浙江理工大学学报(自然科学版),第 37 卷,第 2 期,2017 年 3 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 36, No. 2, Mar. 2017

DOI:10.3969/j.issn.1673-3851.2017.03.022

基于高通量测序半夏珠芽转录组研究

叶　德,杨支力,李东海,徐　涛
(浙江理工大学生命科学院,杭州 310018)

摘　要:利用高通量测序技术,对半夏珠芽和茎进行测序分析,共获得 6.5 Gb 原始数据,122 282 962 条有效 reads,平均长度为 103.84 bp。经 Denovo 组装,获得 90 175 条 unigenes,平均长度为 751 bp。对长度分布、GC 含量、表达水平等方面的评价显示,测序数据质量好且可信度高。将 unigenes 比对到 NR、KOG 和 GO 数据库,分别获得了 55 197、46 237 和 31 622 条注释信息。通过分析半夏珠芽和茎的基因表达量,共筛选出 19 个与玉米素、吲哚乙酸、茉莉酸、脱落酸和赤霉素代谢相关的差异表达基因。基因差异表达分析结果表明:玉米素、生长素、茉莉酸、脱落酸 4 种激素合成代谢相关基因在半夏珠芽中的表达水平高于茎中,而赤霉素则相反。

关键词:半夏;珠芽;茎;转录组;激素
中图分类号:Q946.2　　**文献标志码**:A　　**文章编号**:1673-3851 (2017) 02-0282-07

0　引　言

中药半夏为天南星科多年生草本植物半夏 *Pinellia ternate* (*Thunb.*) *Breit.* 的干燥块茎,始载于《神农草本经》,列为下品。其性温、味辛、有小毒,归脾、胃、肺[1]。具有燥湿化痰、降逆止呕、消痞散结的功效,为治痰、止吐之药。近年来,发现半夏蛋白凝集素具有抗肿瘤的功效。半夏原产于亚洲,现今美洲地区也有种植。半夏繁殖有三种方式:种子繁殖、珠芽繁殖和块茎繁殖。在自然界和农业生产中,由于珠芽繁殖具有效率高、速度快、产量大等优点,大部分半夏都通过珠芽繁殖。因此,珠芽是构成半夏产量的重要因素,也对半夏的种植起决定性作用。

半夏珠芽着生于叶柄下部[2],起始于幼嫩叶叶柄的腹面最外轮维管束外周壁细胞,恢复分裂的薄壁细胞分裂形成珠芽原基细胞团,在原基细胞生长突破叶柄表皮后分化形成具有生长点的珠芽结构[3]。半夏珠芽在形成的过程中,叶柄上端吲哚乙酸、脱落酸、玉米素和茉莉酸激素的含量基本呈上升趋势,而赤霉素在珠芽形成初期急剧下降[4]。

近年来,新一代高通量测序技术发展迅速,并已应用于生物学研究。相比于第一代测序技术,新一代高通量测序技术具有通量高、速度快、准确度高等优点[5]。本实验以半夏为研究对象,以新一代高通量测序为技术手段,通过 Illumina 测序平台研究半夏珠芽和茎的转录组信息,筛选出 5 种激素代谢途径相关的基因,分析差异表达情况,为研究半夏珠芽发育和激素的关系奠定基础。

1　材料与方法

1.1　实验材料

三叶半夏,采集自甘肃天水,种于浙江理工大学试验田。选取长势良好、健康的半夏,在珠芽形成的初期采集至实验室。刮取珠芽为样品 PZY,取珠芽上下部分的茎为样品 PJ,取样后迅速将样品置于液氮中冷冻,保存于超低温冰箱中。

1.2　实验方法

取半夏珠芽和茎各 1 g,分别置于研钵中,加入液氮迅速研磨至粉末状,加入 2 mL 的 Trizol 裂解细胞,再用氯仿异丙醇分离纯化出总 RNA,Agilent

收稿日期:2016－10－30　　网络出版日期:2017－01－19
基金项目:国家自然科学基金项目(30772712)
作者简介:叶　德(1990－),男,浙江温州人,硕士研究生,主要从事分子生物学方面的研究。
通信作者:徐　涛,E-mail:pkuxt@yahoo.com.cn

2100 BioanaLyzer 分析 RNA 提取质量。将总 RNA 经带有 Oligo(dT)的磁珠富集,再进行洗脱获得半夏的总 mRNA。将分离纯化的 mRNA 随机打断成片段,并通过随机引物逆转录合成 cDNA。cDNA 经过纯化、粘性末端修复为平末端、加 polyA 和测序接头,最后经文库质检合格后上机测序。

将测序所得的原始 reads(raw reads)经过过滤处理得到高质量的 reads(valid reads),然后用 Trinity[6]程序对 valid reads 进行从头组装,组装成两端不能再延长的 unigenes。

将组装所得的 unigenes 通过 BlastX[7]程序比对到蛋白质数据库,经 NR[8]、Swiss-Prot[9]、Pfam[10]、和 KOG[11]的软件进行分析,并取值 $E<1\times10^{-5}$ 的注释。

根据两个本转录本的表达丰度值进行差异表达分析,包括差异倍数、显著水平和 FDR 值等差异表达分析。将筛选出来的差异表达的基因比对到 GO 和 KEGG 等数据库进行注释和分类。最后筛选出与各类激素代谢相关的差异表达基因并分析。

2　结果与分析

2.1　测序数据的组装

两个样品经 Illumina Hiseq2500 高通量测序平台测序,珠芽获得 61 454 192 条 raw reads,平均长度为 114.0 bp,茎获得 88 371 236 条 raw reads,平均长度为 93.7 bp。除去接头和低质量的 reads 后,valid reads 中 Q20 的百分数分别为 95.93%和 96.99%(大于 90%),质量合格,能够用于后续分析。采用 Trinity 软件对序列进行组装,共获得 90 175 条 unigenes,平均长度和 N50(N50 指将 unigene 从长到短排序,一次累加 unigenen 碱基数,当累计碱基数达到 unigene 总碱基数的 50%时的 unigene 的长度)分别为 751 bp 和 1 145 bp。unigenes 长度主要分布在 200 bp 到 2 000 bp,数量随着长度的增加而减少,如图 1 所示。

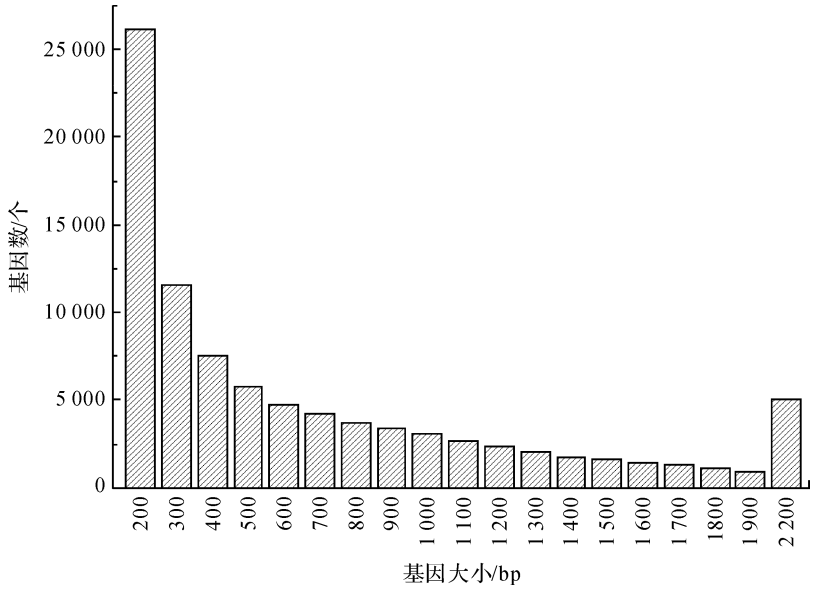

图 1　unigenes 长度分布

GC 含量能直观的反应出基因的结构、功能和进化信息,是基因组碱基序列的重要特征之一,GC 含量分布用于分析是否因测序或建库所带来的 GC 分离现象,以免影响后续样品的定量分析。本次测序样品的 GC 含量为 43%,其中没有 GC 含量过低(小于 20%)或过高(大于 80%)的 unigenes,GC 含量总体上呈正态分布,这从另一方面反映了测序质量较好。

2.2　unigenes 功能注释及相关分析

2.2.1　序列同源相似性分析

将半夏 unigenes 通过 BLAST 程序比对到 NR、KOG、Swiss-prot 和 Pfam 等数据库($E<1\times10^{-5}$),进行 unigenes 的序列相似性分析,注释情况如表 1 所示。与 NR 数据库的比对结果显示,55 197 条 unigenes 序列可以找到相似序列。而在相似性序列匹配的近缘物种中,所占的比例最高的是葡萄(*Vitis vinifera*,28.8%),其后依次是水稻(*Oryza sativa*,7.7%)、蓖麻(*Ricinus communis*,6.5%)、小米(*Setaria italica*,5.1%)、野草莓(*Fragaria vesca*,3.6%)和高粱(*Sorghum bicolor*,3.5%),如图 2 所示。

表1 BLAST注释汇总

数据库	unigenes数/条	百分比/%
Swiss-prot	35 616	39.50
NR	55 197	61.21
Pfam	38 967	43.21
KEGG	22 961	25.46
KOG	46 237	51.27

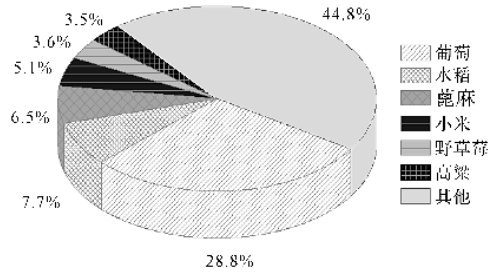

图2 NR注释物种分布

2.2.2 unigenes的KOG功能分类研究

真核生物蛋白相邻类的聚簇(eukaryotic ortholog group,KOG)是通过直系同源的关系预测基因的产物,从而进行分类统计的数据库[12]。将半夏unigenes比对到KOG数据库里,总共有46 237条unigenes(占总数的51.27%)被注释到25种KOG分类中(表2)。被注释的unigenes涵盖大部分生命活动,涉及KOG功能类别也比较全面;其中,"一般功能基因"是最大的分类,包含6 286条(13.60%)unigenes;其次是"信号传导机制"包含了5 223条(11.30%)unigenes;而最少的是"细胞运动",只有13条(0.08%)unigenes;其他类别的基因表达丰度都各不相同。

表2 KOG功能分类

编号	分类	数量/条	编号	分类	数量/条
A	RNA加工与修饰	1 543	N	细胞运动	13
B	染色质结构与变化	536	O	蛋白质翻译后修饰与转运,分子伴侣	3 350
C	能量产生与转化	1 076	P	无机离子运输与代谢	854
D	细胞周期调控与分裂,染色体重排	819	Q	次生产物合成,运输及代谢	1 001
E	氨基酸运输与代谢	1 288	R	一般功能基因	6 286
F	核苷酸运输与代谢	414	S	功能未知	1 837
G	碳水化合物运输与代谢	1 635	T	信号传导机制	5 223
H	辅酶运输与代谢	313	U	膜泡运输与胞内分泌	1 546
I	脂类运输和代谢	1 148	V	防御机制	229
J	翻译,核糖体结构和生物合成	1 661	W	胞外结构	128
K	转录	1 836	Y	核结构	136
L	复制、重组与修复	931	Z	细胞构架	822
M	胞壁/膜生物发生	425			

2.2.3 unigenes的GO功能分类研究

基因本体论(gene ontology,GO)通过建立特定语言,对基因和蛋白质的功能进行限定和描述,从而进行分类[13]。将半夏unigenes通过Blast2GO[14]程序比对到GO数据库,然后用WEGO[15]程序将比对结果进行功能分类统计。GO数据库中包含3个相对独立的本体,各自描述基因产物所处的细胞组分(cellular component)、基因产物的分子功能(molecular function)和基因产物所参与的生物学过程(biological process)。注释结果如图3所示,有31 622条(35.07%)unigenes被注释到GO数据库,分类成57个功能组。其中注释最多的几个分别是:cell(GO:0005623)、cell part(GO:0044464)和cell part(GO:0044464)属于细胞组分;binding(GO:0005488)和catalytic(GO:0003824)属于分子功能;cellular process(GO:0009987)和metabolic process(GO:0008152)属于生物学过程。

2.2.4 差异表达基因的筛选

为了分析差异表达基因的整体情况,以差异倍数和显著水平两个水平绘制火山图进行评估。如图4所示,纵坐标表示基因表达量变化差异的统计学显著性,横坐标表示基因在不同的样本中表达倍数变化。黑色的点表示有显著性差异表达的基因,灰色点表示非显著性差异表达基因。在差异比较中,

log2 fold change 值大于 0 表明该基因上调,log2 fold change 小于 0 表明该基因下调。在本组对比中,以半夏茎为对照样本,半夏珠芽为实验样本,差异基因中有 5 869 个基因上调,有 9 615 个基因下调。unigenes 做聚类分析有利于生物体基因功能、基因调控、细胞过程及细胞亚型等方面的研究[16]。

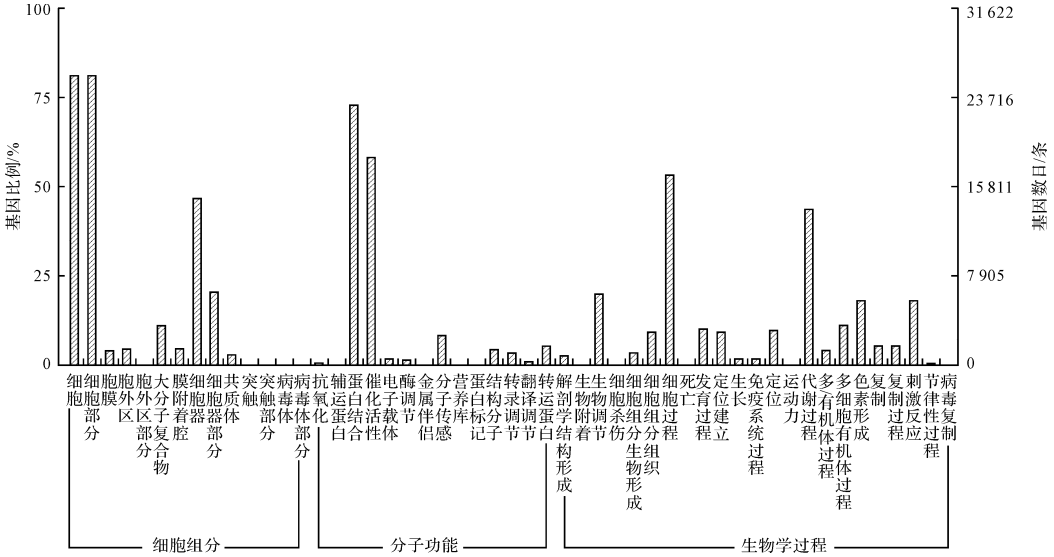

图 3　unigenes 的 GO 分类

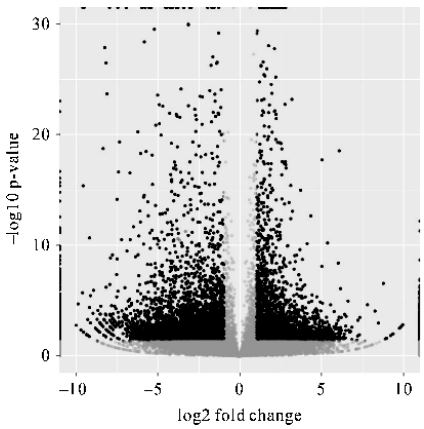

图 4　差异表达基因火山图

2.2.5　差异基因的注释和代谢通路的富集

筛选出差异表达基因后,对差异基因做 GO 富集分析。分析结果如图 5 所示,涉及细胞组分(cellular component)有 18 766 条 unigenes,是最多一个分类;生物学过程(biological process)有 11 628 条 unigenes;最少的是分子功能(molecular function)有 8 285 条 unigenes。

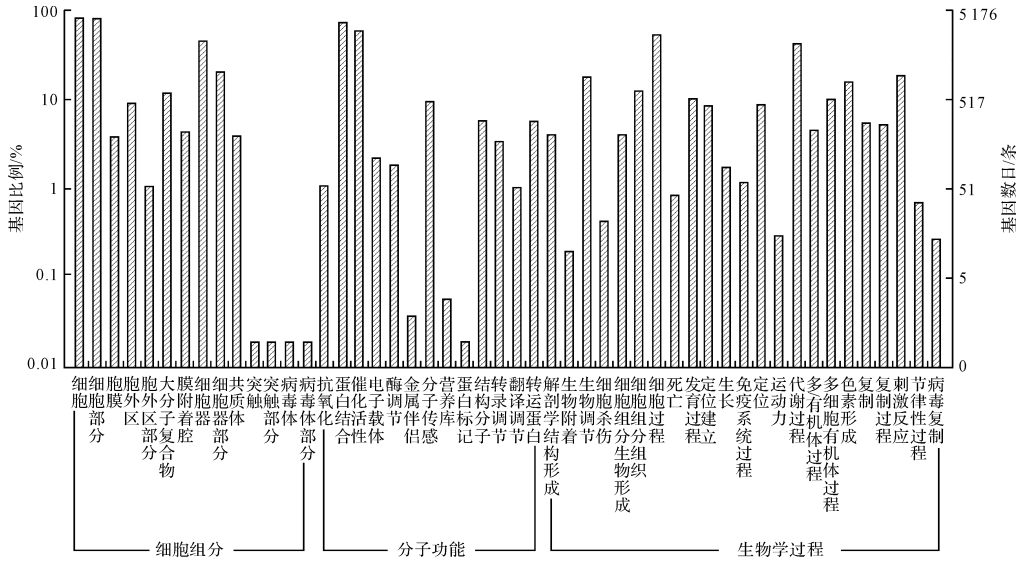

图 5　差异基因的 GO 分类

在生物体内，行使各种生命活动需要不同基因共同协调来完成，为了进一步分析差异表达基因的生物学功能，利用 KEGG（Kyoto encyclopedia of genes and genomes）[17]数据库对筛选出来的差异表达基因进行功能分类和 Pathway 注释，能够通过代谢路径从整体上分析差异基因。结果显示有 4 266 条 unigenes 被注释到 192 个代谢通路中。其中注释最多的是核糖体（ko03010），有 162 条基因；其次是淀粉和蔗糖代谢（ko00500），有 156 条基因。

2.2.6 5 种激素代谢相关基因的筛选

植物激素化学结构简单，含量少，却在植物生长发育中起着重要的作用。其中玉米素（zeatin-riboside，ZR）是一种细胞分裂素，在分析半夏珠芽和茎 pathway 注释的过程中，发现了 ZR 的生物合成途径（ko00908，如图 6 所示）中几个关键基因（如表 3 所示）的差异表达。玉米素的生物合成有两条：第一条是从头合成途径。在这条途径中，磷酸戊烯基转移酶（adenylate isopentenyltransferase，AtIPT）催化二甲基丙烯基二磷酸（dimethylallyl pyrophosphate，DMAPP）的异戊烯基团转移到磷酸腺苷上；该反应为合成玉米素的第一步，AtIPT 则是玉米素合成的限速酶。第二步为细胞分裂素反式羟化酶催化（CYP735A）催化上一步的产物形成反式玉米素。上述分析结果表明，这两个基因在半夏珠芽中均有表达，但是在茎中表达量极低或者不表达。细胞分裂素脱氢酶是玉米素的负调节酶，它通过分解玉米素合成第一步的产物来调节植物体内玉米素的含量。此基因在茎中有表达，在珠芽中未检测到表达。玉米素的第二条合成途径是 tRNA 分解合成途径[18]，该途径在珠芽和茎中的基因表达差异不明显，其原因可能是在半夏中 tRNA 分解合成玉米素途径是次要途径。

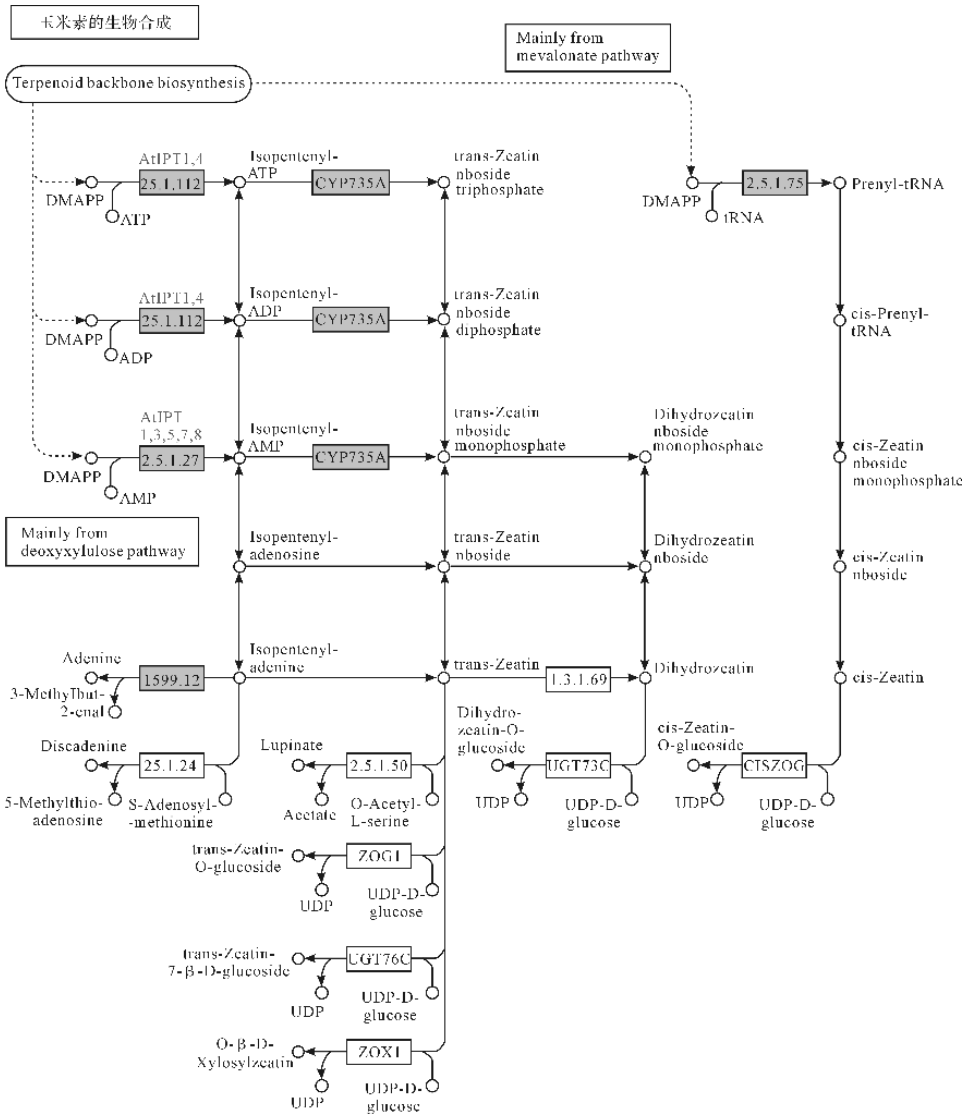

图 6 玉米素的生物合成途径（ko00908）注释

在分析吲哚乙酸(indole-3-cetic acid, IAA)代谢途径过程中发现有 5 个基因差异表达,它们被富集在色氨酸代谢(ko00380)这条途径中(表 3)。吲哚乙酸的生物合成都是从色氨酸开始,其中一条途径是:色氨酸在色氨酸脱羧酶的作用下形成色胺[19]。被注释的另一个酶是吲哚-3-乙醛氧化酶,吲哚-3-乙醛在此酶的作用下转化成吲哚乙酸[20]。

表 3　激素代谢相关的差异表达基因

Gene ID	长度/bp	PZY_FPKM	PJ_FPKM	注释
comp44878_c0_seq5	1683	10.42	1.37	磷酸戊烯基转移酶
comp133640_c0_seq1	696	6.63	0	细胞分裂素反式羟化酶
comp47752_c1_seq2	838	0	4.92	细胞分裂素脱氢酶
comp216751_c0_seq1	263	5.98	0	氨基酸脱羧酶
comp211781_c0_seq1	212	5.62	0	氨基酸脱羧酶
comp34804_c0_seq1	414	11.93	0.66	吲哚-3-乙醛氧化酶
comp24334_c0_seq1	298	9.80	1.42	吲哚-3-乙醛氧化酶
comp42922_c1_seq1	301	16.41	4.85	吲哚-3-乙醛氧化酶
comp49877_c0_seq1	822	23.16	4.27	磷脂酶 A
comp49877_c0_seq4	844	40.49	9.81	磷脂酶 A
comp49877_c0_seq6	821	19.75	2.89	磷脂酶 A
comp49877_c0_seq7	845	46.93	0.97	磷脂酶 A
comp53749_c0_seq1	3284	86.19	28.00	脂肪氧合酶
comp39838_c0_seq6	965	19.96	0.97	丙二烯氧化物合酶
comp48844_c0_seq5	737	10.07	1.19	12-氧-植物二烯酸还原酶

有 3 个差异表达基因被 GO 数据库分类到脱落酸(Abscisic acid, ABA)生物合成途径(GO:0009688),包括 comp15580_c0_seq1、comp24334_c0_seq1 和 comp17177_c0_seq1。其中 comp17177_c0_seq1 被定义为 9-顺式-环氧类胡萝卜素双加氧酶,是 ABA 生物合成的关键酶[21]。此基因珠芽的 FPKM 值为 4.58,茎的 FPKM 值则为 0.85,可以看出在珠芽当中的表达量大于在茎中的表达量。

有 7 个差异表达基因被 pathway 富集到 α-亚麻酸的代谢途径中(ko00592),而亚麻酸是合成茉莉酸(jasmonic acid, JA)的起始物质[22]。这 7 个基因中,有 4 个是催化软磷脂转化成亚麻酸的酶,3 个是茉莉酸合成途径中的酶,并且这 7 个基因在珠芽中的表达量高于茎中,如表 3 所示。

有关赤霉素(gibberellin, GA_3)的差异表达基因只找到一个(comp36256_c0_seq1,707 bp,GO:0009686),注释为内根-贝壳杉烯酸羧化酶。此酶催化内根-贝壳杉烯酸转化成赤霉素 12 醛,为赤霉素合成的中间步骤[23],在珠芽的 FPKM 值为 7.6,在茎的 FPKM 值为 17.45,该基因在茎中的表达量大于在珠芽中的表达量。

3　结　论

本文利用高通量测序技术,对形成初期的半夏珠芽和珠芽旁边的茎进行测序分析,获得大量基因注释信息,为半夏珠芽发育的分子生物学研究提供重要信息。分析 IAA、ABA、ZR、JA 和 GA_3 5 种激素代谢相关基因的差异表达情况,筛选出与它们代谢途径相关的差异表达基因,分析后发现:ZR、IAA、ABA 和 JA 这 4 种激素合成代谢的相关基因在珠芽中的表达量大于茎中的表达量;而 GA_3 合成代谢的相关基因在珠芽的表达量小于茎中的表达量。

本文从转录水平研究分析半夏珠芽转录组的表达情况,90 175 条 unigenes 有 55 197 条得到同源性注释,剩下的 34 975 条 unigene 可能由于属于非编码序列或者是新基因未能注释上[24]。两个样本高通量测序所获得的信息量极大,巨大的信息量将导致后期数据挖掘难度增大,这正是功能基因挖掘和代谢途径分析广泛存在的难题。本文通过分析与 5 种激素代谢相关的基因,为半夏珠芽生长发育与激素的关系奠定基础。但是激素代谢基因转录水平的差异并不代表组织内激素水平的差异,这是由于从基因转录到产物还有一系列步骤,在这些步骤中可能会受到调控等影响。后续将继续分析这些基因受体和上游调控基因的差异表达情况,从整体上阐明激素对半夏珠芽生长发育的影响。另外,凝集素作为目前半夏研究的热点,分析凝集素基因的表达情况,从而揭示凝集素在半夏珠芽中的代谢情况,将有助于高产量凝集素的半夏研究。

参考文献:

[1] 国家药典委员会.中华人民共和国药典[M].北京:化学工业出版社,2005:398.

[2] XU T, WANG B, LIU X F, et al. Microarray-based identification of conserved microRNAs from *Pinellia ternate*[J]. Gene,2012,493:267-272.

[3] 罗睿,杜禹珊,孙莹莹,等.半夏珠芽发育过程的形态学和解刨学研究[J].西北植物学报,2014,34(9):1776-1781.
[4] 常莉,徐有明,薛建平.离体培养条件下半夏叶柄形成珠芽过程中内源激素的变化[J].华中农业大学学报,2007,26(5):612-615.
[5] SHENDURE J, JI H. Next-generation DNA sequencing [J]. Nature Biotechnology,2008,26(10):1135-1145.
[6] GANFRED M G, HAAS B J, YASSOUR M, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome[J]. Nature Biotechnology, 2011,29(7):644-652.
[7] ALTSCHUL S F, GISH W, Miller W, et al. Basic local alignment search tool[J]. Molecular Biology,1990, 215(3):403-410.
[8] PRUITT K D, TATUSOVA T, MAGLOTT D R. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins [J]. Nucleic Acids Research,2005,33:501-504.
[9] BAIROCH A, APWEILER R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000[J]. Nucleic Acids Research,2000,28:45-48.
[10] PUNTA M, COGGILL P C, EBERHARDT R Y, et al. The Pfam protein families database[J]. Nucleic Acids Research,2012, 40:290-301.
[11] LI L, STOECKERT C J, Roos D S. OrthoMCL: identification of ortholog groups for eukaryotic genomes[J]. Genome Research,2003,13(9):2178-89.
[12] ZHOU Y, GAO F, LIU R, et al. De novo sequencing and analysis of root transcriptome using 454 pyrosequencing to discover putative genes associated with drought tolerance in *Ammopiptanthus mongolicus* [J]. BMC Genomics,2012,13(2):266-279.
[13] 杨楠,赵凯歌,陈龙清.蜡梅花转录组数据分析及次生代谢产物合成途径研究[J].北京林业大学学报,2012, 34(1):104-107.
[14] CONESA A, Götz S, García-Gómez J M, et al. Blast2GO a universal tool for annotation, visualization and analysis in functional genomics research [J]. Bioinformatics,2005,21(18):3674-3676.
[15] YE J, FANG L, ZHENG H, et al. WEGO:a web tool for plotting GO annotations [J]. Nucleic Acids Research,2006,34:293-297.
[16] MCLACHLAN G J, BEAN R W, PEEL D. A mixture model-based approach to the clustering of microarray expression data[J]. Bioinformatics,2002,18(3):413-422.
[17] KANEHISA M, GOTO S, KAWASHIMA S, et al. The KEGG resource for deciphering the genome[J]. Nucleic Acids Research,2004,32:277-280.
[18] 张红梅,王俊丽,廖祥儒.细胞分裂素的生物合成、代谢和受体[J].植物生理学通讯,2003,3(3):267-272.
[19] KUO T T, KOSUGE T. Role of aminotransferase and indole-3-pyruvic acid in the synthesis of indole-3-acetic acid in pseudomonas savastanoi[J] Journal of General and Applied Microbiology,1970,16:191-204.
[20] 王家利,刘冬成,郭小丽,等.生长素合成途径的研究进展[J].植物学报,2012,47(3):292-301.
[21] 杨洪强,接玉玲.高等植物脱落酸的生物合成及其调控[J].植物生理学通讯,2001,37(5),457-462.
[22] 蒋科技,皮妍,侯嵘,等.植物内源茉莉酸类物质的生物合成途径及其生物学意义[J].植物学报,2010,45(2): 137-148.
[23] 石琰璟,沙广利,束怀瑞.赤霉素生物合成及其分子机理研究进展[J].西北植物学报,2006,26(7):1482-1489.
[24] KONISHIT, OHNISHI O. A linkage map of common buckwheat based on microsatellite and AFLP markers [J]. Fagopyrum,2006,23(2):1-6.

Transcriptome Analysis of *Pinellia* Bulbil Based on High-throughput Sequencing Technology

YE De, YANG Zhili, LI Donghai, XU Tao
(College of Life Science, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The transcripton of bulbil and stem from *Pinellia* was sequenced by high-throughput sequencing technology. A RNA-seq library was established, which contained 6.5 GB raw data, as well as 122 282 962 pieces of valid reads with an average length of 103.84 bp. After Denono assembly, 90 175 unigenes were obtained with an average length of 751 bp. In addition, the length distribution, GC content and expression level indicated that the data had good quality and high credibility. A total of 55 179 unigenes were annotated in NR database, and 46 237 unigenes were categorized into 25 KOG classifications. 31 622 unigenes were assigned to 57 GO terms which contain three main categories. Furthermore, sequenced results between bulbil and stems from *Pinellia* were compared to study differentially expressed genes and then 19 genes were found related to anabolism genes about ZR, IAA, JA, ABA and GA_3. The analysis results showed that the expression level of genes of ZR, IAA, JA and ABA in bulbil were more high than in stem, except GA_3.

Key words: *Pinellia*; bulbil; stem; transcriptome; hormone

(责任编辑:唐志荣)