

• 药材与资源 •

刺五加转录组和差异性表达分析

宋 菊, 国红玉, 李志栋, 尤鹏升, 龙月红, 邢朝斌*

华北理工大学生命科学学院, 河北 唐山 063000

摘要: **目的** 获得刺五加 *Eleutherococcus senticosus* 转录组数据库和差异表达基因。**方法** 采用皂苷高含量组和低含量组 2 个样本作为受试材料, 采用二代测序方法中的 Illumina HiSeq 4000 进行转录组测序, 并进行系统的生物信息学分析。**结果** 共获得 8.34 Gb 数据, 拼接得到 77 087 条 Unigenes, 与 5 个基因数据库进行比对, 可归类于 55 个 Geneontology (GO) 分类中, 涉及到 116 个 KEGG 标准代谢通路。通过差异性分析发现, 差异性表达基因共 530 条, 其中上调基因占 42.08%, 下调基因占 57.92%, 相差较大。进行 GO 和 Pathway 富集, 得到 408 个 GO 注释和 40 个代谢通路。**结论** 对刺五加转录组进行拼接、组装和功能注释, 得到大量转录本信息, 为刺五加分子生物学研究提供了宝贵的基因组数据库资源。

关键词: 刺五加; 转录组; 表达分析; 皂苷; 代谢通路

中图分类号: R282.12 文献标志码: A 文章编号: 0253-2670(2016)22-4049-05

DOI: 10.7501/j.issn.0253-2670.2016.22.020

Analysis of transcripiotomes and differentially expressed genes in *Eleutherococcus senticosus*

SONG Ju, GUO Hong-yu, LI Zhi-dong, YOU Peng-sheng, LONG Yue-hong, XING Zhao-bin

College of Life Sciences, North China University of Science and Technology, Tangshan 063000, China

Abstract: **Objective** To obtain the transcriptome database and differentially expressed genes of *Eleutherococcus senticosus*. **Methods** We choose the high content group and the low content group of saponin as experimental materials, and use the high-throughput sequencing technology (Illumina HiSeq 4000) to sequence the transcriptome of *E. senticosus*, then we systematically analyze the sequencing results in the bioinformatic way. **Results** We have assembled 8.34 Gb database, after assembly steps, we get 77 087 of *E. senticosus* unigenes, then blasting them with five data banks. All unigenes are involved in 55 GO-terms and 116 metabolic pathways. Though the differentially expressed analysis of two materials, we get 530 differentially expressed genes, the up-regulated genes account for 42.08%, the down-regulated genes account for 57.92%. After GO and Pathway enrichment analysis, we get 408 GO-natations and 40 metabolic pathways. **Conclusion** These data represent the abundant messages about transcripts and provide the valuable genome data sources in molecular biology of *E. senticosus*.

Key words: *Eleutherococcus senticosus* (Rupr. et Maxim.) Maxim.; transcriptome; expression andysis; saponins; metabolic pathways

刺五加 *Eleutherococcus senticosus* (Rupr. et Maxim.) Maxim. 是我国传统的珍贵药用植物, 具有多种生理活性及药理作用, 三萜皂苷类化合物是其主要活性成分之一^[1-3]。目前利用回流浸提法、超声法、微波法等^[4]分离、提纯出刺五加皂苷, 本实验室又针对三萜皂苷合成途径中的关键酶基因进行克隆研究, 成功克隆出 β -香树酯醇合成酶 (β -amyrin synthase, bAS)、3-羟基-3 甲基戊二酰辅酶 A 还原

酶 (3-hydroxy-3-methylglutaryl coenzyme A reductase, HMGR)、法尼基焦磷酸合酶 (farnesyl diphosphate synthase, FPS)、鲨烯合酶 (squalene synthase, SS) 等^[5]基因, 而对于整个基因组的表达规律研究不多, 其基因数据库的资源也较少。

近几年来, 转录组学技术在揭示细胞生理活动规律和生物代谢机制的研究中起着越来越重要的作用, 因此对刺五加转录组的研究势在必行。转录组

收稿日期: 2016-06-20

基金项目: 国家自然科学基金项目 (31570683); 河北省教育厅资助科研项目 (QN2014102); 华北理工大学培育基金 (SP201508)

作者简介: 宋 菊 (1992—), 女, 在读硕士研究生, 研究方向为生物信息在药用植物中的应用。

*通信作者 邢朝斌 (1975—), 男, 教授, 研究方向为分子生药学、药用植物细胞工程。

Tel: (0315)3725859 Fax: (0315)3726341 E-mail: xzbheuu@126.com

是指特定生物体在某种状态下所有基因转录产物的总和, 转录组研究属于功能基因组学研究的范畴, 是连接基因组与蛋白质组的纽带^[6]。通过对同一物种不同样品进行转录组分析, 在能够得到大量功能基因的基础上, 还可获得样品间的表达差异, 这对于揭示基因表达与植物表型的关系具有重要意义^[7]。在植物分子学研究中, 最新的 RNA 测序技术已经取代基因芯片技术, 成为目前从全基因组水平研究基因表达的主流方法。对此, 本实验采用转录组测序技术, 对皂苷量不同的刺五加进行高通量测序、拼接组装, 建立刺五加的转录组数据库, 并针对差异性表达基因进行分析。

1 材料与方法

1.1 材料

以皂苷高含量(质量分数为 3.6%)和低含量(质量分数为 2.1%)组的刺五加叶片[经华北理工大学生命科学院邢朝斌教授鉴定为五加科植物刺五加 *Eleutherococcus senticosus* (Rupr. et Maxim.) Maxim.] 为材料提取总 RNA, 再逆转录为 cDNA, 构建转录组数据库。植物总 RNA 提取试剂盒和逆转录试剂盒购自北京天根生化科技(北京)有限公司。

1.2 转录组数据的组装与分析

1.2.1 转录组数据的组装 采用 Illumina HiSeq 4000 测序技术平台的 PE150 技术进行测序, 将得到的 reads 进行除杂和冗余处理, 获得 clean reads, 利用段序列组装程序 SOAPdenovo^[8]进行拼接, 得到 contig, 再经过双末端测序法组装得到 Unigene, 并通过序列聚类软件 TGICL^[9]做进一步的拼接和去冗余, 得到最终的 Unigene。

1.2.2 转录组功能预测 将最终得到的 Unigene 进行功能注释, 将 Unigene 序列与蛋白数据库 COG、GO、KEGG、Swiss-Prot 和 nr 比对, 得到具有同源信息的 Unigene。COG 是对蛋白质直接进行同源性分类并储存蛋白序列的一个数据库, 通过与之比对, 可预测刺五加可能存在的功能并进行分类统计。

Geneontology (简称 GO), 是一个国际标准化了的基因功能分类体系, 可以对基因和蛋白属性进行全面的描述和限定。为从宏观上认识刺五加的基因功能分布特征, 需根据 NCBI 中的注释信息, 利用 Blast2GO 软件^[10]对 nr 注释信息进行分类, 获得 Unigene 的 GO 注释, 并利用 WEGO 软件^[11]对所有 Unigene 做 GO 功能分类统计。

KEGG 是系统分析基因产物在细胞中的代谢途

径和基因产物功能的数据库, 将刺五加的 Unigene 比对到 KEGG 数据库中, 获得 Unigene 的注释信息, 而后进行 Pathway 分析。

1.3 基因差异性表达分析

1.3.1 差异性表达基因的筛选 利用华大基因研究开发的两个样本间差异性表达基因筛选算法, 算出错误发现率(FDR 值), 并通过基因的表达量(FPKM 值), 计算出该基因在不同样本间的差异表达倍数。FDR 值与差异表达倍数呈反比关系, FDR 值越小, 差异表达倍数越大, 结果的可信度越高。分析时, 将 $FDR \leq 0.001$ 和差异倍数 ≥ 2 作为筛选差异表达基因的标准。

1.3.2 差异性分析 对筛选出的基因进行 GO 功能分析和 KEGG Pathway 分析, 得出 2 个样本间差异表达基因的 GO 功能分类注释和显著性富集分析, 以及 Pathway 富集分析, 可以更好地定量这些基因在刺五加生命活动中行使的主要生物学功能, 显示它们参与的生化代谢途径及信号传导途径。

2 结果与分析

2.1 组装结果

总共得到 8.34 Gb 数据, 组装后得到 contig 3 387 282 条, 总长度 2.08×10^8 nt, 平均长度为 62 nt, N50 长度为 65 nt。进一步去冗余、拼接, 得到最终的 Unigene 77 087 条, 总长度为 52 940 804 nt, 平均长度为 687 nt, 大于 500 nt 的有 26 969 条, 占总数的 34.99%, N50 长度为 1 207 nt。

2.2 功能预测

将最终的 Unigene 通过 Blastx 和 Blastn 与已知的公共数据库 COG、GO、KEGG、Swiss-port、nr 5 个数据库进行比对, 统计获得同源比对信息的 Unigene 分别有 8 084、15 835、6 197、19 450、28 937 条, 分别占总量的 10.47%、20.54%、8.04%、25.23%、37.54%, 总共获得同源比对信息的 Unigene 有 29 348 条, 占 38.07%, 没有获得同源比对信息的有 47 739, 占 61.93%, 获得同源信息的 Unigene 长度大都在 300 nt 以上, 见表 1。

Unigene 与 COG 比对的结果显示, 刺五加数据库中有 8 084 个 Unigene 分布于 25 个 COG 分类中, 其中聚类到一般功能预测, 复制、重组和修复, 转录, 碳水化合物转运和代谢, 翻译、核糖体结构和生物合成等功能基因分布较多。通过与 GO 数据库进行比对, 共得到 71 883 个 GO 功能注释并归类于 52 个 GO 分类中, 其中生物学过程注释最多, 共有 34 002 条 Unigene, 占全部的 47.30%, 其次是分子

功能有 24 212 条，占 33.68%，最少的是细胞组分，有 13 669 条，占 19.02%。这 3 大类又可细分为 52 个功能亚类，生物学过程 20 个亚类，分子功能和细胞组分各 16 个亚类。在生物学过程中，涉及代谢过程、细胞过程和单一有机体过程的基因较多；在分子功能中，涉及催化活性和锚定的基因较多；在细胞组分中，涉及细胞组分、细胞、细胞器和细胞膜

的基因较多。具体种类与数量见图 1。
将测序得到的 Unigene 与 KEGG 数据库进行比对，获得 Pathway 注释信息，比对结果显示，有 6 197 条 Unigene 获得 6 755 个注释（在这之中有多条 Unigene 涉及 2 个及以上的通路），共涉及 116 个 KEGG 标准代谢通路。其中获得注释量较多的前 5 个通路分别是：核糖体途径，共有 341 条，占全部

表 1 Unigene 注释数目统计表
Table 1 Number of annotated Unigene

数据库	注释量/个			占比/%
	总量	300 nt≤长度<1 000 nt	长度≥1 000 nt	
COG	8 084	1 998	5 085	10.47
GO	15 835	5 343	7 231	20.54
KEGG	6 197	2 030	2 950	8.04
Swiss-prot	19 450	6 636	9 163	25.23
nr	28 937	10 818	12 133	37.54
总量	29 348	10 978	12 157	38.07

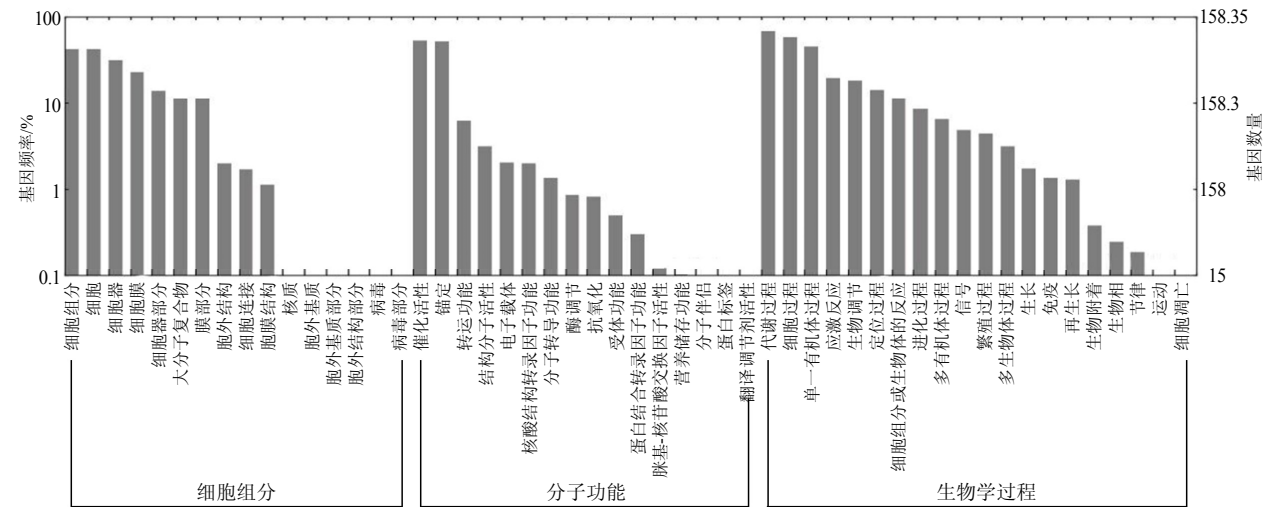


图 1 Unigene GO 功能分类图
Fig. 1 Unigene GO classification

的 5.05%；内质网蛋白加工途径，共有 230 条，占 3.39%；植物激素信号转导途径，有 228 条，占 3.37%；氧化磷酸化途径，有 226 条，占 3.34%；RNA 转运途径，有 212 条，占 3.13%。其余的主要富集于糖酵解、剪接体、嘌呤代谢等代谢途径。

2.3 基因差异性表达分析

利用差异表达基因的筛选方法，筛选出 2 个样本间的差异性表达基因，共 530 条，其中高含量组相对于低含量组上调基因有 223 条，占差异表达基因的 42.08%，下调基因 307 条，占 57.92%（图 2）。根据结果可知，上调和下调的基因相差较大，这些上调和

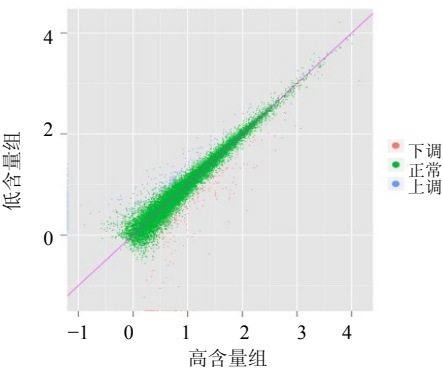


图 2 差异性表达基因
Fig. 2 Differentially expressed genes

下调的基因,为揭示刺五加关键调控基因,以及这些关键调控基因的代谢途径分析提供了可靠的数据。将 530 条差异性表达基因与数据库 COG、GO、KEGG、Swiss-prot 和 nr 进行比对,结果显示有 434 条基因获得注释。其中获得 GO 注释的差异表达基因有 220 条,共获得 408 个注释,分为生物学过程、细胞组分和分子功能 3 类,其中生物学过程得到 211 个功

能注释,占 51.72%,细胞组分得到 55 个功能注释,占 13.48%,分子功能得到 142 个功能注释,占 34.80%。

与 KEGG 数据库进行比对,共有 82 条基因获得注释(在这之中有多条基因涉及两个及以上的通路),共涉及 40 个代谢通路,其中大部分基因归类于遗传信息转导和新陈代谢途径。分类情况见图 3。

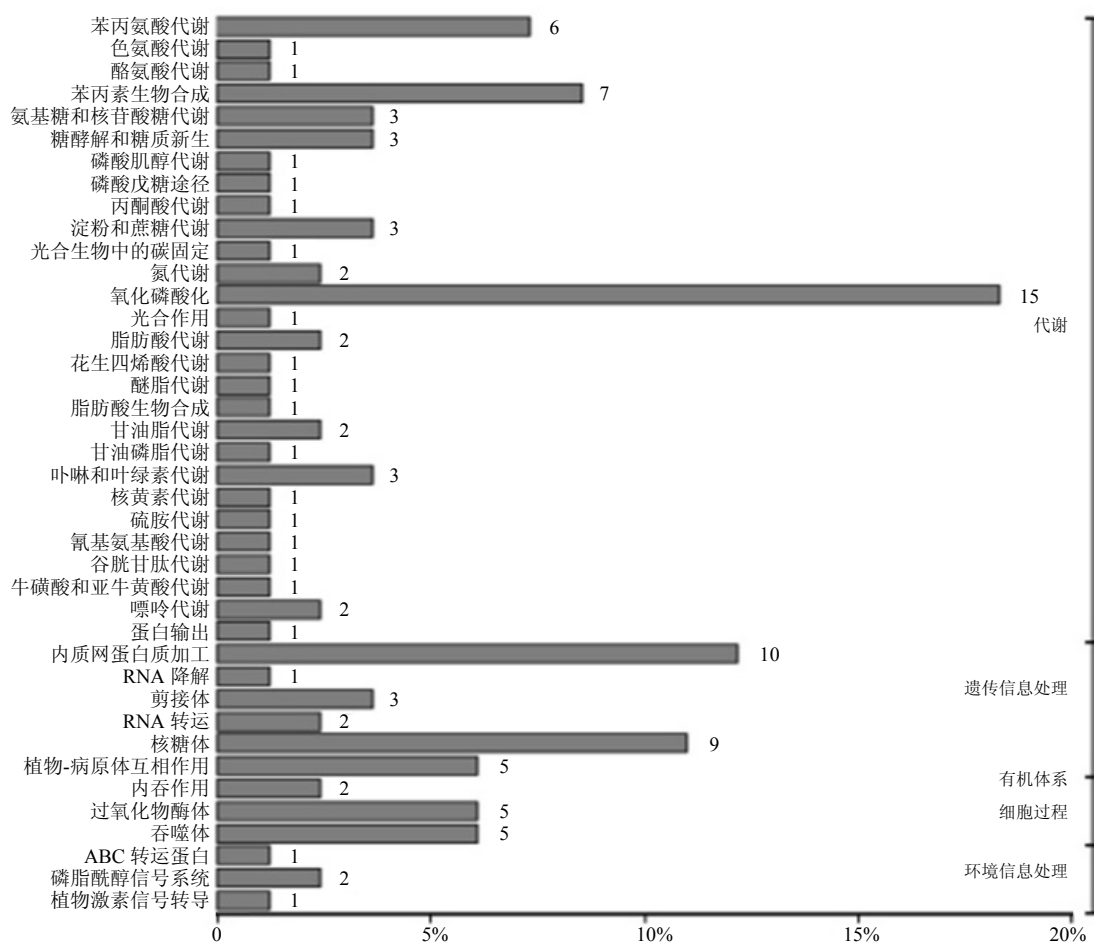


图 3 差异性表达基因 KEGG 分类图

Fig. 3 KEGG classification of differentially expressed genes

显著富集的通路有 3 条,最为富集的是氧化磷酸化途径,有 15 条基因,其次是内质网蛋白加工和核糖体途径,分别有 10 条和 9 条基因。根据皂苷的合成途径^[12]可知,与三萜皂苷合成相关的途径有磷酸戊糖途径、糖酵解和糖质新生、甘油磷脂代谢等途径(表 2),差异性基因共 18 条。

在皂苷合成途径中还需要一些关键酶的参与,这些酶基因的表达量在 2 个样本间存在一定差异,但对皂苷合成量具有重大的影响。法尼基焦磷酸合酶(c44724)催化 2 分子异戊烯基焦磷酸和 1 分子二甲基烯丙基焦磷酸生成法尼基二焦磷酸,为三萜

表 2 刺五加三萜皂苷合成相关基因

Table 2 Saponin biosynthesis related genes in *E. senticosus*

编号	代谢通路	Unigene 数量
Ko00030	磷酸戊糖代谢途径	1
Ko00061	脂肪酸生物合成	1
Ko00380	色氨酸代谢	1
Ko00590	花生四烯酸代谢	1
Ko00360	苯丙氨酸代谢	6
Ko00561	甘油脂代谢	2
Ko00010	糖酵解和糖质新生	3
Ko00564	甘油磷脂代谢	1
Ko00071	脂肪酸代谢	2

皂苷合成提供C15骨架。其表达量分别为高含量组: 263.308 3(相对值); 低含量组: 212.644 2(相对值), 是典型上调基因。鲨烯环氧酶(c49398)、细胞色素P450酶(c18630)以及 β -香树酯醇合成酶(c52011)等关键酶也表达量较高, 都属于上调基因。

3 讨论

利用RNA-Seq技术, 本研究首次获得了刺五加的转录组数据, 得到Unigene 77 087条, 平均长度为687 nt, 组装结果较好。根据GOC、GO和KEGG等的生物信息学结果, 揭示了刺五加转录组的整体表达特征, 并表明本次转录组测序获得较多可用的基因序列。但未获得功能注释的Unigene有47 739条, 占61.93%。造成这种现象的原因可能有2种: (1) 序列太短, 难以获得同源信息, 本转录组组装得到的Unigene长度在500 nt以下的占65.01%, 增加了功能注释和分类的难度; (2) 数据库资源的暂时缺乏, 注释信息不够全面, 使一些表达不够丰富的基因无法得到相应的注释。此现象在已经公布的芝麻^[13]、茶树^[14]和水稻^[15]等5种植物的转录组分析中已经出现过。两个样本间的530条差异性表达基因, 通过GO注释和Pathway富集分析, 证实皂苷高含量组和低含量组的代谢过程转录水平的差异较大。与皂苷合成相关的表达差异较大的基因有18条, 且皂苷合成关键酶基因的表达量存在一定差异, 说明基因表达水平的细微差别即可对生物体产生巨大的影响。

参考文献

- [1] 邢朝斌, 劳风云, 田春迎, 等. 刺五加叶柄的体细胞胚胎发生研究 [J]. 中草药, 2009, 40(8): 1302-1305.
- [2] 王子灿, 乔善义, 马安德, 等. 高效液相色谱-质谱联用技术分析刺五加抗疲劳化学成分 [J]. 第一军医大学学报, 2003, 23(4): 355-357.
- [3] 涂正伟, 周渭渭, 单淇, 等. 刺五加的研究进展 [J].

药物评价研究, 2011, 34(3): 213-216.

- [4] 赵萍, 田晶, 费旭, 等. 刺五加皂苷提取工艺 [J]. 大连工业大学学报, 2013, 32(5): 316-319.
- [5] 邢朝斌, 吴鹏, 修乐山, 等. 刺五加功能基因密码子偏好性分析 [J]. 中草药, 2013, 44(5): 598-603.
- [6] 王东, 曹玲亚, 高建平. 党参转录组中SSR位点信息分析 [J]. 中草药, 2014, 45(16): 2390-2394.
- [7] Jewett M C, Oliveira A P, Patil K R, *et al.* The role of high-throughput transcriptome analysis in metabolic engineering [J]. *Biotechnol Bioprocess Eng*, 2005, 10(5): 385-399.
- [8] Li R Q, Zhu H M, Ruan J, *et al.* De novo assembly of human genomes with massively parallel short read sequencing [J]. *Genome Res*, 2010, 20(2): 265-272.
- [9] Pertea G, Huang X Q, Liang F, *et al.* TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets [J]. *Bioinformatics*, 2003, 19(5): 651-652.
- [10] Conesa A, Gotz S, Garcia-gomez J M, *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research [J]. *Bioinformatics*, 2005, 21(18): 3674-3676.
- [11] Ye J, Fang L, Zheng H, *et al.* WEGO: a web tool for plotting GO annotations [J]. *Nucl Acids Res*, 2006, 34 (Web Server issue): W293-W297.
- [12] 邢朝斌, 王一曼, 陈正恒, 等. 三萜皂苷的生物合成 [J]. 生命化学, 2005, 25(5): 420-422.
- [13] 魏利斌, 苗红梅, 张海洋. 芝麻发育转录组分析 [J]. 中国农业科学, 2012, 45(7): 1246-1256.
- [14] Shi C Y, Yang H, Wei C L, *et al.* Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds [J]. *BMC Genomics*, 2011, 12: 131-149.
- [15] Lu T T, Lu G J, Fan D L, *et al.* Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq [J]. *Genome Res*, 2010, 20(9): 1238-1249.