基于转录组测序挖掘商陆皂苷甲生物合成相关基因

赵 乐[1, 2], 朱昀昊[1, 2], 张 莉[1, 2], 马利刚[1, 2], 冯卫生[1, 2], 郑晓珂[1, 2*]

(1. 河南中医药大学药学院, 河南 郑州 450046;
2. 呼吸疾病诊疗与新药研发河南省协同创新中心, 河南 郑州 450046)

摘要: 为研究商陆皂苷甲的生物合成途径, 利用 Illumina HiSeq 4000 高通量测序技术对商陆幼苗进行转录组测序, 得到 9.60 Gb clean data, 经 Trinity 软件组装后获得 63 957 条 unigenes, 平均长度 988.82 bp, 其中 24 517 条 unigenes (38.33%) 能被 Nr、Swiss-Prot、COG、KOG、Pfam、GO、KEGG 等公共数据库注释。对注释得到的 unigenes 进行 KEGG 代谢通路分析, 发现商陆转录组中有 53 个 unigenes 参与萜类骨架合成通路, 有 8 个 unigenes 参与三萜合成通路, 还有 417 个 unigenes 参与商陆其他次生代谢途径。进一步分析参与商陆皂苷甲生物合成后修饰酶相关基因, 发现有 130 个 unigenes 可能具有 CYP450 的功能, 参与商陆次生代谢产物的氧化/羟基化修饰; 有 46 个 unigenes 与糖基转移酶 UGT 相关。商陆转录组数据的获得为研究商陆皂苷甲和其他次生代谢产物的生物合成途径奠定了基础, 也为商陆药材品质的形成提供理论依据。

关键词: 商陆; 商陆皂苷甲; 转录组; 生物合成途径

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870 (2017) 09-1471-10

Transcriptome analysis reveals candidate genes involved in esculentoside A biosynthesis in *Phytolacca americana*

ZHAO Le[1, 2], ZHU Yun-hao[1, 2], ZHANG Li[1, 2], MA Li-gang[1, 2],
FENG Wei-sheng[1, 2], ZHENG Xiao-ke[1, 2*]

(*1. School of Pharmacy, Henan University of Traditional Chinese Medicine, Zhengzhou 450046, China; 2. Collaborative Innovation Center for Respiratory Disease Diagnosis and Treatment and Chinese Medicine Development of Henan Province, Zhengzhou 450046, China*)

Abstract: In order to study the biosynthesis pathway of esculentoside A, the Illumina HiSeq 4000 high-throughput sequencing method was used to analyze the transcriptome of *Phytolacca americana* seedlings. The 9.60 Gb clean data were obtained after the transcriptome of *P. americana* assembled by Trinity software. The total 63 957 unigenes were obtained after assembly and the average length was 988.82 bp, among them 24 517 unigenes (38.33%) were annotated in the public databases Nr, Swiss-Prot, COG, KOG, Pfam, GO and KEGG. According to the assignment of KEGG pathway, 53 unigenes were involved in terpenoid backbone biosynthesis and 8 unigenes involved in triterpenoid biosynthesis. Additionally, there were 417 unigenes assigned to other secondary metabolic pathways in *P. americana*. The post-modification enzyme genes involved in the esculentoside A biosynthesis were also analyzed in the transcriptome of *P. americana*. The results indicated that 130 unigenes may have the function of CYP450 which was involved in oxidation/hydroxylation modification of *P. americana* secondary metabolites. Furthermore, 46 unigenes had the function of glycosyltransferase UGT.

收稿日期: 2017-03-22; 修回日期: 2017-05-17.
基金项目: 中央引导地方科技发展专项 (河南道地大宗药材种质评价及集约化种植与示范); 教育部科学技术研究重点资助项目: 伏牛山中药资源区系分析研究 (DF2003078); 河南省科技攻关计划资助项目 (162102310468).
*通讯作者 Tel / Fax: 86-371-65962746, E-mail: zhengxk.2006@163.com
DOI: 10.16438/j.0513-4870.2017-0246

The transcriptome data of *P. americana* laid a foundation for studying the biosynthesis pathway of esculentoside A and other secondary metabolites, and also provided theoretical basis for formation of medicinal materials quality.

Key words: *Phytolacca americana*; esculentoside A; transcriptome; biosynthesis pathway

垂序商陆 (*Phytolacca americana* L.) 为商陆科 (Phytolaccaceae) 多年生草本植物，干燥根及其炮制品分别为生商陆和醋商陆入药，是我国传统中药，被 2015 年版《中国药典》收录为商陆药材来源[1]。商陆有逐水消肿、通利二便的功效；用于治疗水肿胀满、二便不通，外治痈肿疮毒等[1]，具有重要的药用价值。目前已从商陆中分离得到商陆皂苷类、黄酮类、酚酸类、甾醇类以及多糖类等多种化学成分，其中商陆皂苷是商陆的特征性化学成分，是其主要的药效物质基础，具有利尿、免疫抑制、抗炎等显著的生理活性，已成为研究热点[2]。至今已从商陆中分离得到 33 种商陆皂苷，且均为齐墩果烷型[2]，由于商陆皂苷甲 (esculentoside A, EsA) 在商陆皂苷中含量较高，2015 年版《中国药典》收录的商陆药材项将其列为对照品[1]，通过测定 EsA 在商陆药材中的含量对商陆质量进行控制，所以商陆皂苷甲 (EsA) 是商陆皂苷中的主要成分。现代药理学研究发现商陆具有利尿、抗菌、抗病毒、抗炎、抗肿瘤等活性；临床上多用其治疗乙型肝炎、银屑病、过敏性紫癜等疑难疾病[2]。对于商陆的抗炎作用，主要集中在抗炎活性极强的 EsA 上，而且 EsA 也具有显著的免疫抑制活性。Zhang 等[3]发现 EsA 能够减弱 CCl_4 和 GalN/LPS 诱导的急性肝损伤，而且这种保护机制与 EsA 的抗炎、抗氧化胁迫作用有关，可能是通过抑制细胞产生 TNF 和 IL-1 来实现的。

作为商陆主要药效物质基础的 EsA，从其化学结构分析属于三萜皂苷 (齐墩果烷型五环三萜)，关于三萜皂苷生物合成途径研究较为深入是人参皂苷，目前已克隆了近 50 个与人参皂苷生物合成相关的基因并对其中 20 个基因进行了功能验证[4, 5]。虽然商陆在化学成分、药理作用等方面已有较为深入的研究，但关于 EsA 生物合成途径的报道较少，不仅上游萜类骨架合成的相关酶类未见报道，而且参与商陆皂苷元母核合成后修饰的酶类，如细胞色素 P450 (cytochrome P450, CYP450) 和糖基转移酶 (uridine diphosphate glycosyltransferase, UGT) 等的基因克隆和功能分析也未见报道。高通量测序技术具有测序通量大、价格低、时间短等优势，极大降低了测序所需成本和时间，广泛应用于药用植物次生代谢途径功能基因的挖掘方面[6]。目前在人参[7]、丹参[8]、三七[9]、金银花[10]等药用植物中已经利用高通量测序技术进行转录组研究，获得了一批与药用植物药效成分生物合成及调控相关的基因[6]。本研究利用高通量测序技术对商陆幼苗进行转录组测序，以期获得一批与 EsA 生物合成相关的基因，为初步阐明 EsA 生物合成途径奠定基础，也为今后利用生物技术对商陆进行遗传改良提供候选基因。

材料与方法

材料 商陆种子采自河南省伏牛山区，经过浓硫酸 (98%) 和酒精 (70%) 表面消毒，各 15 min，再用无菌水漂洗 3 次，将种子种到 1/2 MS 培养基上，在人工智能培养箱中生长。培养条件为 16 h、23 ℃光照，8 h、20 ℃黑暗，光照强度 150 μE·m^{-2}·s^{-1}，约 3 周后，种子萌发长成幼苗，经河南中医药大学董诚明教授鉴定为商陆科植物垂序商陆 (*Phytolacca americana* L.)，在无菌条件下取商陆幼苗样品，放入液氮中速冻。

商陆皂苷甲含量测定 根据药典中所收录的 HPLC-ELSD 含量测定方法，稍作调整，对无菌的商陆幼苗根、茎、叶中商陆皂苷甲进行含量测定[1]。仪器与试药：Waters e2695 Separations Module, Waters 2424 ELS Detector, Waters Empower 高效色谱工作站，MILLIPORE 超纯水发生器；商陆皂苷甲 (国家标准物质，供含量测定用，批号：111922-201102，纯度为 92.2%) 购自中国食品药品检定研究院。取商陆幼苗根、茎、叶冷冻干燥 2 天后，称重，分别打粉或研磨粉碎，过 40 目筛，备用。色谱条件及检测条件：色谱柱为 ODS (Inertsil ODS-SP, 5 μm, 4.6 mm×250 mm, GL. Sciences Inc.)；流动相：甲醇−水 (65∶35)；流速：1 mL·min^{-1}；柱温：25 ℃；ELSD 条件：漂移管温度 74 ℃，载气流速：2.98 L·min^{-1}。该条件下，信噪比最佳。商陆皂苷甲色谱峰理论塔板数不低于 2 000。标准品溶液制备：精密称取商陆皂苷甲 5.0 mg，置于 10 mL 容量瓶中，加入 6 mL 甲醇并超声溶解 10 min，定容至 10 mL，制成浓度为 0.5 mg·mL^{-1} 的对照品溶液。从 0.5 mg·mL^{-1} 的对照品溶液中精确吸取 1 mL 至 5 mL 容量瓶中，加入甲醇定容至 5 mL，以 0.22 μm

微孔滤膜过滤两次, 制成浓度为 0.1 mg·mL^{-1}的对照品溶液。供试品溶液制备: 精密称量商陆根粉末 0.5 g, 商陆茎粉末 0.5 g, 商陆叶粉末 0.5 g, 分别放入 3 个 10 mL 容量瓶中, 6 mL 70%甲醇溶解, 超声提取 30 min, 定容至 10 mL, 以 0.22 μm 微孔滤膜过滤两次, HPLC 用。线性关系考察与样品测定: 设定 0.1 mg·mL^{-1}的对照品溶液 5、10、15、20 μL, 0.5 mg·mL^{-1}的对照品溶液 10、20 μL 自动进样, 按照上述色谱条件, 用液相色谱仪进行测定, 所得峰面积与浓度进行线性回归。回归方程为 $y=364\,668x-262\,439$, $R=0.997\,6$。在上述色谱条件下, 设定供试品 20 μL 自动进样, 按外标峰面积法计算样品中商陆皂苷甲的含量。

RNA 提取与检测　从商陆幼苗根、茎、叶不同组织混合样品中提取总 RNA, 分别用 Nanodrop 和 Aglient 2100 检测 RNA 样品的浓度、纯度和完整性等, 保证 RNA 质量满足建库要求进行转录组测序。

RNA 文库构建及文库质控　样品检测合格后, 进行文库构建, ① 用带有 Oligo(dT) 的磁珠富集商陆幼苗 mRNA; ② 加入 fragmentation buffer 将 mRNA 打断成短片段; ③ 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、RNase H 和 DNA polymerase I 合成第二条 cDNA 链, 利用 AMPure XP beads 纯化 cDNA; ④ 纯化的双链 cDNA 先进行末端修复、加 A 尾并连接测序接头, 再用 AMPure XP beads 进行片段大小选择; ⑤ 最后通过 PCR 富集得到 cDNA 文库。文库构建完成后, 先使用 Qubit2.0 进行初步定量, 随后使用 Agilent 2 100 对文库的插入片段大小进行检测, 以保证文库质量。

转录组测序　文库检测合格后, 用 Illumina HiSeq 4000 进行高通量测序, 利用双末端测序 (Paired-End) 法, 测序读长为 PE150。测序得到的原始图像数据文件经碱基识别转化为原始数据 (raw data), 对原始数据进行数据过滤, 去除接头、重复序列、低质量的序列, 获得高质量的 clean data。使用 Trinity 软件[11]对 clean data 进行拼接, 通过序列之间的 overlap 拼接得到重叠群 (contigs), 然后进一步组装得到转录本 (transcripts), 最后获得单基因簇 (unigenes)。

Unigene 功能注释　使用 BLAST 软件将 unigene 序列与 Nr (Non-redundant Protein Sequence Database in GenBank)、Swiss-Prot (Swiss-Prot Protein Sequence Database)、GO (Gene Ontology)、COG (Clusters of Orthologous Groups)、Pfam (Protein family)、KOG (euKaryotic Orthologous Groups) 、KEGG (Kyoto Encyclopedia of Genes and Genomes) 数据库比对, 获得 unigene 的注释信息, GO 功能注释基于 Nr 和 Pfam 两部分的蛋白注释结果, 在 Blast2GO 软件上进行分析, 通过上述数据库对商陆的转录组数据进行功能基因注释, 找到参与 EsA 生物合成途径的关键基因。使用 MEGA5 软件相邻连接法 (neighbor-joining) 构建系统进化树, bootstrap 检验的重复次数为 1 000 次

商陆转录组 SSR 分析　利用 MISA 软件对筛选得到的 1 kb 以上的 unigenes 做 SSR 位点分析, 筛选的标准为: 单核苷酸重复≥10 bp、二核苷酸重复≥12 bp、三核苷酸重复≥15 bp、四核苷酸重复≥20 bp、五核苷酸重复≥25 bp、六核苷酸重复≥30 bp, 同时单核苷酸、二核苷酸、三核苷酸、四核苷酸、五核苷酸和六核苷酸最少重复次数分别为 10、6、5、5、4、4, 复合型 SSR 至少包含 2 个 SSR 位点, 且 2 个位点之间距离小于 100 bp, 对获得的 SSR 种类、数量等基本信息进行统计分析。

结果与分析

1　商陆皂苷甲含量测定

取商陆幼苗根、茎、叶的样品, 按“材料与方法”项下的操作要求制备供试样品溶液, 每个样品平行制备 3 份溶液, 按“材料与方法”中的色谱条件, 供试样品 20 μL 自动进样, 按外标峰面积法计算样品中商陆皂苷甲的含量, HPLC 色谱图见图 1。从图 1 中可看出, 商陆幼苗不同组织部位中 EsA 的含量差异较大, 根中含量较高为 0.77%, 茎中较低为 0.37%, 而叶中未检出。

2　商陆幼苗 RNA 提取

商陆幼苗 RNA 样品经 Nanodrop、Agilent 2100 检测, OD_{260}/OD_{280}、OD_{260}/OD_{230} 等各项指标均合格 (表 1), 所提取的 RNA 质量较好, 符合转录组测序文库构建要求。

3　商陆转录组测序结果与数据组装

采用 Illumina HiSeq 4000 高通量测序平台对商陆转录组进行测序, 共获得 32 470 764 个 reads 片段, 包含了 9 602 387 872 bp 的序列信息, 经过测序质量控制, 得到 9.60 Gb clean data, Q30 碱基百分比到达 90.12%, GC%含量平均值为 46.28%。商陆转录组的原始数据 (Raw data)已上传至 NCBI 的 SRA 数据库, 登录号为 SRP105831。使用 Trinity 软件对商陆转录组进行组装, 共得到 6 859 608 条 contigs, 其中长度 200～300 bp 的 contigs 序列有 6 717 497 条, 占总数的

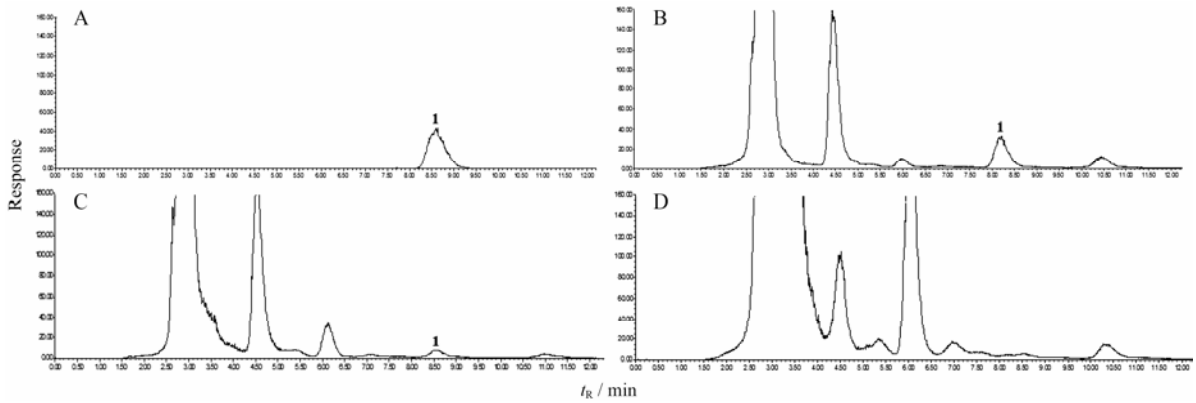

Figure 1 HPLC chromatograms of esculentoside A (EsA) standard and different tissues of *P. americana*. A: EsA standard; B: Roots; C: Stems; D: Leaves; 1: EsA

Table 1 The quality of the RNA sample

Sample name	Concentration/ng·μL^{-1}	Volume/μL	Total content/μg	OD_{260}/OD_{280}	OD_{260}/OD_{230}	RIN value	28S/18S
Pa	415	24	10	2.18	2.05	7.7	2.2

97.93%, 300～2 000 bp 的占 1.94%, 2 000 bp 以上的占 0.13% (表 2), 可见 contigs 序列的分布以 200～300 bp 为主, 其分布特征符合 Illumina 测序的预期结果, 可为后续数据组装提供良好的数据。在 contigs 数据的基础上, 再进行组装获得 128 964 条 Transcripts, 序列总长度为 176 102 380 bp, 平均长度为 1 365.52 bp, N50 为 2 097 bp (表 2)。对得到的 Transcripts 序列进一步组装获得 63 957 条 unigenes, 序列总长度为 63 241 729 bp, 平均长度为 988.82 bp, N50 为 1 541 bp (表 2)。

Table 2 Summary of tanscriptome data assembly from *P. americana*

Length range / bp	Contig	Transcript	Unigene
200−300	6 717 497 (97.93%)	12 763 (9.90%)	9 956 (15.57%)
300−500	67 803 (0.99%)	22 024 (17.08%)	15 969 (24.97%)
500−1 000	45 582 (0.66%)	32 254 (25.01%)	18 711 (29.26%)
1 000−2 000	19 751 (0.29%)	33 214 (25.75%)	11 614 (18.16%)
2 000+	8 975 (0.13%)	28 709 (22.26%)	7 707 (12.05%)
Total number	6 859 608	128 964	63 957
Total length	412 722 676	176 102 380	63 241 729
N50 length	66	2 097	1 541
Mean length	60.17	1 365.52	988.82

4 Unigene 功能注释

通过选择 BLAST 参数 E-value 不大于 1×10^{-5} 和 HMMER 参数 E-value 不大于 1×10^{-10}, 最终获得 24 517 个 (38.33%) 有注释信息的 unigenes (表 3), 其中能被 Nr 数据库注释的有 24 176 个 (37.80%); 能被 Swiss-Prot 数据库注释的有 15 161 个 (23.70%); 被 Pfam 数据库注释的有 16 428 个 (25.69%); 能被 KOG 数据库注释的有 13 755 个 (21.51%); 能被 KEGG 数据库注释的有 8 252 个 (12.90%); 能被 GO 数据库注释的有 13 172 个 (20.60%); 能被 COG 数据库注释的有 6 945 个 (10.86%); 有 39 440 个 unigenes (61.7%) 不能被已有的数据库注释。

Table 3 Summary statistics of unigenes functional annotation for *P. americana* tanscriptome

Annotated databases	Number of unigenes	Annotation percentage/%
Nr	24 176	37.80
Swiss-Prot	15 161	23.70
Pfam	16 428	25.69
KOG	13 755	21.51
KEGG	8 252	12.90
GO	13 172	20.60
COG	6 945	10.86
All	24 517	38.33
Total	63 957	100

Unigenes 在 Nr 数据库相似序列匹配的近缘物种中, 甜菜 (*Beta vulgaris*) 所占比例最高 (14 654 条, 60.64%), 其次是葡萄 (*Vitis vinifera*, 1 492 条, 6.17%)、北美云杉 (*Picea sitchensis*, 485 条, 2.01%)、美花烟草 (*Nicotiana sylvestris*, 341 条, 1.41%)、莲 (*Nelumbo nucifera*, 329 条, 1.36%)、绒毛烟草 (*Nicotiana tomentosiformis*, 290 条, 1.20%)、可可 (*Theobroma cacao*, 286 条, 1.18%)、甜橙 (*Citrus sinensis*, 223 条, 0.92%)、桃 (*Prunus persica*, 201 条, 0.83%)、蓖麻 (*Ricinus communis*, 200 条, 0.83%) 和其他物种 (5 663 条, 23.44%)。

5　GO 分类

GO (Gene Ontology) 数据库是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表来全面描述生物体中基因和基因产物的功能属性，为了对商陆的转录组数据进行功能分析，将 unigenes 进行 GO 注释，然后将注释成功的 unigenes 再进行 GO 分类。共有 13 172 个 unigenes 获得至少一个注释结果，这些 unigenes 被分为细胞组份 (cellular component, CC)、分子功能 (molecular function, MF)、生物学过程 (biological process, BP) 3 个大类，51 个小类 (图 2)，其中 25 542 个 unigenes 被注释为细胞组份，15 767 个 unigenes 被注释为分子功能；36 872 个 unigenes 被注释为生物学过程。

6　COG 相关功能分类

COG (clusters of orthologous groups) 数据库是对基因产物进行同源分类的数据库，是一个较早的识别直系同源基因的数据库，通过对多种生物的蛋白质序列大量比较而来。将商陆 unigenes 与 COG 数据库进行比对，预测 unigenes 功能并进行分类统计。研究结果表明，商陆 unigenes 根据其功能大致可分为 25 类。Unigenes 涉及的 COG 功能类别比较全面，其中，一般功能预测类基因最多 (1 806 个)；其次是复制、重组和修复类基因 (1 042 个)，翻译、核糖体结构和生物合成类基因 (878 个)，转录类基因 (871 个)，信号转导类基因 (771 个)，翻译后修饰、分子伴侣类基因 (706 个)；而细胞运动类基因 (3 个) 和细胞核结构类基因 (1 个) 较少；未发现胞外结构类基因。

7　KEGG 分类

为了分析商陆转录组数据中 unigenes 所参与的代谢路径，将获得 KO 注释的 unigenes 进行 KEGG 代谢通路分析。共有 8 252 个 unigenes 归入 127 个代谢通路，包含 unigenes 最多的代谢通路是核糖体途径 (ko03010)，共有 679 个；其次是氨基酸生物合成途径 (ko01230)，共有 285 个。在商陆的转录组数据中有 53 个 unigenes 映射到萜类骨架合成通路上 (ko00900)，有 8 个 unigenes 映射到倍半萜和三萜的生物合成通路上 (ko00909)，共编码三萜类化合物合成途径中的 19 个关键酶 (表 4)，包括羟甲基戊二酰辅酶 A 合成酶 (HMGS)、羟甲基戊二酰辅酶 A 还原酶 (HMGR) 等 6 个甲羟戊酸途径 (MVA 途径) 的酶；1-脱氧-*D*-木酮糖-5-磷酸合成酶 (DXS)、1-脱氧-*D*-木酮糖-5-磷酸还原异构酶 (DXR) 等 6 个甲基赤藓醇磷酸途径 (MEP 途径) 的关键酶；以及牻牛儿基焦磷酸合成酶 (GPPS)、法尼基焦磷酸合酶 (FPPS)、牻牛儿基牻牛儿基焦磷酸合酶 (GGPPS) 等 3 个催化萜类骨架直接前体生成的异戊烯基转移酶；鲨烯合酶 (SQS)、鲨烯环氧酶 (SE)、β-香树脂合成酶 (β-AS) 等 3 个催化生成齐墩果烷型五环三萜前体的关键酶，这 19 个关键酶参与 EsA 可能的生物合成途径如图 3 所示。

根据 KEGG 代谢通路分析结果，有 13 条代谢通

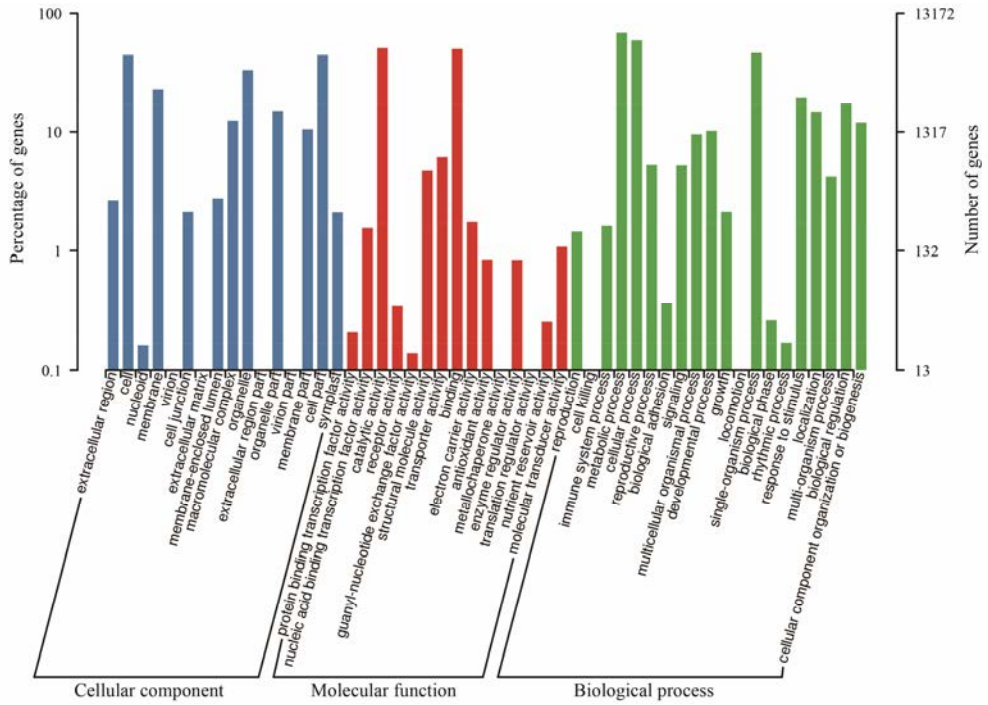

Figure 2　Gene ontology classification of unigenes

Table 4 Candidate genes involved in esculentoside A biosythesis pathway

No.	Gene name	Abbreviation	KO No.	Number of unigenes
1	Acetoacety1-CoA transferase	AACT	K00626	2
2	Hydroxy methylglutaryl-CoA synthase	HMGS	K01641	3
3	Hydroxy methylglutaryl-CoA reductase	HMGR	K00021	5
4	Mevalonate kinase	MK	K00869	1
5	Phosphomevalonate kinase	PMK	K00938	1
6	Mevalonate 5-diphosphate decarboxylase	MPD	K01597	2
7	1-Deoxy-*D*-xylulose-5-phosphate synthase	DXS	K01662	2
8	1-Deoxy-*D*-xylulose-5-phosphate reductoisomerase	DXR	K00099	3
9	2-*C*-Methyl-*D*-erythritol-4-phosphate cytidyltransferase	MCT	K00991	1
10	4-(Cytidine-5-diphospho)-2- *C*-methylerythritol kinase	CMK	K00919	1
11	2-*C*-Methylerythritol-2,4-cyclodiphosphate synthase	MCS	K01770	2
12	4-Hydroxy-3-methyl-but-2-enyl diphosphate reductase	HDR	K03527	1
13	Isopentenyl diphosphate isomerase	IDI	K01823	2
14	Geranyl diphosphate synthase	GPPS	K14066	3
15	Geranylgeranyl diphosphate synthase	GGPPS	K13789	3
16	Farnesyl diphosphate synthase	FPPS	K00787	4
17	Squalene synthase	SQS	K00801	3
18	Squalene epoxidase	SE	K00511	1
19	*β*-Amyrin synthase	*β*-AS	K15813	1

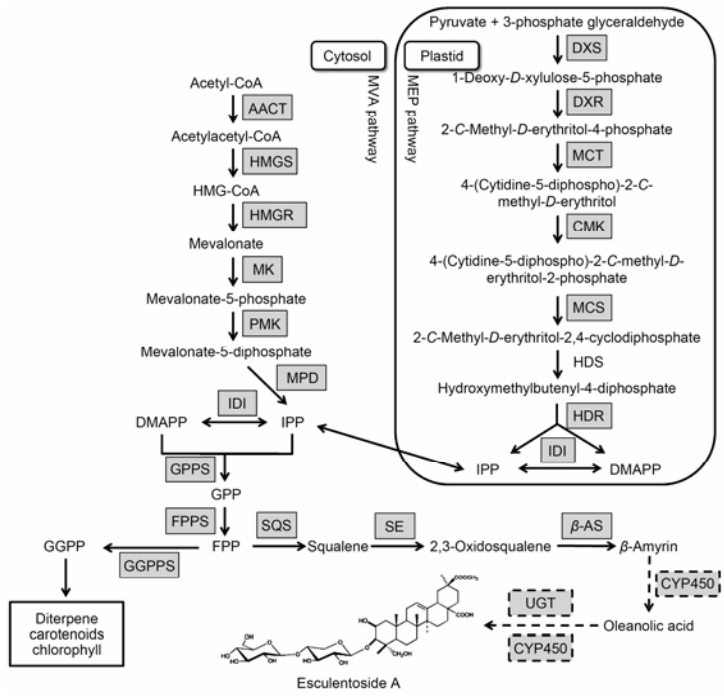

Figure 3 The possible biosynthesis pathway of esculentoside A (The dashed arrows indicate the supposed reaction, the functions of enzymes in dashed boxes are not yet elucidated)

路中 409 条 unigenes 可能参与商陆其他次生代谢途径 (图 4)。其中苯丙素类生物合成途径 (ko00940) 所占比例最大，达 34.72%；其次是类黄酮生物合成 (ko00941) 和 N 聚糖生物合成 (ko00510) 分别占 11.25%；泛醌和萜醌类物质生物合成 (ko00130) 占 7.82%；二萜生物合成 (ko00904) 占 6.60%；而单萜生物合成 (ko00902) 和花青素生物合成 (ko00942) 最少，分别占 0.98% 和 0.24%。商陆中多条次生代谢途径及相关基因的发现，表明其次生代谢生物合成途径的复杂性，以及商陆化学成分的多样性，为后续从商陆中分离新化合物提供了线索，也为阐明商陆功效物质基础提供了理论依据。

8 系统进化树分析

根据推导的商陆 MVA 途径基因的氨基酸序列，

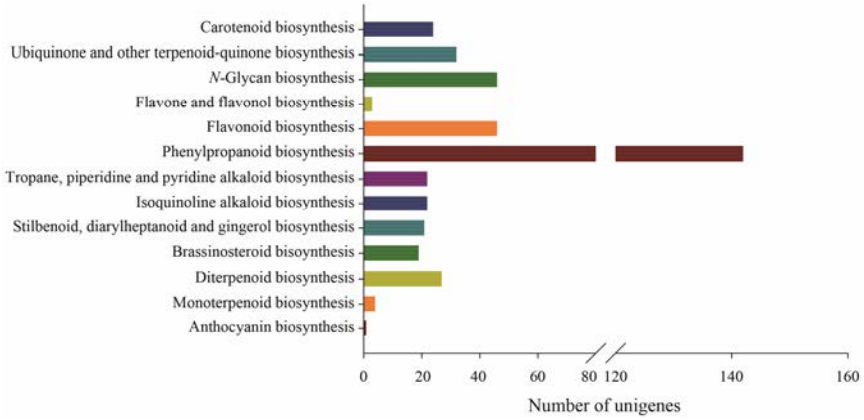

Figure 4　Unigenes related to other secondary metabolism from tanscriptome of *P. americana*

和来源于不同物种相同基因家族的氨基酸序列，采用相邻连接法 (neighbor-joining) 构建系统进化树(图 5)。从系统进化树中可以看出，商陆 MVA 途径的蛋白都归属被子植物的双子叶植物分支，与甜菜(*Beta vulgaris*) 的亲缘关系较近，处于同一分支上。这一结果也与商陆转录组 unigenes 功能注释时，在 Nr 数据库相似序列匹配的近缘物种中甜菜占的比例最高相一致。

9　EsA 生物合成的后修饰酶相关基因

目前已从商陆中分离得到 33 种商陆皂苷，都属于三萜皂苷 (齐墩果烷型五环三萜)[2]，植物合成三萜化合物的骨架后，都会在三萜骨架上进行后修饰，如氧化、羟基化、糖基化、甲基化、乙酰化等多种反应，形成多种三萜皂苷，这些后修饰反应大幅增加了萜类化合物的种类及其结构的多样性[12]。

EsA 生物合成途径中的后修饰反应，推测主要包括骨架的氧化/羟基化和糖基化，分别由不同超基因家族编码的 CYP450 和 UGT 进行催化。氧化是植物次生代谢产物后修饰中最常见的方式，其中绝大部分依赖细胞色素 P450 (CYP450) 的催化，植物 CYP450 具有广泛的催化活性，其作用特点是在底物分子中加入一个氧原子，从而参与萜类、苯丙烷类、生物碱类等多种次生代谢产物的生物合成。在三萜皂苷的生物合成中，细胞色素 P450 主要催化三萜骨架惰性甲基和亚甲基的氧化[13]。通过分析商陆转录组数据在 Swiss-Prot 数据库的注释结果，共找到 130 条可能的 *CYP450* 基因，隶属于 24 个 CYP450 家族 (表 5)。属于 CYP71 家族的 unigenes 最多，有 17.69%；其次是 CYP72、CYP86 和 CYP90，分别为 9.23%、8.46% 和 7.69%；而 CYP73、CYP77、CYP83、CYP84、CYP85 和 CYP703 家族成员最少，仅各有 1 个 unigene。

糖基化是植物次生代谢过程中广泛的一种修饰反应，尿苷二磷酸−糖基转移酶 (UDP-glycosyltransferases，UGT) 能够催化尿苷二磷酸上连接的活性糖转移到多种受体，如可以转移到三萜皂苷的苷元上，增加三萜的水溶性，改善其化学稳定性和生物活性，因此糖基化在植物三萜皂苷的生物合成中非常重要[14]。根据商陆转录组数据在 Swiss-Prot 数据库的注释结果，共找到属于 14 个 UGT 亚家族的 46 个 UGTs，其中包括 1 个 UGT73、2 个 UGT74、2 个 UGT75、6 个 UGT76、1 个 UGT79、6 个 UGT80、5 个 UGT85、6 个 UGT86、1 个 UGT87、8 个 UGT89、1 个 UGT90、1 个 UGT91、2 个 UGT92 和 4 个 UGT709。商陆转录组中大量 CYP450 和 UGTs 的发现为进一步分析具体 CYP450 和 UGT 的催化功能，阐明商陆 EsA 生物合成途径的下游阶段后修饰的氧化和糖基化反应奠定基础。

10　SSR 分析

利用 MISA 软件对筛选得到的 1 kb 以上的 unigenes 做 SSR 分析，共检出含 7 924 个 SSR。如表 6 所示，商陆转录组 SSR 种类丰富，单核苷酸重复至六核苷酸重复类型均存在，但各类型出现的频率具有较大的差异。商陆转录组 SSR 种类中单核苷酸重复最多，占 SSR 总数的 50.92%，其次是二核苷酸重复和三核苷酸重复，分别占 SSR 总数的 24.41% 和 19.13%。在检出 SSR 中，共发现 76 种重复基元，其中 A/T 在单核苷酸重复基元出现最多，有 3 638 个，占 SSR 总数的 45.91%；AG/TC、CT/GA、AT/TA 等 3 种类型在二核苷酸重复基元出现最多，分别有 775 个 (9.78%)、727 个 (9.17%) 和 303 个 (3.82%)；在三核苷酸重复基元中，CCA/GGT 和 CTT/GAA 出现最多，分别为 111 个 (1.40%) 和 96 个 (1.21%)。对这些 SSR 的鉴定，将为进一步筛选和开发商陆 SSR 标记奠定了基础，对商陆遗传多样性分析、分子标记辅助育种

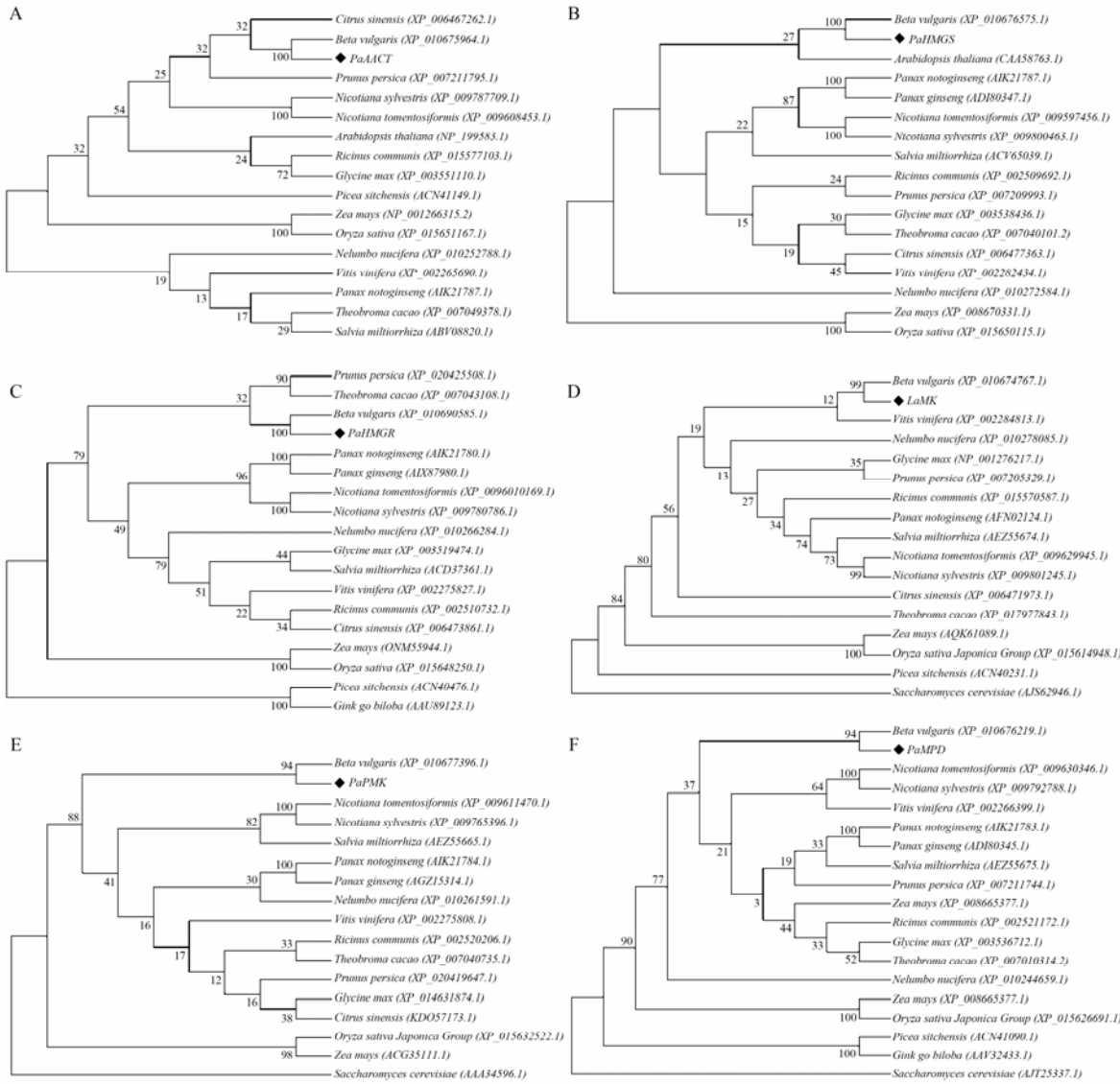

Figure 5 Phylogenetic trees analysis of MVA pathway enzymes. A: Phylogenetic tree of AACTs; B: Phylogenetic tree of HMGSs; C: Phylogenetic tree of HMGRs; D: Phylogenetic tree of MKs; E: Phylogenetic tree of PMKs; F: Phylogenetic tree of MPDs

Table 5 Summary of *CYP450* genes in the transcriptome of *P. americana*

Gene family name	Number of unigenes	Percentage /%	Gene family name	Number of unigenes	Percentage /%
CYP71	23	17.69	CYP87	2	1.54
CYP72	12	9.23	CYP89	3	2.31
CYP73	1	0.77	CYP90	10	7.69
CYP76	3	2.31	CYP94	9	6.92
CYP77	1	0.77	CYP98	5	3.85
CYP78	9	6.92	CYP102	4	3.08
CYP81	6	4.62	CYP703	1	0.77
CYP82	9	6.92	CYP704	2	1.54
CYP83	1	0.77	CYP710	2	1.54
CYP84	1	0.77	CYP714	2	1.54
CYP85	1	0.77	CYP734	5	3.85
CYP86	11	8.46	CYP736	7	5.38

等提供帮助。

讨论

商陆作为我国传统中药，EsA 是商陆的主要药效成分，关于其化学成分和药理作用方面的研究已有良好的基础，但是商陆的基因组信息未知，严重限制了 EsA 生物合成途径及相关功能基因的研究。Neller 等[15]对茉莉酸 (jasmonate acid, JA) 处理前后商陆叶片的转录组变化进行研究，发现差异表达基因多集中在逆境相关基因和抗病毒蛋白基因方面，没有发现次生代谢途径相关的基因。本研究利用 Illumina HiSeq 4000 高通量测序技术对商陆幼苗进行转录组测序，得到 9.60 Gb 的 clean data，组装后获得 63 957

Table 6　Distribution of SSRs with the numbers of repeat motifs in transcriptome of *P. americana*

Type of repeat	Number of repeat									Total	Percentage /%
	5	6	7	8	9	10	11	12	>12		
Mono-nuleotide	0	0	0	0	0	1 856	832	501	846	4 035	50.92
Di-nuleotide	0	887	471	297	182	78	19	0	0	1 934	24.41
Tri-nucleotide	1 013	373	121	8	1	0	0	0	0	1 516	19.13
Tetra-nucleotide	40	5	0	0	0	0	0	0	0	45	0.57
Penta-nucleotide	6	0	0	0	0	0	0	0	0	6	0.08
Hexa-nucleotide	5	3	0	0	0	0	0	0	0	8	0.10
Compound formation	0	0	0	0	0	19	29	33	299	380	4.80
Total	1 064	1 268	592	305	183	1 953	880	534	1 145	7 924	100
Percentage /%	13.43	16.00	7.47	3.85	2.31	24.65	11.11	6.74	14.45	100	

条 unigenes, 平均长度 988.82 bp, N50 为 1 541 bp, 其中 24 176 条 unigenes 能被 Nr 数据库注释, 转录组测序数据饱和度检验显示, 检测到的基因数目趋于饱和。这些结果表明本次商陆转录数据组装效果较好, 获得了商陆大量的基因序列信息, 可满足后续数据分析的要求, Illumina HiSeq 4000 高通量测序技术可作为批量挖掘 EsA 生物合成途径功能基因的有效工具。

由于在商陆幼苗的根、茎、叶中 EsA 含量差异较大, EsA 主要存在于商陆的根中, 这与商陆以根作为入药部位一致。根据商陆转录组数据的 KEGG 分析结果结合其他植物中三萜皂苷的生物合成途径, EsA 的生物合成途径大致可分为萜类前体物质的合成、三萜皂苷骨架的合成和后修饰 3 个阶段。上游阶段通过位于细胞质的 MVA 途径[16]和位于质体中的 MEP 途径[17], 合成萜类共同的前体物质异戊烯焦磷酸 (IPP) 和 IPP 的双键异构体二甲基烯丙基焦磷酸 (DMAPP); 中游阶段通过 SQS、SE、*β*-AS 等关键酶催化生成 *β*-香树脂 (齐墩果烷型五环三萜); 下游阶段主要是经 CYP450 和 UGT 催化各种复杂的修饰反应生成不同类型的商陆皂苷, 但是下游阶段有哪些 CYP450 和 UGT 参与 EsA 的生物合成目前并不清楚。在商陆的转录组数据中有 53 个 unigenes 映射到萜类骨架合成通路上 (ko00900), 有 8 个 unigenes 映射到三萜合成通路上 (ko00909), 包括 6 个 MVA 途径的酶, 6 个 MEP 途径的酶, 以及 GPPS、FPPS、GGPPS 等 3 个催化萜类骨架生成的的异戊烯基转移酶, SQS、SE、*β*-AS 等 3 个催化生成 *β*-香树脂 (齐墩果烷型五环三萜) 的关键酶。

三萜皂苷的生物合成途径研究较为深入的是人参皂苷, 目前关于人参皂苷生物合成途径的基本框架及关键酶的研究取得较大进展, 已从人参属植物中 (人参、西洋参、竹节参、三七等) 克隆得到了近 50 个参与人参皂苷生物合成途径的基因, 并对其中 20 个基因进行了功能研究, 为今后利用合成生物学生产人参皂苷奠定了基础[5]。人参皂苷的生物合成途径由 20 多步连续的酶促反应组成, 其中的关键酶有羟甲基戊二酰辅酶 A 还原酶 (HMGR)、法尼基焦磷酸合酶 (FPPS)、鲨烯合酶 (SQS)、鲨烯环氧酶 (SE)、达玛烯二醇 II 合成酶 (DS)、*β*-香树脂合成酶 (*β*-AS)、CYP450 和 UGT 等。HMGR 是人参皂苷生物合成途径的第一个限速酶, 与动物中只含有 1 个 ***HMGR*** 基因不同, 在植物中 ***HMGR*** 基因属于多基因家族, 在人参中有 2 个 ***HMGR*** 基因 (***PgHMGR1*** 和 ***PgHMGR2***), 使用 HMGR 特异性的抑制剂美维诺林 (mevinolin) 竞争性抑制 HMGR 的酶活性, 会显着降低人参不定根中人参皂苷的总含量, 而超表达 ***PgHMGR1*** 基因则会在人参中积累较多三萜类成分[18]。在商陆的转录组数据中有 5 个 unigenes 被注释为 HMGR, 其中 3 个具有完整的开放阅读框 (ORF), 其中 c62237 与 PgHMGR1 氨基酸序列一致性为 47.09%, 可能具有 HMGR 的活性, 所以可使用 HMGR 抑制剂或者超表达商陆 ***HMGR*** 基因, 研究商陆皂苷含量变化和 HMGR 活性及表达量变化之间的关系, 最终确定商陆 ***HMGR*** 基因在商陆皂苷生物合成途径中的功能。人参中的 DS 催化 2,3-氧化鲨烯生成达玛烯二醇 II, 然后在 CYP450 (CYP716A47、CYP716A53 等) 和 UGT (UGT71A27、UGT74AE2 等) 的催化下生成达玛烷型人参皂苷 (Rb1、Rb2、Rc、Rd、Rg1、Rg2 等)[19], 这类人参皂苷属于四环三萜型皂苷, 是人参皂苷的主要成分; 而 *β*-AS 催化 2,3-氧化鲨烯生成 *β*-香树脂, 然后在 CYP450 (CYP716A52) 催化下生产齐墩果酸, 进一步在 UGT 的催化下生成齐墩果烷型人参皂苷 (Ro)[20], 这类人参皂苷属于五环三萜型皂苷。虽然在人参皂苷生物合成途径中已有部分 ***CYP450*** 和 ***UGT*** 基因进行了功能研究, 但 ***CYP450*** 和 ***UGT*** 都属于超基因家族, 目前仍有大量基因家族成员功能未知。从化学结构分析 EsA 属于齐墩果烷型

五环三萜，根据商陆转录组数据的分析结果，在商陆中 2,3-氧化鲨烯经 *β*-AS 催化生成 *β*-香树脂后，*β*-香树脂可能在 CYP450 的催化下生产齐墩果酸，进一步经 UGT 的催化生成齐墩果烷型商陆皂苷 EsA，但是有哪些 *CYP450* 和 *UGT* 基因参与这一过程，目前并不清楚，可通过与已知功能的 *CYP450* 或 *UGT* 基因进行序列比对，设计简并引物扩增商陆的 *CYP450* 和 *UGT* 基因，或者通过转录组分析对茉莉酸甲酯 (MeJA) 处理前后的商陆根进行比较转录组学研究，筛选参与 EsA 生物合成途径的 *CYP450* 和 *UGT* 基因[21]。根据本研究获得的商陆转录组数据在 Swiss-Prot 数据库的注释结果，共获得 130 条 unigenes 可能具有 CYP450 催化功能和 46 条 unigenes 可能具有 UGT 糖基化功能，以及参与商陆其他次生代谢途径的 417 条 unigenes。本研究获得的参与商陆次生代谢途径的 unigenes，为研究 EsA 生物合成途径及调控机制、克隆关键基因及功能分析奠定了基础，也为商陆药材品质的形成提供理论依据。

References

[1] Chinese Pharmacopoeia Commission. Chinese Pharmacopoeia: Vol 1 (中国药典: 一部) [S]. Beijing: China Medical Science Press, 2015: 324−325.

[2] Wang PC, Wang QH, Zhao S, et al. Research progress on chemical constituents, pharmacological effects, and clinical applications of Phytolaccae Radix [J]. Chin Tradit Herb Drugs (中草药), 2014, 45: 2722−2731.

[3] Zhang F, Wang X, Qiu X, et al. The protective effect of esculentoside A on experimental acute liver injury in mice [J]. PLoS One, 2014, 9: e113107.

[4] Yang JL, Gao LL, Zhu P. Advances in the biosynthesis research of ginsenosides [J]. Acta Pharm Sin (药学学报), 2013, 48: 170−178.

[5] Kim YJ, Zhang D, Yang DC. Biosynthesis and biotechnological production of ginsenosides [J]. Biotechnol Adv, 2015, 33: 717−735.

[6] Wang XY, Song JY, Xie CX, et al. RNA-Seq and genuine traditional Chinese medicine [J]. Acta Pharm Sin (药学学报), 2014, 49: 1650−1657.

[7] Chen S, Luo H, Li Y, et al. 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax ginseng* [J]. Plant Cell Rep, 2011, 30: 1593−1601.

[8] Hua WP H, Zhang Y, Song J, et al. *De novo* transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients [J]. Genomics, 2011, 98: 272−279.

[9] Luo HM, Sun C, Sun YZ, et al. Analysis of the transcriptome of *Panax notoginseng* root uncovers putative triterpene saponin-biosynthetic genes and genetic markers [J]. BMC Genomics, 2011, 12: 1−15.

[10] He L, Xu XL, Li Y, et al. Transcriptome analysis of buds and leaves using 454 pyrosequencing to discover genes associated with the biosynthesis of active ingredients in *Lonicera japonica* Thunb [J]. PLoS One, 2013, 8: e62922.

[11] Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome [J]. Nat Biotechnol, 2011, 29: 644−652.

[12] Li JL, Luo XD, Zhao PJ, et al. Post-modification enzymes involved in the biosynthesis of plant terpenoids [J]. Acta Bot Yunnan (云南植物研究), 2009, 31: 461−468.

[13] Tu J, Zhu P, Cheng KD. Heterologous expression systems of plant cytochrome P450 [J]. Chin Biotechnol (中国生物工程杂志), 2003, 23: 32−37.

[14] Luo Y, Liu XG, Zhou ZQ. Research progress on methods for isolating the gene of plant glycosyltransferase, and its biological functions [J]. Biotechnol Bull (生物技术通报), 2016, 32: 34−39.

[15] Neller KCM, Klenov A, Hudak KA. The pokeweed leaf mRNA transcriptome and its regulation by jasmonic acid [J]. Front Plant Sci, 2016, 7: 283.

[16] Newman JD, Chappell J. Isoprenoid biosynthesis in plants: carbon partitioning within the cytoplasmic pathway [J]. Crit Rev Biochem Mol Biol, 1999, 34: 95−106.

[17] Lichtenthaler HK. The 1-deoxy-*D*-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants [J]. Annu Rev Plant Physiol Plant Mol Biol, 2003, 50: 47−65.

[18] Kim YJ, Lee OR, Oh JY, et al. Functional analysis of 3-hydroxy-3-methylglutaryl coenzyme a reductase encoding genes in triterpene saponin-producing ginseng [J]. Plant Physiol, 2014, 165: 373−387.

[19] Han JY, Kim HJ, Kwon YS, et al. The Cyt P450 enzyme CYP716A47 catalyzes the formation of protopanaxadiol from dammarenediol-II during ginsenoside biosynthesis in *Panax ginseng* [J]. Plant Cell Physiol, 2011, 52: 2062−2073.

[20] Han JY, Kim MJ, Ban YW, et al. The involvement of *β*-amyrin 28-oxidase (CYP716A52v2) in oleanane-type ginsenoside biosynthesis in *Panax ginseng* [J]. Plant Cell Physiol, 2013, 54: 2034−2046.

[21] Liang HC, Wang QH, Gong T, et al. The basic strategies and research advances in the studies on glycosyltransferases involved in ginsenoside biosynthesis [J]. Acta Pharm Sin (药学学报), 2015, 50: 148−153.