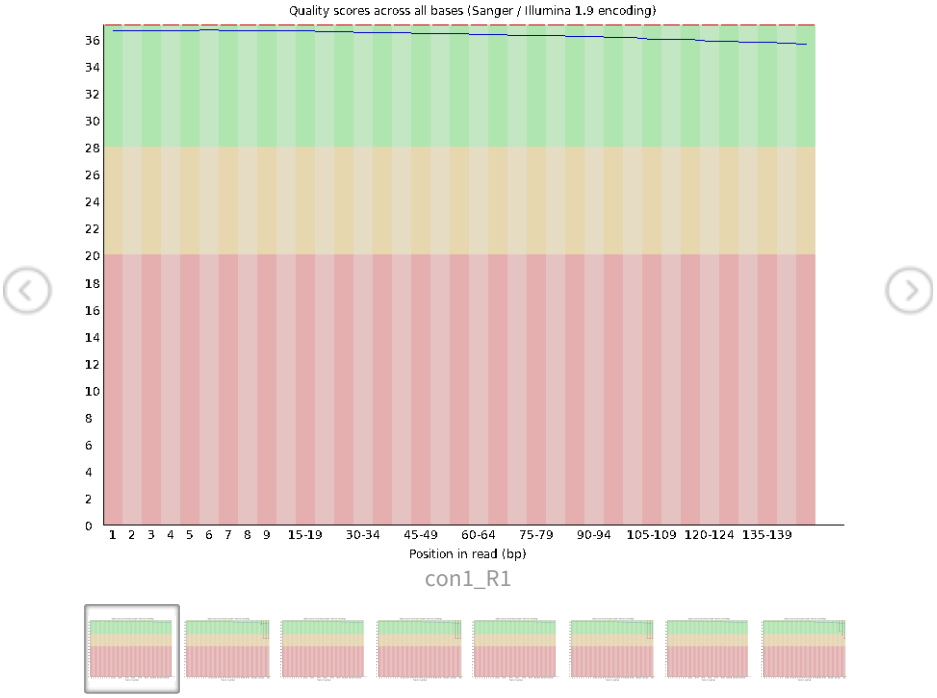


图2.5 过滤数据碱基质量分析

2.3.2 过滤数据碱基分布图

横坐标为read中的碱基位置，纵坐标为对应位点上单个碱基所占的比例。不同颜色代表不同的碱基类别。正常情况下四种碱基的出现频率应该是接近的，而且没有位置差异。因此好的样本中四条线应该平行且接近。当部分位置碱基的比例出现bias时，即四条线在某些位置纷乱交织，往往提示我们有overrepresented sequence的污染。当所有位置的碱基比例一致的表现出bias时，即四条线平行但分开，往往代表文库有bias (建库过程或本身特点)，或者是测序中的系统误差。

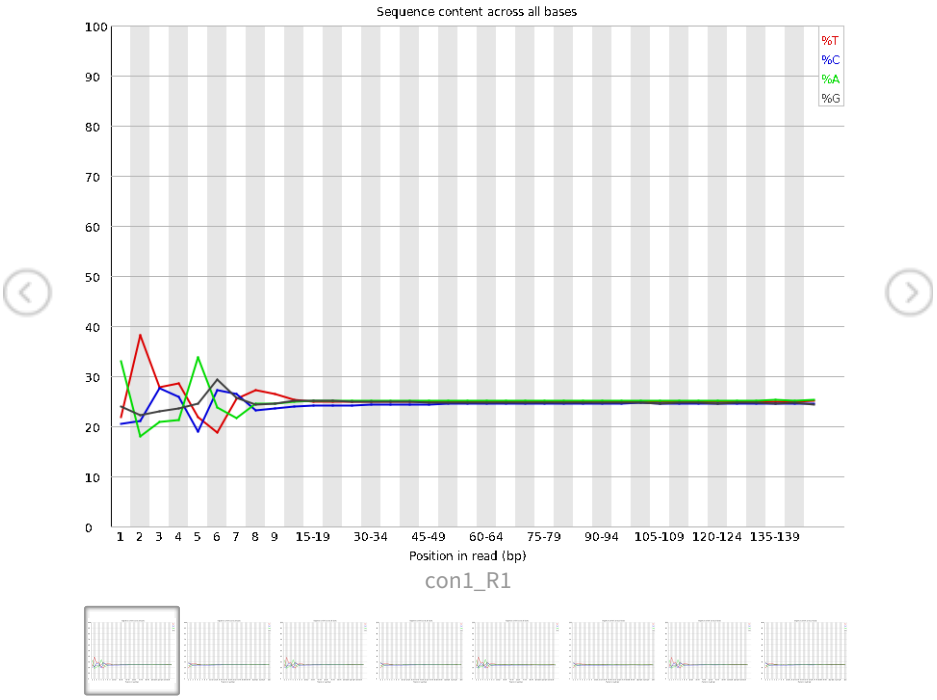


图2.6 原始数据碱基分布图

2.3.3 过滤数据N碱基含量图

横坐标为序列长度，纵坐标为N碱基的比例。当测序仪无法识别具体是哪种碱基时，就会给出N, N比例越小越好。当某个位点N碱基的比例大于5%时，会给出警告信息，大于20%时，会给出错误信息。

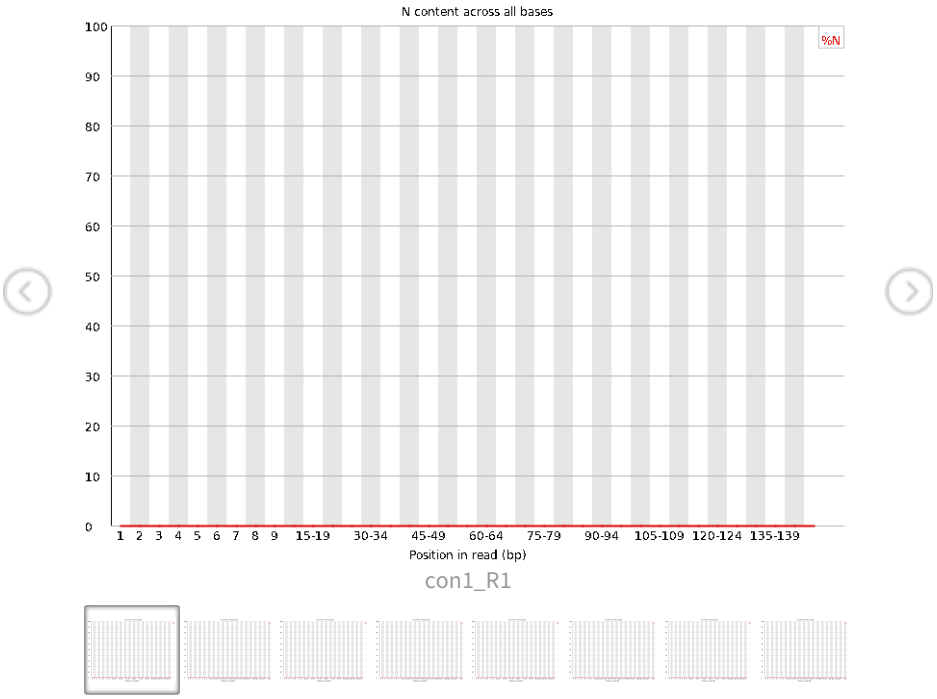


图2.7 过滤数据N碱基含量图

2.4 Clean Data 数据统计

对原始数据和过滤处理得到高质量的数据（clean reads）进行数据统计，得到数据基本信息。统计结果如下表：

表2.2 clean data数据统计表

sample	rawReads	rawBase	cleanReads	cleanBase	cleanQ20	cleanQ30	cleanGC
test1	45,901,520	6.885G	45,280,568 (98.65%)	6.392G(92.84%)	98.12%	94.59%	48.94%
test2	45,929,174	6.889G	45,299,946 (98.63%)	6.392G(92.78%)	98.08%	94.56%	49.47%
con1	46,482,426	6.972G	45,862,608 (98.67%)	6.480G(92.94%)	98.06%	94.44%	49.57%
con2	46,749,712	7.012G	46,125,318 (98.66%)	6.513G(92.88%)	98.07%	94.49%	49.55%

统计说明：

1. sample：样品名称
2. rawReads：原始测序 reads数量
3. rawBase：原始测序数据的总碱基数
4. cleanReads：将 Raw Reads过滤得到的 reads数量
5. cleanBase：过滤得到的数据的总碱基数
6. cleanQ20：测序错误率小于 1%的碱基数目占总碱基数比例
7. cleanQ30：测序错误率小于 0.1%的碱基数目占总碱基数比例
8. cleanGC：碱基G和C的数量占总的碱基数量的百分比

3 基因组信息

3.1 基因组区域信息

根据gff的注释情况，将基因组区域划分为exon,intron,intergenic,promoter(transcript start site upstream 2k) 等区域，并统计各个区域的长度。

表3.1 基因组各区域长度统计表

region	length	ratio
exon	165,867,428	5.08%
intron	1,676,558,350	51.32%
promoter	62,693,546	1.92%
intergenic	1,361,998,028	41.69%

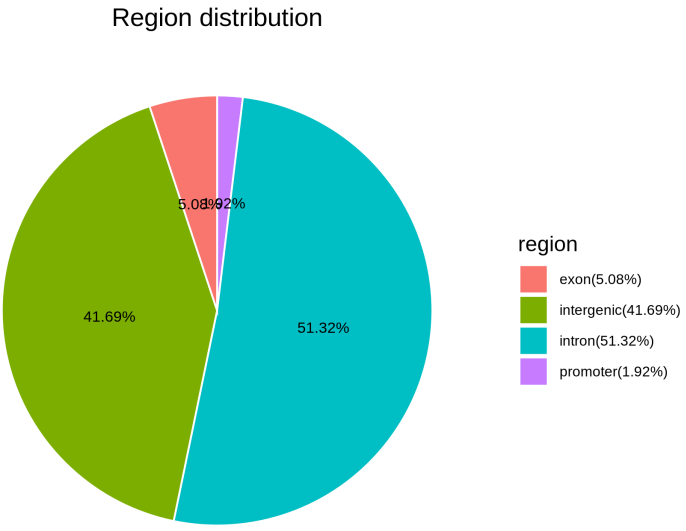


图3.1 基因组各区域长度

4 基础数据分析

4.1 基因组比对分析

将Clean reads数据使用hisat2软件（version:version 2.2.1）[3],默认参数比对到参考基因组。对比对得到的reads进行唯一比对筛选和去重处理，用于后续分析。比对情况如下表：

表4.1 比对结果统计表

sample	totalReads	mapReads	uniqMapReads
test1	45,280,568	43,594,788 (96.28%)	40,568,796 (89.59%)
test2	45,299,946	43,507,516 (96.04%)	40,451,602 (89.30%)
con1	45,862,608	44,205,788 (96.39%)	41,172,742 (89.77%)
con2	46,125,318	44,432,072 (96.33%)	41,393,677 (89.74%)

统计说明：

1. sample：样品名称
2. total：clean reads总数
3. mapped：比对上的 reads总数及比例
4. uniqMap：唯一比对的 reads数及比例

4.2 插入片段长度

根据pair-end 的 reads 的比对情况，统计插入片段的长度，绘制bar图。

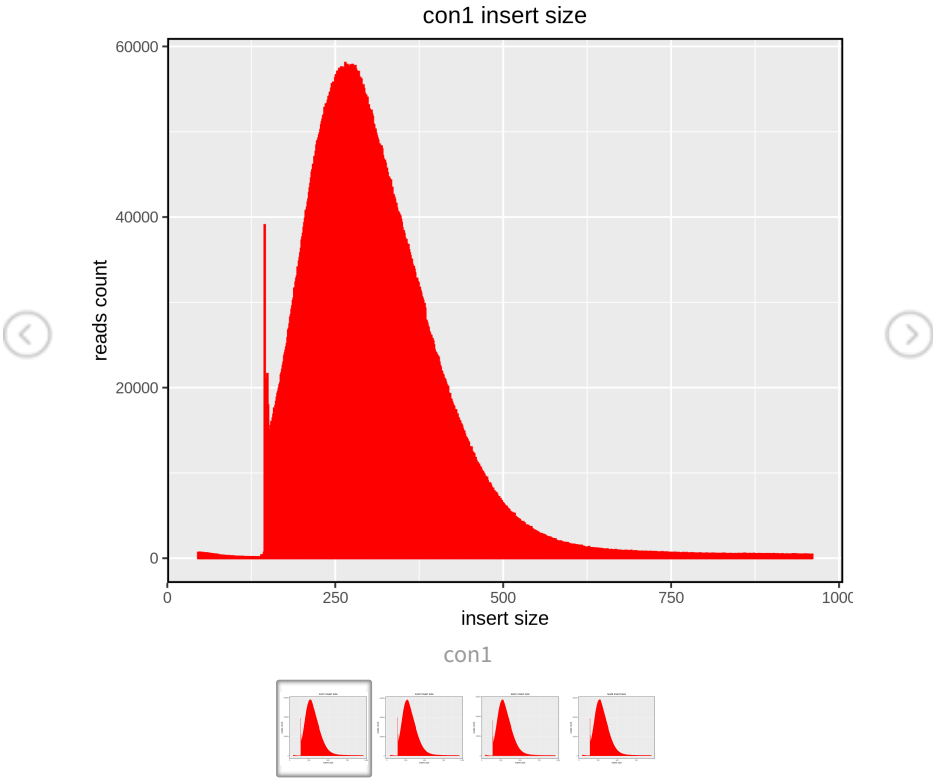


图4.1 insert size

4.3 测序饱和度分析

以10% 为步长，随机提取10% ~ 100%的测序量进行定量分析。计算 FPKM (Fragments per kilo base of a gene per million reads)的概念对每一个基因进行定量。计算每一个基因的FPKM， 将以100%测序量分析得到的为最终定量水平。用各个百分比的数据量得到的各个window的定量水平和和最终定量水平进行比较，如果差异小于15%，则认为该window在该数据量条件下被准确定量。

定量饱和曲线检查反映了测序定量对数据量的要求。富集较多的基因，容易被准确定量，而富集较低的基因，则需要较大的测序量才能被准确定量。结果如图所示。

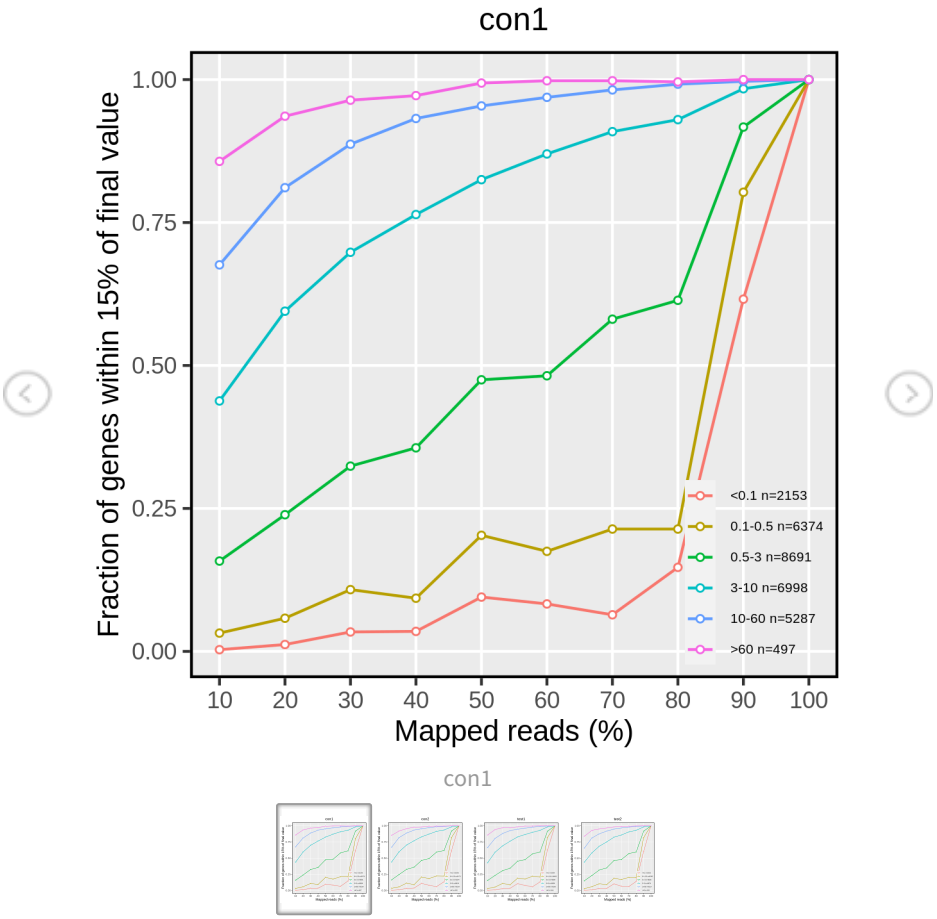
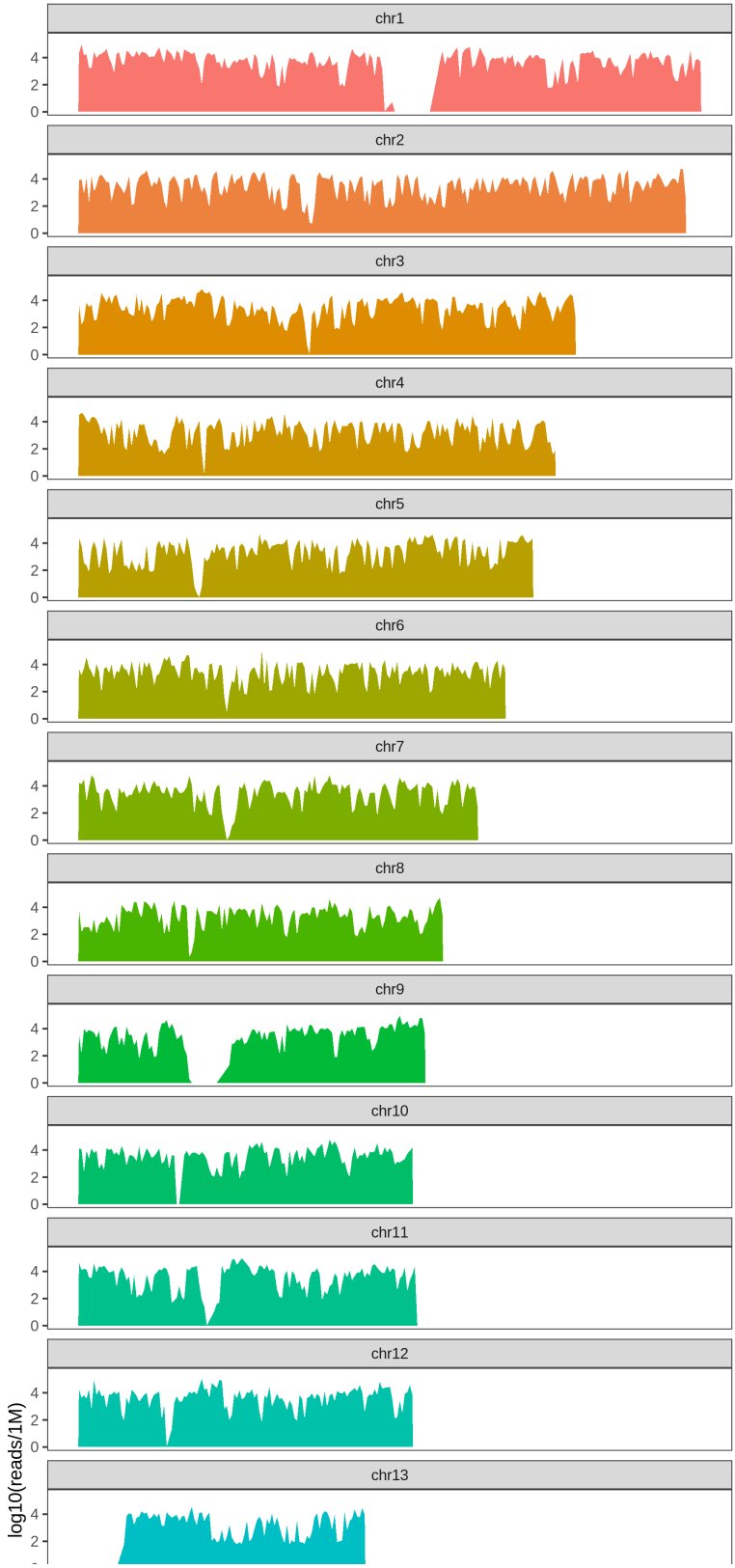


图4.2 reads saturation

4.4 reads在基因组染色体上分布

为了直观的展示染色体长度和Reads总数以及Reads覆盖情况，将比对到基因组上各条染色体的Reads进行唯一比对和去重处理后进行统计，以1M bin计算各个bin内部比对到碱基位置上的Reads数目的均值（作图时取对数）。通常情况下染色体越长定位到该染色体的Reads数目越多。reads 在整个基因组上的分布情况如图。



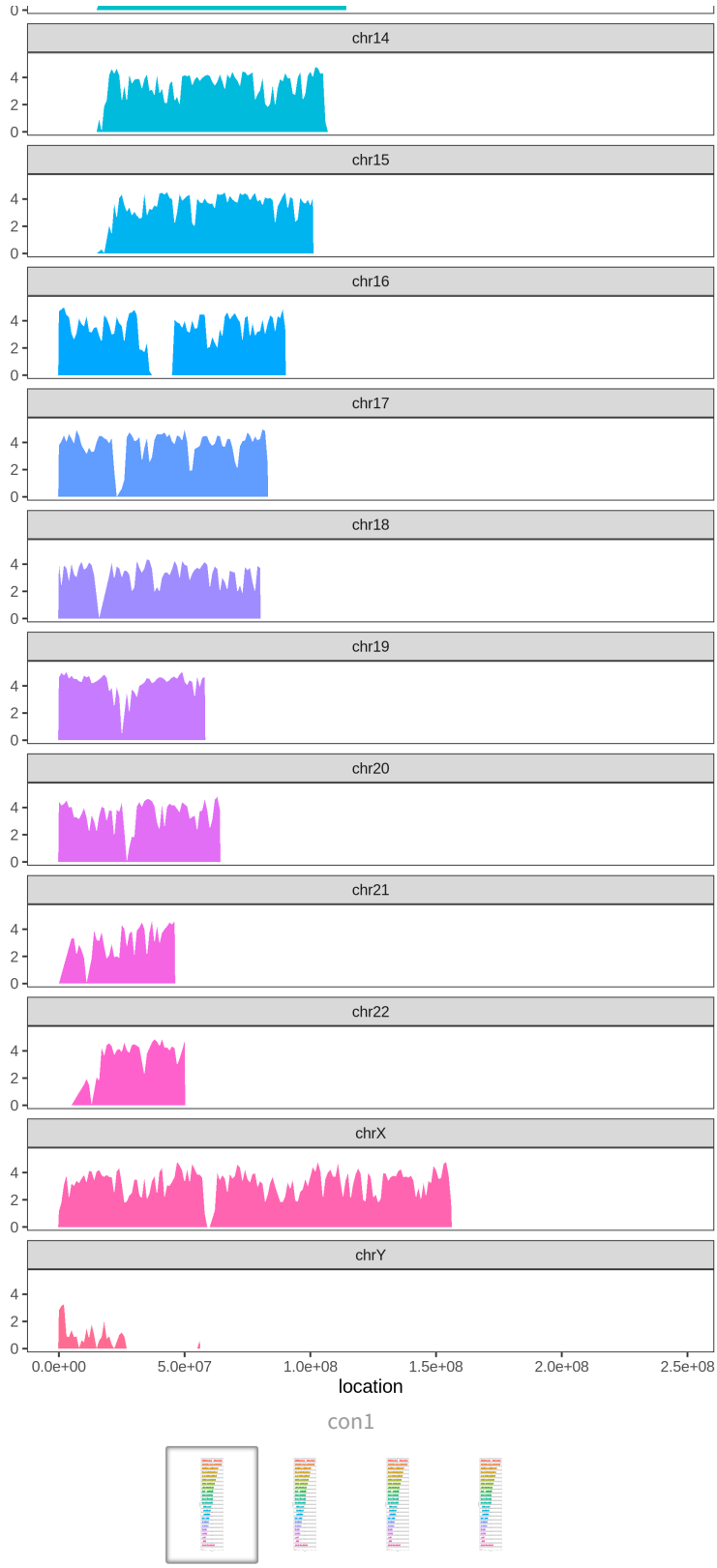


图4.3 reads on genome

4.5 reads距离TSS分布

以转录起始位点（TSS，transcription start site）为中心，统计其上下游2kb范围内reads的分布情况。

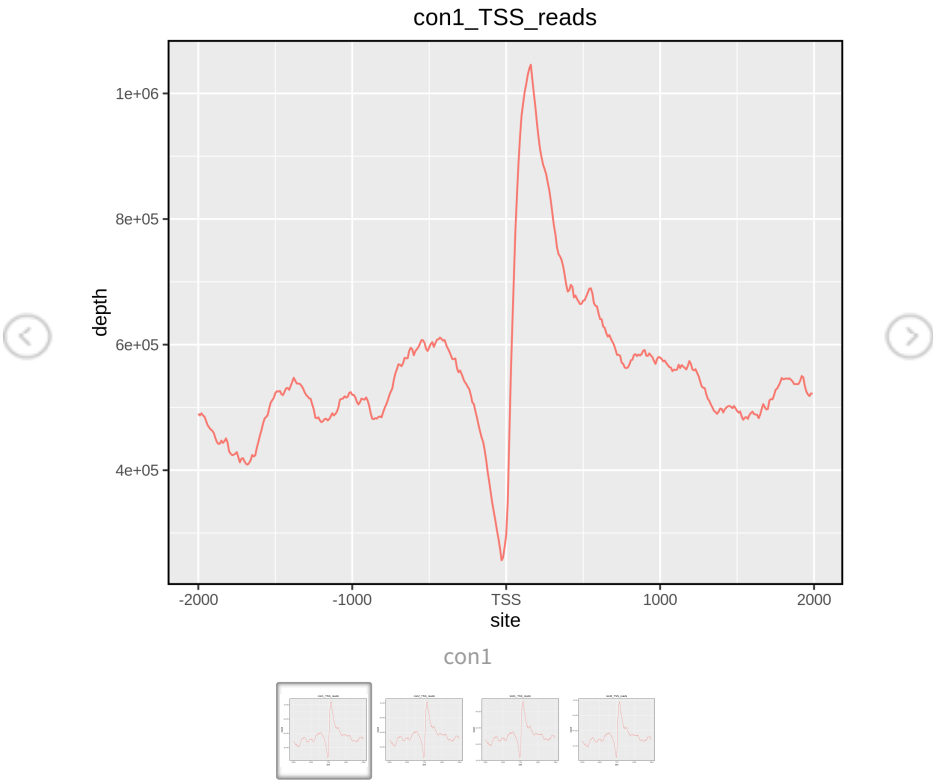


图4.4 reads nearby TSS

4.6 reads距离TES分布

以转录终止位点（TES，transcription end site）为中心，统计其上下游2kb范围内reads的分布情况。

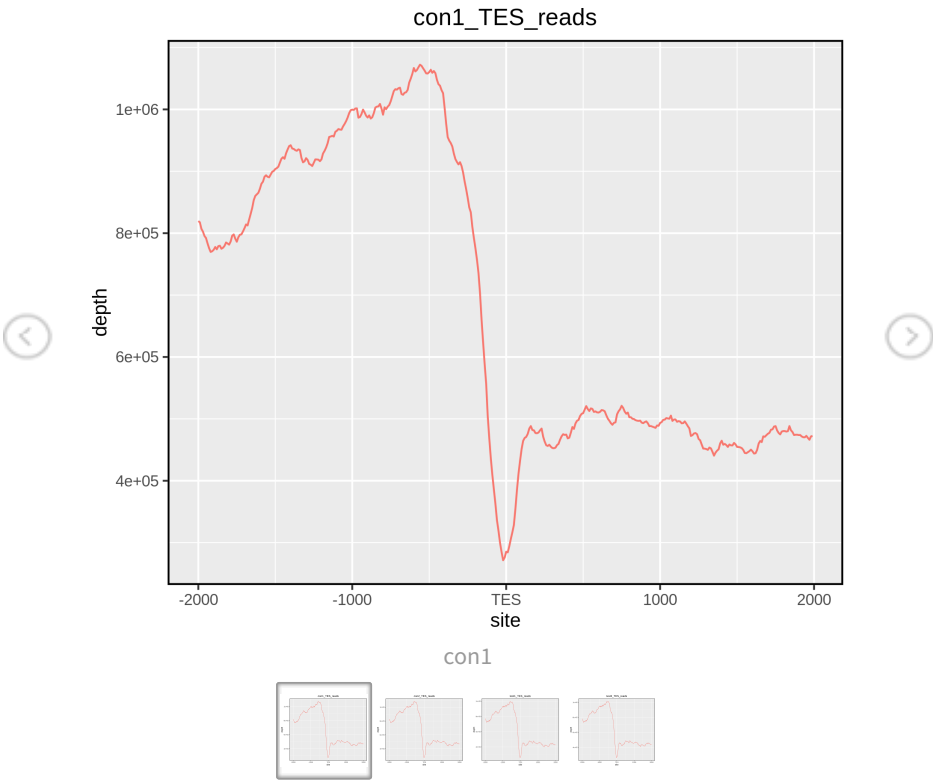


图4.5 reads nearby TES

4.7 reads region分布

统计reads在基因组上各类元件的分布，IP项目中reads主要分布在intergenic（基因间隔区域）和intron区,RNA项目reads主要分布在CDS和intron区。

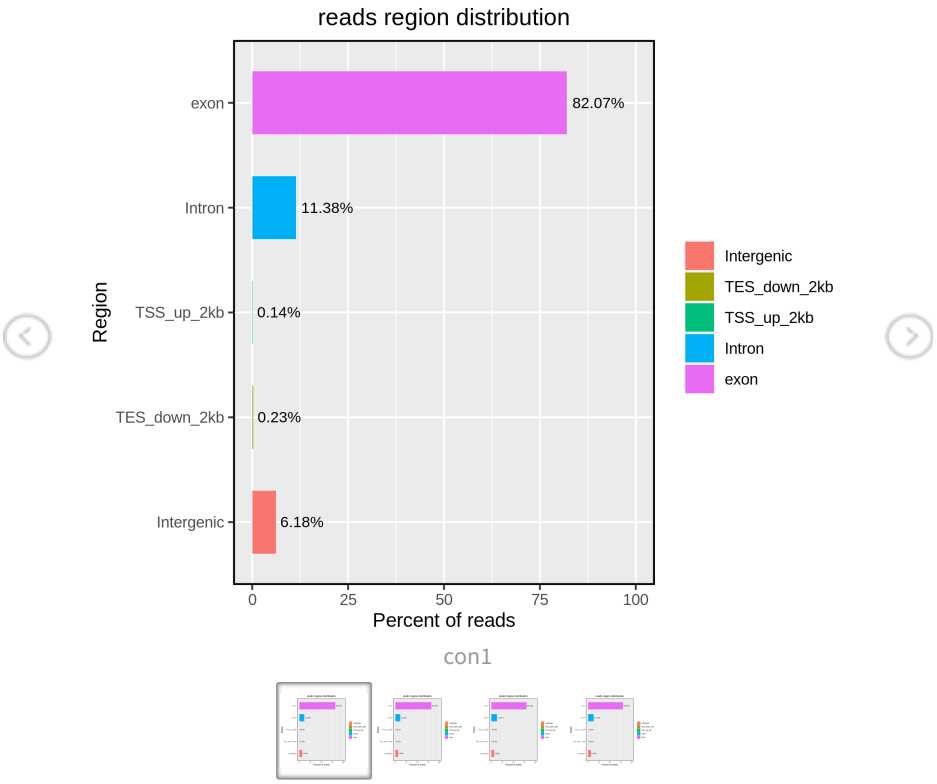


图4.6 reads region

4.8 reads在基因组区域分布

统计reads在基因组上各类元件的分布的总表和总的bar图展示。

表4.2 基因组各区域分布统计

Sample	exon	Intron	TSS_up_2kb	TES_down_2kb	Intergenic
test1	80.43%	12.91%	0.13%	0.24%	6.29%
test2	80.54%	12.79%	0.14%	0.23%	6.29%
con1	82.07%	11.38%	0.14%	0.23%	6.18%
con2	82.30%	11.31%	0.12%	0.22%	6.04%

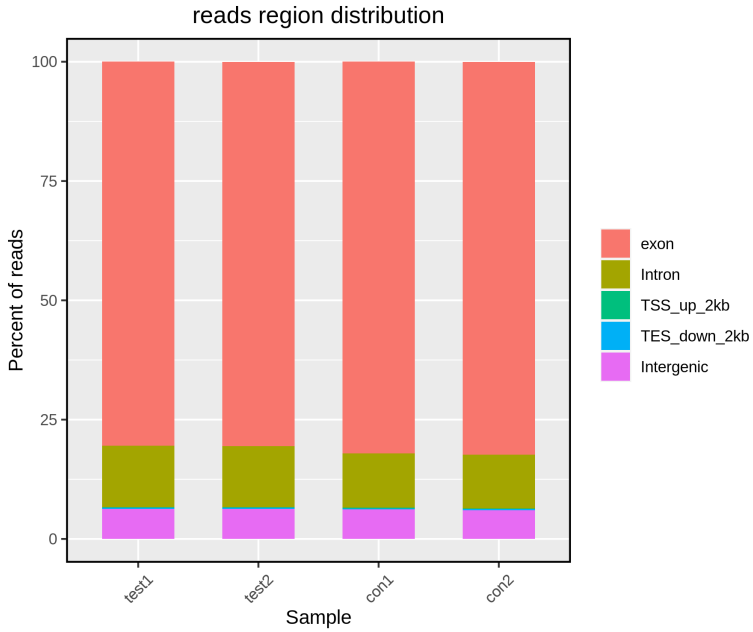


图4.7 基因组各区域分布

5 表达量分析

5.1 样本表达量分析

根据比对的reads，保留uniq比对（在基因组上仅比对到一个位置），去掉重复的reads(reads 片段的序列相同)，然后使用featureCounts(v2.0.1)[4]对样本进行表达量分析。

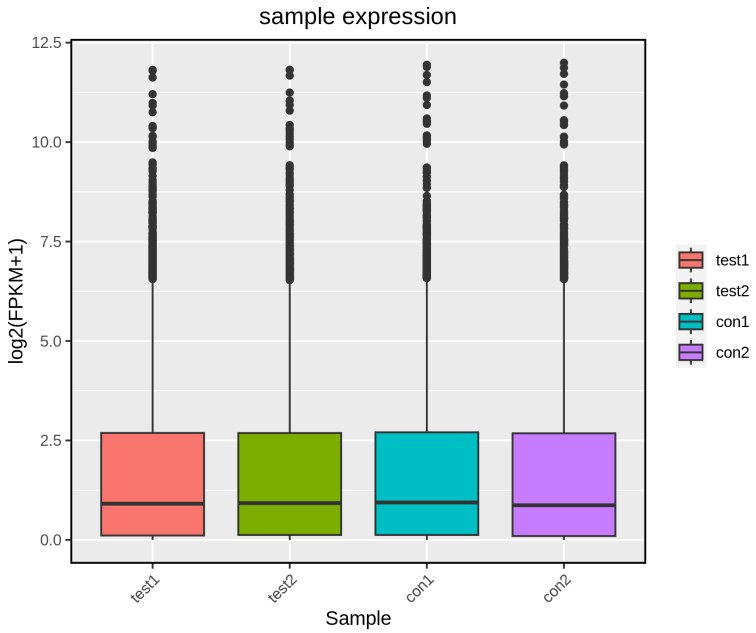


图5.1 样本表达量分布

5.2 样本相关性

根据基因在各个样本中的FPKM (Fragments per kilo base of a gene per million reads)的概念对每一个基因进行定量，通过FPKM计算样本的相关性分析。

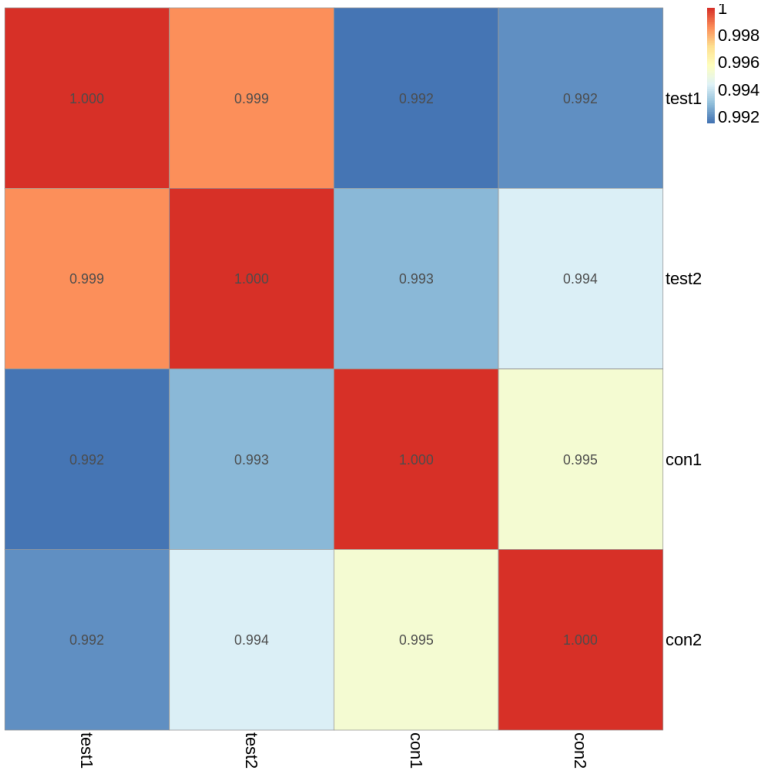


图5.2 样本相关性

5.3 PCA分析

根据基因在各个样本中的FPKM (Fragments per kilo base of a gene per million reads)的概念对每一个基因进行定量，通过FPKM计算样本的PCA分析。

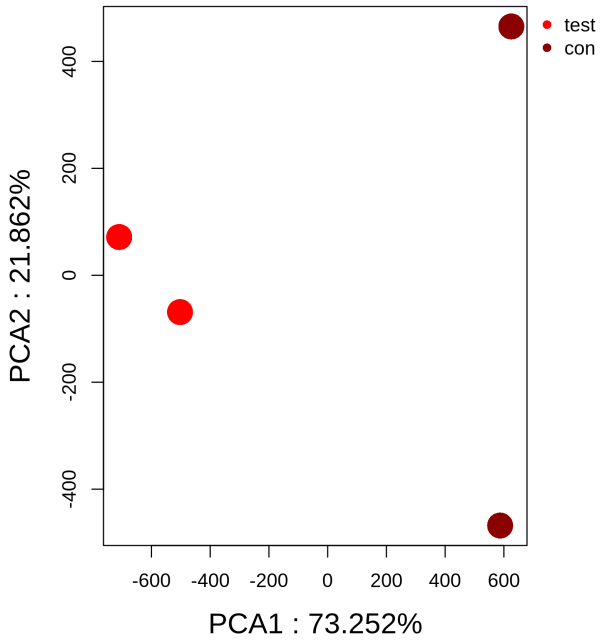


图5.3 样本PCA分析

6 差异表达分析

统计各个基因的reads数，根据reads,使用edgeR[5]方法，分析各组的差异表达基因。参数一般选择FC(fold change) = 2，pValue = 0.05。

6.1 火山图

各组样本差异表达分析的火山图

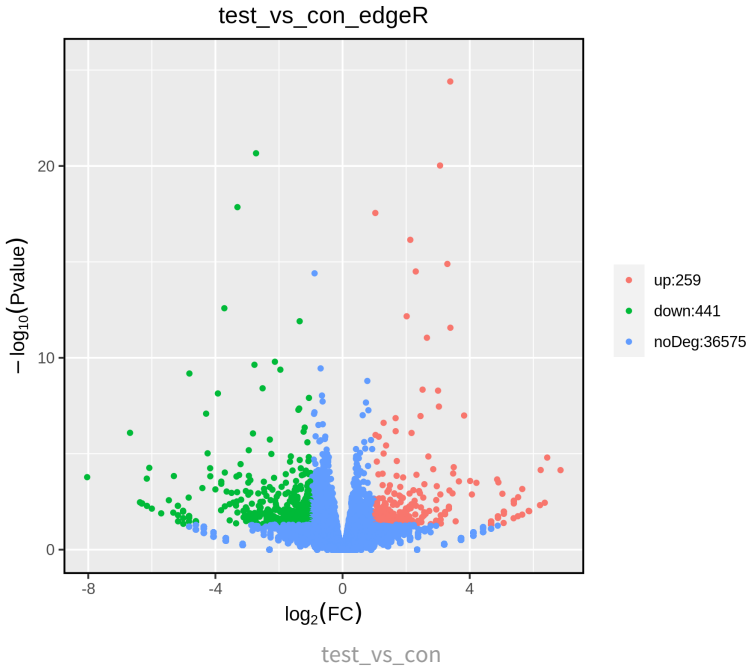


图6.1 火山图

6.2 差异基因统计

差异基因统计

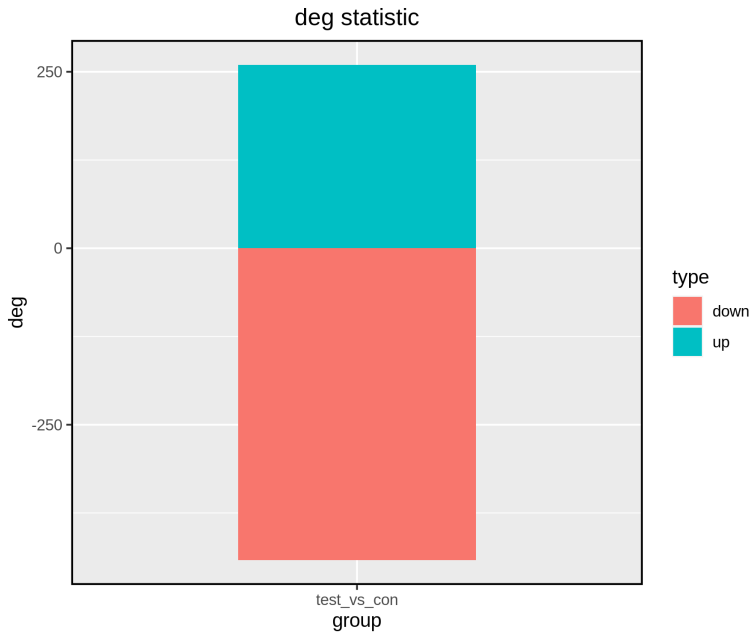


图6.2 差异基因统计

6.3 差异基因样本相关性分析

根据差异表达的基因在各个样本中的FPKM (Fragments per kilo base of a gene per million reads)的概念对每一个差异表达基因进行定量，通过FPKM计算样本的相关性分析。

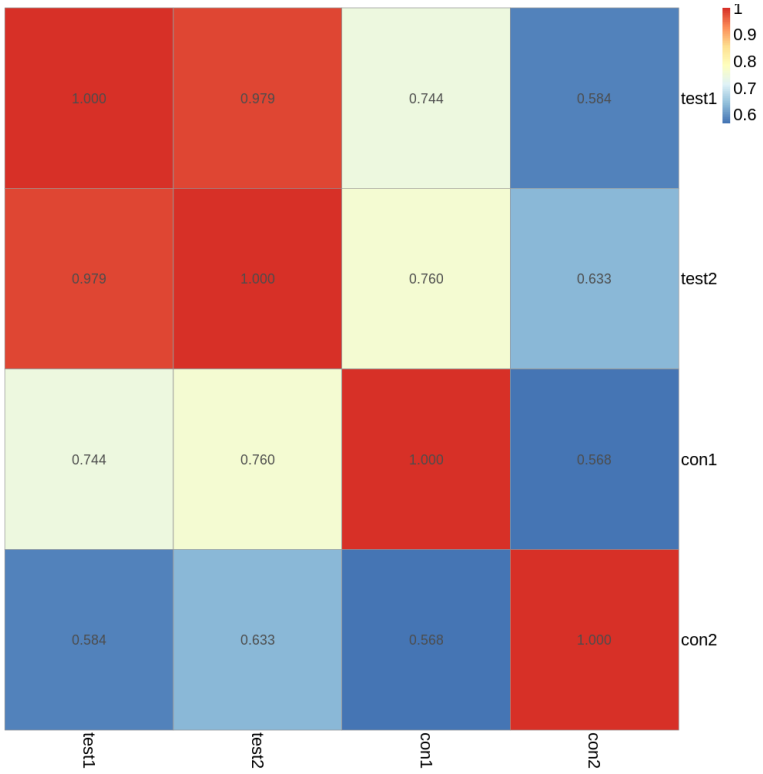


图6.3 差异基因样本相关性

6.4 差异基因样本PCA分析

根据差异表达的基因在各个样本中的FPKM (Fragments per kilo base of a gene per million reads)的概念对每一个差异表达基因进行定量，通过FPKM计算样本的PCA分析。

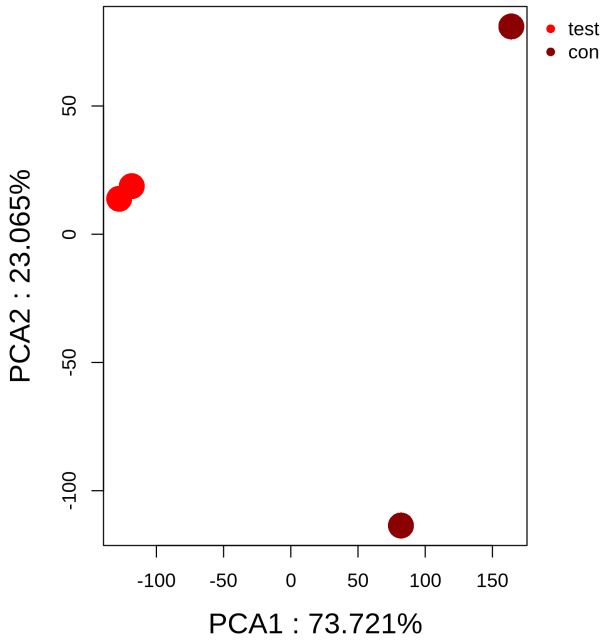


图6.4 差异基因样本PCA分析

6.5 差异基因样本heatMap分析

差异基因样本heatMap分析

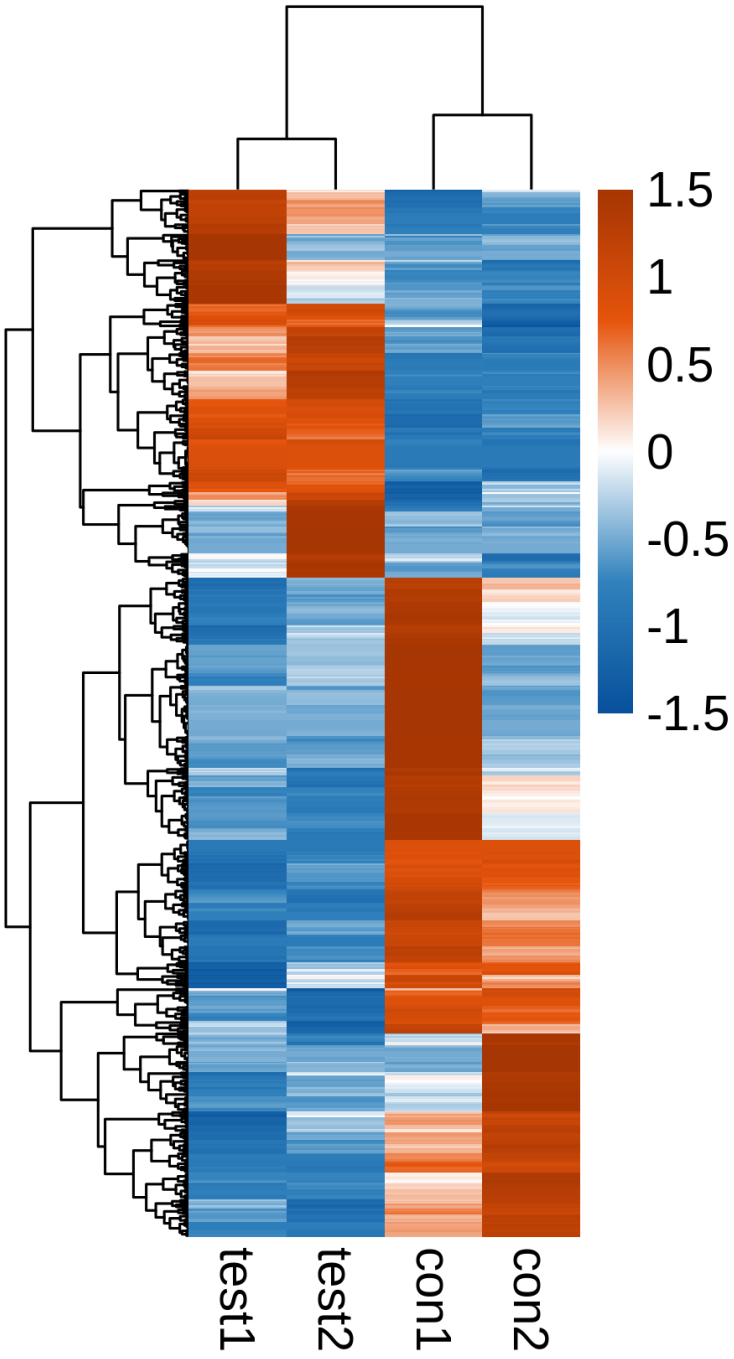


图6.5 差异基因样本heatMap分析

7 功能聚类分析

基因本体论（Gene Ontology）简称GO，是一个国际标准化的基因功能分类体系，包含生物学领域知识体系本质的表示形式，本体通常由一组类（或术语或概念）组成，分为3大类：分别是分子功能（Molecular Function）、细胞组分（Cellular Component）、生物过程（Biological Process）。

京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes）简称KEGG，是系统分析基因功能、 基因组信息的数据库，它整合了基因组学、生物化学以及系统功能组学的信息，有助于研究者把基因及表达信息作为一个整体网络进行研究。

使用 peaks 关联的基因进行GO和KEGG富集分析，研究的物种使用相关GO、KEGG注释数据库或利用blast，得到每个基因对应的GO term 或KEGG pathway。根据结合峰相关基因注释信息，统计每个基因所在的GO term 或KEGG pathway，根据每个GO term 或 KEGG pathway的基因数目，以及背景中此GO term 或KEGG pathway的基因数目，用Fisher Exact Test分析每个GO term 或KEGG pathway的显著性。根据校正p-value和百分比作图进行展示前20 的GO term 或KEGG pathway。

如果基因较少，部分图可能无法绘制或p-value不会很显著。

7.1 GO 功能聚类气泡图

GO 功能聚类气泡图展示。纵坐标是GO Term 名称，横坐标是对应GO Term 中检出的基因占背景基因的个数，颜色代表显著性p-value。

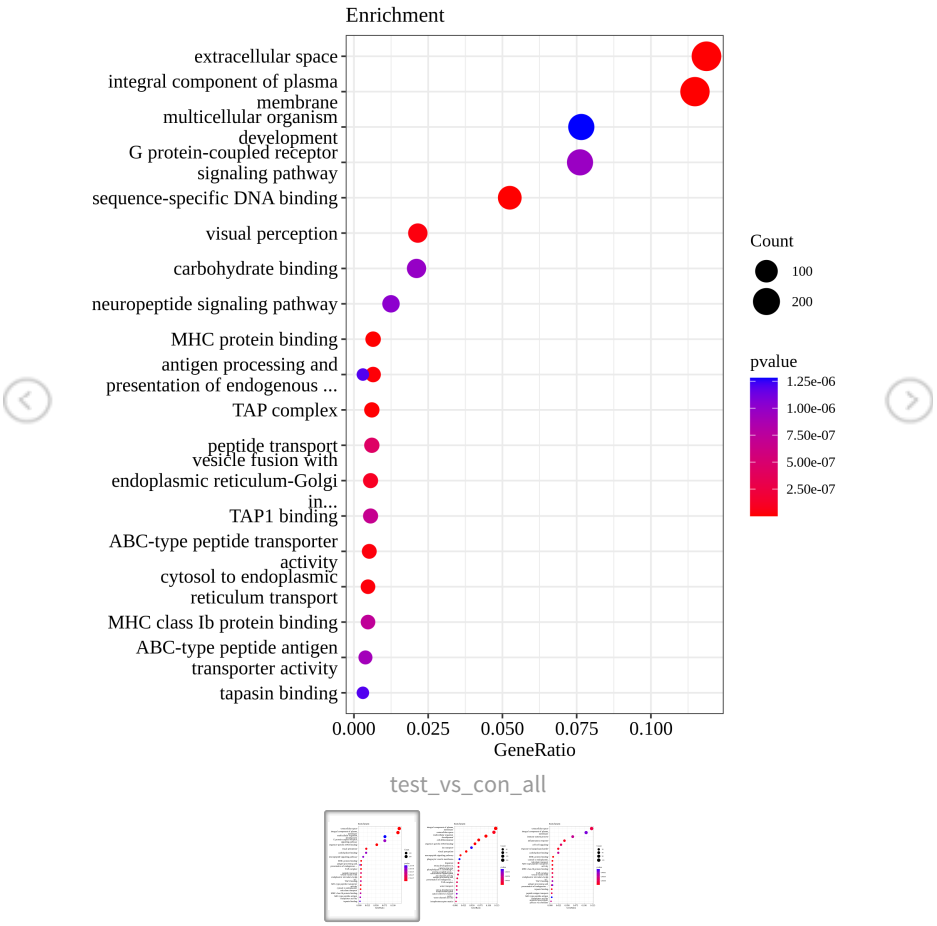


图7.1 GO功能聚类气泡图

7.2 GO 功能聚类bar图

GO 功能聚类bar图展示。纵坐标是GO Term 名称，横坐标是对应GO Term 中检出的基因个数，颜色代表显著性p-value。

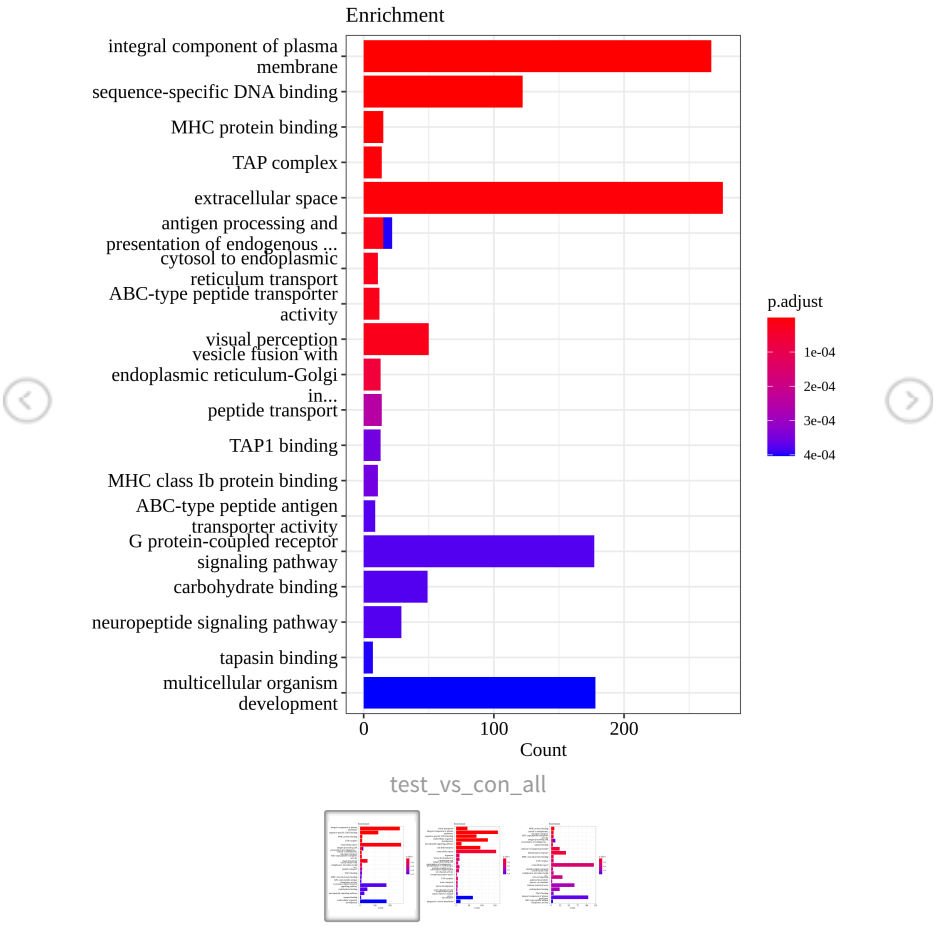


图7.2 GO功能聚类bar图

7.3 GO 功能聚类wego图

GO 功能聚类wego图展示。横坐标坐标是GO Term 名称，纵坐标是对应GO Term 中检出的基因个数，取对数。根据GO分成了3大类，分别是分子功能（Molecular Function，MF）、细胞组分（Cellular Component，CC）、生物过程（Biological Process，BP）。

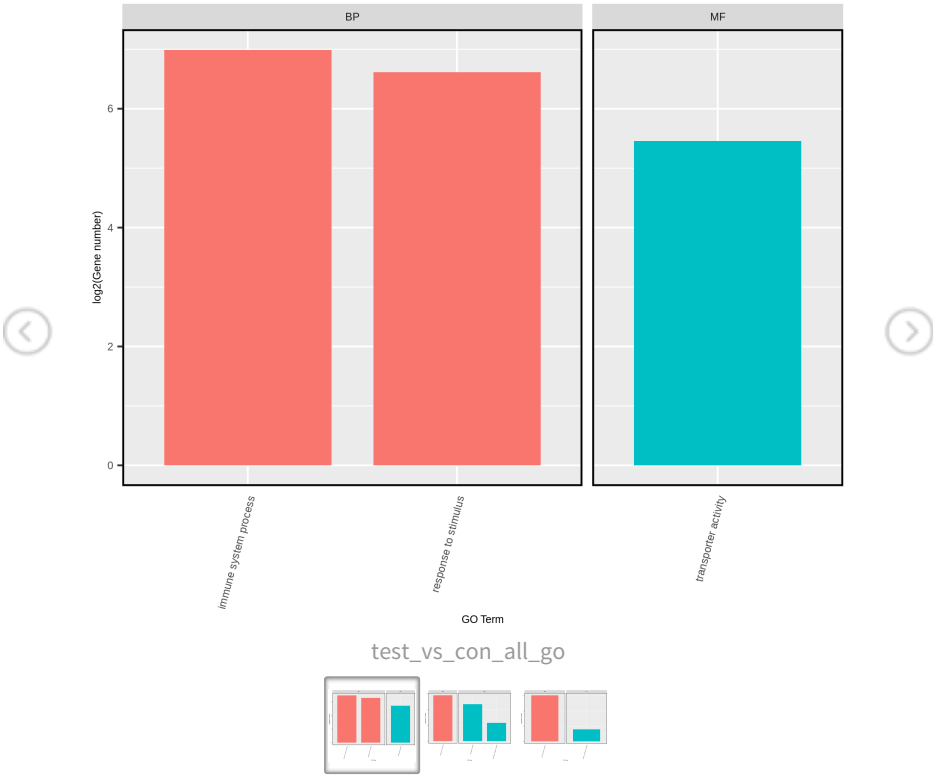


图7.3 GO功能聚类wego图

7.4 KEGG 功能聚类气泡图

KEGG功能聚类气泡图展示。纵坐标是KEGG pathway名称，横坐标是对应KEGG pathway中检出的基因占背景基因的个数，颜色代表显著性p-value。

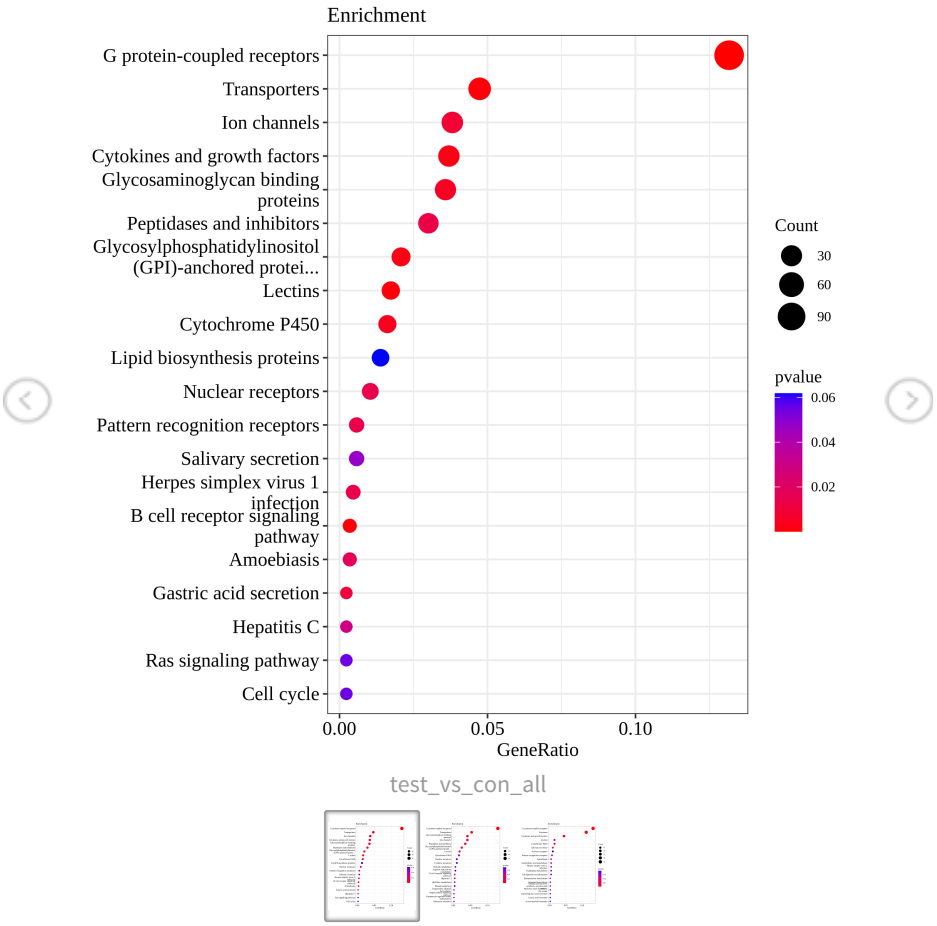


图7.4 功能聚类气泡图

7.5 KEGG 功能聚类bar图

KEGG功能聚类bar图展示。纵坐标是KEGG pathway名称，横坐标是对应KEGG pathway中检出的基因个数，颜色代表显著性p-value。

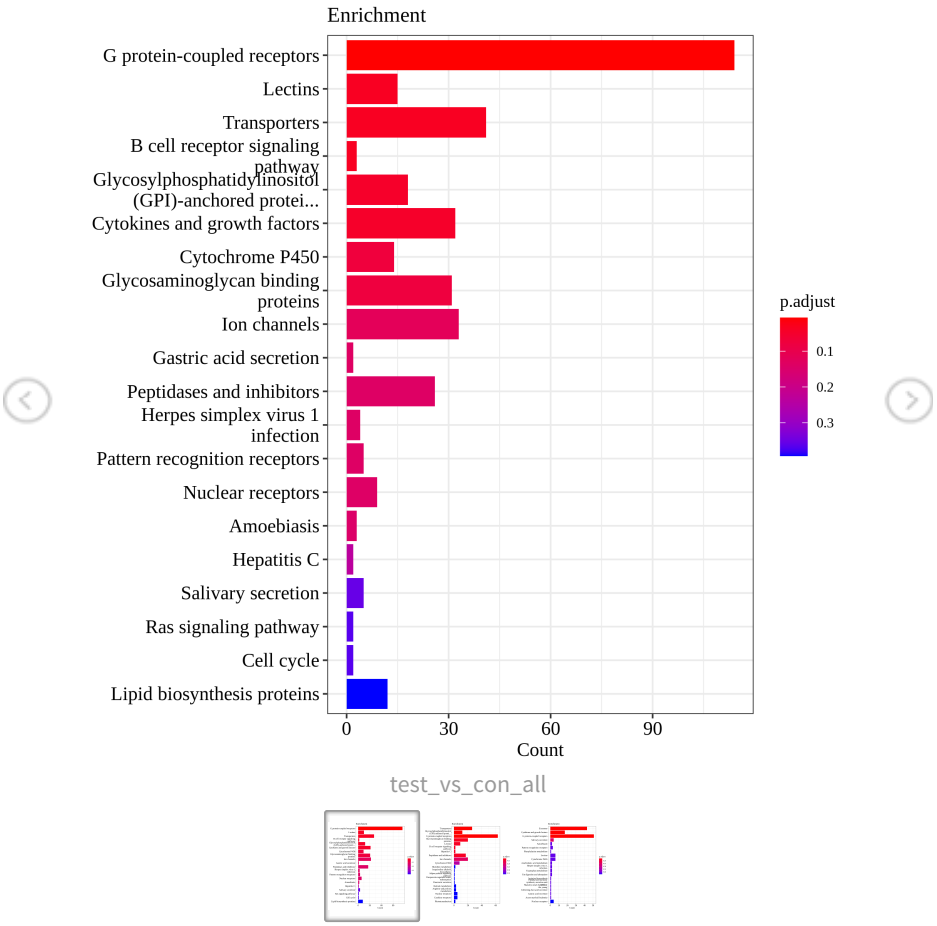


图7.5 功能聚类气泡图

7.6 KEGG 功能聚类分类图

KEGG Pathway主要划分为7类：分别为Metabolism, Genetic information Processing等。其中每类又分为二、三、四级子条目。功能聚类分类图课展示pathway所属一级条目和二级条目的情况。

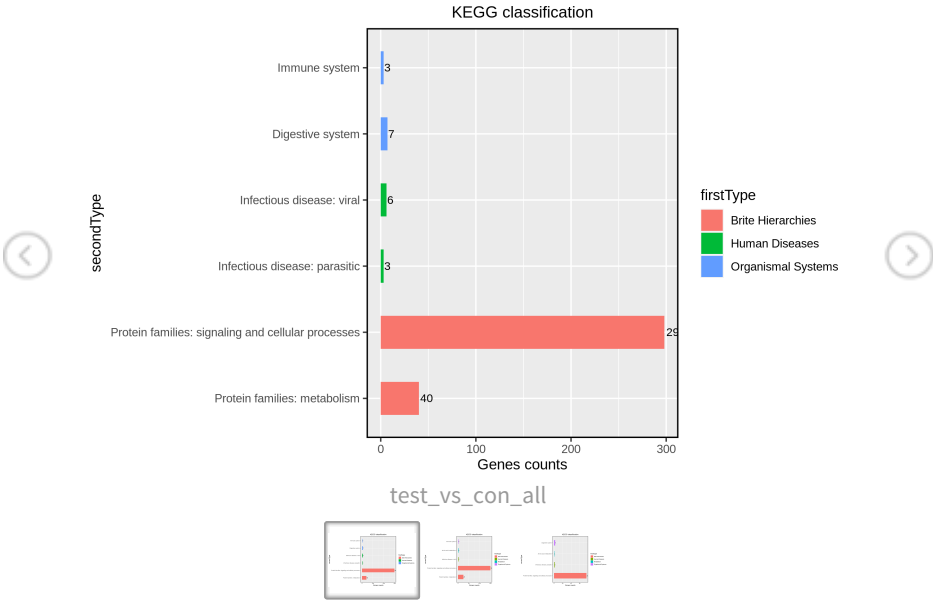


图7.6 功能聚类分类图

8 参考文献

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

[2]. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884-i890 (2018).

[3]. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907-915 (2019).

[4]. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923-930 (2014).

[5]. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140 (2010).