利用转录组测序挖掘掌叶大黄蒽醌类生物合成相关基因

李　欢[1], 张　娜[1,2], 李依民[1], 黑小斌[1], 李元敏[1], 邓　翀[1],
颜永刚[1], 刘蒙蒙[2*], 张　岗[1*]

(1. 陕西中医药大学药学院/陕西省中药基础与新药研究重点实验室, 陕西 西安 712046;
2. 江苏理工学院电气信息工程学院生物信息与医药工程研究所, 江苏 常州 213001)

摘要: 蒽醌为中药大黄的主要活性成分, 也是大黄质量控制的指标成分。为研究大黄蒽醌类生物合成通路, 用 Illumina HiSeq™ 2000 150PE 对掌叶大黄幼苗转录组文库进行高通量测序, 得到 11.04 G 数据, 736 309 74 条高质量 reads (SRA 数据库注册号 SRP160030)。Trinity *do novo* 组装产生 93 646 个 unigenes, 平均长度 1 108 nt。功能注释表明所有 unigenes 在 NR、NT、Swiss-port、PFAM、KOG 等数据库得到注释, 可归为 GO 分类的生物过程、细胞组分和分子功能 3 大类 57 分支, KEGG 分析发现 1 107 条 unigenes 参与 19 个次生代谢标准通路。172 条 unigenes 编码蒽醌类生物合成相关的 MVA、MEP、莽草酸及聚酮途径 28 个关键酶。125 条 CYP450 基因可能参与次生代谢物的修饰, 73 条与糖基转移酶相关。RT-PCR 和测序成功验证 7 个蒽醌及黄酮类候选全长 unigenes。MISA 还发现 18 885 个 SSRs。该研究首次获得掌叶大黄幼苗转录组基因表达特征及蒽醌类生物合成通路基因, 为后续基因功能鉴定、次生代谢途径解析及蒽醌类生物合成与调控分子机制研究提供基础资料。

关键词: 掌叶大黄; 转录组; 蒽醌; 基因; 代谢通路

中图分类号: R931　　**文献标识码**: A　　**文章编号**: 0513-4870 (2018) 11-1908-10

High-throughput transcriptomic sequencing of *Rheum palmatum* L. seedlings and elucidation of genes in anthraquinone biosynthesis

LI Huan[1], ZHANG Na[1,2], LI Yi-min[1], HEI Xiao-bin[1], LI Yuan-min[1], DENG Chong[1],
YAN Yong-gang[1], LIU Meng-meng[2*], ZHANG Gang[1*]

(*1. College of Pharmacy and Shaanxi Provincial Key Laboratory for Chinese Medicine Basis & New Drugs Research, Shaanxi University of Chinese Medicine, Xi'an 712046, China;　2. Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, China*)

Abstract: Anthraquinones are not only the main active constituents but also the index components for the quality control of Rhei Radix et Rhizoma. To study the anthraquinone biosynthesis, *Rheum palmatum* L. seedlings were subjected to a high-throughput transcriptomic sequencing analysis by Illumina HiSeq™ 2000 150PE. The Illumina sequencing generated a total of 11.04 G clean data resulting in 736 309 74 clean reads, deposited in the sequence read archive (SRA accession SRP160030). Trinity *do novo* assembly yielded 93 646 unigenes, with an average of 1 108 nt. Functional annotation revealed that all unigenes were successfully annotated in the NR, NT, Swiss-port, PFAM, and KOG databases. GO enrichments showed that 57 subgroups were involved in biological process, cellular component, and molecular function. KEGG analysis indicated that

收稿日期: 2018-06-11;　修回日期: 2018-09-13.

基金项目: 陕西省高校青年杰出人才支持计划项目; 江苏省自然科学基金资助项目 (BK20170311); 咸阳市中青年科技领军人才项目; 陕西中医药大学新进博士科研启动经费 (104080001).

*通讯作者 Tel / Fax: 86-29-38185165, E-mail: jay_gumling2003@aliyun.com; mmliu1987@sina.com

DOI: 10.16438/j.0513-4870.2018-0547

1 107 unigenes were implicated in 19 standard secondary metabolic pathways. 172 unigenes were analyzed to encode 28 key enzymes during the MVA, MEP, shikimic acid, and polyketide pathways related to anthraquinone biosynthesis. 125 CYP450 and 73 UGTs unigenes were related the modification of secondary metabolites in *R. palmatum* L. Furthermore, seven unigenes with full length cDNAs were successfully verified by RT-PCR and sequencing analyses. Then, MISA prediction produced a number of 18 885 simple sequence repeats (SSRs). Herein, the transcriptomic gene expression profiles of *R. palmatum* L. and candidate genes during the anthraquinone biosynthesis pathway were obtained for the first time. The results provided basic information for subsequent gene function characterization, secondary metabolic pathway analysis, and anthraquinone biosynthesis and regulation elucidation in *R. palmatum* L.

Key words: *Rheum palmatum* L.; transcriptome; anthraquinone; gene; metabolism pathway

掌叶大黄 *Rheum palmatum* L.、唐古特大黄 *R. tanguticum* Maxim. ex Balf. 和药用大黄 *R. offcinale* Baill.为蓼科大黄属多年生高大草本，为 2015 版《中华人民共和国药典》收录的大黄三基源，其干燥根及根茎药用，始载于《神农本草经》，性寒、味苦，具有泻下攻积、清热泻火、凉血解毒、逐瘀通经、利湿退黄之功效，常用于实热积滞便秘、血热吐衄、目赤咽肿等症的临床治疗[1]。现代研究揭示大黄主要含有蒽醌类、酚类、萘苷类、酰基糖苷类、二苯乙烯类及鞣质等多种成分，蒽醌类化合物为其主要活性成分，具有抗肿瘤[2]、抗炎[3]及保肝[4]等药理活性。以大黄药材或蒽醌类成分为原料生产的药物就有数百种，大黄的临床应用非常广泛。

作为正品大黄的一个重要来源，掌叶大黄资源丰富，质量最佳，产量、市场份额显著高于其他两种。当前，大量研究已在掌叶大黄分子鉴别、质量评价、活性成分等方面取得重要进展。叶绿体基因 *matK* 序列的差异可在分子水平鉴别掌叶大黄真伪[5]。掌叶大黄质量与产地、生长年限、气压、湿度或温度等生态因素密切相关[6,7]。在生长发育进程中，蒽醌类在掌叶大黄根中的累积含量大于叶柄和叶片[8]。这些研究说明植物生理生态调控作用在大黄活性成分合成与积累方面起重要作用。然而，目前尚不清楚大黄蒽醌类代谢通路及调控机制，就无法科学阐明大黄药材品质形成的内在生物学规律。蒽醌类主要成分又是大黄药材质量控制的关键指标，利用现代分子生物学技术解析大黄蒽醌类生物合成通路及表达调控就显得更为迫切。

转录组测序技术是基因组与蛋白组的桥梁，能够从整体水平研究特定生物样本在某一生理状态下所有基因转录本全局信息，揭示特定条件下生物体生长发育、次生代谢及生理适应等方面的分子机制[9]。随着本草基因组学的迅速发展，基于高通量测序技术的转录组分析策略广泛用于药用植物功能基因组研究，现已解析柴胡[10]、人参[11]和商陆[12]等众多大宗药材的转录组特征，为阐明中药种质资源遗传基础奠定了重要基础。本研究在明确掌叶大黄幼苗活性成分含量的基础上，利用二代高通量测序平台 Illumina HiSeq^{TM} 2000 150PE 进行幼苗转录组测序分析，挖掘蒽醌类的生物合成通路及调控方面的遗传信息，为掌叶大黄品质形成机制及大黄质量控制研究提供科学依据。

材料与方法

植物材料　掌叶大黄种子系课题组 2017 年 11 月采自甘肃省宕昌县阿坞乡麻界村，东经 104°10'6.5"、西经 34°16'51.98"，海拔 2 377 m，经陕西中医药大学张岗教授鉴定为掌叶大黄 *R. palmatum* L.。将大黄种子用沙土 (1∶4) 栽种至花盆中，置于 20±2 ℃的温室，光照 16 h，黑暗 8 h，培养 2 个月取幼苗备用。

仪器和试药　Waters 2695 高效液相色谱仪，包括四元超高压溶剂系统、自动进样恒温样品管理器，Waters 2998 PAD 检测器，Empower 3 色谱工作站 (Waters, USA); GB204 型电子分析天平 (北京赛多利斯); KQ-200KED 超声波清洗机 (江苏昆山); GZX-9140MBE 电热鼓风干燥箱 (上海博迅)。

对照品没食子酸 (批号 122811)、儿茶素 (11c15)、番泻苷 B (11z15) 购自天津西玛科技有限公司；大黄素 (110795-200505)、大黄酸 (0757-200206)、大黄素甲醚 (110758-200610)、大黄酚-8-*O*-葡萄糖苷 (110796-200615)、大黄素-8-*O*-葡萄糖苷 (10756-200110) 购自中国食品药品检定研究院。大黄酚 (B2038) 和芦荟大黄素 (B20772) 购自上海源叶生物科技有限公司。色谱甲醇购自上海泰坦科技有限公司。娃哈哈纯净水购自杭州娃哈哈集团有限公司。其他试剂均为国产分析纯。

掌叶大黄 HPLC (high performance liquid chromatography) 含量测定 取掌叶大黄幼苗若干株, 随机选择 3 株混合, 3 次平行取样。常规方法烘干, 参照课题组前期构建大黄 HPLC 含量测定方法[13]进行。色谱柱为武本 C18 (5 μm, 4.6 mm×250 mm) 色谱柱; 流动相由甲醇 (A) 和 0.2%磷酸水 (B) 组成, 梯度洗脱 (0～5 min, 5%～15% A; 5～15 min, 15%～30% A; 15～25 min, 30%～35% A; 25～31 min, 35%～42% A; 31～46 min, 42%～53% A; 46～66 min, 53%～68% A; 66～75 min, 68%～100% A; 75～85 min, 100% A), 检测波长 260 nm, 柱温 30 ℃, 体积流量 1.0 mL·min^{-1}。进样量为 10 μL。在上述色谱条件下分析, 理论板数按各个成分计算均不低于 5 000, 与相邻组分峰的分离度均大于 1.5, 色谱峰对称因子均在 0.95～1.05。

对照品制备: 精密称取没食子酸、儿茶素、番泻苷 B、大黄酚-8-*O*-葡萄糖苷、大黄素-8-*O*-葡萄糖苷、芦荟大黄素、大黄酸、大黄素、大黄酚和大黄素甲醚等 10 种对照品适量, 分别置于 10 mL 容量瓶中, 用甲醇溶解并稀释至刻度, 摇匀, 质量浓度分别为 0.224、0.71、0.45、0.172、0.234、0.079、0.077、0.028、0.048、0.027 mg·mL^{-1} 的对照品储备液。分别精密量取各对照品储备液 1 mL, 甲醇稀释 10 倍, 得到相应质量浓度的混合对照品储备液。4 ℃保存备用。

供试品制备: 取掌叶大黄幼苗烘干, 称重, 分别研磨粉碎混匀, 取 0.1 g, 精密称定, 置于 50 mL 锥形瓶中, 精密加入甲醇 4.5 mL, 称重。超声处理 30 min (功率 500 W, 频率 40 kHz) , 放至室温, 补足失量, 10 500 r·min^{-1} 离心 12 min, 取上清液, 过 0.22 μm 微孔滤膜, 待测。

转录组测序分析 委托北京诺禾致源科技股份有限公司利用 Illumina HiSeq™ 2000 150PE 进行掌叶大黄幼苗转录组高通量测序分析。

取掌叶大黄新鲜幼苗全株, 采用 EASYspin 植物 RNA 快速提取试剂盒 (Aidlab, China) 制备总 RNA, NanoDrop™ 2000 分光光度计 (Thermo Fisher, USA) 检测完整性。用带有 Oligo (dT) 的磁珠富集 mRNA, 加入 fragmentation buffer 将 mRNA 打断成短片段, 用六碱基随机引物 (random hexamers) 合成 cDNA 第一链; 然后加入缓冲液、dNTPs、RNase H 和 DNA polymerase I 合成 cDNA 第二链; 再经过 QiaQuick PCR 试剂盒 (QIAGEN, Germany) 纯化并加 EB 缓冲液洗脱之后做末端修复、加 poly (A) 并连接测序接头, 然后用琼脂糖凝胶电泳进行片段大小选择, 最后进行 PCR 扩增构建测序文库。

测序原始图像数据经 base calling 转化为序列数据 raw reads, 经数据评估、过滤除杂和冗余处理等质控得到 clean reads, 再利用 Trinity 做转录组 *de novo* 组装分析。Trinity 首先将具有一定长度 overlap 的 reads 连成更长的片段, 这些通过 reads overlap 关系得到的不含 N 的组装片段作为组装出来的 unigenes。

利用 BLAST 将 unigenes 序列与蛋白数据库 NR、NT、Swiss-port、PFAM、KOG (Cluster of Orthologous Groups of proteins、蛋白相邻类的聚簇) 和 KEGG (京都基因与基因组百科全书) 进行比对 (E 值<1×10^{-5}), 得到与相应 unigenes 具有最高序列相似性的蛋白, 进而得到 unigenes 注释信息。根据 NR 注释信息, 使用 Blast2GO 软件得到 unigenes 的 GO (gene ontology) 注释, 用 WEGO 软件对所有 unigenes 做 GO 功能分类统计, 从宏观上认识该物种的基因功能分布特征。使用 MISA 检测掌叶大黄转录组 unigenes, 搜索 SSRs (simple sequence repeats) 并进行统计分析。

蒽醌生物合成基因筛选 蒽醌类生物合成主要由甲羟戊酸 (MVA)、2-*C*-甲基-*D*-赤藓醇 4-磷酸 (MEP)、莽草酸 (shikimate) 和聚酮 (polyketide) 途径介导的骨架基因, 与衍生化修饰基因经细胞色素 (cytochrome P450, CYP450)、尿苷二磷酸−糖基转移酶 (UDP-glycosyltransferases, UGT) 等共同参与完成[14, 15]。对这些基因的筛选主要根据 unigenes 在 Nr 数据库中的注释结果手动检索, 以 FPKM (expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) 计算基因表达量。

全长 ungienes 基因的 RT-PCR 验证 选择蒽醌及黄酮类生物合成 7 个全长候选 unigenes, 包括乙酰-CoA 乙酰转移酶 (acetyl-CoA *C*-acetyltransferase, AACT)、MVA 激酶 (mevalonate kinase, MK)、1-脱氧-*D*-木酮糖 5-磷酸合成酶 (1-deoxy-*D*-xylulose-5-phosphate synthase, DXS)、2-甲基-*D*-赤藓糖醇-4-磷酸胞苷酰基转移 (2-*C*-methyl-*D*-erythritol 4-phosphate cytidylyltransferase, MCT)、苯丙氨酸解氨酶 (phenylalanine ammonialyase, PAL)、肉桂酸-4-羟基化酶 (cinnamate 4-hydroxylase, C4H)、黄烷酮-3-羟化酶 (flavanone-3-hydroxylase, F3H) 等, 利用 Primer 3 分别设计扩增 ORF 的特异引物 (表 1), 采用 RT-PCR 验证。标准 PCR 体系包含以下组分: 10× *Ex Taq* Buffer (20 mmol·L^{-1}) 2.0 μL, dNTP Mixture (各 2.5 mmol·L^{-1}) 1.6 μL, cDNA Template 0.5 μL, 正反向引物各 0.5 μL,

Table 1　Seven candidate genes and ORF amplification primers

Unigene	Size/bp	Gene symbol	Primer sequences 5'−3'	ORF size/bp
Cluster-5019.54375	1 754	AACT	S: ATGGCGGTAGAGAATTCTTCCAG AS: TCATAGCTTTGAGTGACCTGACCA	1 245
Cluster-3447.0	1 568	MK	S: ATGGAGGTAACTGCGAGAGCTC AS: TCAGGAGCCAAAGCAGATTTC	1 158
Cluster-21445.0	4 502	DXS	S: ATGAGCGCTGCTCCTATCGA AS: TCAGCACATCAAGAGAAGTGCTT	2 163
Cluster-16015.0	1 239	MCT	S: ATGGGAGTATTAGGAATGGAGCAG AS: TTATGAGCTTAAATTCAGTATCCTCTCA	930
Cluster-8002.0	2 635	PAL	S: ATGGAGATCGCAAACGG AS: CTAGCAAATAGGAAGAGGAGCA	2 178
Cluster-5019.18786	1 774	C4H	S: ATGGATTTGGTTCTGCTCC AS: TTAGAAGCTCCTGGGCTTC	1 518
Cluster-5019.36107	1 485	F3H	S: ATGGCGCCGGCAGCA AS: TCAAGCAAGGATATCATCAATGGT	1 095

TaKaRa *Ex Taq* (5 U·μL^{-1}) 0.1 μL, 补 ddH_2O 至 25 μL。PCR 程序为: 95 ℃ 3 min, 95 ℃ 30 s, 60 ℃ 30 s, 72 ℃ 2.0 min, 36 个循环; 72 ℃ 10 min, 12 ℃保温。PCR 产物经电泳分析, 送武汉金开瑞生物工程有限公司测序。

结果与分析

1　掌叶大黄幼苗 HPLC 含量测定

掌叶大黄幼苗 HPLC 色谱分析结果 (图 1) 表明, 10 种标准品在掌叶大黄幼苗中均能检测到, 各标准品含量存在明显差异。全株植物中各标品含量依次为: 儿茶素 > 大黄素-8-*O*-葡萄糖苷 > 大黄酚-8-*O*-葡萄糖苷 > 大黄素 > 大黄酚 > 番泻苷 B > 大黄素甲醚 > 没食子酸 > 芦荟大黄素 > 大黄酸。其中, 大黄特征性成分大黄素、大黄酚、大黄素甲醚、芦荟大黄素、大黄酸的含量分别为 0.021%、0.016%、0.014%、0.002% 和 0.001%, 说明掌叶大黄生长发育过程中次生代谢途径比较活跃。

2　转录组数据组装与质量分析

掌叶大黄幼苗转录组测序得到 74 638 394 条 raw reads, 过滤产生了 73 630 974 条高质量 clean reads, 包含 11.04 Gb 个核苷酸信息, 将原始数据提交 SRA 数据库获得注册号 SRP160030。Q20 和 Q30 分别为 96.43%、90.85%, GC 量为 49.12%, 说明测序质控良好, clean reads 质量合格。Trinity 无参组装获得 93 646 个 unigenes, 平均长度 1 108 nt, 最长达到 14 605 nt, 最短序列为 201 bp, N50 为 1 794 nt。Unigenes 长度分布显示, 51 280 条 unigenes 长度超过 1 000 nt, 205 86 条序列大于 2 000 nt。

3　转录组 unigenes 的功能注释

使用 BLAST 将所有 unigenes 与 NR、Swiss-port、KOG、KEGG 等数据库进行一致性比对分析, 对各数据库注释的 unigenes 数目进行统计, 进而获得掌叶大黄转录组 unigenes 的功能注释信息。结果表明,

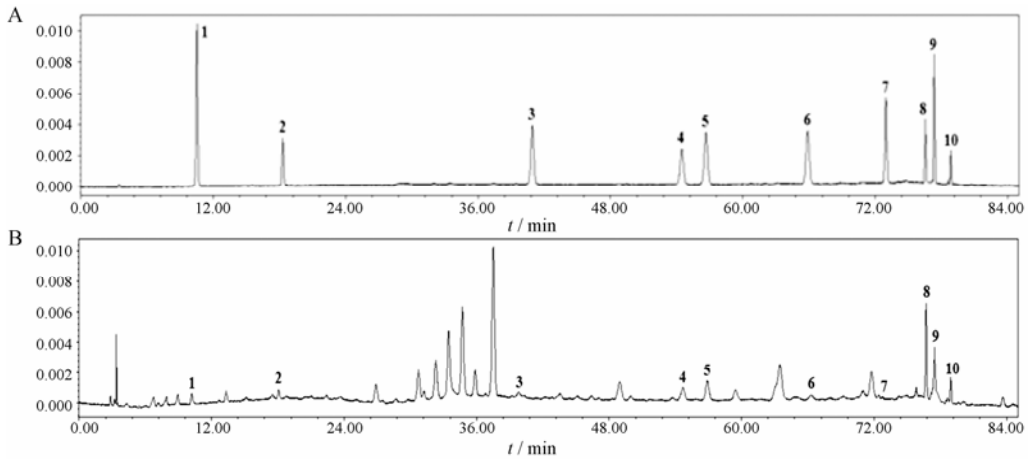

Figure 1　HPLC analyses of mixed reference substances (A) and *R. palmatum* L. (B) seedlings.　1: Gallic acid; 2: Catechin; 3: Senna glycoside B; 4: Chrysophanol-8-*O*-glucoside; 5: Emodin-8-*O*-glucoside; 6: Aloe-emodin; 7: Rhein; 8: Emodin; 9: Chrysophanol; 10: Physcion

67 076 条 unigenes (71.62%) 在 NR 数据库比对成功得到注释，在 KEGG 中注释 28 158 个 (30.06%)，在 GO、KOG 等数据库获得注释的 unigenes 数目依次为 49 885 (53.26%)、20 362 (21.74%)。10 444 条 unigenes 同时在所有数据库中注释，至少有一种数据库注释成功的 unigenes 共 71 304 条 (76.14%)。

以 NR 数据库为例进行分析，unigenes 注释同源基因的物种分布如图 2 所示，在相似序列匹配度较高的物种中，甜菜 *Beta vulgaris* L. 所占比例最高，18 390 条 (27.6%)；其次为葡萄 *Vitis vinifera* L.，9 256 条 (13.9%)，可可豆 *Theobroma cacao* L. 2 899 条 (3.6%)，麻风树 *Jatropha curcas* L. 2 397 条 (3.6%)，莲 *Nelumbo nucifera* Gaertn. 2 395 条 (3.6%)，其余匹配物种比例在 4.3% 以下，unigenes 小于 1 000 条比例小于 3.2%，16 616 条占 27.4%。

根据 NR 注释信息得到 GO 功能分类 (图 3)，

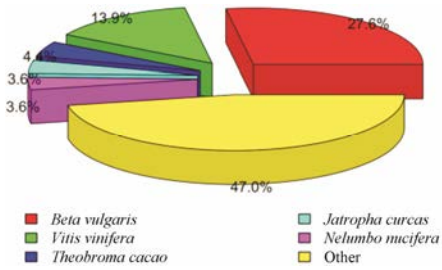

Figure 2　Species distribution of transcriptomic unigenes against NR database

49 885 个 unigenes 被注释到生物过程、细胞组分和分子功能 3 个 GO 类别的 57 个小组。细胞组分中细胞 (cell) 和细胞部分 (cell part) 相关基因丰度最高，达 15 483 和 15 480 条；其次是细胞器 (organelle)，有 10 348 条；共质体 (symplast) 基因较少，有 11 条。生物过程主要聚集在代谢过程 (metabolic process) 和细胞过程 (cellular process)，涉及的基因分别有 26 898 条和 29 119 条；单组织过程 (single-organism process) 和生物调节 (biological regulation) 基因数量分别为 21 056、9 917 条。分子功能中具有催化活性 (catalytic activity) 和结合功能的基因 (binding) 数量较高，分别为 22 942 和 29 184 条，其他类别基因数目普遍较少。

进一步进行 KOG 功能分类 (图 4)，共得到 25 个不同的 KOG 功能类群种类比较全面，包括大多数的生命活动；翻译后修饰，蛋白反转、伴侣最多，有 2 708 条；一般功能预测的基因数量次之，为 2 585 条；翻译、核糖体结构和生源 unigene 数目 1 704 条；其他种类基因丰度不尽相同。掌叶大黄转录组 unigenes 参与 KEGG 代谢通路 (图 5) 分为五大分支：细胞过程 A (cellular processes) 1 520 条、环境信息处理 B (environmental information processing) 1 060 条、遗传信息处理 C (genetic information processing) 6 086 条、代谢 D (metabolism) 11 941 条和有机系统 E (organismal systems) 832 条；19 条子通路，其中碳水化合物代谢最多 1 097 条，单萜生物合成最少为 1 条。

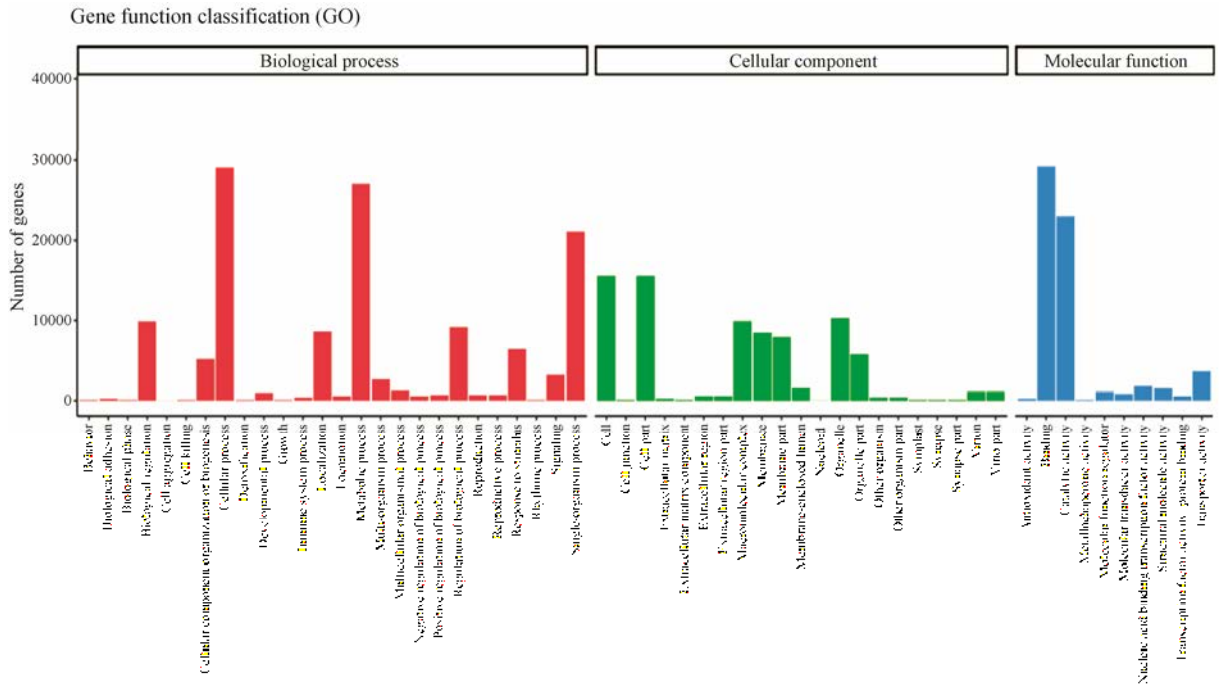

Figure 3　GO classification of transcriptomic unigenes

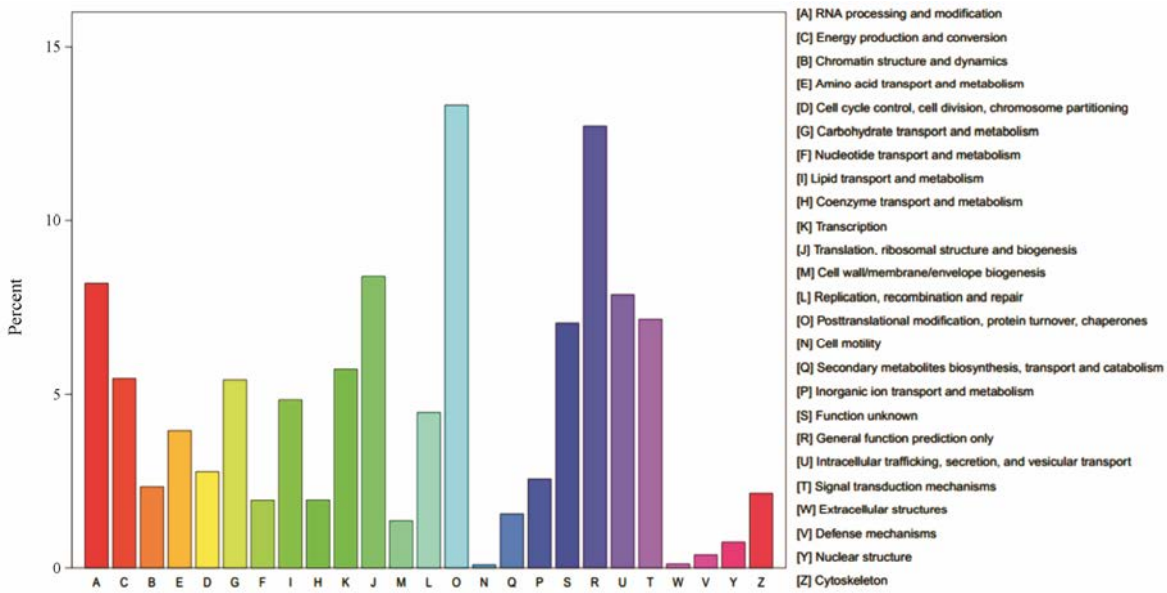

Figure 4　KOG annotation distribution of transcriptomic unigenes

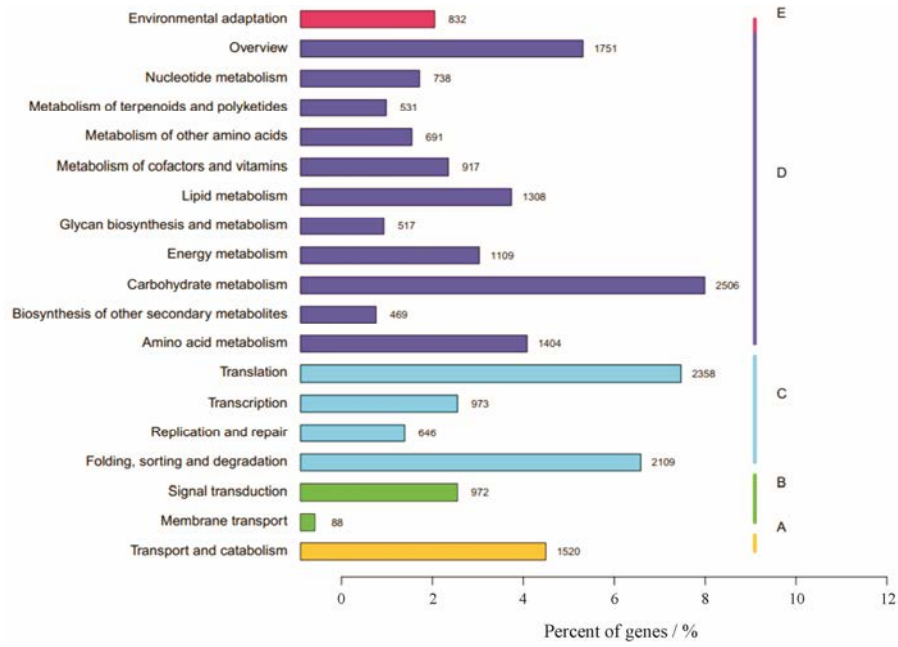

Figure 5　KEGG classification of assembled unigenes

KEGG 代谢通路分析还发现 1 107 条 unigenes 参与类胡萝卜素、苯丙素类、玉米素、生物碱、黄酮类、花青素等生物合成相关的 19 个次生代谢标准通路 (表 2)。其中, 苯丙素的生物合成 (ko00940) 基因数量最多, 为 233 个; 萜类化合物骨架生物合成 (ko00900) 基因数量次之, 为 201 条; 与类胡萝卜素的生物合成代谢通路 (ko00906) 有关的基因有 183 条; 有 77 个 unigenes 与玉米素的生物合成 (ko00908) 相关; 类黄酮的生物合成 (ko00941) 基因 70 条; 异喹啉生物碱 (ko00950) 基因有 55 条; 花青素生物合成、咖啡因的代谢、黄酮和黄酮醇的生物合成、甜菜红色素、吲哚生物碱的生物合成、异黄酮类的生物合成通路基因数较少。

4　蒽醌合成相关基因

4.1　蒽醌骨架合成基因　参与大黄蒽醌类骨架生物合成 MVA、MEP、莽草酸及聚酮途径候选基因见表 3。从表 3 可以看出, 这 4 个途径关键酶基因表达数量和表达量存在差异。MVA 途径关键酶有乙酰-CoA 乙酰转移酶 (AACT)、HMG-CoA 合酶 (HMGS)、HMG-CoA 还原酶 (HMGR)、MVA 激酶 (MK)、MVP

Table 2 Secondary metabolism KEGG pathway analysis of transcriptomic unigenes

No.	KEGG pathway	No. of unigenes	Percentage/%	KEGG ID
1	Phenylpropanoid biosynthesis	233	21.04	ko00940
2	Ternenoid backbone biosynthesis	201	18.16	ko00900
3	Carotenoid biosynthesis	183	16.53	ko00906
4	Zeatin biosynthesis	77	6.95	ko00908
5	Flavonoid biosynthesis	70	6.32	ko00941
6	Tropane, piperidine and pyridine alkaloid biosynthesis	65	5.87	ko00960
7	Isoquinoline alkaloid biosynthesis	55	4.97	ko00950
8	Stilbenoid, diarylheptanoid and gingerol	49	4.42	ko00945
9	Limonene and pinene degradation	46	4.15	ko00903
10	Diterpenoid biosynthesis	45	4.07	ko00904
11	Sesquiterpeniod and triterpenoid biosynthesis	36	3.25	ko00909
12	Brassinosteroid biosynthesis	15	1.35	ko00905
13	Anthocyanin biosynthesis	10	0.90	ko00942
14	Caffeine metabolism	8	0.72	ko00232
15	Flavone and flavonol biosynthesis	6	0.54	ko00944
16	Indole alkaloid biosynthesis	3	0.27	ko00901
17	Betalain biosynthesis	2	0.02	ko00965
18	Flavonoid biosynthesis	2	0.02	ko00943
19	Monoterpenoid biosynthesis	1	0.01	ko00902

Table 3 Unigenes involved in anthraquinone biosynthesis

Pathway	Gene name	Enzyme symbol	EC	No. of unigenes	Average FKPM
MVA	Acetyl-CoA acetyltransferase	AACT	2.3.1.9	11	4.14
	HGM-CoA synthase	HMGS	2.3.1.10	9	3.89
	HGM-CoA reductase	HMGR	1.1.1.88	3	14.38
	MVA kinase	MK	2.7.1.36	16	1.91
	MVP kinase	PMK	2.7.4.2	6	2.19
	MVPP decarboxylase	MPD	4.1.1.33	9	2.48
	IPP isomerase	IPPs	5.3.3.2	3	29.15
MEP	1-Deoxy-*D*-xylulose-5-phosphate synthase	DXS	2.2.1.7	5	1.13
	1-Deoxy-*D*-xylulose-5-phosphate reductoisomerase	DXR	1.1.1.267	1	0.85
	2-*C*-Methyl-*D*-erythritol 4-phosphate cytidylyltransferase	MCT/ispD	2.7.7.60	2	0.09
	4-Diphosphocytidyl-2-*C*-methyl-*D*-erythritol kinase	CMK/ispE	2.7.1.148	6	4.10
	2-*C*-Methyl-*D*-erythritol 2,4-cyclodiphosphate synthase	MDS/ispF	4.6.1.12	2	48.85
	4-Hydroxy-2-methylbut-2-en-1-yl diphosphate synthase	HDS/ispG	1.17.7.3	6	3.62
	4-Hydroxy-3-methylbut-2-enyl diphosphate reductase	HDR/ispH	1.17.7.4	5	87.91
Shikimate	3-Deoxy-*D*-arabino-heptulosonate 7-phosphate synthase	DAHPS	2.5.1.54	10	4.26
	3-Dehydroquinate synthase	DHQS	4.2.3.4	2	1.63
	Shikimate dehydrogenase	SDH	1.1.1.25	13	2.91
	Shikimate kinase	SMK	2.7.1.71	9	5.08
	3-Phosphoshikimate 1-carboxyvinyltransferase	EPSPs	2.5.1.19	2	14.64
	Chorismate synthase	CS	4.2.3.5	1	34.24
	Isochorismate synthase	IS/MenF	5.4.4.2	7	1.07
	O-succinylbenzoic acid coa ligase	MenE	6.2.1.26	15	5.50
	Naphthoate synthase	MenB	4.1.3.36	2	3.24
Polyketide	Chalcone synthase	CHS	2.3.1.74	12	5.49
	Polyketide synthase	PKS	−	1	2.68
	Polyketide synthase Ⅲ	PKSⅢ	−	3	1.71
	Polyketide cyclase	PKC	−	4	16.61

激酶 (PMK)、MVPP 脱羧酶 (MPD) 和 IPP 异构酶 (IPPs), 共 57 条 unigenes。其中, 编码 MK unigenes 数量最多, 为 16 条, 但平均表达量不高, 仅 1.91; 3 条 unigenes 编码 IPPs, 平均表达量最高, 达到 29.15。

33 条 unigenes 编码 MEP 途径关键酶, 包括 1-脱氧-*D*-葡萄糖-5-磷酶合成酶 (DXS)、1-脱氧-*D*-葡萄糖-5-磷酶还原异构酶 (DXR)、2-甲基-*D*-赤藓糖醇-4-磷酸胞苷酰基转移酶 (MCT/ispD)、4-二磷酸胞苷-2-甲基-*D*-赤藓糖醇激酶 (CMK/ispE)、4-羟基-3-甲基丁烯-2-烯基-1-二磷酸合酶 (ispG/HDS)、2-*C*-甲基-*D*-赤藓糖醇-1,4-环磷酸二磷酸合成酶 (MDS/ispF) 和 4-羟基-3-甲基丁烯-2-烯基-1-二磷酸还原酶 (HDR/ispH)。HDR/ispH 序列 5 条, 平均表达量最高, 为 87.91; MDS/ispF 平均表达量次之, 为 48.85; 其余基因表达水平较低, MCT/ispD 平均表达量仅 0.09。

61 条 unigenes 编码莽草酸途径关键酶, 包括 3-脱氧-*D*-阿拉伯庚酮糖-7-磷酸合成酶 (DAHPS)、3-脱氧奎尼酸合成酶 (DHQS)、莽草酸脱氧酶 (SDH)、莽草酸激酶 (SMK)、EPSP 合成酶 (EPSPs)、分支酸合成酶 (CS)、异分支酸合成酶 (IS/MenF)。其中, CS 表达量为 34.24, IS/MenF 平均表达量仅 1.07。2-琥珀酰-5-烯醇丙酮酸-6-羟基-3-环己烯-1-羧酸合酶 (MenD)、2-琥珀酰-6-羟基-2,4-环己二烯-1-羧酸合酶 (MenH)、2-琥珀酰苯甲酸合酶 (MenC) 没有被注释。

聚酮途径 unigenes 共 21 条, 分别编码聚酮合成酶 (PKS)、Ⅲ型聚酮合成酶 (PKSⅢ)、聚酮环化酶 (PKC) 及查尔酮合酶 (CHS)。PKC 平均表达量最高, 为 16.61, PKSⅢ平均表达量最低仅 1.71, CHS unigenes 数量多达 12 条, 仅 1 条 PKS 基因序列。

4.2　蒽醌修饰相关基因　根据 NR 注释结果, 共找到 125 条 CYP450 基因, 隶属于 22 个 CYP450 家族。属于 CYP71 家族的 unigenes 最多, 有 25.40%; 其次是 CYP94、CYP87 和 CYP76, 分别为 15.08%、10.32% 和 7.2%。而 CYP89、CYP81 家族成员最少, 仅各有 1 个 unigene。共找到属于 13 个 UGT 亚家族的 73 个 UGTs, 其中包括 21 个 UGT80、2 个 UGTPg19、10 个 UGT71、3 个 UGT70、15 个 UGT92、2 个 UGT84、3 个 UGT7、2 个 UGT89、4 个 UGT72、1 个 UGTPg26、4 个 UGT88、1 个 UGTPg35、4 个 UGT85。

5　SSRs 分析

利用 MISA 软件对转录组 unigenes 进行 SSRs 分析 (表 4), 93 646 个 unigenes 中共计 18 885 个 SSRs。其中, 单碱基重复 SSR 数量最丰富, 有 7 542 个 (39.94%), 其中 A 重复最多。二碱基重复 SSRs 数量次之, 有 5 714 个, 占 SSRs 总量的 30.26%, 其中 AG/CT 重复类型数量最多。四碱基和五碱基重复分别为 249 和 82 个, 各占 1.32%、0.43%; 六碱基重复相对较少, 仅占 0.41%。此外, 还发现 SSRs 重复单元数量存在一定变化。

6　全长 unigenes 基因的 RT-PCR 验证

利用表 1 中的 7 对 ORF 扩增引物进行 RT-PCR 分析, 图 6 结果表明, 7 个基因特异引物均产生目标条带。进一步 PCR 产物直接测序分析表明, 这 7 个基因与原 unigenes ORF 一致, 说明这些候选 unigenes 为全长基因。

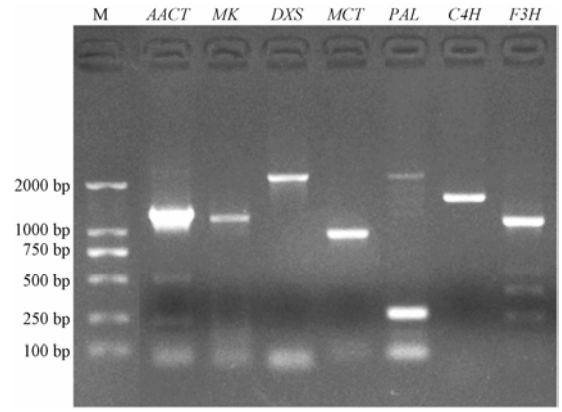

Figure 6　Agarose gel electrophoresis of seven candidate genes amplified by RT-PCR

Table 4　SSRs analysis of transcriptomic unigene

Type	Repeat number												Total	Percentage/%
	5	6	7	8	9	10	11	12	13	14	15	≥16		
mono	0	0	0	0	0	3 614	1 352	635	427	234	273	1 001	7 542	39.94
di	0	1 996	1 199	717	490	259	245	182	107	75	69	375	5 714	30.26
tri	3 007	1 224	458	285	77	63	28	14	12	19	2	32	5 221	27.65
tetra	170	49	23	5	1	0	0	1	0	0	0	0	249	1.32
path	55	26	1	0	0	0	0	0	0	0	0	2	82	0.43
hex	35	23	12	6	0	1	0	0	0	0	0	0	77	0.41
Total	3 267	3 318	1 693	1 013	568	3 937	1 625	832	546	328	344	1 410	18 885	100.0

讨论

随着高通量测序技术的不断进步，其广泛应用在药用植物转录组分析中，并取得重大进展[16]。本研究在明确掌叶大黄幼苗主要有效成分的基础上，采用 Illumina HiSeq™ 2000 100PE 测序平台，首次进行掌叶大黄转录组测序分析，填补了大黄转录组信息的空白。高通量测序数据约 11.04 G，利用 Trinity 组装共得到 93 646 个 unigenes，测序质量良好、质控严格，序列长度与 reads 分布区域对应合理。转录组序列信息量庞大，数据基本涵盖全转录组信息，能够清晰反映传统中药材大黄的基因表达特征，为深入研究大黄生长发育、次生代谢、转录调控等生物学过程功能基因的批量发掘提供支持。

基于高通量测序的转录组数据通常采用生物信息学分析策略进行基因注释和功能分类。本研究利用多种生物信息软件，对掌叶大黄转录组进行注释和功能分类。基于 BLAST 分析，将所有 unigenes 与 NR、Swiss-port、PFAM、KOG 等数据库比对，这与已报道的柴胡[10]、人参[11]和商陆[12]等物种转录组测序注释比例类似，说明掌叶大黄转录组中存在大量序列特征及功能尚未知的 unigenes。GO 分类揭示掌叶大黄转录组特性与生物过程、细胞组分和分子功能相关；KOG 功能分析从基因组水平寻找直系同源体，预测未知 ORF 的生物学功能，可大大提高基因功能注释的准确性。本研究共得到 25 个不同的 KOG 类群，说明掌叶大黄转录组 KOG 种类比较全面。130 个 KEGG 标准代谢通路基因可能参与掌叶大黄水分吸收、矿质营养、光合作用和呼吸作用等生命代谢活动；此外，还发现大量 unigenes 参与萜类、芪类、生物碱、黄酮类、花青素等生物合成相关的 19 个次生代谢标准通路，有助于研究大黄次生代谢物生物合成途径。

目前，蒽醌类成分主要分为大黄素型和茜草素型，通常认为前者由聚酮途径合成，后者由 MVA/MEP 偶联莽草酸途径介导，但是植物中蒽醌生物合成通路仍不清楚。茜草科短小蛇根草 *Ophiorrhiza pumila*[14]和豆科决明 *Cassia obtusifolia*[15]转录组研究发现大量 MVA、MEP、莽草酸及聚酮途径基因，提示蒽醌生物合成与这 4 个途径相关。本研究以掌叶大黄为对象，在明确蒽醌类特征性成分含量的基础上，利用转录组高通量测序技术系统筛选，获得 57、32、61、21 条 unigenes 分别编码 MVA、MEP、莽草酸及聚酮途径的 28 个关键酶，其中 RT-PCR 验证 7 个 unigenes 为全长基因，为下一步研究大黄素型蒽醌生物合成的基因分子功能提供基础数据。拟南芥 MenD/MenH/MenC 基因融合为一个基因[17]，该基因在短小蛇根草和决明中均有转录[14, 15]，本研究未检测到 MenD/MenH/MenC 的表达，也未发现聚酮途径 polyketide hydrolase 转录本，可能与样品的生长发育状态有关。此外，基于 FPKM 统计这些关键酶基因的表达特征存在一定差异，推测它们通过不同的表达调控机制参与了蒽醌类生物合成。

在萜类、黄酮等次生代谢物的衍生化修饰过程中，CYP450 和 UGT 主要起催化氧化/羟基化和糖基化的重要作用[10]。微生物中有这两类基因修饰蒽醌成分的研究。黄枝孢霉 *Cladosporium fulvum* CYP450 能够催化蒽醌类 nataloe-emodin 二聚化[18]，芽孢杆菌 *Bacillus licheniformis* DSM13 UGT 能够在体内外糖基化大黄素和芦荟大黄素而不影响抗癌活性，稳定性显著提高[19]。本研究发现转录组中大量 CYP450、UGTs 序列，可能在大黄蒽醌类衍生修饰中起作用，哪些 unigenes 如何参与蒽醌类衍生化修饰仍需深入研究。

SSR 标记包括 EST-SSR、基因组 SSR。基于 cDNA 文库和转录组测序的 EST-SSR 在植物遗传多样性、分子标记等研究方面应用广泛[20]。本研究 MISA 发现 SSR 从单核苷酸类型到六核苷酸类型均具备，表明掌叶大黄基因组内具有较高丰度的 SSR。重复类型以三核苷酸为主，双核苷酸所占比例次之。这与以三核苷酸重复类型为主的主要作物水稻、大麦等的研究结果相同[21]。在双核苷酸重复 SSRs 中 AG/CT 最多，三核苷酸重复中 AAG/CTT 最多，与人参[22]等的情况一致，说明 SSRs 重复类型可能存在一定的保守性。

本研究首次运用二代高通量测序技术开展了掌叶大黄幼苗转录组测序研究，获得了大量掌叶大黄遗传信息和基因表达特征，发掘了蒽醌类次生代谢物的生物合成途径关键基因，为后期进一步研究蒽醌类生物合成基因克隆及功能分析提供基础资料，有助于深入解析大黄蒽醌生物合成通路及调控机制，为科学阐释大黄生长发育、生理适应及品质形成提供理论基础。

References

[1] Chinese Pharmacopoeia Commission. Pharmacopoeia of the People's Republic of China (中华人民共和国药典) [S]. Part Ⅰ. 2015 ed. Beijing: China Medical Science Press, 2015.

[2] Yi J, Yang JR, Gao F, et al. Emodin enhances arsenic trioxide-induced apoptosis via generation of reactive oxygen species

and inhibition of survival signaling [J]. Cancer Res, 2015, 64: 108−116.

[3] He ZH, Zhou R, He M F, et al. Anti-angiogenic effect and mechanism of rhein from Rhizoma Rhei [J]. Phytomedicine, 2011, 18: 470−478.

[4] Yang F, Xu Y, Xiong A, et al. Evaluation of the protective effect of Rhei Radix et Rhizoma against α-naphthylisothiocyanate induced liver injury based on metabolic profile of bile acids [J]. J Ethnopharmacol, 2012, 144: 599−604.

[5] Zhang XQ, Liu CS, Yan XL, et al. Sequence analysis and identification of a chloroplast matK gene in Rhei Rhizoma from different botanical origins [J]. Acta Pharm Sin (药学学报), 2013, 48: 1722−1728.

[6] Lu GD, Li CY, Wang HZ, et al. Analysis on quality of ***Rheum palmatum*** L. from Gansu province based on multicriteria method [J]. Chin J Inf TCM (中国中医药信息杂志), 2017, 24: 57−63.

[7] Wei WL, Zeng R, Huang LF. Correlation analysis between quality of ***Rheum palmatum*** L. and ecological factors [J]. World Sci Technol−Mod of Tradit Chin Med (世界科学技术−中医药现代化), 2015, 17: 1849−1854.

[8] Liu J, Liu P, Duan JA, et al. Main components analysis in different parts of ***Rheum palmatum*** [J]. Chin Tradit Herb Drugs (中草药), 2017, 48: 567−572.

[9] Qi YX, Liu YB, Rong WH. RNA-Seq and its applications: a new technology for transcriptomics [J]. Hereditas (遗传), 2011, 33: 1191−1202.

[10] Sui C, Zhang J, Wei J, et al. Transcriptome analysis of ***Bupleurum chinense*** focusing on genes involved in the biosynthesis of saikosaponins [J]. BMC Genomics, 2011, 12: 539.

[11] Jung I, Kang H, Kim JU, et al. The mRNA and miRNA transcriptomic landscape of ***Panax ginseng*** under the high ambient temperature [J]. BMC Syst Biol, 2018, 12 (Suppl 2): 27.

[12] Zhao L, Zhu YH, Zhang L, et al. Transcriptome analysis reveals candidate genes involved in esculentoside A biosynthesis in ***Phytolacca americana*** [J]. Acta Pharm Sin (药学学报), 2017, 52: 1471−1480.

[13] Yan YG, Yin LM, Wang HY, et al. HPLC method for simultaneous determination of nine components from leaves of ***Rheum officinale*** by HPLC [J]. Chin Tradit Herb Drugs (中草药), 2016, 47: 2360−2364.

[14] Yamazaki M, Mochida K, Asano T, et al. Coupling deep transcriptome analysis with untargeted metabolic profiling in ***Ophiorrhiza pumila*** to further the understanding of the biosynthesis of the anti-cancer alkaloid camptothecin and anthraquinones [J] Plant Cell Physiol, 2013, 54: 686−696.

[15] Rama Reddy NR, Mehta RH, Soni PH, et al. Next generation sequencing and transcriptome analysis predicts biosynthetic pathway of sennosides from senna (***Cassia angustifolia*** Vahl.), a non-model plant with potent laxative properties [J]. PLoS One, 2015, 10: e0129422.

[16] Zhang ZB, Hou L, Pan Q, et al. Advances in high-throughput transcriptome research of traditional Chinese medicines [J]. China J Chin Mater Med (中国中药杂志), 2014, 39: 1553−1558.

[17] Gross J, Cho WK, Lezhneva L, et al. A plant locus essential for phylloquinone (vitamin K1) biosynthesis originated from a fusion of four eubacterial genes [J]. J Biol Chem, 2016, 281: 17189−17196.

[18] Griffiths S, Mesarich CH, Saccomanno B, et al. Elucidation of cladofulvin biosynthesis reveals a cytochrome P450 monooxygenase required for anthraquinone dimerization [J]. Proc Natl Acad Sci U S A, 2016, 113: 6851−6856.

[19] Ghimire GP, Koirala N, Pandey RP, et al. Modification of emodin and aloe-emodin by glycosylation in engineered ***Escherihia coli*** [J]. World J Microbiol Biotechnol, 2015, 31: 611−619.

[20] Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications [J]. Trends Biotechnol, 2005, 23: 48−55.

[21] Cardle L, Ramsay L, Milbourne D, et al. Computational and experimental characterization of physically clustered simple sequence repeats in plants [J]. Genetics, 2000, 156: 847−854.

[22] Li C, Zhu Y, Guo X, et al. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in ***Panax ginseng*** C. A. Meyer [J]. BMC Genomics, 2013, 14: 245.