



October 29, 2018

生物信息分析报告

华大基因生物技术有限公司

目录

1 项目概况	4
2 建库测序流程	5
2.1 Total RNA 样品检测	5
2.2 文库构建	5
3 生物信息分析流程	6
4 结果展示及说明	7
4.1 原始序列数据	7
4.2 测序数据质量评估	7
4.2.1 测序错误率分布检查	7
4.2.2 测序数据过滤	8
4.3 Denovo 组装	9
4.4 Unigene 功能注释	11
4.4.1 NT 和 NR 注释	11
4.4.2 KOG 注释	13
4.4.3 GO 注释	14
4.4.4 KEGG 注释	15
4.4.5 SwissProt 注释	17
4.5 Unigene 的 CDS 预测	18
4.6 Unigene 的 SSR 检测	19
4.7 SNP 检测	21
4.8 Unigene 的 TF 编码能力预测	23
4.9 基因表达量计算	23
4.9.1 基因表达量水平	23
4.9.2 样品中基因表达量的分布	24
4.9.3 PCA 分析	27
4.10 时间序列分析	28
4.11 差异表达基因检测	30
4.12 差异表达基因 GO 功能分析	34
4.13 差异表达基因 Pathway 功能分析	37
4.14 差异基因蛋白互作分析	40
4.15 植物抗病基因预测	41
4.16 真菌致病基因预测	42



5 报告补充说明	43
5.1 文件目录链接	43
5.2 结果文件	43
6 参考文献	44
7 常见问题	45
8 联系方式	46



1 项目概况

项目名称: rice mRNASeq

客户名称: yueyao

客户单位: 华大基因

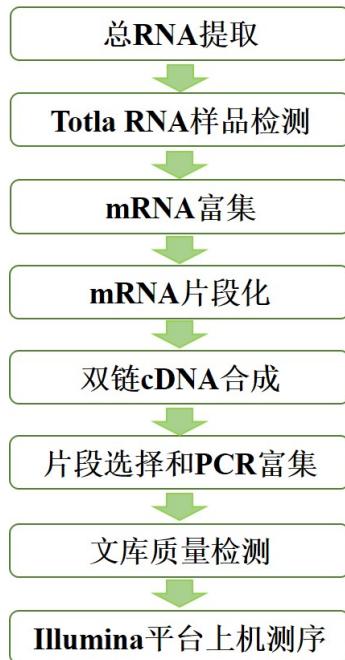
项目编号: 000001

物 种: rice

样本数量: EUlef

2 建库测序流程

从 RNA 样品到最终测序数据的分析，样品检测、建库、测序每一个环节都会对数据质量和数量产生影响，而数据质量又会直接影响后续信息分析的结果。为了从源头上保证测序数据的准确性、可靠性，华大对样品检测、建库、测序每一个生产步骤都严格把控，从根本上确保了高质量数据的产出。流程图如下：



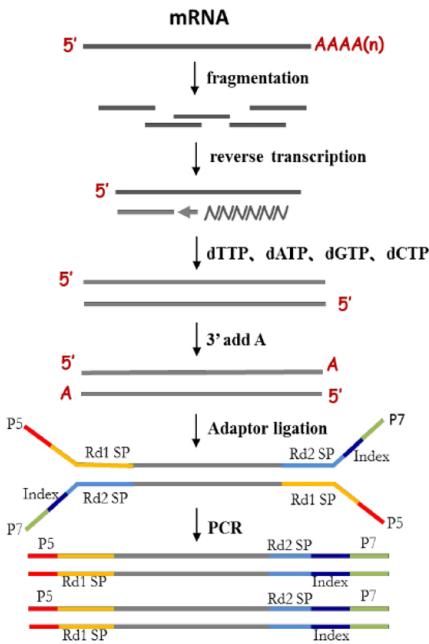
2.1 Total RNA 样品检测

对 RNA 样品的检测主要包括 4 种方法：

- 琼脂糖凝胶电泳分析 RNA 降解程度以及是否有污染
- Nanodrop 检测 RNA 的纯度（OD₂₆₀/280 比值）
- Qubit 对 RNA 浓度进行精确定量
- Agilent2100 精确检测 RNA 的完整性

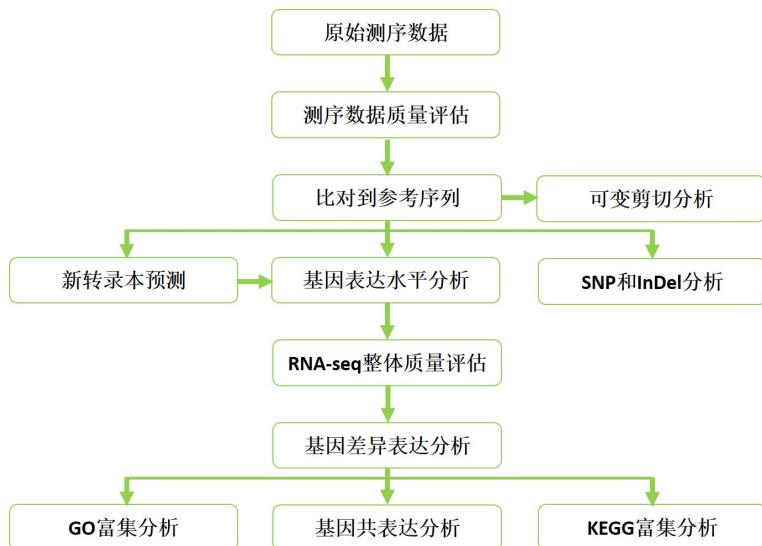
2.2 文库构建

样品检测合格后，用带有 Oligo (dT) 的磁珠富集真核生物 mRNA（若为原核生物，则通过试剂盒去除 rRNA 来富集 mRNA）。随后加入 fragmentation buffer 将 mRNA 打断成短片段，以 mRNA 为模板，用六碱基随机引物（random hexamers）合成一链 cDNA，然后加入缓冲液、dNTPs 和 DNA polymerase I 合成二链 cDNA，随后利用 AMPure XP beads 纯化双链 cDNA。纯化的双链 cDNA 再进行末端修复、加 A 尾并连接测序接头，然后用 AMPure XP beads 进行片段大小选择，最后进行 PCR 富集得到最终的 cDNA 文库。构建原理图如下：



3 生物信息分析流程

首先对原始下机数据（raw data）进行过滤，将过滤后得到的高质量序列（clean data）比对到该物种的参考基因组上。根据比对结果，计算每个基因的表达量。在此基础上，进一步对样品进行表达差异分析、富集分析和聚类分析。



4 结果展示及说明

4.1 原始序列数据

高通量测序(如 Illumina HiSeqTM 2500/MiseqTM)得到的原始图像数据文件经 CASAVA 碱基识别(Base Calling)分析转化为原始测序序列(Sequenced Reads), 我们称之为 Raw Data 或 Raw Reads, 结果以 FASTQ(简称为 fq) 文件格式存储, 其中包含测序序列(reads)的序列信息以及其对应的测序质量信息。

FASTQ 格式文件中每个 read 由四行描述, 如下:

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT
```

```
NAAGAACACGTTGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
```

```
+
```

```
#55???BBBBB?BA@DEEFFcffHHFFCFFHHHHHHHFAE0ECFFD/AEHH
```

其中第一行以'@'开头, 随后为 Illumina 测序标识符(Sequence Identifiers)和描述文字(选择性部分);

第二行是碱基序列;

第三行以'+'开头, 随后为 Illumina 测序标识符(选择性部分);

第四行是对应碱基的测序质量, 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。

测序样本序列见文件夹 [/BGI_result/1.CleanData](#) 文件夹

4.2 测序数据质量评估

4.2.1 测序错误率分布检查

每个碱基测序错误率是通过测序 Phred 数值(Phred score, Q phred)通过公式(公式 1 : $Q_{phred} = -10\log_{10} e$)转化得到, 而 Phred 数值是在碱基识别(Base Calling)过程中通过一种概率模型计算得到, 这种模型可以准确地预测碱基判别的错误率。Phred 分值, 不正确的碱基识别率, 碱基正确识别率以及 Q-score 的对应关系如下表所显示:

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

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

测序错误率与碱基质量有关, 受测序仪本身、测序试剂、样品等多个因素共同影响。对于 RNA-seq 技术, 测序错误率分布具有两个特点:

(1) 测序错误率会随着测序序列(Sequenced Reads)长度的增加而升高, 这是由于测序过程中化学试剂的消耗导致的, 并且为 illumina 高通量测序平台都具有的特征。

(2) 前 6 个碱基的位置也会发生较高的测序错误率，而这个长度也正好等于在 RNA-seq 建库过程中反转录所需要的随机引物的长度。所以前 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。

4.2.2 测序数据过滤

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

数据处理的步骤如下：

- (1) 去除带接头 (adapter) 的 reads;
- (2) 去除 N(N 表示无法确定碱基信息) 的比例大于 10% 的 reads;
- (3) 去除低质量 reads(质量值 $sQ \leq 5$ 的碱基数占整个 read 长度的 50% 以上的 reads)

过滤后 reads 的质量指标见表1，碱基含量分布以及质量分布见图1和图2

测序样本序列比对到基因组结果见文件夹 [./BGI_result/1.CleanData](#) 文件夹

表 1: 过滤后的 reads 质量统计

Sample	Total Raw Reads(Mb)	Total Clean Reads(Gb)	Total Clean Bases(Gb)	Clean Reads Q20(%)	Clean Reads Q30(%)	Clean Reads Ratio(%)
EUlef	40.99	38.76	3.49	97.19	91.36	94.56
EUski	62.29	58.65	5.28	97.03	91.02	94.16

¹ Q20: 质量值大于 20 的碱基数目占总碱基数目的比例。

¹ Total Clean Reads(Mb): 过滤后的 reads 数

² Total Clean Bases(Gb): 过滤后的碱基总数

³ Clean Reads Q20(%): 过滤后的 reads 中质量值大于 20 的碱基数占总碱基数的百分比

⁴ Clean Reads Q30(%): 过滤后的 reads 中质量值大于 30 的碱基数占总碱基数的百分比

⁵ Clean Reads Ratio(%): 过滤后的 reads 的比例

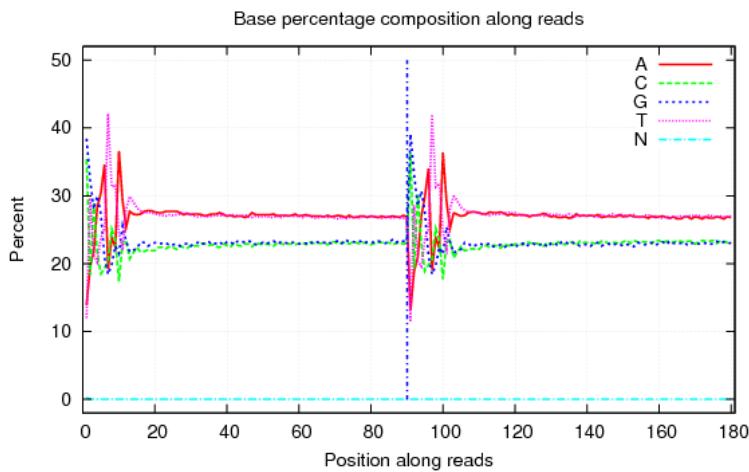


图 1: Clean reads 的碱基含量分布图

X 轴代表碱基在 read 中的位置，Y 轴代表此类碱基的含量比例。正常情况下，reads 每个位置的碱基含量分布稳定，无 AT 或 GC 分离现象。由于 Illumina 平台在 RNA-Seq 测序中，反转录成 cDNA 时所用的 6bp 随机引物会引起前 6 个位置的碱基组成存在偏好性，故图中前 6bp 碱基比例的波动为正常现象。

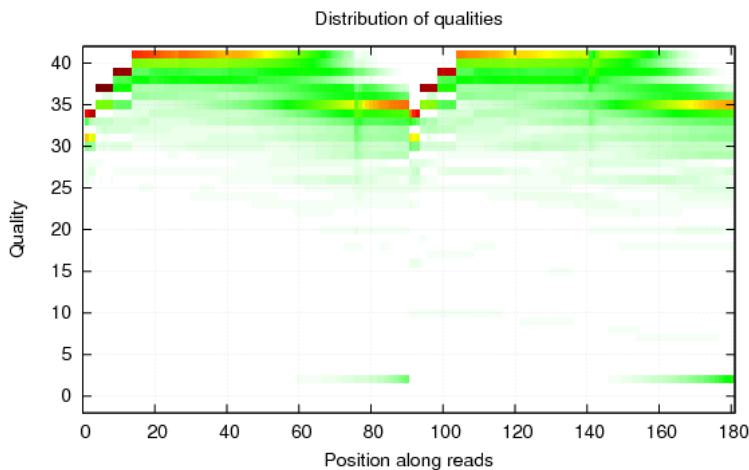


图 2: Clean reads 的碱基质量分布图

X 轴代表碱基在 read 中的位置，Y 轴代表碱基质量值，图中每个点表示此位置达到某一质量值的碱基总数，颜色越深表示数目越多。碱基质量分布反映了测序 reads 的准确性，测序仪、测序试剂、样品质量等均能影响碱基质量。正常情况下，reads 中的前几个碱基质量值不高，是因为反转录时随机引物不能很好地结合 RNA 模板；随着测序长度的增加，高质量碱基的比例会有所提高；但长度达到一定阈值后，由于测序试剂的消耗，高质量碱基的比例会降低。从整体上看，如果低质量 (Quality<20) 的碱基比例较低，说明测序质量较好。

4.3 Denovo 组装

对于无参考基因组的项目，获得 clean reads 后，需要对 clean reads 进行拼接以获取后续分析的参考序列。我们使用 Trinity[1] 对 clean reads 进行组装，将 Trinity 拼接得到的转录本序列，作为后续分析的参考序列。取每条基因中最长的转录本作为 Unigene，以此进行后续的分析。对转录本及 Unigene 的长度分别进行统计，结果见表2，表3，图3和图4 转录组组装的详细结果见文件夹 [./BGI_result/2.Assembly](#) 文件夹

表 2: 拼接长度分布情况一览表

Sample	Min Length	Mean Length	Max Length	N50	N70	N90	Total
Unigenes	251	645	8893	809	481	302	20354233
Transcripts	251	709	8893	963	542	314	25890420

¹ N50/N90 的定义为：将拼接转录本按照长度从长到短排序，累加转录本的长度，到不小于总长 50%/90% 的拼接转录本的长度就是 N50/N90

表 3: 拼接长度频数分布情况一览表

Transcript length interval	200-500bp	500-1kb ¹	1k-2kb	2kb ¹	Total
Number of Unigenes	18933	7769	3575	1275	31552
Number of transcripts	20517	9135	4878	1973	36503

¹ transcripts 表示转录本, unigenes 表示 Unigene

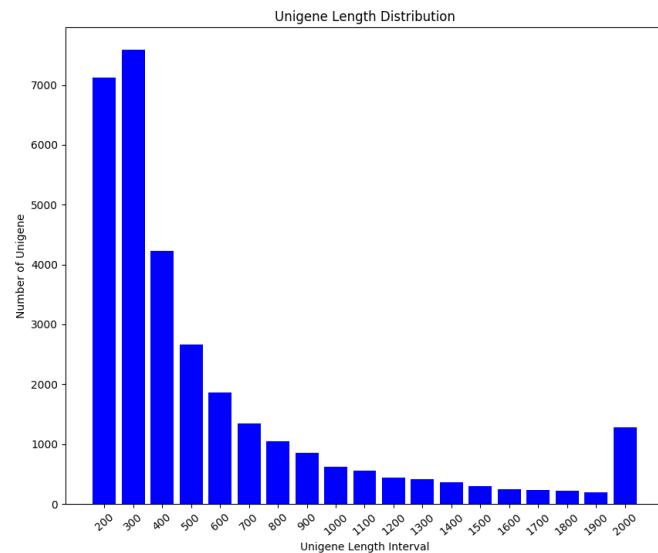


图 3: Unigene 的长度分布图。

X 轴代表 Unigene 长度, Y 轴代表相应 Unigene 的数目。

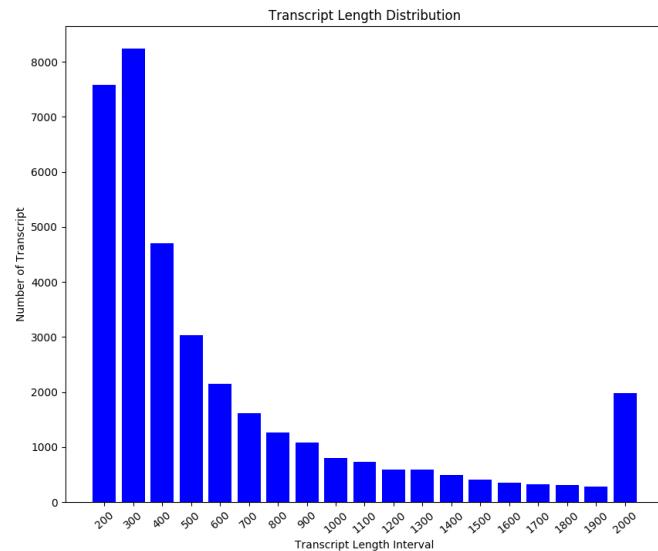


图 4: Transcript 的长度分布图。

X 轴代表 Transcript 长度, Y 轴代表相应 Transcript 的数目。

4.4 Unigene 功能注释

组装完毕后, 我们将对组装得到的 Unigene 进行七大功能数据库注释 (NR、NT、GO、KOG、KEGG、SwissPprot 和 InterPro), 注释结果见表4

表 4: 功能注释结果统计

Values	Total	Nr	Nt	SwissProt	KEGG	KOG	InterPro	GO	Overall
Number	60,740	37,136	32,509	24,806	27,146	14,481	25,746	9,798	41,004
Percentage	100%	61.14%	53.52%	40.84%	44.69%	23.84%	42.39%	16.13%	67.51%

¹ Intersection: 被七大数据库中所有数据库注释上的 Unigene 总数及比例

² Overall: 被七大数据库中任意一个数据库注释上的 Unigene 总数及比例

4.4.1 NT 和 NR 注释

NT 是 NCBI 官方的核酸序列数据库, NR 是官方的蛋白序列数据库, 具有全面、非冗余的特点, 我们将 NT、NR 数据库分为动物、植物、真菌、细菌等几大类, 并将 Unigene 序列注释到相应的分类, 得到相应的功能注释。Unigene 与 NT 注释部分结果展示见表5, Unigene 与 NR 注释部分结果展示见表6, NR, NT 完整注释结果见文件夹 [./BGI_result/3.Annotation](#)

表 5: NT 功能注释部分结果示例

Query_id	Subject_id	Align_length	E_value	Score
CL4308.Contig1_All	gi 57634409 gb AC141866.11	42	2e-06	60.0
CL4308.Contig1_All	gi 147855979 emb AM445472.2	29	8e-06	58.0
CL4308.Contig1_All	gi 147854678 emb AM430380.2	29	8e-06	58.0
CL4308.Contig1_All	gi 147785102 emb AM486977.2	29	8e-06	58.0
CL4308.Contig2_All	gi 57634409 gb AC141866.11	42	1e-06	60.0

¹ Query_id: Unigne 的 ID(查询 ID)

² Subject_id: 数据库中的序列 ID, 以” | ” 分割, 第二个域是 GenBank ID, 第四个域是 NT 数据库中的序列 ID

³ Align_length: 比对上的序列长度

⁴ E_value: 比对的期望值, 值越小代表比对结果越可信

⁵ Score: 比对得分, 值越大代表比对结果越可信

表 6: NR 功能注释部分结果示例

Query_id	Subject_id	Align_length	E_value	Score
CL4309.Contig1_All	gi 568881138 ref XP_006493446.1	91	2.25E-30	135.191
CL4309.Contig1_All	gi 567870155 ref XP_006427699.1	91	3.84E-30	134.42
CL4309.Contig1_All	gi 593700811 ref XP_007150829.1	106	1.12E-29	132.88
CL4309.Contig1_All	gi 118489278 gb ABK96444.1	91	1.46E-29	132.494
CL4309.Contig1_All	gi 225443215 ref XP_002269716.1	91	9.78E-26	119.783

¹ Query_id: Unigne 的 ID(查询 ID)

² Subject_id: 数据库中的序列 ID, 以” | ” 分割, 第二个域是 GenBank ID, 第四个域是 NR 数据库中的序列 ID

³ Align_length: 比对上的序列长度

⁴ E_value: 比对的期望值, 值越小代表比对结果越可信

⁵ Score: 比对得分, 值越大代表比对结果越可信

根据 NR 注释结果, 统计 Unigene 注释上不同物种的比例, 并绘制物种分布图, 见图5

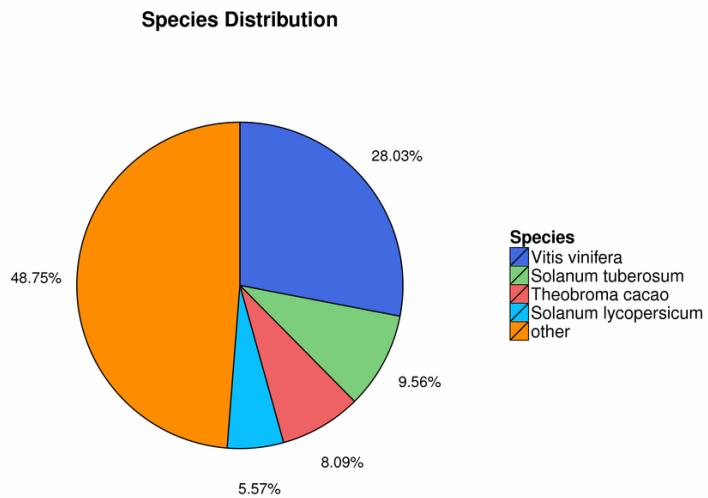


图 5: NR 库注释物种分布图

4.4.2 KOG 注释

我们将 Unigene 注释到 KOG 数据库，并对比对上 KOG 数据库 25 个功能组的 Unigene 进行分类统计，KOG 注释部分结果示例见表7，KOG 功能分布统计图见图6。

表 7: KOG 部分结果示例

Query_id	Subject_id	Align_length	E_value	Score
CL4310.Contig1_All	YAR019c	166	2e-12	70.9
CL4310.Contig1_All	ECU02g0550	148	2e-11	67.8
CL4310.Contig1_All	SA1063_1	179	7e-11	65.9
CL4310.Contig1_All	ECU01g0630	146	2e-10	64.7
CL4310.Contig1_All	CAC1728_1	175	5e-10	63.2

¹ Query_id: Unigene 的 ID(查询 ID)

² Subject_id: 数据库中的序列 ID, 以”|”分割, 第二个域是 GenBank ID, 第四个域是 NR 数据库中的序列 ID

³ Align_length: 比对上的序列长度

⁴ E_value: 比对的期望值, 值越小代表比对结果越可信

⁵ Score: 比对得分, 值越大代表比对结果越可信

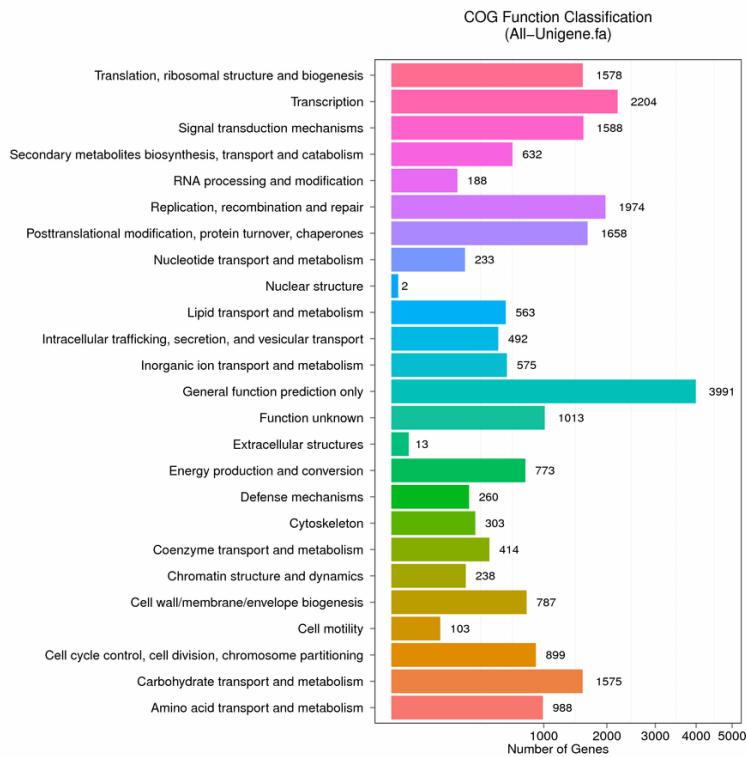


图 6: KOG 功能分布图

X 轴代表相应的 Unigene 数目，Y 轴代表 KOG 功能分类名称。

4.4.3 GO 注释

我们使用 Blast2GO[3] 软件将所有比对上 NR 数据库的 Unigene 结果注释到 GO 数据库，并统计注释到 GO 三个方面生物过程 (Biological Process)、细胞组成 (Cellular Component)、分子功能 (Molecular Function) 的分类图。Unigene GO 注释部分结果示例见表8，GO 功能分布统计图见图7。

GO 完整注释结果见文件夹 [./BGI_result/3.Annotation](#)

表 8: Unigene GO 注释部分结果示例

Ontology	GO term	Number of Genes	Gene members
biological_process	biological adhesion	1,551	CL1000.Contig1_All;...
cellular_component	cell	23,717	CL1.Contig2_All;...
molecular_function	antioxidant activity	82	CL1122.Contig1_All;...

¹ Ontology: GO 的三个大类

² GO term: GO 三个大类的下一级

³ Number of Genes: 注释上某一 GO Term 的基因数

⁴ Gene members: 注释上某一 GO Term 的基因名称, 以“;”分割

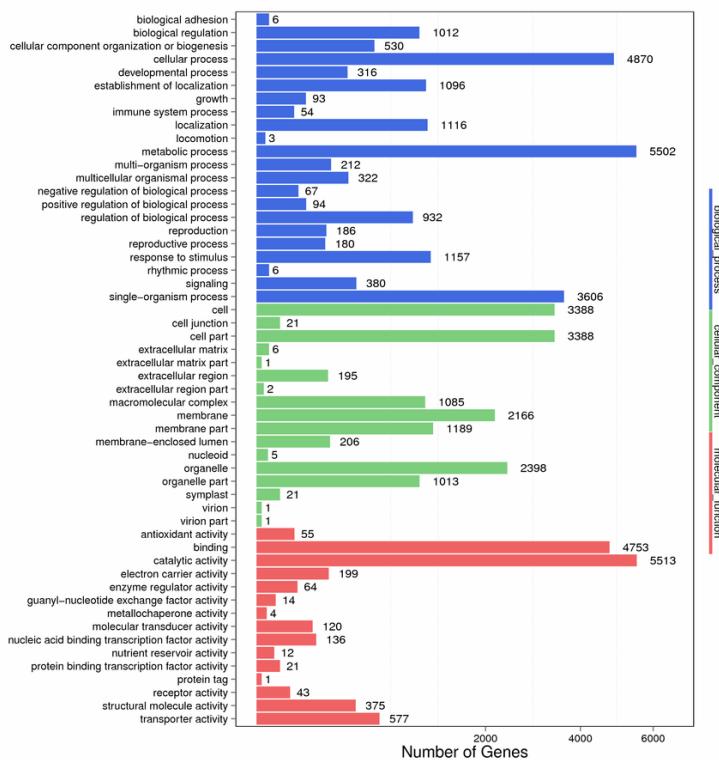


图 7: GO 功能分布图

X 轴代表相应的 Unigene 数目, Y 轴代表 GO 功能分类名称。

4.4.4 KEGG 注释

将所有 Unigene 比对到 KEGG 数据库, 并统计注释上 KEGG 数据库 level1 层级和 level2 层级的 Unigene, 绘制 KEGG 功能分布统计图, 见图8。比对上 KEGG 数据库的部分表格示例见表9, 完整表格下载见表 24。

KEGG 完整注释结果见文件夹 [./BGI_result/3.Annotation](#)

表 9: Unigene KEGG 注释部分结果示例

Pathway_level1	Pathway_level2	Number of Genes	Genes
Cellular Processes	Cell growth and death	776	CL1015.Contig3_All;...
Environmental Information Processing	Membrane transport	142	CL1807.Contig1_All;...
Genetic Information Processing	Folding, sorting and degradation	1,274	CL101.Contig1_All;...
Human Diseases	Cancers: Overview	3,047	CL10.Contig2_All;...

¹ Pathway_level1: KEGG 第一级

² Pathway_level2: KEGG 第二级

³ Number of Genes: 注释 KEGG 的 Unigene 基因个数

⁴ Genes: 注释上 KEGG 对应的基因

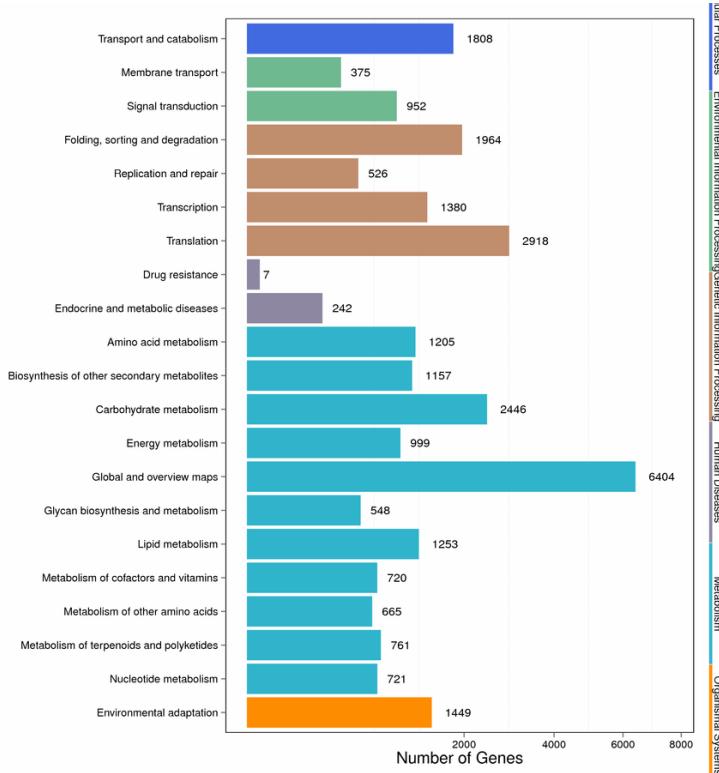


图 8: KEGG 功能分布图

X 轴代表相应的 Unigene 数目，Y 轴代表 KEGG 功能分类。将基因根据参与的 KEGG 代谢通路分为 7 个分支：细胞过程 (CellularProcesses)、环境信息处理 (EnvironmentalInformationProcessing)、遗传信息处理 (GeneticInformation Processing)、人类疾病 (HumanDiseases)（仅限动物）、代谢 (Metabolism)、有机系统 (OrganismalSystems)、药物开发 (DrugDevelopment)。

4.4.5 SwissProt 注释

SwissProt 数据库是检查过的、手工注释的蛋白数据库，我们将 Unigene 注释到 SwissProt 数据库，以得到更加高质量的注释结果，注释结果部分表格见表10。

表 10: Swissprot 功能注释部分结果示例

Query_id	Subject_id	Align_length	E_value	Score
CL4310.Contig1_All	sp O81906 B120_ARATH	209	5e-74	274
CL4310.Contig1_All	sp O81833 SD11_ARATH	197	2e-70	262
CL4310.Contig1_All	sp O64783 Y1137_ARATH	188	6e-70	260
CL4310.Contig1_All	sp Q9LPZ3 Y1141_ARATH	239	3e-69	258
CL4310.Contig1_All	sp Q9ZT07 RKS1_ARATH	199	2e-68	255

¹ Query_id: Unigene 的 ID(查询 ID)

² Subject_id: SwissProt 数据库中的序列 ID

³ Align_length: 比对上的序列长度

⁴ E_value: 比对的期望值，值越小代表比对结果越可信

⁵ Score: 比对得分，值越大代表比对结果越可信

同时，使用维恩图来展示 NR、KOG、KEGG、SwissProt 以及 InterPro 的注释结果，见图9。

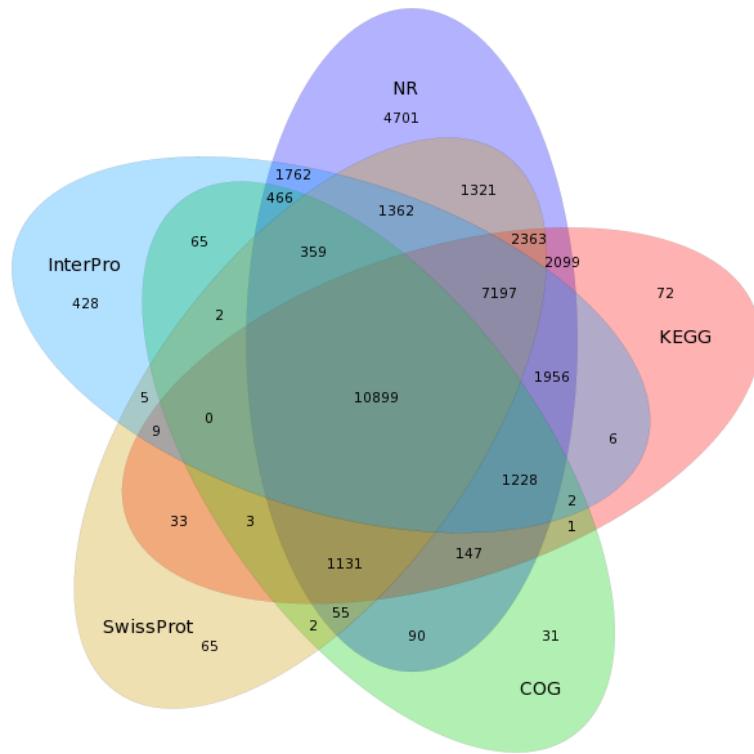


图 9: NR、KOG、KEGG、SwissProt 以及 InterPro 功能注释维恩图

4.5 Unigene 的 CDS 预测

我们使用 TransDecoder 软件识别 Unigene 中的候选编码区域，首先提取最长的开放阅读框，然后通过 Blast 比对 SwissProt 数据库和 Hmmscan 搜索 Pfam 蛋白同源序列，从而预测编码区域。预测结果见表11，预测的 CDS 长度分布见图10

CDS 详细结果见文件夹 [./BGI_result/4.Structure/CDSpredict](#)

表 11: CDS 的质量指标

Software	Total number	Total length	N50	N90	Max length	Min length	Sequence GC
Blast	37003	28537023	771	1143	753	354	46.22
ESTScan	2884	984288	341	342	267	219	49.09
Overall	39887	29521311	740	1110	714	327	46.32

¹ Total number: CDS 数目

² Total length: CDS 的总长度

³ N 50 : 用于衡量序列的连续性, 数值越大说明连续性越好, 计算方法为: 按 CDS 长度从大到小排序后逐个累加至所有 CDS 总长度的 50% 时, 最后一个累加的数值大小即为 N50

⁴ N90: 参考 N50

⁵ Max length: 最大长度

⁶ Min length: 最小长度

⁷ GC (%): 碱基 G 和 C 的比例

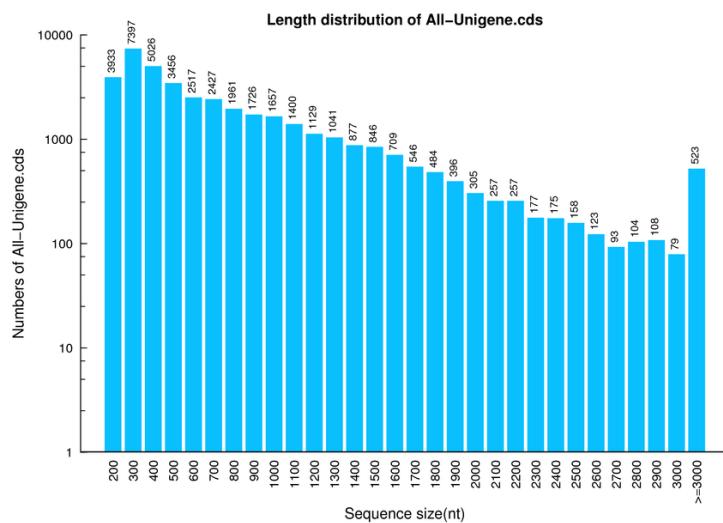


图 10: CDS 长度分布图

X 轴代表 CDS 长度, Y 轴代表相应 CDS 的数目。

4.6 Unigene 的 SSR 检测

根据组装结果, 我们对 Unigene 的 SSR 进行检测, 同时为每个 SSR 设计引物。SSR 长度特征见表12和图11。引物设计结果见表13。

SSR 详细结果见文件夹 [./BGI_result/4.Structure/SSR](#)

表 12: SSR 长度统计

Number	Mono-nucleotide	Di-nucleotide	Tri-nucleotide	Quad-nucleotide	Penta-nucleotide	Hexa-nucleotide
4	0	0	0	0	133	358
5	0	0	1,658	88	25	40
6	0	3,005	652	18	3	33
7	0	1,954	358	4	1	4
8	0	1,457	206	2	4	4

¹ Number: SSR 数目

² Mono-nucleotide: 单碱基重复的 SSR 数目

³ Di-nucleotide: 二碱基重复的 SSR 数目

⁴ Tri-nucleotide: 三碱基重复的 SSR 数目

⁵ Quad-nucleotide: 四碱基重复的 SSR 数目

⁶ Penta-nucleotide: 五碱基重复的 SSR 数目

⁷ Hexa-nucleotide: 六碱基重复的 SSR 数目

表 13: SSR 引物设计表部分结果示例

Unigene_ID	FORWARD PRIMER(5'-3')	Tm	REVERSE PRIMER(5'-3')	Tm
CL1010.Contig1_All_544_1	CCTTTTAGATGATCTCCCCAGTT	59.854	AAGTCGGAGTCATTCTCTTCCT	59.780
CL1010.Contig1_All_544_2	TCCCCTCTTCTCTAGCTACCAT	59.769	CAAAGATGAAGATGACGAACACA	60.161
CL1010.Contig1_All_544_3	CCTTTAGATGATCTCCCCAGTT	59.854	AGTCGGAGTCATTCTCTTCCTC	60.253
CL1010.Contig1_All_544_5	CCTTTAGATGATCTCCCCAGTT	59.854	ACGAACACAGATAAGTCGGAGTC	59.696
CL1015.Contig1_All_548_1	TCTTATCTCCTCTTCCACGTTTG	59.770	ACCACAGAGACTTATTGGAGCAT	59.183
CL1015.Contig1_All_548_2	TCTTATCTCCTCTTCCACGTTTG	59.770	TACCACAGAGACTTATTGGAGCA	58.917

¹ Unigene_ID: 基因 ID

² FORWARD PRIMER(5' '-3'): 前端引物序列

³ Tm: 前端引物退火温度

⁴ FORWARD PRIMER(5' '-3'): 后端引物序列

⁵ Tm: 后端引物退火温度

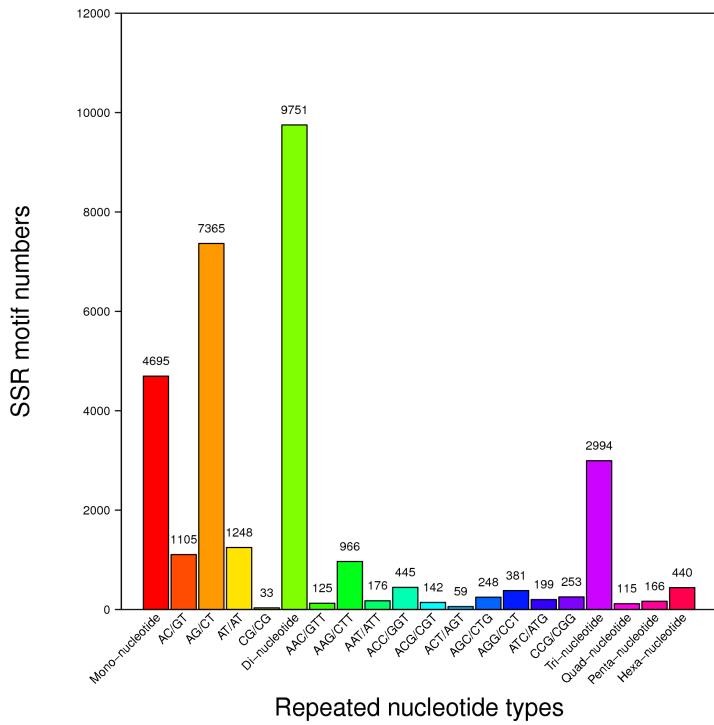


图 11: SSR 长度分布图

X 轴代表 SSR 类型，Y 轴代表相应的 SSR 数目

4.7 SNP 检测

根据组装结果，我们使用 GATK [4] 对每个样品进行 SNP 检测，结果存储为以 VCF 格式。SNP 检测结果见表14和图12。

SNP 详细结果见文件夹 [./BGI_result/4.Structure/SNP](#)

表 14: SNP variant type summary

Sample	A-G	C-T	Transition	A-C	A-T	C-G	G-T	Transversion	Total
EULEf	14499	14258	28757	3695	3901	3379	3860	14835	43592
EUSki	15547	15223	30770	3935	4047	3500	4054	15536	46306

¹ Sample: 样品名

² A-G: A-G 变异 (包括 A->G、G->A) 的 SNP 数目

³ C-T: C-T 变异的 SNP 数目

⁴ Transition: A-G 和 C-T 变异的 SNP 数目, 嘧呤和嘌呤之间的替换, 或嘧啶和嘧啶之间的替换。

⁵ A-C: A-C 变异的 SNP 数目

⁶ A-T: A-T 变异的 SNP 数目

⁷ C-G: C-G 变异的 SNP 数目

⁸ G-T: G-T 变异的 SNP 数目

⁹ Transversion: A-C,A-T,C-G 和 G-T 变异的 SNP 数目, 嘙呤和嘧啶之间的替换

¹⁰ Total: 所有变异类型的总数目

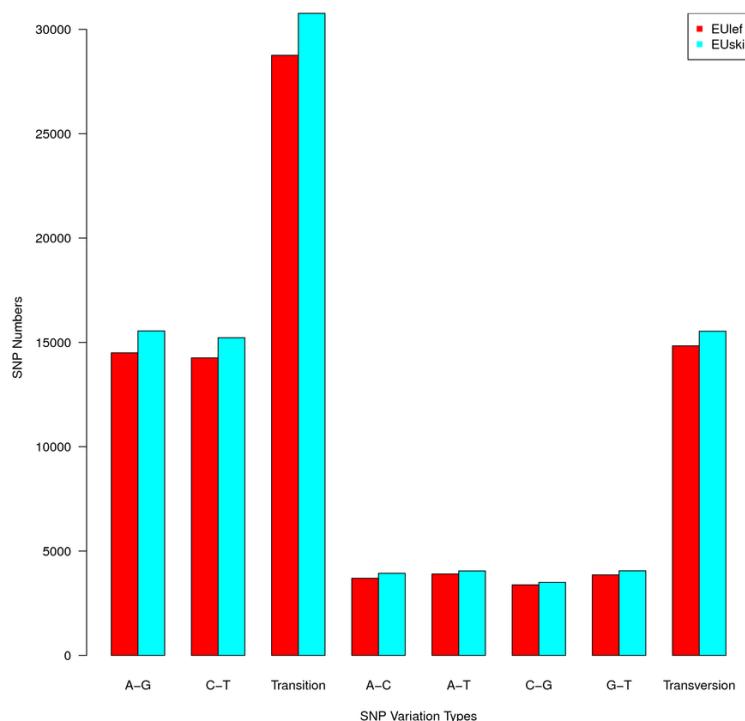


图 12: SNP 变异类型分布

X 轴代表变异类型，Y 轴代表相应的 SNP 数目

4.8 Unigene 的 TF 编码能力预测

做动物、植物研究的时候，根据组装结果，我们对具有编码转录因子（TF）能力的 Unigene 进行预测。我们的项目中，具有编码转录因子能力的 Unigene 见表，同时对预测的转录因子进行家族分类，见图13。

TF 详细结果见文件夹 [./BGI_result/4.Structure/TFpredict](#)

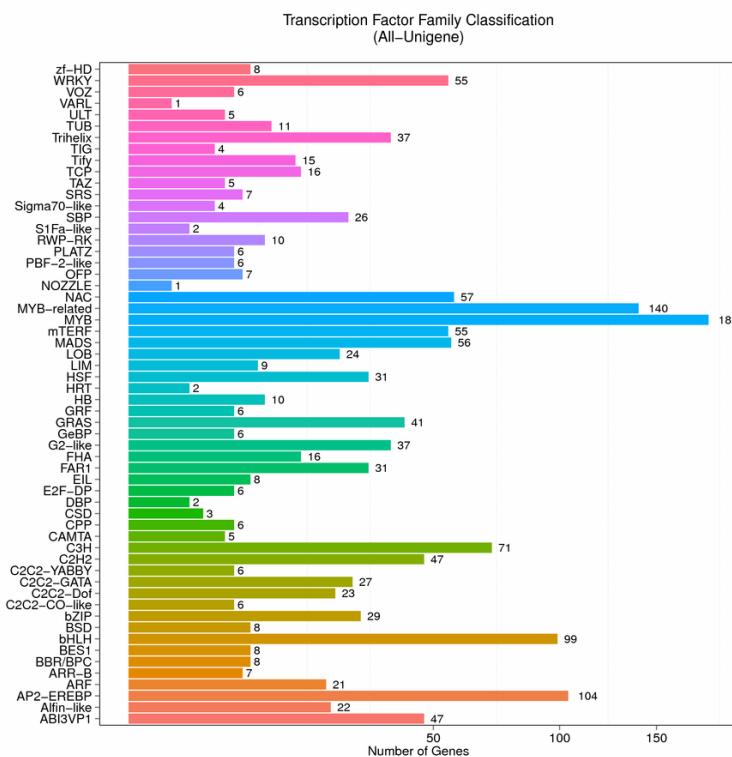


图 13: 转录因子家族分类

X 轴代表相应的 Unigene 数目，Y 轴代表转录因子家族分类。

4.9 基因表达量计算

4.9.1 基因表达量水平

根据组装结果，我们使用 Bowtie2 软件把每个样本的 clean reads 比对到 Unigene，之后使用 RSEM 计算每个样品的基因表达水平。比对结果统计表见表16，基因表达水平统计表部分结果见表15

表达量详细结果见文件夹 [./BGI_result/5.Quantify/GeneExpression](#)

表 15: 基因表达水平统计表部分结果示例

gene_id	length	expected_count	FPKM
CL1.Contig1_All	653.00	29.01	6.40
CL1.Contig2_All	450.00	6.27	2.20
CL10.Contig1_All	434.00	2.50	0.92
CL10.Contig2_All	438.00	2.50	0.91
CL100.Contig1_All	827.00	53.00	8.87

¹ gene_id: Unigene ID

² length: 基因长度

³ expected_count: 基因对应的 count 值

⁴ FPKM: 基因对应的 FPKM 值

表 16: 比对结果统计表

Sample	Total Bases	Total Reads	Total Mapped Reads	Unique Mapped Reads
EUlef	3488310540	38759006	31255650(80.64%)	24294886(62.68%)
EUski	5278771080	58653012	48502970(82.69%)	38759086(66.08%)

¹ Sample: 样品名

² Total Bases: 样品对应的碱基数

³ Total Reads: 样品对应的 Reads 数

⁴ Total Mapped Reads: 所有比对上参考序列的 Reads 数

⁵ Unique Mapped Reads: 比对到参考序列唯一位置的 reads 数

4.9.2 样品中基因表达量的分布

根据表达量信息，本项目采取箱线图展示各样品基因表达水平的分布情况，可以观察到数据分布的分散程度，如图14所示。密度图能够展示样品中基因丰度随着表达量变化的趋势，可以清晰地反映样本中基因表达量集中的区间，如图15所示。为了更直观地展示每个样品在不同 FPKM 区间的基因数目，我们对 FPKM (FPKM<=1、FPKM>10、FPKM>=10) 的三种情况进行了基因数目的统计，如图16所示。

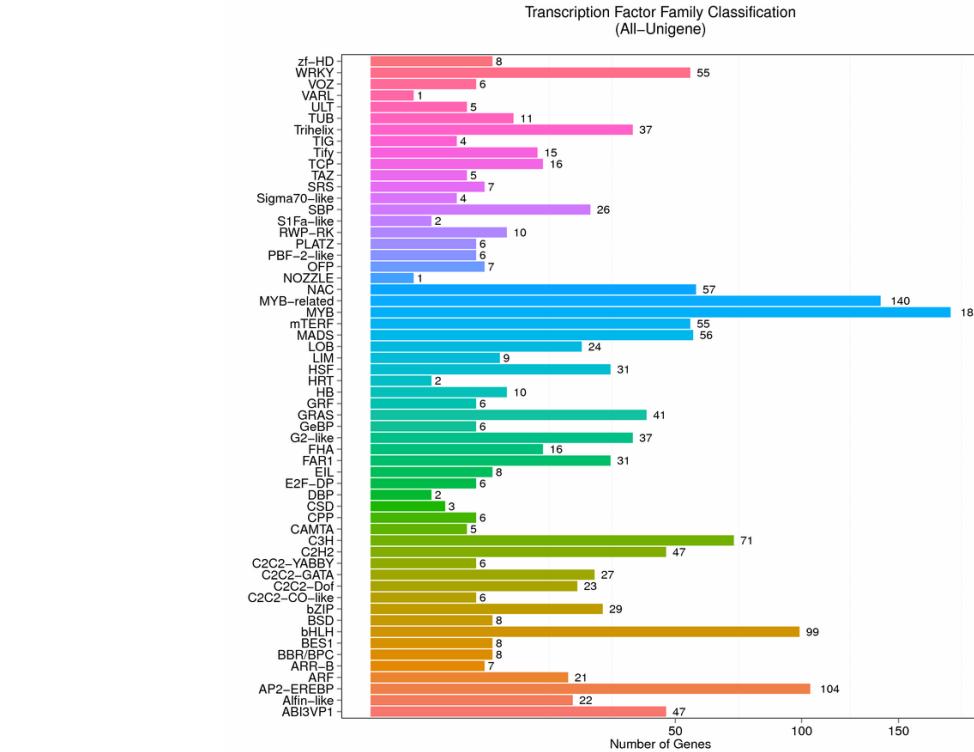


图 14: 表达量箱线图

X 轴为样品名称，Y 轴为 log₁₀FPKM，每个区域的箱线图对应五个统计量（自上而下分别为最大值，上四分位数，中值，下四分位数和最小值）。

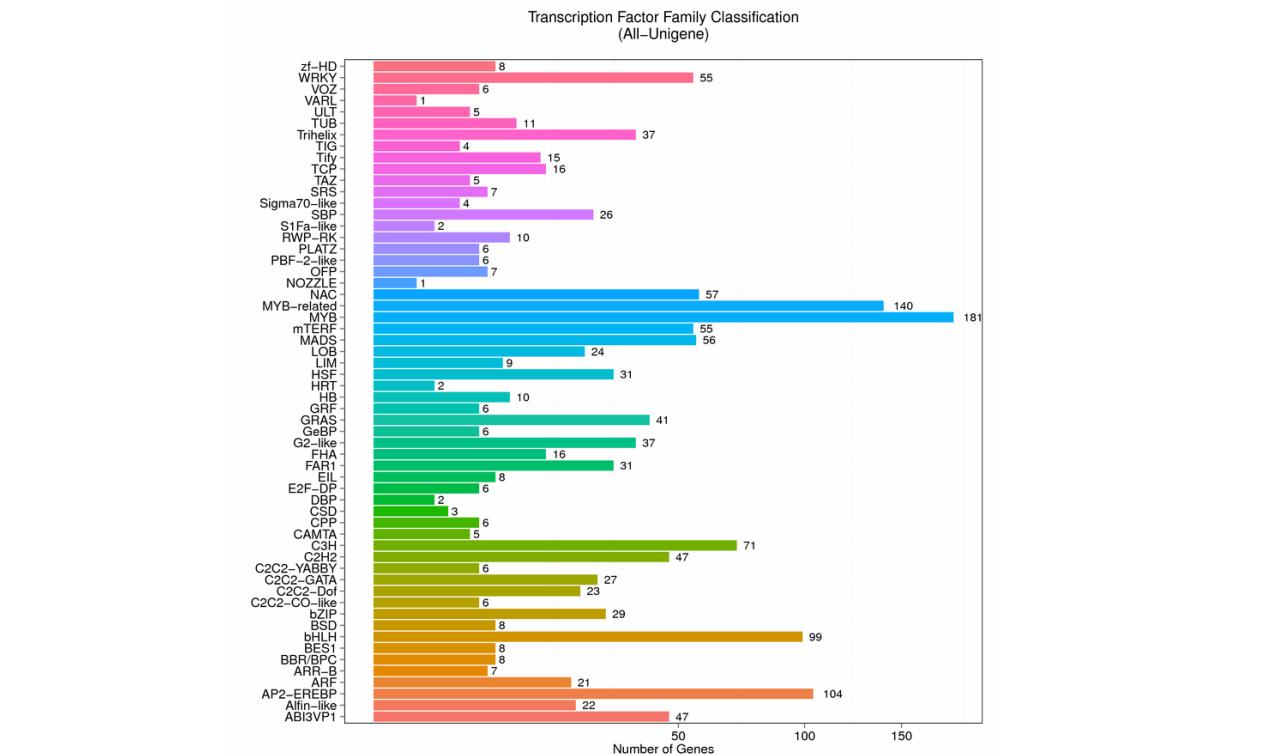


图 15: 表达量密度图

X 轴为 log₁₀FPKM，Y 轴为基因的密度，即该表达量下的基因数与表达基因总数的比例

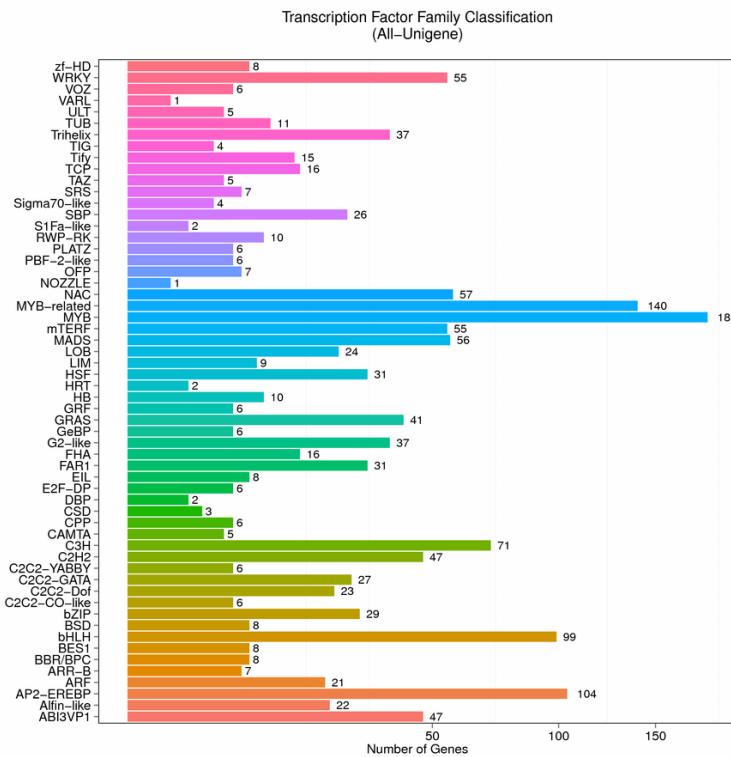


图 16: 基因表达量分布图

X 轴表示样品名称，Y 轴表示基因数目，颜色深浅表示不同表达量水平：FPKM \leq 1 的为极低表达水平的基因，FPKM 在 1 ~ 10 之间的为较低表达水平的基因，FPKM \geq 10 为中高表达水平的基因

4.9.3 PCA 分析

主成分分析（PCA）是将多个变量通过降维为少数几个相互独立的变量（即主成分），同时尽可能多地保留原始数据信息的一种多元统计分析方法。在转录组的分析中，PCA 将样本所包含的大量基因表达量信息降维为少数几个互相无关的主成分，以进行样本间的比较，方便找出离群样品、判别相似性高的样品簇等。本项目的 PCA 分析结果见图17。

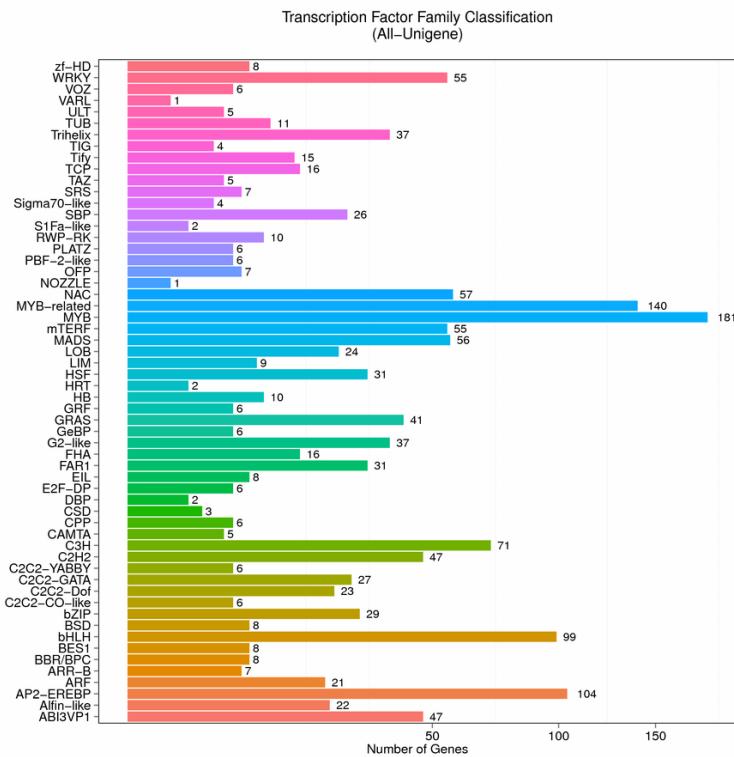


图 17: PCA 分析

X、Y 轴表示样品表达量经过降维处理后得到的对应主成分的新的数据集，用来表示样品之间的差距；坐标轴标签括号中的数值代表对应主成分解释总体方差的百分比。点代表每个样品，同一个颜色代表同一个样品组。

4.10 时间序列分析

样品在不同时间阶段一些基因会有相似的表达模式，根据基因的表达量信息，可以聚类成与时间相关联的基因簇，表达模式一致的基因会被聚到同一个簇，目前也有文章用该方法分析样品间的组织特异性。图18显示的是基因聚类成各种基因簇的情况，同时下载文件中也有每个簇的单个大图供研究人员查看，每个基因簇具有的基因详见下表17：

表达量详细结果见文件夹 [./BGI_result/5.Quantify/GeneExpression](#)

表 17: 基因表达水平统计表部分结果示例

gene_id	cluster	UHRR1_fpkm	UHRR2_fpkm	...
CL1006.Contig1_All	1	0	0	...
CL100.Contig3_All	1	0	0	...
CL1014.Contig3_All	1	118.28	115.68	...
CL1015.Contig4_All	1	3.09	3.27	...

¹ gene_id: Unigene ID

² cluster: 基因所对应的基因簇

³ Sample1_fpkm: 各样品基因对应的 FPKM 值

⁴ Sample2_fpkm: 各样品基因对应的 FPKM 值

⁵ ...: 其他各个样品基因对应的 FPKM 值

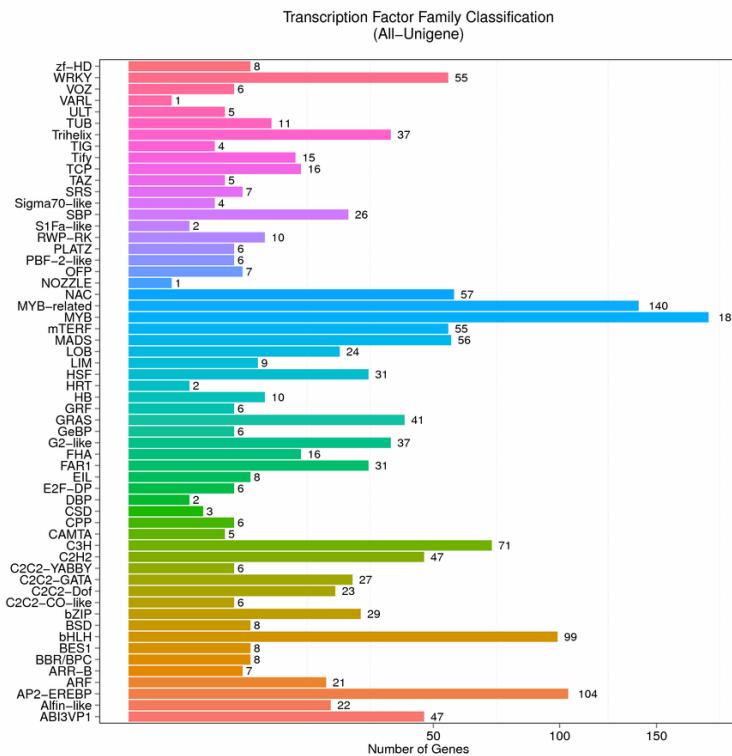


图 18: 时间序列分析的 Mfuzz 图

X 轴代表各个时间点, Y 轴代表均一化后的表达量值

4.11 差异表达基因检测

根据各个样品基因表达水平结果，我们可以检测样品（或者样品组）之间的差异表达基因（DEG），本项目使用 DEGseq、DEseq2、EBseq、NOIseq 和 PossionDis 算法进行 DEG 检测，差异表达结果如下：

差异表达详细结果见文件夹 [./BGI_result/5.Quantify/DifferentExpressedGene](#)

表 18: 差异表达部分结果示例

GeneID	HBRR2-Expression	UHRR2-Expression	log2FoldChange(UHRR2/HBRR2)	FDR	Pvalue
Unigene28509_All	0.01	6,462.36	19.30	0.00e+00	0.00e+00
CL1804.Contig1_All	0.01	524.53	15.68	0.00e+00	0.00e+00
Unigene35357_All	0.01	453.05	15.47	0.00e+00	0.00e+00
Unigene20863_All	0.01	426.05	15.38	0.00e+00	0.00e+00

¹ GeneID: Unigene ID

² Sample1-Expression: Sample1 表达量

³ Sample2-Expression: Sample2 表达量

⁴ log2FoldChange(样品 2/样品 1): 经过 log2 转换后的样品（组）间的差异表达倍数

⁵ FDR: FDR 校正后的统计值

⁶ Pvalue: 显著性统计值

DEG 检测结果见图19。同时使用 MA plot、Scatter plot、heatmap plot 和 Volcano plot 展示 DEG 的分布，见图20，图21，图22和图23

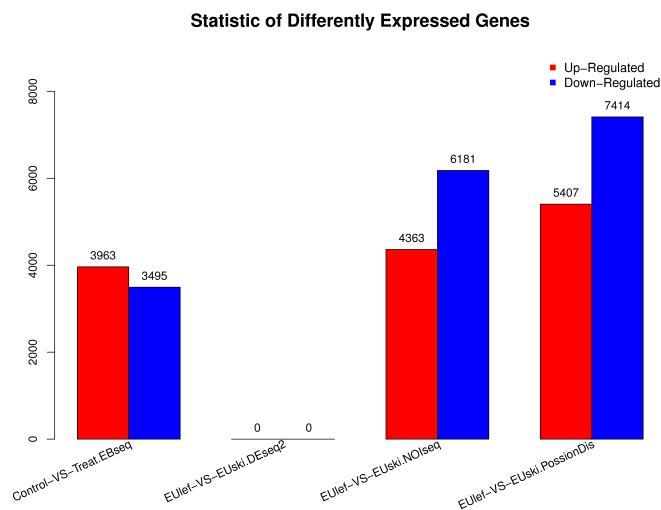


图 19: DEG 数量统计图

X 轴代表每组差异比对方案，Y 轴代表相应的 DEG 数目。红色代表上调的 DEG 数目，蓝色代表下调的 DEG 数目

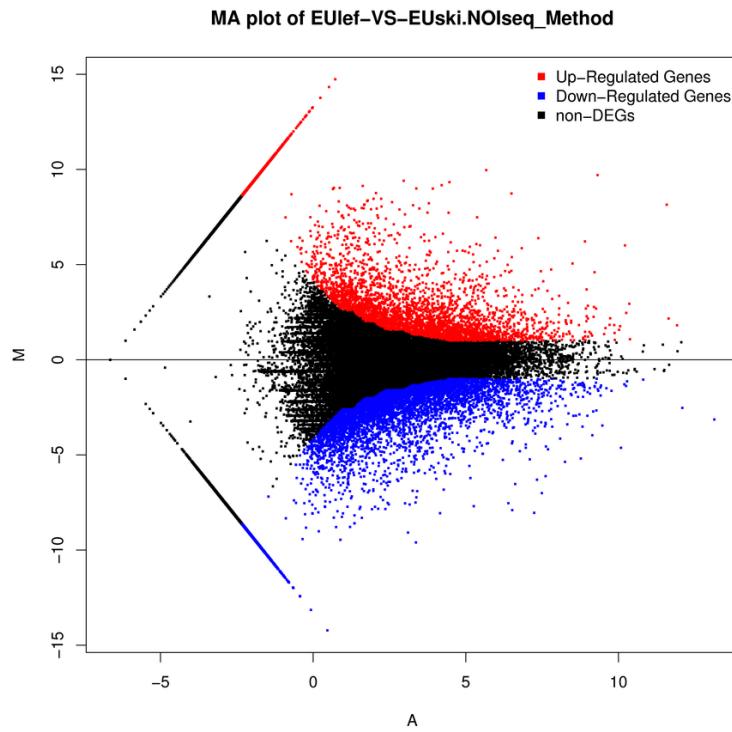


图 20: DEG 的 MA-plot 分布图

X 轴代表 A 值 (\log_2 转换后的平均表达水平), Y 轴代表 M 值 (\log_2 转换后的差异倍数)。红色代表上调的 DEG, 蓝色代表下调的 DEG, 灰色代表非 DEG。

MA plot of EUlef-VS-EUski.NOseq_Method

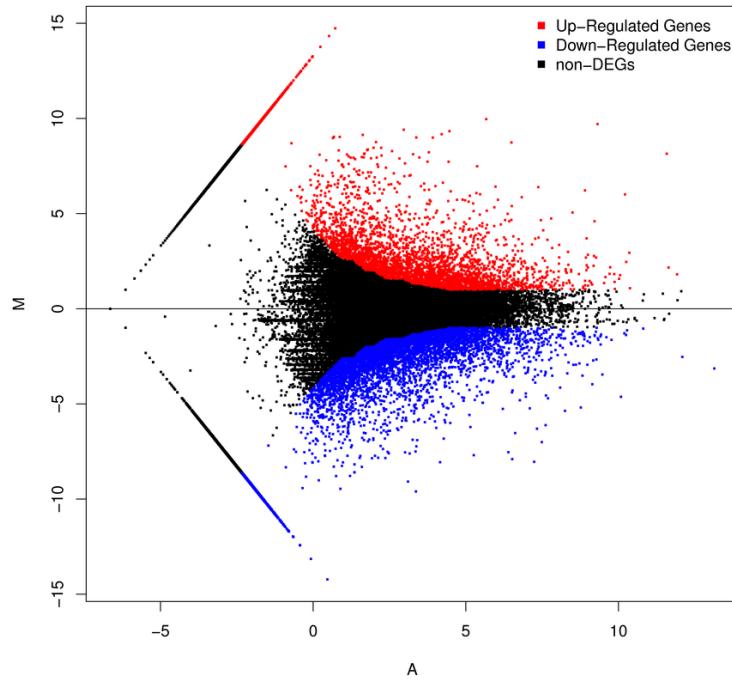


图 21: 差异对中所有表达基因散点图

X、Y 坐标轴均取基因表达量的对数值，蓝色表示下调基因，红色表示上调基因，灰色则是非显著差异基因

MA plot of EUlef-VS-EUski.NOseq_Method

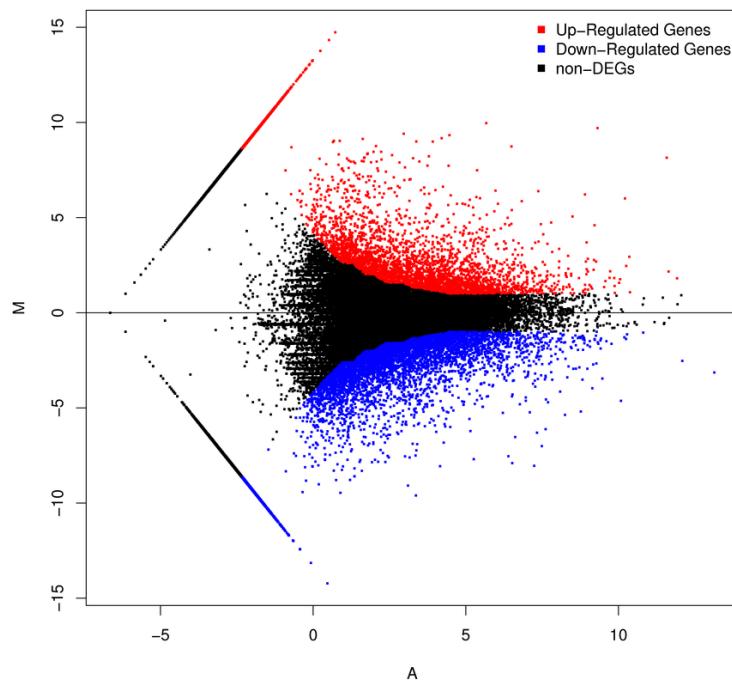


图 22: 差异基因表达量热图

X 轴表示不同样品，Y 轴表示差异基因，颜色越深表示表达量越高，越浅表示表达量越低

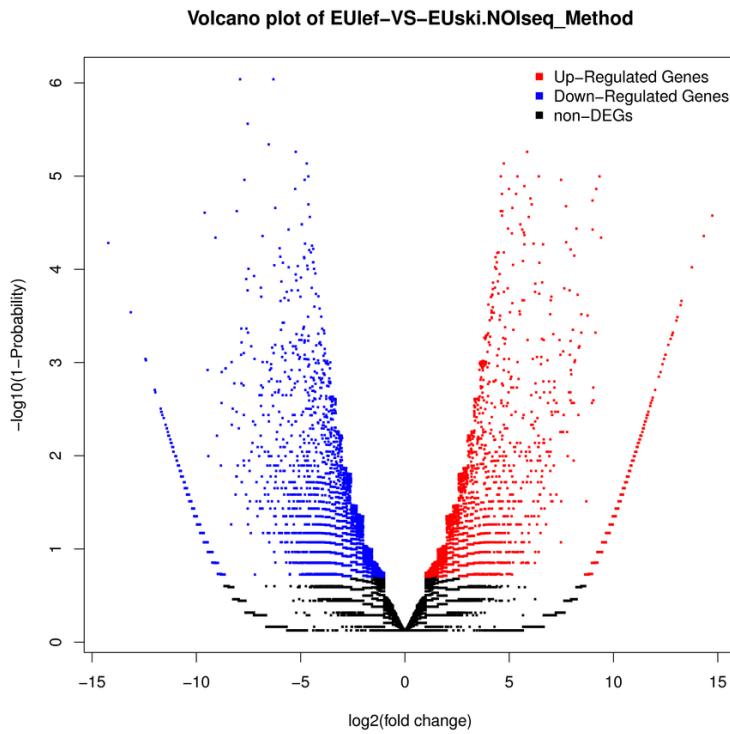


图 23: DEG 的 Volcano-plot 分布图

X 轴代表 log2 转换后的差异倍数值，Y 轴代表-log10 转换后的显著性值。红色代表上调的 DEG，蓝色代表下调的 DEG，黑色代表非 DEG。

4.12 差异表达基因 GO 功能分析

根据差异基因检测结果，我们对其 Gene Ontology (GO) 功能进行分类以及富集分析。GO 分为分子功能 (Molecular Function)、细胞组分 (Cellular Component) 和生物过程 (Biological Process) 三大功能类，我们将对三大功能类进一步分类以及进行富集分析。GO 功能分类结果见图24，GO 功能对应的差异表达基因上下调统计见图25，GO 功能富集结果见图26。

GO 富集分析详细结果见文件夹

[./BGI_result/5.Quantify/DifferentExpressedGene/Functional_Enrichment/GO](#)

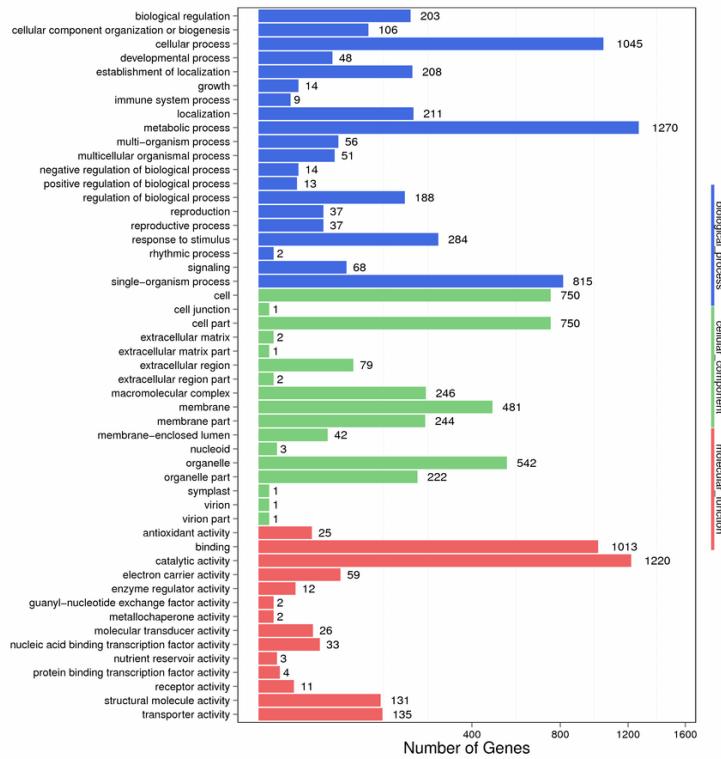


图 24: 差异基因 GO 功能分类图

X 轴代表 DEG 数目, Y 轴代表 GO 功能分类。

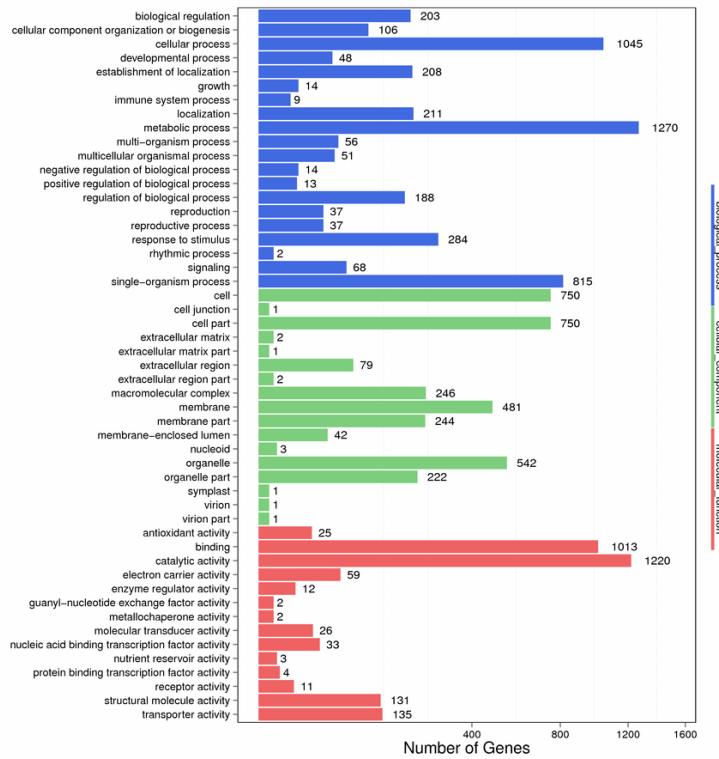


图 25: 差异基因上下调 GO 功能分类图

X 轴代表 GO 功能分类,Y 轴表示对应 GO Term 上下调基因数目

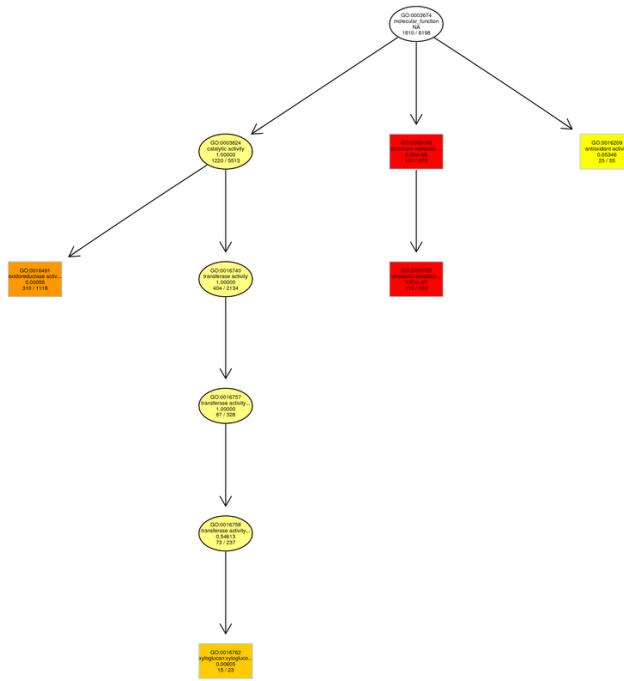


图 26: 差异基因 GO 功能富集结果

有向无环图 (DirectedAcyclicGraph, DAG) 为差异基因 GO 富集分析结果的图形化展示方式。图中，分支代表包含关系，从上至下所定义的功能范围越来越小，方框代表的是每个分类富集程度排名前 5 的 GO Term，并通过包含关系，将相关联的 GO Term 一起展示。每个节点上展示了该 Term 的名称及富集分析校正后的 p-value，颜色越深（红）表示 pvalue 越小、富集程度越高。

4.13 差异表达基因 Pathway 功能分析

根据差异基因检测结果，我们对其进行 KEGG 生物通路分类以及富集分析。通路分类结果见图27，通路富集结果见图29。富集通路对应的差异表达基因上下调见图28。

KEGG 富集分析详细结果见文件夹

[./BGI_result/5.Quantify/DifferentExpressedGene/Functional_Enrichment/Pathway](#)

表 19: 差异表达部分结果示例

#Pathway	HBRR2-VSUHRR2.PossionDis_Method(10597)	All-Unigene(27624)	Pvalue	Qvalue	Pathway ID
Cell adhesion molecules (CAMs)	0.01	6,462.36	19.30	0.00e+00	0.00e+00
Ribosome	0.01	524.53	15.68	0.00e+00	0.00e+00
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.01	453.05	15.47	0.00e+00	0.00e+00
ECM-receptor interaction	0.01	426.05	15.38	0.00e+00	0.00e+00

¹ #Pathway: 代谢通路名称

² group_Method: 注释上某一通路的差异基因个数

³ All-Unigene: 注释上某一通路的 Unigene 个数

⁴ Pvalue: 显著性统计值

⁵ Qvalue: 校正后的 pvalue。qvalue 越小, 表示差异越显著

⁶ Pathway ID: 代谢通路 ID 号

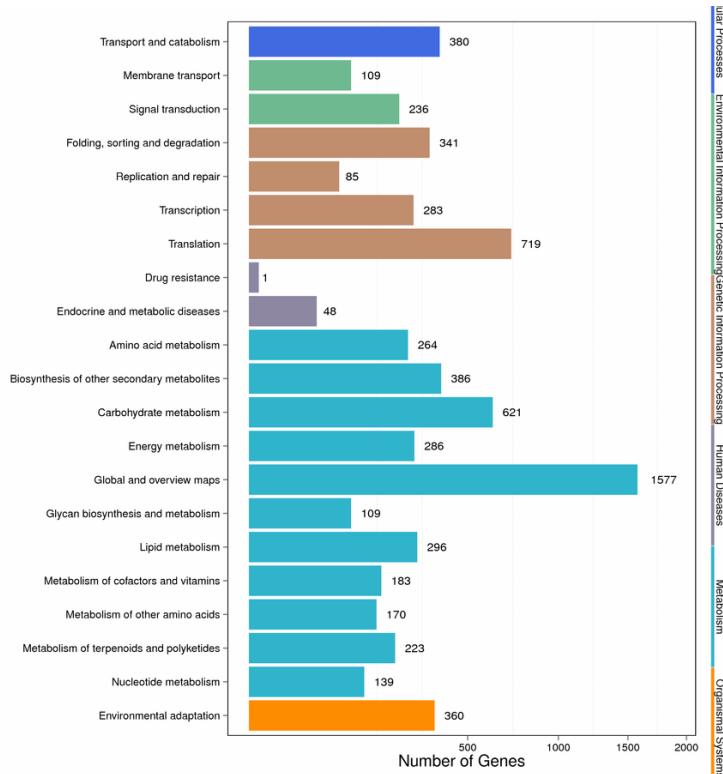


图 27: 差异基因 Pathway 分类图

X 轴代表相应的 Unigene 数目, Y 轴代表 KEGG 功能分类。将基因根据参与的 KEGG 代谢通路分为 7 个分支: 细胞过程 (CellularProcesses)、环境信息处理 (EnvironmentalInformationProcessing)、遗传信息处理 (GeneticInformation Processing)、人类疾病 (HumanDiseases) (仅限动物)、代谢 (Metabolism)、有机系统 (OrganismalSystems)、药物开发 (DrugDevelopment)。

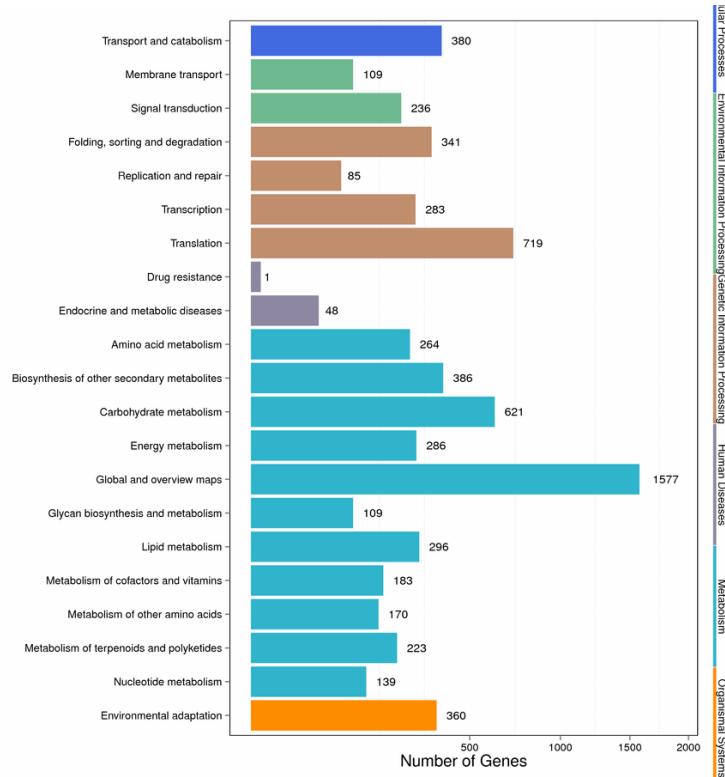


图 28: 差异基因 Pathway 富集结果气泡图

X 轴代表富集因子值，Y 轴代表通路名称。颜色代表 qvalue(颜色越白值越大，越蓝值越小)，值越小代表富集结果越显著。点的大小代表 DEG 数目 (点越大代表数目越大，越小代表数目越少)。RichFactor 指的是富集因子值，是注释上某一通路的前景值 (差异基因个数)、与注释上某一通路的背景值 (所有基因个数) 之商，数据越大，说明富集结果越明显

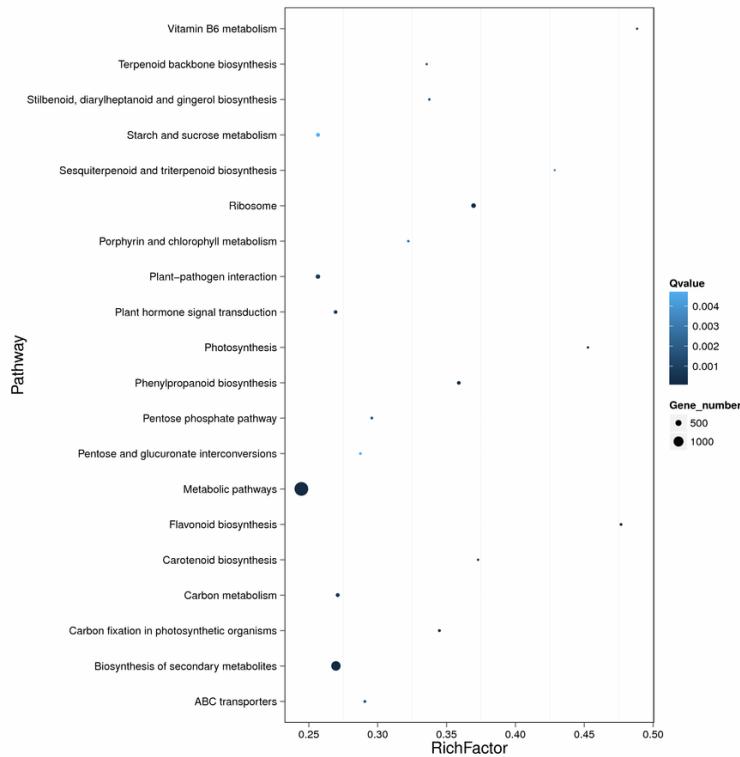


图 29: 富集通路差异基因上下调

X 轴表示 Pathway 条目, Y 轴表示对应 Pathway 条目上下调基因数目。

4.14 差异基因蛋白互作分析

我们使用 STRING[5] 蛋白互作数据库, 对每组差异表达基因进行蛋白互作分析, 构建互作网络图, 我们抽取可信度最高的前 100 个关系进行画图, 蛋白互作关系部分结果见表20蛋白互作网络见图30。

PPI 分析详细结果见文件夹

[./BGI_result/5.Quantify/DifferentExpressedGene/PPI](#)

表 20: 差异表达基因蛋白互作关系部分结果示例

#Pathway	HBR2-VSUHRR2.PossionDis_Method(10597)	All-Unigene(27624)	Pvalue	Qvalue	Pathway ID
Cell adhesion molecules (CAMs)	0.01	6,462.36	19.30	0.00e+00	0.00e+00
Ribosome	0.01	524.53	15.68	0.00e+00	0.00e+00
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.01	453.05	15.47	0.00e+00	0.00e+00
ECM-receptor interaction	0.01	426.05	15.38	0.00e+00	0.00e+00

¹ #Pathway: 代谢通路名称

² group_Method: 注释上某一通路的差异基因个数

³ All-Unigene: 注释上某一通路的 Unigene 个数

⁴ Pvalue: 显著性统计值

⁵ Qvalue: 校正后的 pvalue, qvalue 越小, 表示差异越显著

⁶ Pathway ID: 代谢通路 ID 号



图 30: 蛋白互作网络图

图中红色表示上调，蓝色表示下调，圆圈大小表示相互作用的关系的个数，圆圈越大表示关系越密集

4.15 植物抗病基因预测

根据 PRG 数据库，对所有 Unigene 基因进行植物抗病基因分析，列表如下：

PRG 分析详细结果见文件夹 [./BGI_result/5.Quantify/DifferentExpressedGene/PRG](#)

表 21: 植物抗病基因的注释部分结果示例

gene_id	PRGID	Name	Species	Class	GenBank_ID	GenBank_Locus
Unigene43208_All	PRGDB00078803	Vvi.12916	Vitis vinifera	N	225452756	XM_002283015
Unigene23022_All	PRGDB00078803	Vvi.12916	Vitis vinifera	N	225452756	XM_002283015
Unigene27073_All	PRGDB00078803	Vvi.12916	Vitis vinifera	N	225452756	XM_002283015
Unigene24071_All	PRGDB00194209	Pin.5041	Phytophthora infestans	N	301091447	XM_002895863

¹ gene_id: Unigene ID

² PRGID: PRG 数据库的登录号

³ Name: 比对上的基因名称

⁴ Species: 物种名

⁵ Class: 抗病基因结构域的类型

⁶ GenBank_ID: GenBank 数据库 ID

⁷ GenBank_Locus: GenBank 序列名称

4.16 真菌致病基因预测

根据 PHI 数据库，对所有 Unigene 基因进行真菌致病基因分析，列表如下：

PHI 分析详细结果见文件夹 [./BGI_result/5.Quantify/DifferentExpressedGene/PHI](#)

表 22: 真菌致病基因注释部分结果示例

gene_id	Identity	Query_coverage	Function
CL1021.Contig2_All	62.07	48.6	PHI:3945 ...
CL1140.Contig1_All	65.67	54.9	PHI:1658 ...
CL1140.Contig2_All	65.67	55.3	PHI:1658 ...
CL1167.Contig1_All	59.07	47.1	PHI:3594 ...

¹ gene_id: Unigene ID

² Identity: 比对相似度

³ Query_coverage: 比对序列的覆盖度

⁴ Function: PHI 数据库功能注释

5 报告补充说明

5.1 文件目录链接

为方便您查看项目结果文件，华大基因为您提供了结果文件的目录列表，项目结果文件可通过此目录列表查看，点击目录列表即可弹出相应结果文件夹（注：请保证 mRNA_denovo_report.pdf 文件夹和 BGI_result 文件夹在同一文件夹下）。

5.2 结果文件

结果文件解析说明：

文件格式	电脑系统	查看程序
*.tar.gz 形式的压缩文件	windows 用户	使用解压缩软件如 WinRAR、7-Zip 等
.gz 形式的压缩文件	windows 用户	使用解压缩软件如 WinRAR、7-Zip 等
.zip 形式的压缩文件	windows 用户	使用解压缩软件如 WinRAR、7-Zip 等

结果文件查看说明；

文件格式	文件说明	查看程序
*.fasta	序列文件， fasta 格式，一般为基因序列或者基因组序列。因文件一般较大，打开较为困难	windows 使用高级文本编辑器 PilotEdit, Notepad++ 等查看。
.fq/fastq	序列文件， fastq 格式，一般为 reads 序列；因文件一般较大，打开较为困难。	windows 使用高级文本编辑器 PilotEdit, Notepad++ 等查看。
.xls, *.txt	结果数据表格文件，文件以制表符（Tab）分隔	用 Microsoft Excel 或文本编辑器打开
.png, *.jpg	结果图像文件	用户可以使用图片浏览器打开，如 photo-shop 等
.pdf	结果图像文件，矢量图，可以放大和缩小而不失真，方便用户查看和编辑处理	用户可以使用 Adobe Reader 或福昕阅读器等打开

6 参考文献

- [1] Cock P., et al.(2010). The SangerFASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38(6): 1767-1771.
- [2] Altschul SF, et al.(1990).Basic local alignment search tool.J Mol Biol. 1990 Oct 5;215(3):403-10.
- [3] Conesa A, et al.(2005).Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.Bioinformatics. 2005 Sep 15;21(18):3674-6.
- [4] Quevillon E, et al.(2005).InterProScan: protein domains identifier.Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W116-20.
- [5] Iseli C, et al.(1999).ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.Proc Int Conf Intell Syst Mol Biol. 1999:138-48.
- [6] Thiel T, et al.(2003).Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare L.*).Theor Appl Genet. 2003 Feb;106(3):411-22.
- [7] UntergrasserA, et al.(2012).Primer3 -newcapabilities and interfaces.Nucl. Acids Res. (2012)40 (15): e115.
- [8] Kim D, et al.(2015).HISAT: a fast spliced alignerwith lowmemory requirements. Nature Methods 2015.
- [9] McKenna N, et al.(2010).The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.Genome Res. 2010 Sep;20(9):1297-303.
- [10] Langmead B, et al.(2012).Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.
- [11] Li B, et al.(2011).RSEM: accurate transcript quantification from RNA-Seq data with orwithout a reference genome.BMC Bioinformatics. 2011 Aug 4;12:323.
- [12] Eisen, M. B., et al. (2001). Clusteranalysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, (1998)95(25): 14863-8. 2001.29: 1165-1188.
- [13] M. J. L. de Hoon, et al. (2004). Open Source Clustering Software.Bioinformatics, 20(9): 1453-1454.
- [14] Saldanha, A. J. (2004). Java Treeview-extensible visualization of microarray data. Bioinformatics, 20(17): 3246-8.
- [15] GrabherrMG, et al.(2011).Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52.
- [16] Pertea G, et al.(2002).TIGRGene Indices clustering tools (TGICL): a software system forfast clustering of large EST datasets.Bioinformatics (2003)19 (5): 651-652.



7 常见问题



8 联系方式

联系我们

服务热线：400-706-6615

邮箱：info@bgitechsolutions.com

网址：www.bgitechsolutions.com

地址：广东省深圳市盐田区北山工业区 B2 栋（邮编：518083）