转录组结题报告

合同编号：	Contract No
项目编号：	Project No
客户单位：	Customer
报告单位：	康圣序源生物科技(武汉)有限公司
报告人员：	Analyser
审核人员：	Reviewer
报告日期：	Date

目 录

一、 摘要

1.1 分析结果概述

完成 6 个样品的真核有参转录组（RNA-seq）分析，共获得 38.04 Gb Clean Data，各样品 Clean Data 均 ≥ 6.00 Gb，Q30 碱基百分比均 ≥ 93.24%。

分别将各样品的 Clean Reads 与指定的参考基因组进行序列比对，比对效率从 84.60% 到 87.60% 不等。基于比对结果，进行可变剪接预测分析、基因结构优化分析以及新基因的发掘，发掘新基因 982 个，其中 285 个得到功能注释。

在本项目中，Fold Change ≥ 2 且 Pvalue < 0.05 作为差异基因筛选标准，在设置的各个比较组中，均获得差异表达基因列表、差异表达基因功能富集分析、GSEA 分析、差异可变剪切、差异基因蛋白互作等结果，详细结果请见正文各章节。

二、 生物信息分析

2.1 生物信息分析流程

生物信息分析流程如下图所示：

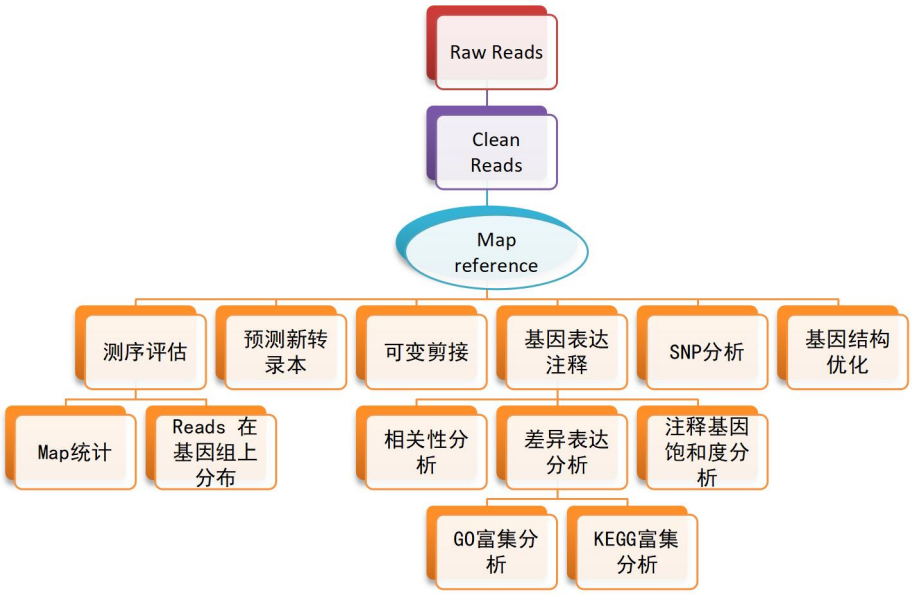


图 1.1 转录组分析流程

高通量测序下机所得的数据称为 raw reads 或 raw data，随后要对 raw reads 进行质控(QC)，以确定测序数据是否适用于后续分析。

质控后，经过滤得到 clean reads，将 clean reads 比对到参考序列。通过比对，统计 reads 在参考序列上的分布情况及覆盖度，判断比对结果是否通过第二次质控 (QC of alignment)。若通过，则进行基因表达、可变剪切、预测新转录本、SNP 检测等一系列后续分析，并从基因表达结果中，筛选出样品间差异表达的基因，基于差异表达基因，进行 GO 功能显著性富集分析和 Pathway 显著性富集分析。

2.2 序列比对

我们采用 hisat2(https://daehwankimlab.github.io/hisat2/) 对过滤后的 reads 进行参考基因组的比对分析。参考基因组链接：https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000240135.3/

HISAT2 是 TopHat2/Bowti2 的继任者，使用改进的 BWT 算法，实现了更快的速度和更少的资源占用，作者推荐 TopHat2/Bowti2 和 HISAT 的用户转换到 HISAT2。

正常情况下，如果参考基因组选择合适，而且相关实验不存在污染，实验样品测序所产生的 reads 定位的百分比会高于 70% (Total Mapped Reads or Fragments)。

所有样品比对结果如下：

表 2.2.1 基因组比对结果统计

Sample	Total Reads	Mapped Reads	Mapped Reads	Mapping Rate	Unique Mapped	Unique Mapped	Unique Mapped Rate
A1	37,057,408	32,456,242	32,462,289	87.60%	23,612,200	23,612,980	63.72%
A2	34,517,398	30,122,986	30,133,688	87.30%	22,127,218	22,125,652	64.10%
A3	47,950,374	40,582,327	40,566,016	84.60%	27,825,149	27,825,602	58.03%
B1	38,942,482	32,456,242	33,373,707	85.70%	23,612,200	24,849,198	63.81%
B2	36,124,566	30,122,986	31,175,500	86.30%	22,127,218	23,293,120	64.48%
B3	39,710,466	40,582,327	33,714,186	84.90%	27,825,149	24,243,239	61.05%

注：比对结果提供的文件为 bam 格式，可通过 IGV 软件（http://www.broadinstitute.org/igv/），并结合打开查看效果图。

2.2.1 Reads 均一性分布

根据转录组建库实验的特点，转录本其产生的测序序列实际覆盖度分布特点：距离转录本

的 5’端和 3’端越近，平均测序深度越低，但总体的均一化程度比较高。

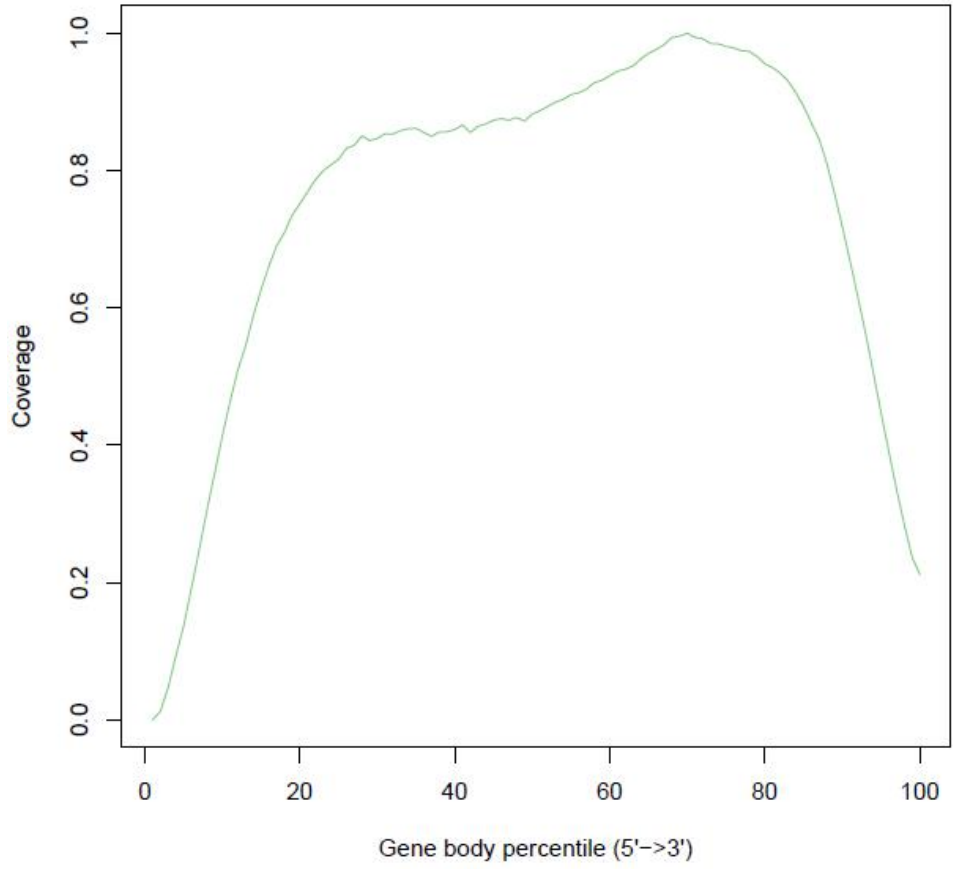


图 1.2.1 Reads 均一化分布（A1）

2.2.2 Reads 在基因组上的分布

我们使用 ANNOVAR 对 Reads 比对上的位置进行注释，注释优先级为外显子区(UTR5/UTR3, 基因上游/基因下游)>剪切区 >内含子区>基因间隔区，然后统计 Reads 在基因组上的分布。示例如下：

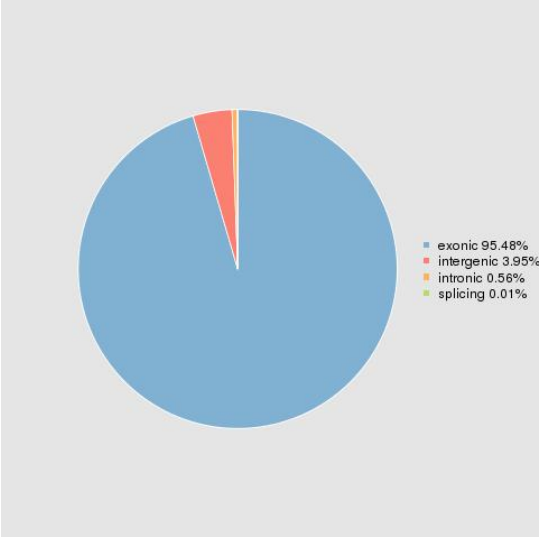


图 1.2.2 Reads 在基因组成分中的分布(A1)

注：图中主要统计比对到 exonic(外显子)、splicing(剪切区) 、intronic(内含子)、intergenic(基因间隔区域)的 Reads 所占的百分率。

从图中可以看出，所有样品 90%以上比对 exon 区域，比对效果很好。

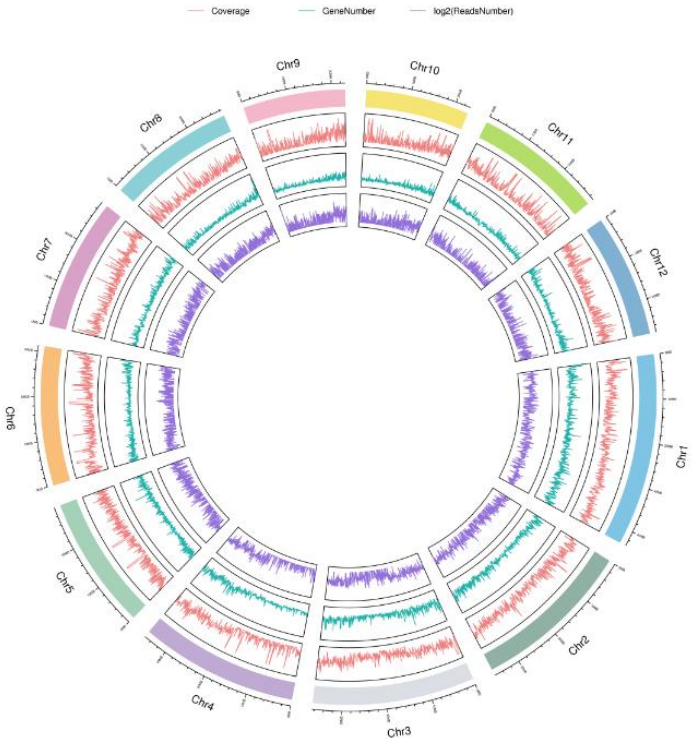


图 1.2.3 Reads 在基因组成分中的分布

- **结果文件**

比对结果 bam 文件：results/ 02_mapping/样品名/accepted_hits.bam

比对结果 bam 的 index 文件：results/ 02_mapping/样品名/accepted_hits.bam.bai

均一化分布：results/ 02_mapping/样品名/uniform.geneBodyCoverage.curves.png

Reads 在基因组成分中的分布: results/02_mapping/样品名/reads_map.png

- **参考文献**

1）doi:10.1038/nmeth.3317 (**hisat2**)

2.3 转录本拼接

Stringtie 使用的是流神经网络算法，Cufflinks 则是吝啬算法；从组装效果上来看 Stringtie 在灵敏度和准确度上表现较好，能够拼接出更完整、更准确的基因；从定量上来说，两者相差不大，但是 cufflinks 在一些特殊情况下会有异常的表达量；从运行速度上来说，Stringtie 远远快了 cufflinks。

- **结果文件**

Strintie 拼接结果文件：results/ 02_stringtie/样品名/样品名.*.gtf.(按需提供)

表 1.3 序列拼接结果展示

Chr	Source	Feature	Start	End	Score	Strand	Frame	Atributes
Supercontig_8.1	Cufflinks	transcript	42074	42397	1	+	.	gene_id "MGG_15984"; transcript_id "MGG_15984T0"; FPKM "0.0000000000"; …;
Supercontig_8.1	Cufflinks	exon	42074	42076	1	+	.	gene_id "MGG_15984"; transcript_id "MGG_15984T0"; exon_number "1"; FPKM "0.0000000000";…;
Supercontig_8.1	Cufflinks	exon	42212	42397	1	+	.	gene_id "MGG_15984"; transcript_id "MGG_15984T0"; exon_number "2"; FPKM "0.0000000000";…;
Supercontig_8.1	Cufflinks	transcript	42735	45926	1	-	.	gene_id "MGG_01949"; transcript_id "MGG_01949T0"; FPKM "0.0000000000"; frac "0.000000"; …;
Supercontig_8.1	Cufflinks	exon	42735	44381	1	-	.	gene_id "MGG_01949"; transcript_id "MGG_01949T0"; exon_number "1"; FPKM "0.0000000000";…;

注：

1) Chr: chromosome 或 scaffold 的名称;

2) Source：数据来源信息，为'Cufflinks'

3) Feature：序列类型描述，为“transcript” 或 “exon”

4) Start：起始坐标

5) End：终止坐标

6) Score：对应位置拼接得分

7) Strand：序列正负链信息

8) Frame：序列编码起始位点的信息，Cufflinks 不进行起始和中止密码子的预测，这一列为'.'

9) Attributes：序列的其他描述信息，如基因 ID，转录本 ID 以及表达值信息等

2.3.1 基因表达水平

一个基因表达水平的直接体现就是其转录本的丰度情况，转录本的丰度程度越高，则基因的表达水平越高。在 RNA-seq 中，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。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、测序深度成正相关。

FPKM（expected number of Fragments Per KiloR88002 of transcript sequence per Millions R88002 pairs sequenced）是每百万 fragments 中来自某一基因每千碱基长度的 fragments 数目，其同时考虑了测序深度和基因长度对 fragments 计数的影响，是目前最为常用的基因表达水平估算方法（Trapnell, Cole, et al., 2010）。

RPKM（Reads Per Kilo R88002s 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，也是目前最为常用的基因表达水平估算方法(Mortazavi et al., 2008)。

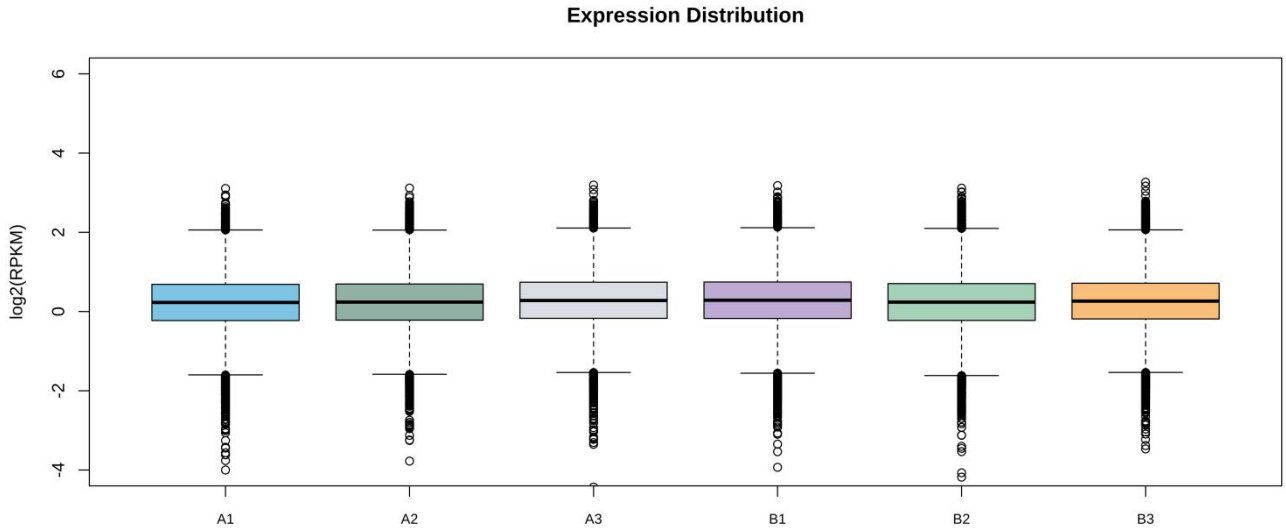
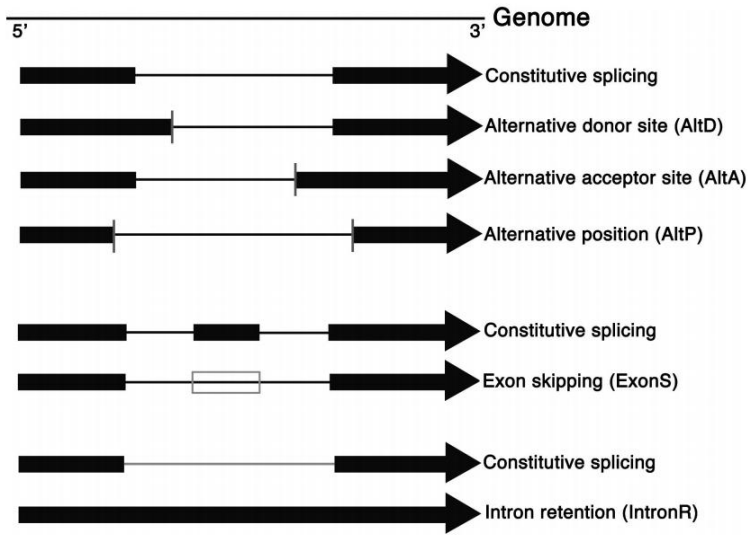


图 2.3.1 各样品表达量分布图（mRNA）

2.3.2 可变剪切分析

可变剪接（alternative splicing）使一个基因产生多个 mRNA 转录本，不同 mRNA 可能翻译成不同蛋白。因此，通过可变剪接一个基因可能产生多个蛋白，极大地增加了蛋白多样性。

在生物体内，主要存在以下几种可变剪切类型，A) Exon skipping(SE/ExonS); B) Intron retention(IR/IntronR)； C) Alternative 5' splice site(AltD/A5SS)； D) Alternative 3' splice site(AltA/A3SS)；E) Alternative posion(AltP)。具体如下：



根据拼接结果，我们使用 Astalavista 软件对前四种可变剪切进行分析，统计结果如下：

表 2.3.2 可变剪接统计表

Sample	SE	A3SS	A5SS	RI
A1	31	255	80	636
A2	29	259	75	578
A3	31	295	101	632
B1	34	265	78	625
B2	28	238	69	598
B3	34	304	123	641

注：各列分别表示发生相应可变剪切类型的基因数目的统计值。

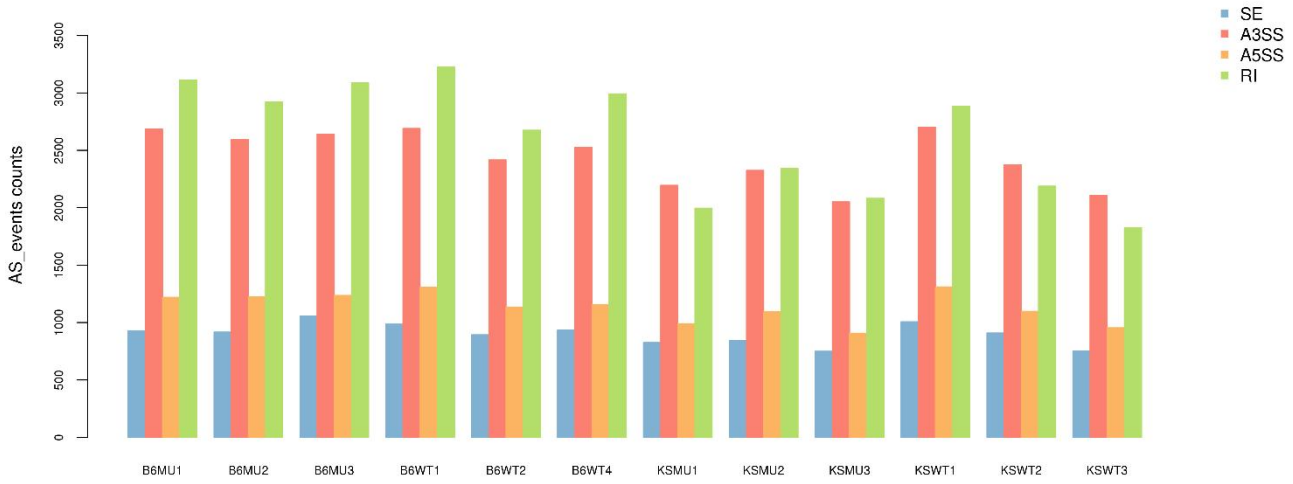


图 2.3.1 可变剪切统计图

➢ **结果文件**

Cufflinks 拼接结果文件：results/03_cufflinks/样品名/transcripts.gtf
所有样品的基因 FPKM 表达值结果文件：results/03_cufflinks/all.genes.FPKM.xls
所有样品的转录本 FPKM 表达值结果文件：results/03_cufflinks/all.isoform.FPKM.xls
样品间可变剪接详细结果：results/03_cufflinks/01_AS/AStalavista/样品名/transcripts.astalavista.gtf
样品间可变剪接事件统计结果：results/03_cufflinks/01_AS/AStalavista/all.sample.AS.stat.xls
样品间可变剪接事件统计图：results/03_cufflinks/01_AS/AStalavista/AS.barplot.tiff

➢ **参考文献**

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

2) Foissac S, Sammeth M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets[J]. Nucleic acids research, 2007, 35(suppl 2): W297-W299.（Astalavista）

3) Shen S, Park J W, Lu Z, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data[J]. Proceedings of the National Academy of Sciences, 2014, 111(51): E5593-E5601.（rMATS）

2.3.3 基因结构优化

通过比较测序结果和现有基因注释结果，对基因的 5'端或 3'端进行延长。

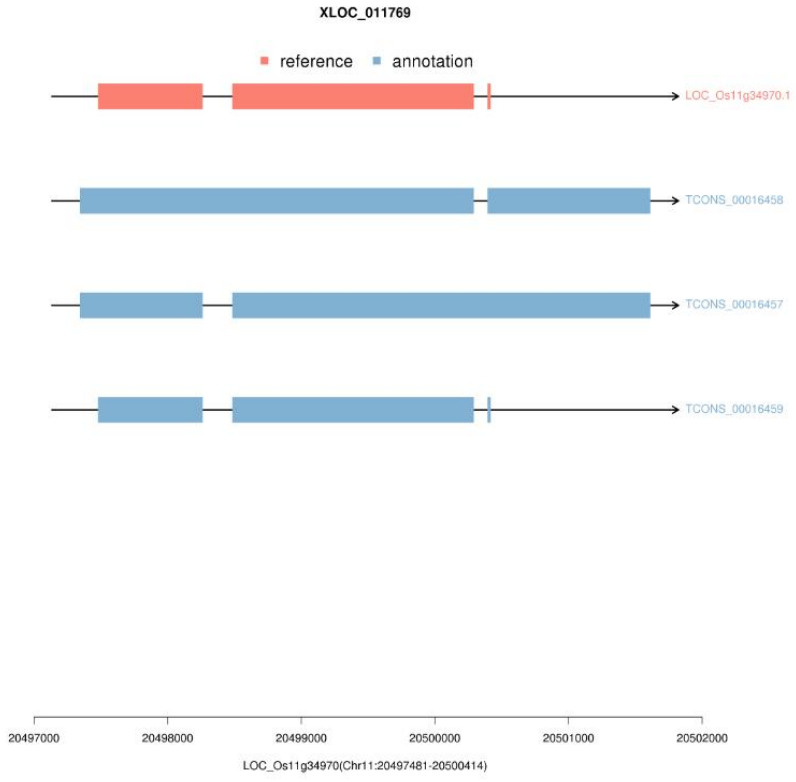


图 2.3.3 某基因的基因结构优化示例

注：从图中可以看出，此基因的 5’和 3’端均有一定程度的延伸。

表 2.3.3 基因 3' or 5'延长结果展示

Gene	isoform	chr	Strand	5' or 3'	OriginalRegion	FinalRegion
LOC_Os01g01830	LOC_Os01g01830.1	Chr1	+	3'	437613-449137	437613-449303

LOC_Os01g02360	LOC_Os01g02360.1	Chr1	+	5' and 3'	745452-748945	745100-749027
LOC_Os05g34914	LOC_Os05g34914.1	Chr5	+	5' and 3'	20737930-20741084	20737841-20746179
LOC_Os12g39440	LOC_Os12g39440.5	Chr12	-	5'	24273868-24275502	24273868-24275536

注： gene:发生结构优化区域的基因；
isoform:发生结构优化的基因对应的转录本；
5' or 3' ：延长端；
Chr：染色体号；
strand：链的方向；
original_region：原始参考基因组的位置；
final_region：延长后的 exon 区域。

2.3.2 新转录本预测

比对上的 reads 或拼接后的 reads，因为未注释到已知参考基因上，故可能是发现的新基因。

为找到新转录本区域，我们将组装的转录本与参考序列注释的转录本进行比较。成为新转录本的组装转录本必须满足以下条件：

1) 距离现有的注释 gene 200bp 以上；

2) 长度不短于 180bp;

3) 测序深度不小于 2。

我们对新转录本个数进行了统计，得到下表：

表 1.3.5.1 新转录本个数统计表

Sample	Number of Novel_Prediction Transcripts
B6MU1	1,383
B6MU2	1,347
B6MU3	1,465
B6WT1	1,479
B6WT2	1,278
B6WT4	1,389
KSMU1	1,189
KSMU2	1,280
KSMU3	988
KSWT1	1,508
KSWT2	1,411
KSWT3	1,167

新转录本预测结果展示如下：

表 1.3.4.2 新转录本预测结果展示表

Novel_TU_ID	Chromosome	+/-	Blocks	Sizes	Starts	Length	Counts	RPKM
newGene_1	Chr1	+	3	134, 126, 864	596166, 596430, 596931	1124	78	59.5907
newGene_2	Chr1	+	2	143, 518	686013, 686285	661	137	177.979
newGene_3	Chr1	+	2	143, 513	686013, 686290	656	135	176.718
newGene_4	Chr1	+	2	155, 513	686013, 686290	668	136	174.829
newGene_5	Chr1	-	3	257, 107, 259	1001070, 1001459,	620	34	47.0909

从左往右，每一列含义如下：

1) 新转录本 ID；
2) 所属染色体；
3) 新转录本所在染色体的正负链信息；
4) 新转录本外显子数目；
5) 新转录本每个外显子大小（逗号分隔）；
6) 每个外显子起始位点在染色体上的位置（逗号分隔）；
7) 新转录本的长度；
8) 新转录本的 counts 值，即支持转录本的 reads 条数；
9) 新转录本的 RPKM 值。

后续可根据研究需要，可对相关新基因做编码潜能预测，以及相关功能注释。

- **结果文件**

所有样品的基因 FPKM 表达值结果文件：results/03_cufflinks/all.genes.FPKM.xls
所有样品的转录本 FPKM 表达值结果文件：results/03_cufflinks/all.isoform.FPKM.xls
样品间可变剪接详细结果：results/03_cufflinks/01_AS/AStalavista/样品名/transcripts.astalavista.gtf
样品间可变剪接事件统计结果：results/03_cufflinks/01_AS/AStalavista/all.sample.AS.stat.xls
样品间可变剪接事件统计图：results/03_cufflinks/01_AS/AStalavista/AS.barplot.tiff
基因结构优化结果：results/03_cufflinks/02_optimization/样品名/optimization_information.out.xls
反义转录本列表：results/03_cufflinks/03_other_structure/oppo_strand.out.xls
所有基因注释结构信息图：results/03_cufflinks/structure/*.tiff
新转录本详细文件：results/06_novel/样品名/*Novel.information.xls
新转录本序列文件：results/06_novel/样品名/*Novel.xls.fa

2.4 样品间表达相关性检查

生物学重复是任何生物学实验和测序技术所必需的。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。

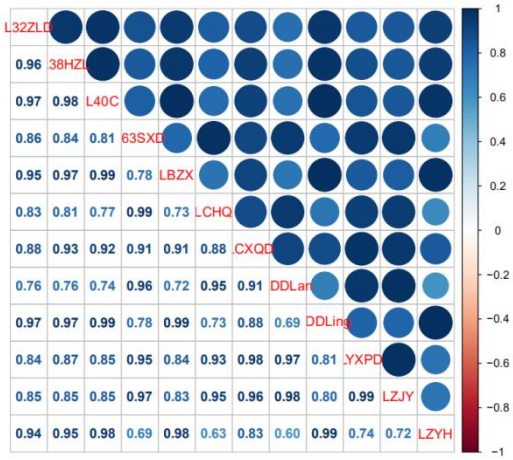


图 2.4 各样品表达量相关性图

➢ **结果文件**

相关性表格：results/04_cor/sample.cor.xls

2.5 SNP 分析

SNP 全称为单核苷酸多态性(Single-nucleotide polymorphism)，一般是指在 DNA 或 RNA 水平上，发生于样本或个体间的单核苷酸突变。

我们使用比对后的序列，用 samtools mpileup 等进行 call SNP 变异检测。

统计结果如下：

表 2.5.1 染色体上 SNP-Indel 数目统计表

chromosome	Number-of-SNP	Number-of-Indel
Chr1	1029	38
Chr10	304	8
Chr11	931	21
Chr12	474	29
Chr2	519	28
Chr3	509	24
Chr4	680	62
Chr5	221	16
Chr6	2086	81
Chr7	933	41
Chr8	268	8

Chr9	277	11
ChrSy	13	2
ChrUn	1	0

对基因的 exon 区域发生 snp 的突变率统计，结果如下：

表 1.5.2 基因突变率统计

Chromosome	Gene	Gene_Region	Gene_Length	CDS_Length	Number_of_SNP	Mutation_Rate
Chr9	LOC_Os09g30418	18535746-18541109(	5364	2493	2	0.000802246289611
Chr6	LOC_Os06g33120	19281458-19282361(	904	687	2	0.00291120815138
Chr10	LOC_Os10g07040	3696797-3698774(+)	1978	1197	1	0.000835421888053
Chr7	LOC_Os07g45300	27030234-27035093(	4860	1062	3	0.002824858757006
Chr7	LOC_Os07g35720	21391473-21393560(	2088	815	3	0.003680981595009
Chr2	LOC_Os02g06480	3227508-3233934(-)	6427	1026	4	0.003898635477558

注：chromosome：基因所在染色体

Gene:发生 snp 的基因

Gene_Region：基因区间

Gene_Length: 基因的总长度

CDS_Length：基因的 exon 的总长

Number_of_SNP: gene 的 exon 区域发生 snp 的个数

Mutation rate：发生在 gene 的 exon 区域的突变率

图 2.5.1 基因突变率统计图

从结果来看，发生 SNP 的基因，基本在千分之一内。

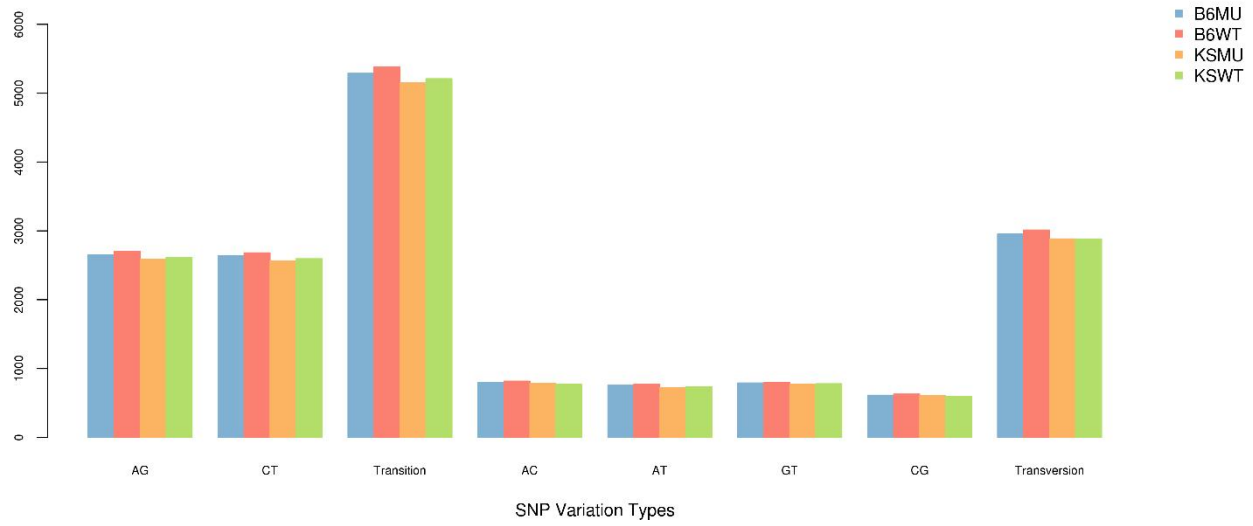


图 2.5.2 SNP 类型统计图

小注：碱基颠换(transversion): 是指在碱基置换中嘌呤与嘧啶之间的替代；而转换(transition)则是一个嘌呤被另一个嘌呤，或者是一个嘧啶被另一个嘧啶替代。

- **结果文件**

SNP vcf 结果：/results/05_SNP_Indel/SNP.vcf
Indel vcf 结果：/results/05_SNP_Indel/INDEL.vcf
染色水平突变统计：/results/05_SNP_Indel/SNP-Indel.chr.stat.xls
基因水平 SNP 统计：/results/05_SNP_Indel/Gene_mutation.stat.xls
基因 SNP 突变率统计图：/results/05_SNP_Indel/Gene_mutation.stat.tiff

SNP 类型统计图：/results/05_SNP_Indel/SNP_type.stat.tiff

2.6 差异表达分析

通过差异分析，我们能够得到在不同处理或不同表型样品之间表达显著差异的转录本 (基因)，它们可能在不同处理或不同表型中发挥功能。

差异算法选择：

Cuffdiff：cufflinks 配套软件，输入表达值为 FPKM；

DEGseq：R 语言 package，输入表达值为 RPKM 或 FPKM；

DESeq、DESeq2：R 语言 package，输入表达值为 Counts 值；

edgeR：R 语言 package，输入表达值为 Counts 值。

…

RNAseq 的软件的差异算法通常符合负二项分布或者泊松分布，这也是与 RNAseq 的 reads 分布相符合的。选择合适的算法，对差异基因的选择起到事半功倍的效果。

差异表达分析本次使用 Cuffdiff (http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html) 软件。

在实际分析过程中，若有生物学重复，则以校正后的 p 值（q_value/fdr）≤ 0.05 为阈值；若无生物学重复，则以 p_value ≤ 0.05 为阈值。同时选择倍数差异在 2 倍以上的基因。

本次分析筛选条件为 p_value ≤0.05 且|log2(foldchange)|≥1。

表 2.6.1 差异结果统计

差异比较组	Up	Down	All
A_VS_B	216	180	396

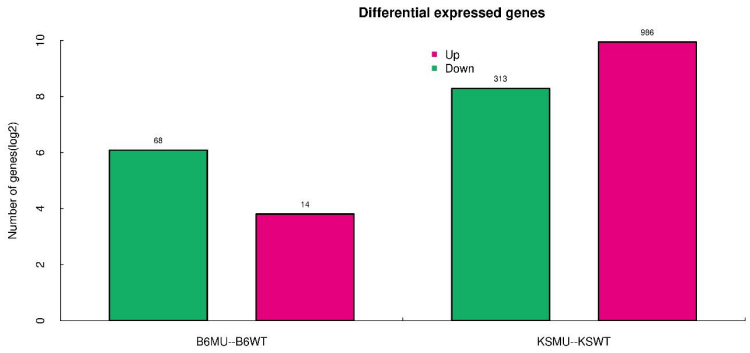


图 2.6.1 差异结果统计图

表 2.6.2 差异结果示例

gene_id	control	treat	log_2(fold_change)	p_value	q_value
MGG_00016	1.97537	0.338241	-2.546	0.04345	0.997423
MGG_00053	12.2691	5.23543	-1.22865	0.0417	0.997423
MGG_00085	6.47688	17.8838	1.46528	0.0162	0.859866
MGG_00087	36.2687	16.2929	-1.15448	0.0455	0.997423

注：fold_change表示treat/control，p_value表示通过统计算法得到的p_value，q_value表示多重显著性差异校正值。

差异火山图示例如下：

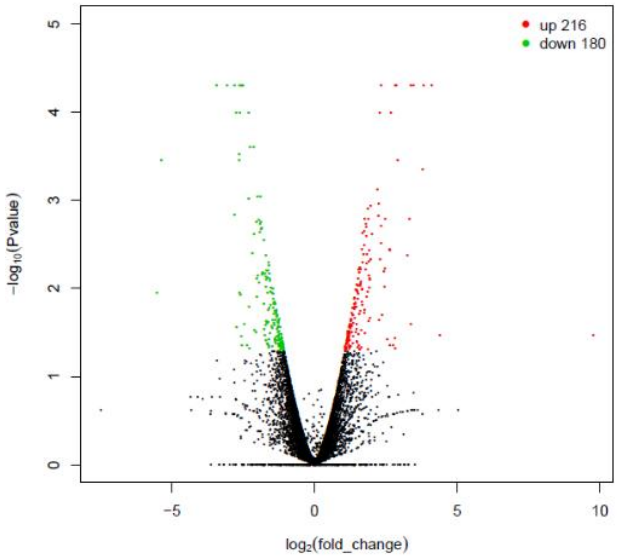


图 2.6.1 差异火山图

差异结果聚类结果如下：

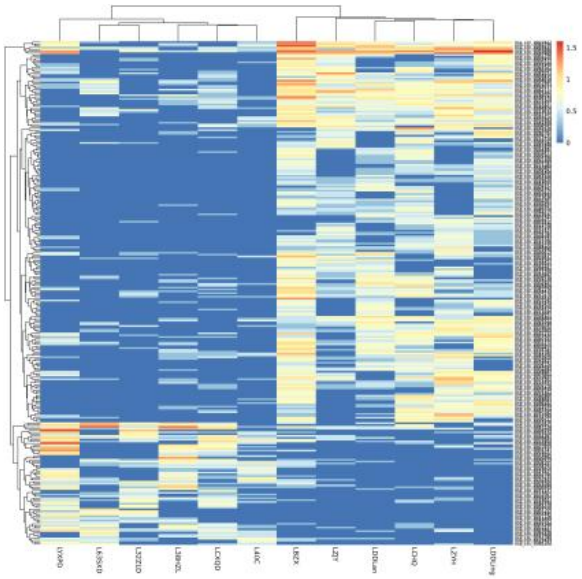


图 2.6.2 差异聚类图

- **结果文件**

基因差异结果总文件：results/06_diff/比较组/gene/比较组.gene.exp.diff.xls

基因差异结果过滤文件：results/06_diff/比较组/gene/比较组.gene.exp.diff.filter.xls

转录本差异结果总文件：results/06_diff/比较组/isoform/比较组.isoform.exp.diff.xls

转录本差异结果过滤文件：results/06_diff/比较组/isoform/比较组.isoform.exp.diff.filter.xls

差异基因火山图：results/06_diff/比较组/gene/比较组.diff.genes.volcano.pdf

差异结果统计图：results/06_diff/比较组/Diffgenes.barplot.tiff

2.7 差异基因富集

2.7.1 GO 富集分析

Gene Ontology（简称 GO）是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）来全面描述生物体中基因和基因产物的属性。

GO 总共有三个 ontology（本体），分别描述基因的分子功能（molecular function）、所处的细胞位置（cellular component）、参与的生物过程（biological process）。GO 的基本单位是 term（词条、节点），每个 term 都对应一个属性。GO 功能分析一方面给出差异表达基因的

GO 功能分类注释；另一方面给出差异表达基因的 GO 功能显著性富集分析。

GO 功能分类注释给出具有某个 GO 功能的基因列表及基因数目统计。GO 功能显著性富集分析给出与基因组背景相比，在差异表达基因中显著富集的 GO 功能条目，从而给出差异表达基因与哪些生物学功能显著相关。该分析首先把所有差异表达基因向 Gene Ontology 数据库（http://www.geneontology.org/）的各个 term 映射，计算每个 term 的基因数目，然后应用卡方检验或者超几何检验，找出与整个基因组背景相比，计算得到的 pvalue 通过 Bonferroni 校正之后，以 corrected-pvalue≤0.05 为阈值，满足此条件的 GO term 定义为在差异表达基因中显著富集的 GO term。通过 GO 功能显著性富集分析能确定差异表达基因行使的主要生物学功能。

富集结果如下：

表 2.7.1 差异基因 GO 富集结果示例

GO_ID	GO_Term	GO_Class	Pvalue	AdjustedPv	x1	x2	n	N	EnrichDirect	GOlevl	GeneID
GO:0005575	cellular_component	CC	0.0031545466474064	0.352945272291744	31	3497	116	8630	Under	1	CE116606_2321,...
GO:0005623	cell	CC	0.0168501433616840	0.352945272291744	21	2458	116	8630	Under	2	CE125468_1574,...
GO:0005622	intracellular	CC	0.0217342470015189	0.352945272291744	21	2420	116	8630	Under	3	CE125468_1574,...
GO:0043226	organelle	CC	0.0248750339291314	0.352945272291744	11	1539	116	8630	Under	2	CE125468_1574,...

从左至右各列含义如下：

1) GO ID 号
2) GO 所属分类
3) GO 所属 Ontology
4) 富集 P value
5) 校正后的 P value
6) 用于富集的差异基因中注释到该 GO 号的基因数
7) 基因组中所有基因中注释到该 GO 号的基因数
8) 用于富集的差异基因注释到 GO 的基因个数
9) 基因组中所有注释到 GO 的基因个数
10) 富集结果：Over 表示为显著富集，Under 表示非显著富集
11) GO ID 对应的 GO level
12) 富集上的基因

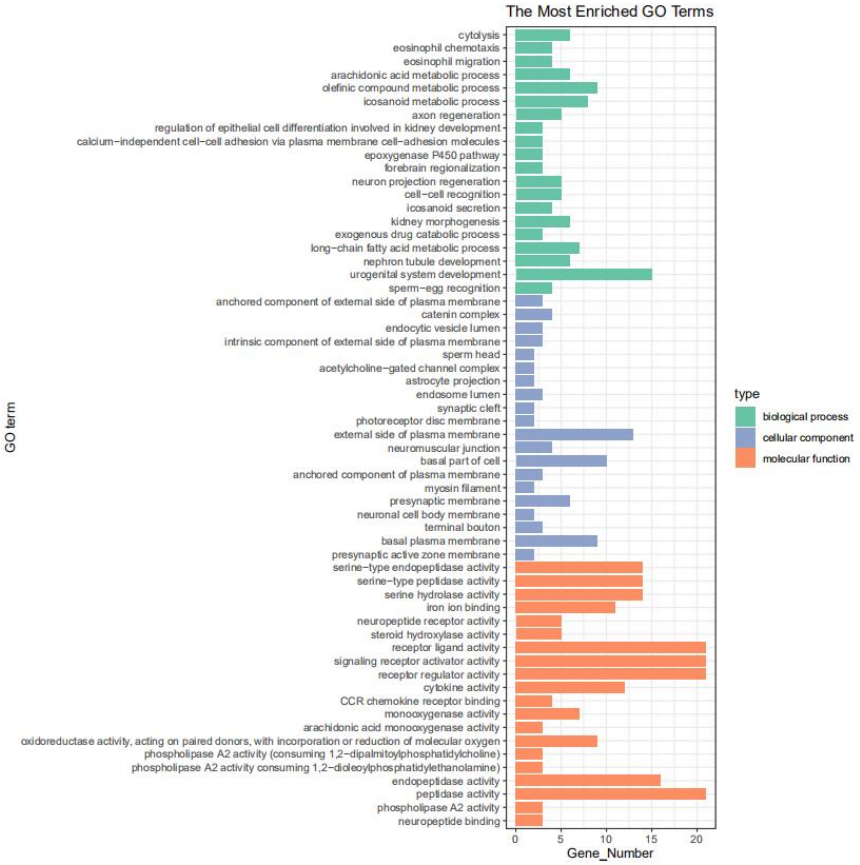


图 2.7.1 GO 富集柱状图

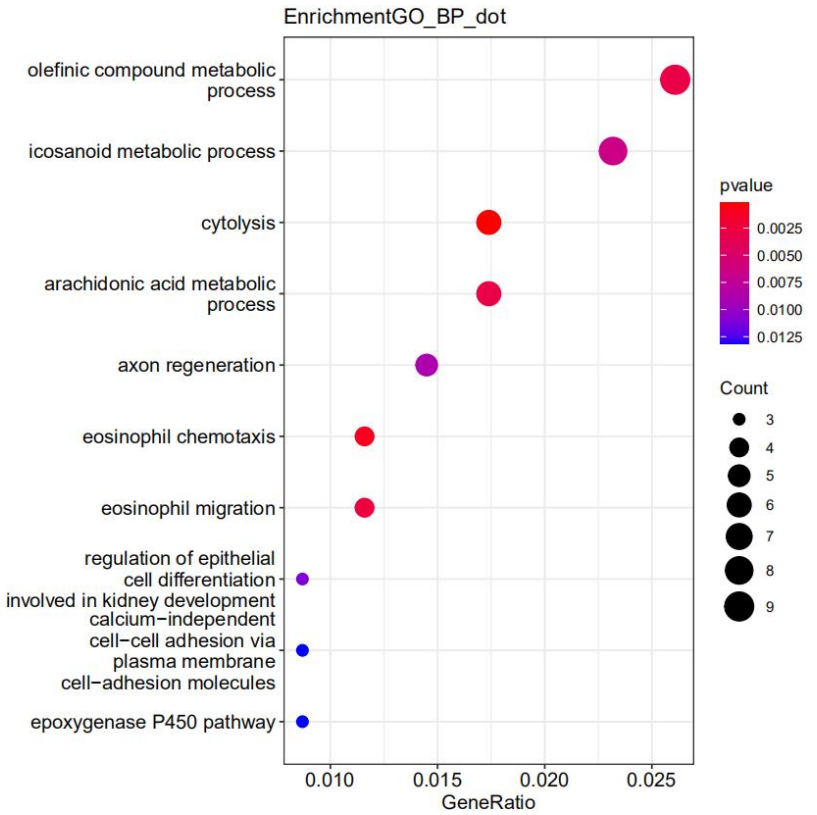


图 2.7.2 GO 富集散点图（BP）

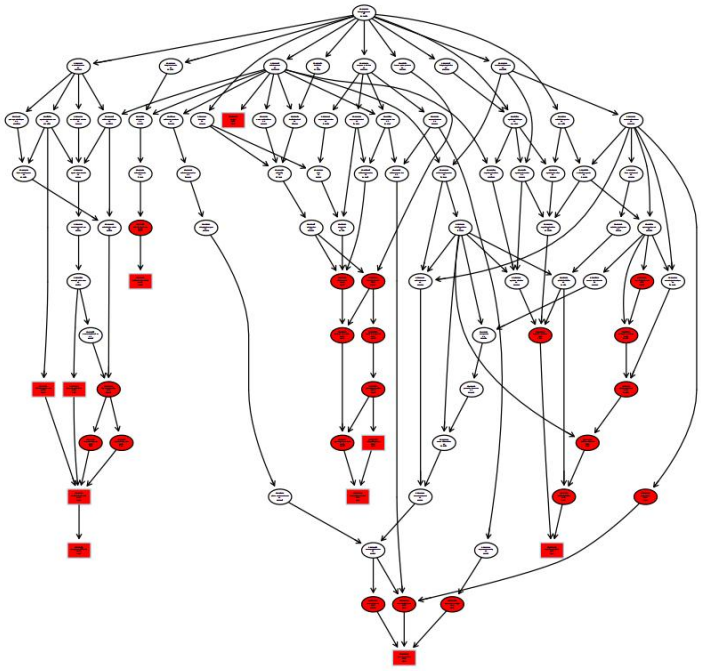


图 2.7.3 GO 富集有向无环图（BP）

2.7.2 KEGG 富集分析

KEGG（Kyoto Encyclopedia of Genes and Genomes）是系统分析基因产物在细胞中的代谢途径（Pathway）以及这些基因产物功能的主要公共数据库，利用 KEGG 可以进一步研究基因在生物学上的复杂行为。

与 GO 富集类似，我们基于差异分析和 KEGG 富集的结果，应用超几何检验，找出显著富集的 KEGG Pathway。

表 2.7.2 差异基因 KEGG 富集结果示例

MapID	MapTitle	Pvalue	AdjustedPv	x	y	n	N	EnrichDirect	GeneIDs
map00410	beta-Alanine metabolism	2.57762253446676e-08	6.03163673065221e-06	19	259	465	21605	Over	CE127_112925,…
map00310	Lysine degradation	1.06936244292537e-07	1.25115405822268e-05	20	226	465	21605	Over	CE100068_467,…
map003	Tryptophan	1.9058497	0.001486562	17	28	46	216	Over	CE252902

80	metabolism	018533e-05	76744558		4	5	05		_142804,…
map04530	Tight junction	6.838961084875 83e-05	0.00315549063862184	12	144	465	21605	Over	CE136342_4729,…

从左至右各列含义如下：

1) KEGG Pathway MapID
2) Map 标题
3) 富集 P value
4) 校正后的 P value
5) 用于富集的差异基因中注释到该 KEGG Pathway 的基因数
6) 基因组中所有基因中注释到该 KEGG Pathway 的基因数
7) 用于富集的差异基因中注释到 KEGG Pathway 的基因数
8) 基因组中所有注释到 KEGG Pathway 的基因数
9) 富集结果：Over 表示为显著富集，Under 表示非显著富集
10) 富集上的基因

KEGG富集统计结果如下：

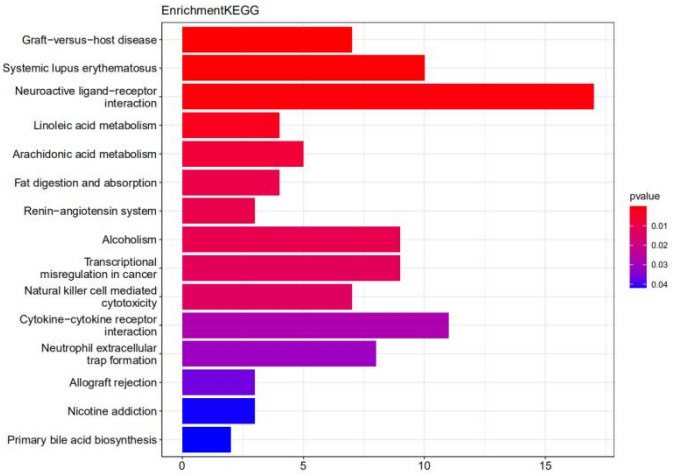


图 2.7.4 KEGG 富集柱状图

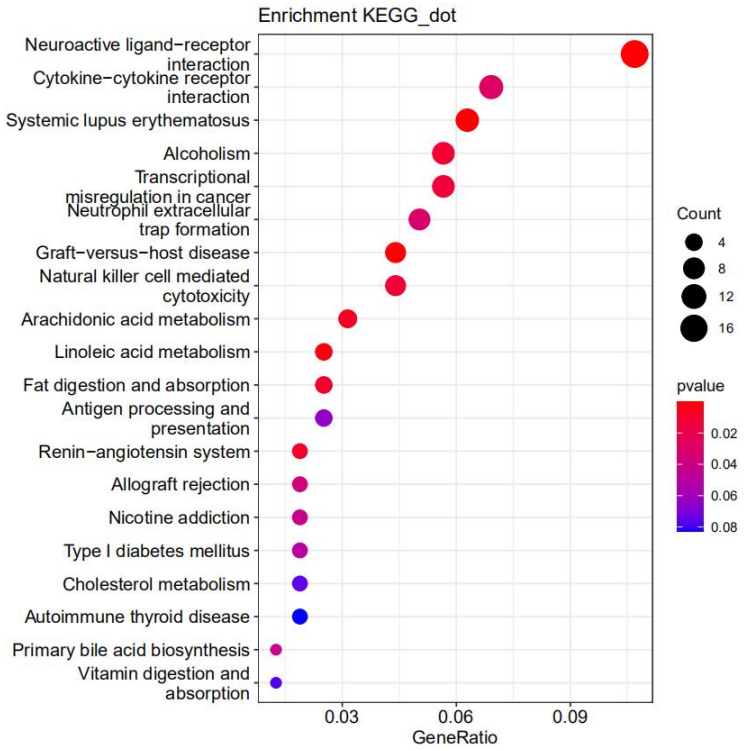


图 2.7.5 KEGG 富集散点图

➢ **结果文件**

差异基因GO富集信息：results/07_Enrich/比较组/GO/比较组.diff.gene.GO.enrich.xls

差异基因GO富集过滤信息：results/07_Enrich/比较组/GO/比较组.diff.gene.GO.enrich.filter.xls

差异基因GO富集统计图：results/07_Enrich/比较组/GO/比较组.enrich.GO.stat.pdf

差异基因KEGG富集信息：results/07_Enrich/比较组/KEGG/比较组.diff.gene.kegg.enrich.xls

差异基因KEGG富集过滤信息：results/07_Enrich/比较组/KEGG/比较组.diff.gene.kegg.enrich.filter.xls

差异基因KEGG富集统计图：results/07_Enrich/比较组/KEGG/比较组.enrich.KEGG.stat.tiff

差异基因KEGG富集通路图：results/07_Enrich/比较组/KEGG/KEGG_graph/*

➢ **参考文献**

1) Young, Matthew D., et al. "goseq: Gene Ontology testing for RNA-seq datasets." (2012).