西藏黄牡丹授粉前后转录组测序和生物信息学分析

张姗姗*
(西藏自治区农牧科学院蔬菜研究所·西藏拉萨·850032)

摘 要:采用 Illumina Hiseq2000 高通量测序技术对西藏黄牡丹授粉前后花蕾转录组测序,运用生物信息学分析基因表达谱研究和差异表达基因功能预测,以期了解西藏黄牡丹授粉前后花蕾生长发育分子机理,并为西藏黄牡丹的基因水平研究奠定基础。结果表明:对过滤得到的高质量序列进行拼装授粉前(A2-1)和授粉后(A3-1)平均得到了 45453 条和 53742 条 unigenes,GC 含量分别为 41.64%和 41.98%,Unigenes 的平均长度为 822bp 和 722bp。两个样品之间差异表达的上调基因有 11321 个,差异表达的下调基因有 5585 个。其中 13288 个差异表达基因(DGEs)分别注释到 GO 数据库,分别涉及到细胞组成成份(cellular component),生物学过程(biological process)和分子功能(molecular function)三大功能。KEGG 数据库成功注释 9842 个 DEGs,占总 DEGs 的 58.22%。共涉及 Cellular Processes(细胞过程)、Environmental Information Processing(环境信息处理)、Genetic Information Processing (遗传信息处理)、Human Diseases (人类疾病)、Metabolism (代谢)、Organismal Systems (生物系统)等 6 个大的功能类别。本研究首次对西藏黄牡丹转录组进行了分析,为黄牡丹的分子生物学研究提供了宝贵的基因组数据来源。

关键词:西藏黄牡丹 转录组 高通量测序 差异表达基因

Sequencing and Analysis of the Transcriptome of Flower bud of Paeonia lutea

Zhang Shan-shan*
(Institute of Vegetables Research,TAAAS, Lhasa, China 850032)

Abstract:The transcriptome of Paeonia lutea was sequenced by Illumina HiSeq 2000 platform that is a new generation of high-throughput sequencing technology to study the expression profiling and predict the functional genes. Through filtering, splicing, assembling and going redundancy, 45453 unigenes and 53742 unigenes were obtained for before pollination (A2-1) and after pollination (A3-1), with an average length of 822bp and 722bp, and a GC percentage of 41.64% and 41.98%. A total of 16906 unigenes were identified as DEGS between A2-1 and A3-1, with 11321 up-regulated and 5585 down-regulated. We assigned 13288 of the 16906 DEGs to three major GO categories (cellular component,biological process,molecular function). To further analyzed the biological functions of annotated DEGs by mapping the DEGs to six main categories in the KEGG database, 9842 DEGs were assigned to six main categories including Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Diseases, Metabolism, Organismal Systems. This is the first time to study the gene associated in Paeonia lutea, and these data will provide a novel insight into the expressed genes and valuable theoretical basis to understand the molecular mechanisms.

Keywords: Paeonia lutea; Transcriptome; High-throughput sequencing; Different express Genes (DGEs)

近年来,包括基因组、转录组、蛋白质组等各种组学技术在揭示细胞生理活动规律和生物代谢机理的研究中起着越来越重要的作用,而转录组学是率先发展起来以及应用最为广泛的技术[1]。转录组是指细胞在特定状态下全部表达的的总和,反映相同基因在不同条件下表达水平的差异,并能揭示不同基因的相互作用及各自功能,转录组测序能全面快速地获得某一物种特定细胞或组织在某一状态下的基因表达情况。用于研究基因结构和功能、可变剪接和新转录本预测等[2-5]。相对于传统的芯片杂交平台,转录组测序无需已知序列设计探针。可对任意物种的整体转录活动进行检测,提供更精确的数字化信号、更高的检测通量以及更广泛的检测范围。对于许多缺乏基因组信息的物种而言,转录组研究已在非模式植物中得到了广泛应用。

*作者简介:张姗姗(1983-),女,助理研究员。主要从事果树育种和示范推广工作。Email:335877593@QQ.com

西藏黄牡丹为中国西南地区特有植物，花黄色，是培育牡丹、芍药等新品种的种质基因，在园艺育种上有重要的科学价值。是全世界濒临灭绝的珍稀植物之一，全世界现仅有林芝有大面积生长。此外，野生黄牡丹还是栽培牡丹的祖先，在科学上有很大的研究价值。

本试验使用 Illumina Hiseq 2000 高通量测序技术应用到西藏黄牡丹转录组的研究中，将测序得到的数据进行拼接、组装，并运用生物信息学方法对拼接得到的 Unigene 进行差异表达基因的功能注释、功能分类、代谢途径分析等，从功能基因组水平上研究西藏黄牡丹授粉前后转录组的差异表达基因，试图找出授粉前后控制花蕾生长变化的差异表达基因，同时为西藏黄牡丹基因水平研究提供重要的理论依据。

1 材料与方法

1.1 试验材料

2016 年 5 月取生长于西藏自治区林芝市巴宜区一片野生黄牡丹群体的花蕾为供试材料，授粉前的样品取于大风铃期(A2-1)，授粉后样品(A3-1)取于人为授粉后八天。样品采集后经液氮速冻后于-80℃贮存备用。

1.2 试验方法

1.2.1 总 RNA 提取，文库构建及测序

利用改良 CTAB 法提取嫩叶花蕾基因组 RNA。在提取的 RNA 中加入 DNase I，置于 37℃水浴 30min，除去其中的 DNA。用微量紫外分光光度计(2100 Bio-analyzer, Agilent Technologies, CA)检测剩余 RNA 的浓度和纯度。用带有 Oligo(dT)的磁珠(QIAGEN)富集 mRNA，然后加入破碎缓冲液(fragmentation buffer)将 mRNA 随机打断成片段，以这些 RNA 片段为模板反转录为第一链 cDNA 链，然后加入缓冲液、dNTPs、RNase H 和 DNA 聚合酶合成互补链。经纯化后的 cDNA 片段加 EB 缓冲液洗脱之后做末端修复 poly(A)，并连接测序接头，然后用琼脂糖凝胶电泳检测文库插入片段大小，最后利用 PCR 技术对连好接头的 cDNA 进行扩增制备文库。建好的文库送往华大基因(武汉)用 Illumina HiSeq 2000，采用 Illumina 双末端测序(Paired-end，PE)方法进行高通量测序。

1.2.2 转录组生物信息学分析

1.2.2.1 测序数据过滤。测序的原始数据包含低质量、接头污染以及未知碱基 N 含量过高的 reads，数据分析之前需要去除这些 reads 以保证结果的可靠性。

1.2.2.2De novo 组装。数据过滤后，使用 Trinity[6]对 clean reads 进行组装，通过序列之间的 overlap 信息组装得到重叠群(Contigs)，然后在局部进行组装得到转录本(Transcripts)，接下来我们使用 Tgicl[7] 对转录本进行聚类去冗余得到非冗余特异基因序列(Unigene)，冗余得到最终的 Unigene 用于后续功能注释分析。

1.2.2.3 分析不同样品的差异表达水平，然后根据差异表达水平的不同检测样品之间的差异表达基因(DGE)。

1.2.2.4 差异表达基因的功能注释。使用 BLAST 程序将组装得到的差异表达基因进行七大功能数据库注释 (NR, NT, GO, COG, KEGG, Swissprot and Interpro)，根据 NR (Non-redundant protein database，Nr)注释结果，我们统计了注释结果的物种分布，根据 COG(Cluster of orthologous groups)、GO 和 KEGG (Kyoto encyclopedia of genes and genomes)注释结果，统计了各自的功能分类[8,9]。

2 结果与分析

2.1 原始序列及拼装

从表 1 可以看出，样品 A2-1 和 A3-1 分别得到 56.67Mb 和 58.29Mb 的原始序列，过滤后的序列分别占原始序列的 78.65%和 77.80%。对过滤得到的高质量序列进行拼装每个样品平均得到了 45453 条和 53742 条 unigenes，GC 含量分别为

表 1 测序后转录组数据统计

Table 1 Summary of transcriptome sequencing and assembly results after Illumina sequencing.

Total raw reads (Mb)	Total clean reads (Mb)	Unigene	Total number	Total length (bp)	Mean length (bp)	N50 (bp)	GC percentage (%)
56.67	44.57		45453	37389176	822	1324	41.64
58.29	45.35		53742	38828973	722	1126	41.98

41.64%和 41.98%,Unigenes 的平均长度为 822bp 和 722bp。样品 A2–1 的序列大小从 201–9750 bp,N50 为 1324 bp。其中,长度在 200 – 500 bp 的 Unigene 有 22530 个,占总体的 49.57%;500 – 1000 bp 的 Unigene 有 10452 个,占总体的 23.00%;1000 –2000bp 的 Unigene 有 8690 个,占总体的 19.12%;2000 –3000bp 的 Unigene 有 2675 个,占总体的 5.89%;>=3000bp 的 Unigene 有 1106 个,占总体的 2.43% (图 1)。样品 A3–1 的序列大小从 201–9750 bp,N50 为 1126 bp。其中,长度在 200 – 500 bp 的 Unigene 有 28533 个,占总体的 53.09%;500 –1000 bp 的 Unigene 有 12190 个,占总体的 22.68%;1000 –2000bp 的 Unigene 有 10230 个,占总体的 19.04%;2000 –3000bp 的 Unigene 有 2336 个,占总体的 4.35%;>=3000bp 的 Unigene 有 453 个,占总体的 0.84%(图 2)。

2.2 差异表达基因(DEGs)分析及注释

2.2.1 根据两个样品基因表达水平结果,将 log2FoldChange(A3/A2)大于 2 小于–2 的定义为差异表达基因 (图 3)。其中差异表达的上调基因有 11321 个,差异表达的下调基因有 5585 个。

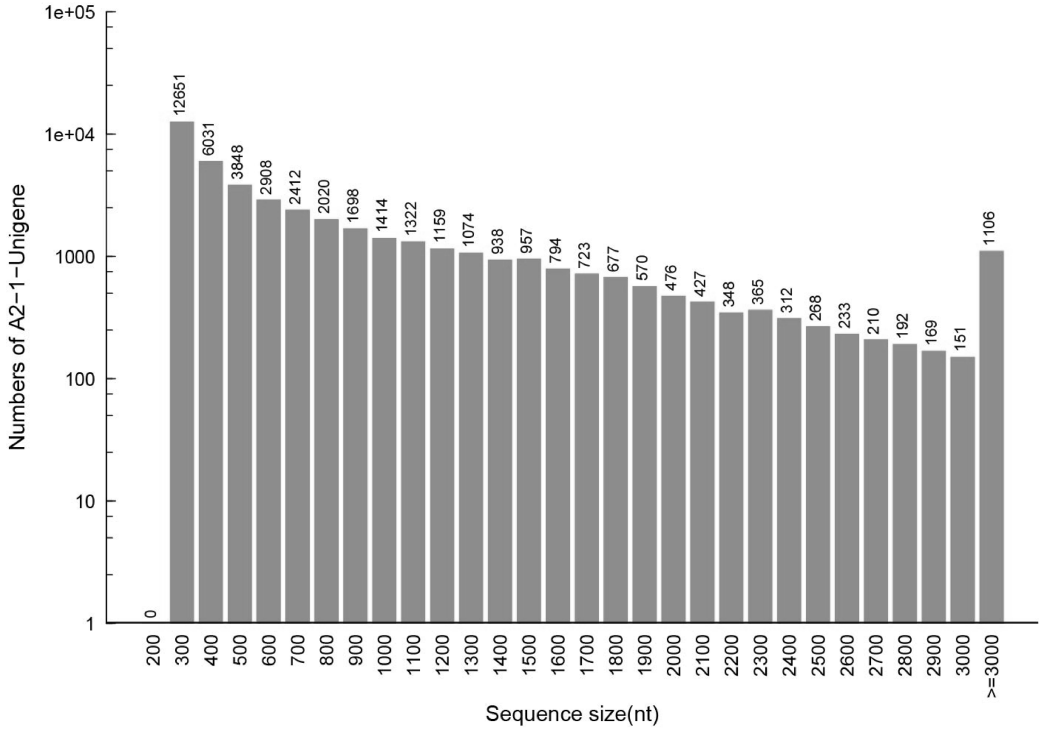

图 1 A2–1 的 Unigene 长度分布

Fig.1 Length distribution of A2–1–Unigene

2.2.2 A2–1 和 A3–1 差异表达基因(DEGs)的 GO 注释 GO 是一个基因功能分类数据库,能系统的描述不同生物中不同基因的生物学功能。结合 GO 数据库对黄牡丹两个时期的差异表达基因进行功能分类,注释结果表明 DEGs 被注释到三个类生物学功能,分别为细胞组成成份(cellular component),生物学过程(biological process)和分子功能(molecular function)。从图 4 可以看出 4754 个 DEGs 归属于细胞组分,涉及细胞、细胞功能、细胞膜、胞外区等 17 个更为具体的功能;3117 个 DEGs 归属于结合、催化活性、转运活性、分子结构等 14 个分子功能;5417 个 DEGs 归属于生物学过程,又可细分为生物调节、细胞转化、结合、应激反应等 22 个更为具体的功能。这一分类结果显示了黄牡丹授粉后生长过程中基因表达变化情况。

2.2.3 A2–1 和 A3–1 差异表达基因 (DEGs)的 KEGG 注释 从图 5 中可知,KEGG 数据库共注释到 9842 个 DEGs,占总 DEGs 的 58.22%。共涉及 Cellular Processes (细胞过程)、Environmental Information Processing (环境信息处理)、Genetic

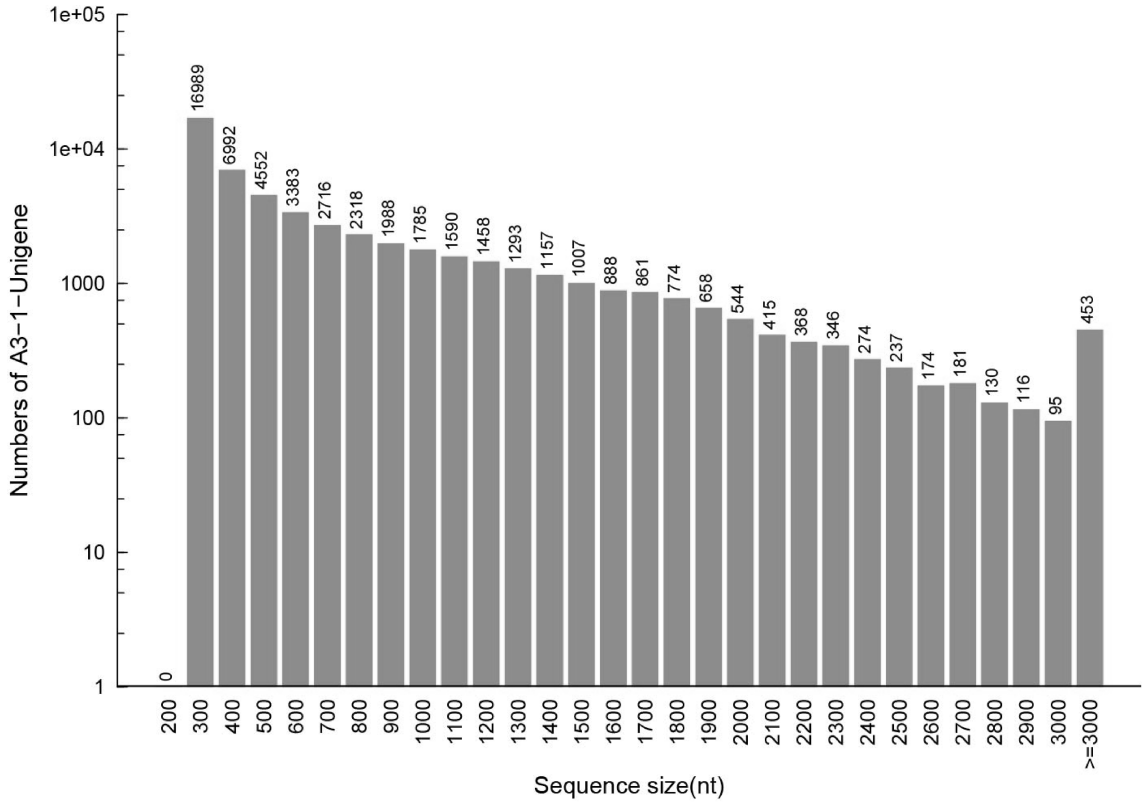

图 2 A3–1 的 Unigene 长度分布

Fig.2 Length distribution of A2–1–Unigene

Information Processing（遗传信息处理）、Human Diseases（人类疾病）、Metabolism（代谢）、Organismal Systems（生物系统）等 6 个大的功能类别，Transport and catabolism（运输和代谢）、Membrane transport（膜转运）、Signal transduction（信号转导）、Folding, sorting and degradation（排序

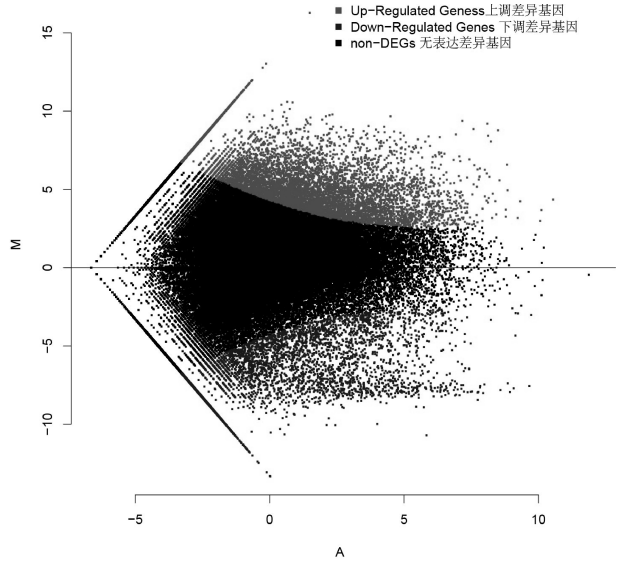

图 3 A2–1 和 A3–1 差异表达基因。

Fig.3 The DGEs between A2–1and A3–1.

和降解)、Replication and repair (复制和修复)、Transcription(翻译)、Translation(转录)等 20 个中等功能类别。其中涉及全球概况 (Global and overview maps)的 DEGs 数量最多,为 2293 条。

3 结论与讨论

近年来,国内外对黄牡丹的研究日益增加,主要集中在种子萌发、花粉活力鉴定、远缘杂交[11,12]等。随着基因组测序技术和生物信息学的广泛应用,植物基因组研究得到快速发展,关于黄牡丹基因组学的研究也越来越多,但利用基因组测序来解释黄牡丹授粉后生长发育的分子机理等方面的研究还相对较少。该研究对西藏黄牡丹花蕾转录组进行测序和生物信息学分析,对过滤得到的高质量序列进行拼装授粉前(A2-1)和授粉后(A3-1)平均得到了45453 条和 53742 条 unigenes,GC 含量分别为 41.64%和 41.98%,Unigenes 的平均长

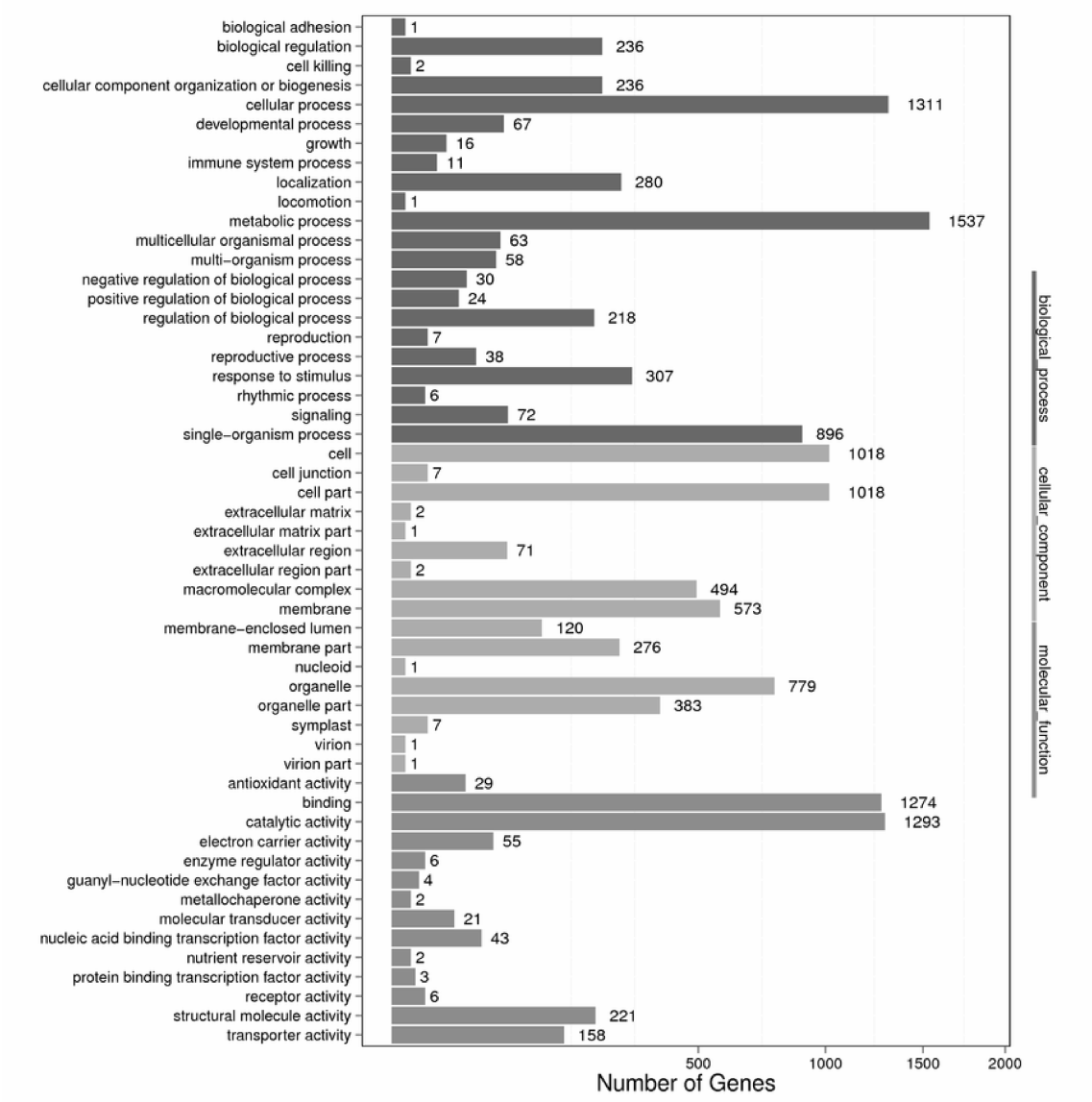

图 4 A2-1 和 A3-1 差异表达基因(DEGs)的 GO注释

Fig.4 GO functional categories of DEGs between A2-1and A3-1

度为 822bp 和 722bp。两个样品之间差异表达的上调基因有 11321 个，差异表达的下调基因有 5585 个。其中 13288 个差异表达基因(DGEs)分别注释到 GO 数据库，分别涉及到细胞组成成份，生物学过程和分子功能三大功能。9842 个 DEGs 成功注释到 KEGG 数据库的 6 个大的功能类别。本研究首次对西藏黄牡丹转录组进行了分析，为黄牡丹的分子生物学研究提供了宝贵的基因组数据来源。将为西藏黄牡丹花的生长发育过程、代谢生理等的分子机理研究及分子育种提供必要的基础信息。

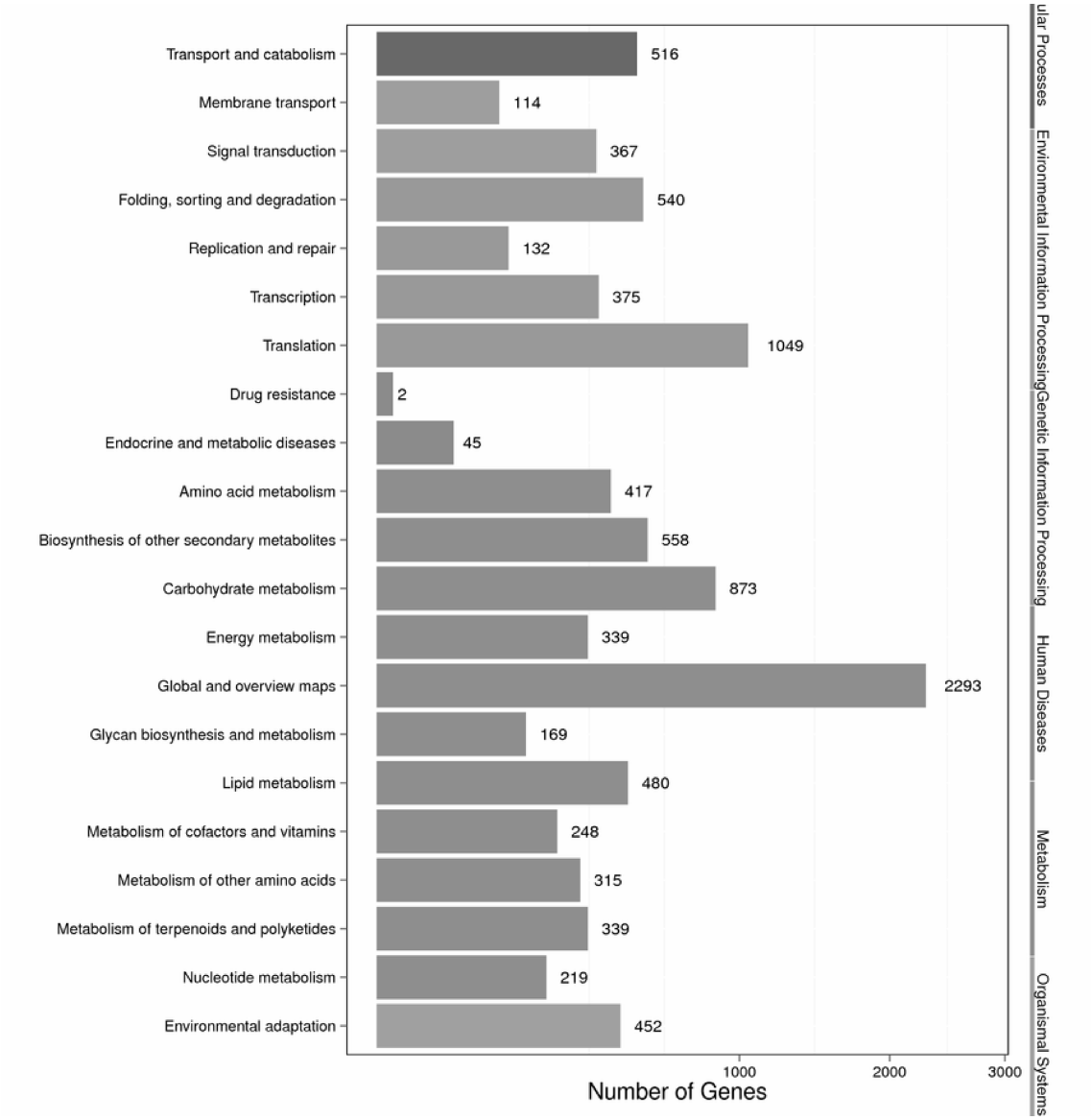

图 5 A2–1 和 A3–1 差异表达基因(DEGs)的 KEGG 注释

Fig.5 KEGG functional categories of DEGs between A2–1and A3–1

参考文献

[1] Lockhart D J, Winzeler E A. 2008. Genomics, gene express and DNA arrays. Nature,405 (6788):827 - 836.

[2] Alagna F D,Agostino N,Torchia L,et al. 2009. Comparative 454 pyrosequencing of transcriptsfrom twoolive genotypes during fruitdevelopment.BMC Genomics,10:399.

[3] Barakat A,DiLoreto D S,Zhang Y,et al. 2009.Comparison of the transcriptomes of Americanchestnut(Castanea dentata)and Chinese chestnut(Castanea mollissima)in response to the chestnutblight infection. BMC Plant Biol,9:51.

[4] Dassanayake M,Haas J S,Bohnert H J,et al.2010. Shedding light on an extremophile life style through transcriptomics. New Phytol,183:764 - 775.

[5] Li R,Fan W,Tian G,et al. 2010. The sequence and denovo assembly of the giant panda genome.Nature,463:311 - 317.

[6] Grabherr MG, Haas BJ, Yassour Moran, et al.2011. Full-length transcriptome assembly fromRNA-Seq data without a reference genome. Nat Biotechnol. 2011 May 15; 29(7): 644-52.

[7] Pertea G, Huang X, Liang F, et al. (2002).TIGR Gene Indices clustering tools (TGICL): a softwaresystem for fast clustering of large EST datasets. Bioinformatics (2003) 19 (5): 651-652.

[8] Roman L T,Michael Y G,Darren A N, et al. 2003. The COG database:A tool for genome-scale analysis of protein functions and evolution.Nucleic Acids Res,28 (1):33 - 36.

[9] Minoru K,Susumu G,Shuichi K, et al. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res, 32:277 - 280.

[10] Iseli C, Jongeneel CV, Bucher Philipp, et al. 1999.ESTS can: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol. 1999:138-48.

[11] 律春燕.中国林业科学研究院, 黄牡丹野生种与牡丹、芍药栽培品种远缘杂交研究,2012.

[12] 赵娜,石颜通,袁涛.黄牡丹远缘杂交后代花粉粒特性,[J]广西植物,2016,36(3):280-288.