山楂果实转录组分析及三萜合成关键酶基因*SQE*的克隆与生信分析

吴君贤[1], 徐 睿[1], 尹旻臻[1], 李 景[1], 余函纹[1], 刘梦丽[1], 彭华胜[2,3*], 查良平[1,4*]

(1. 安徽中医药大学药学院, 安徽 合肥 230012; 2. 中国中医科学院中药资源中心, 道地药材国家重点实验室培育基地, 北京 100700; 3. 中国医学科学院道地药材研究创新单元 (2019RU57), 北京 100700; 4. 安徽省中医药科学院中药资源保护与开发研究所, 安徽 合肥 230012)

摘要: 山楂为我国传统中药, 含有机酸和三萜酸类等活性成分, 具有重要的药用和食用价值。为研究山楂不同发育时期的基因表达差异并挖掘山楂有效成分生物合成的基因, 本研究利用 Illumina Hiseq 2000 高通量测序技术对同一产地不同发育时期的山楂果实进行转录组测序和生物信息学分析。经 Trinity 软件组装后获得 78 496 条 Unigenes, 平均长度为 941 nt, 其中 58 395 条 Unigenes 能被 NR、NT、Swiss-Prot、KEGG、COG、GO 等公共数据库注释。对注释得到的 Unigenes 进行 KEGG 代谢通路分析, 有 52 条 Unigenes 编码 15 个关键酶参与柠檬酸循环, 有 62 条 Unigenes 参与山楂的三萜合成通路。同时, 本研究克隆获得了 2 个鲨烯环氧酶基因 *CpSQE1*、*CpSQE2* 并进行了生物信息学分析, 结果表明 *CpSQE1* 和 *CpSQE2* ORF 全长分别为 1 594 bp、1 597 bp, 分别编码 530、531 个氨基酸, 蛋白质分子质量为 57.6 kDa、57.5 kDa, 生物信息学分析表明 CpSQE1 和 CpSQE2 蛋白保守区都含有一个 PLN02985 superfamily 保守结构域, 均属于鲨烯单加氧酶超家族, 系统进化树显示 CpSQE1、CpSQE2 均与其他植物中具有鲨烯环氧酶功能的 SQE 聚为一支。该研究为进一步挖掘山楂有效成分生物合成过程中的关键基因, 解析调控其有效成分生物合成途径提供研究基础。

关键词: 山楂; 不同时期; 有效成分; 基因表达

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870(2021)12-3313-12

Different development phase of transcriptome analysis from *Crataegus pinnatifida* Bge. and cloning, structure and function prediction of squalene epoxidase involved in triterpenic acid biosynthesis

WU Jun-xian[1], XU Rui[1], YIN Min-zhen[1], LI Jing[1], YU Han-wen[1], LIU Meng-li[1], PENG Hua-sheng[2,3*], ZHA Liang-ping[1,4*]

(*1. School of Pharmacy, Anhui University of Chinese Medicine, Hefei 230012, China; 2. State Key Laboratory Breeding Base of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China; 3. Research Unit of DAO-DI Herbs, Chinese Academy of Medical Sciences, 2019RU57, Beijing 100700, China; 4. Institute of Conservation and Development of Traditional Chinese Medicine Resources, Anhui Academy of Chinese Medicine, Hefei 230012, China*)

Abstract: *Crataegus pinnatifida* is a traditional Chinese medicine, which contains organic acids, triterpenoid acids and other active components, has important medicinal and edible value. In order to study the difference of gene expression level in different developmental stages of hawthorn and explore the genes of active ingredient

收稿日期: 2021-05-05; 修回日期: 2021-06-07.

基金项目: 国家自然科学基金项目 (82073957); 安徽省自然科学基金青年科学基金项目(1808085QH290); 名贵中药资源可持续利用能力建设项目 (2060302).

*通讯作者 E-mail: hspeng@126.com; zlp_ahtcm@126.com

DOI: 10.16438/j.0513-4870.2021-0665

biosynthesis in *Crataegus pinnatifida*, high-throughput Illumina HiSeq 2000 technology were used to conduct transcriptome sequencing and bioinformatics analysis on *Crataegus pinnatifida* fruits from the same origin at different developmental stages. 78 496 Unigenes with an average length of 941 nt were obtained by Trinity software. Among them, 58 395 Unigenes can be annotated by NR, NT, Swiss prot, KEGG, COG, GO and other public databases. KEGG pathway analysis showed that 52 Unigenes encoding 15 key enzymes involved in the citric acid cycle. There are 62 Unigenes were involved in the triterpene biosynthesis pathway of *Crataegus pinnatifida*. Two key enzymes SQE of triterpenoid metabolism pathway in *Crataegus pinnatifida* were cloned and performed bioinformatic analysis. The results showed that ORF of *CpSQE1* and *CpSQE2* were 1 594 bp and 1 597 bp, respectively, encoding 530 and 531 amino acids. The molecular weight of proteins was 57.6 kDa and 57.5 kDa. Bioinformatics analysis showed that both CpSQE1 and CpSQE2 proteins have a PLN02985 superfamily conserved domain, belonging to the squalene monooxygenase superfamily. The phylogenetic tree shows that CpSQE1 and CpSQE2 are clustered together with SQE with squalene epoxidase function in other plants. This study provides a research basis for further exploring the key genes in the biosynthesis process of hawthorn active ingredients and analyzing the regulation pathway of its active ingredient biosynthesis.

Key words: *Crataegus pinnatifida*; different periods; active ingredient; gene expression

山楂始载于《本草纲目》[1], 2020版《中国药典》中记载为蔷薇科植物山里红*Crataegus pinnatifida* Bge. var. *major* N. E. Br.或山楂*Crataegus pinnatifida* Bge.的干燥成熟果实, 具有消食健胃、行气散瘀、化浊降脂等功效, 用于肉食积滞、胃脘胀满等症[2]。山楂广泛分布于我国南北地区, 黑龙江、内蒙古、河北、河南、山东、陕西、江苏等地均有生产[3], 是我国传统的药食两用中药材品种之一[4]。三萜和有机酸是山楂的主要化学成分, 2020版《中国药典》[2]中收录的山楂药材项将枸橼酸 (柠檬酸) 作为对照品, 通过测定枸橼酸在山楂果实中的含量对山楂的质量进行控制, 并且以熊果酸为对照品鉴别山楂真伪。迄今为止, 已从山楂中鉴定出150多种化学成分, 包括黄酮、三萜、甾体、木脂素、有机酸等多种成分[5,6], 熊果酸、齐墩果酸和山楂酸等三萜类化合物在人体内可表现出明显的生物活性, 例如抗癌、抗炎、抗菌、抗病毒和抗氧化等特性[7-9]。

目前对柠檬酸代谢途径的研究多集中于含有机酸的药用植物, 如榔梅[10]、柑橘[11]等。对三萜代谢途径研究较为深入的有甘草[12]、人参[13]、罗汉果[14]、秦艽[15]、三七[16]等药用植物, 鲨烯环氧酶 (squalene eposidases, SQE) 被认为是三萜生物合成途径中的一个关键调控酶。鲨烯环氧酶在甲羟戊酸 (mevalonic acid, MVA) 途径中催化鲨烯生成甾醇及三萜类物质的前体物质2,3-氧化鲨烯, 鲨烯环氧酶的活性和含量与药用植物三萜类活性成分的产量直接相关[17]。在枇杷中克隆得到两个鲨烯环氧酶基因*EjSQE1*和*EjSQE2*, 且鲨烯环氧酶基因的表达量与枇杷悬浮培养细胞中的熊果酸含量呈正相关[18]。将拟南芥的3个SQE和酿酒酵母ERG1等4个SQE基因与羽扇豆醇合成酶基因LUS整合至工程菌株NK2-SQ09的染色体上, 发现拟南芥的AtSQE2能显著提高下游中间体羊毛甾醇的产量, 进一步通过调控达玛烯二醇II合酶基因的表达和发酵工艺优化等策略, 创建出产量能达到15 g·L^{-1}人参皂苷前体达玛烯二醇II的高效酵母细胞工厂[19]。因此, 在三萜代谢途径中SQE的作用至关重要。

转录组测序 (RNA-seq) 是利用高通量测序技术将细胞或组织中全部或部分mRNA、small RNA和no-coding RNA进行测序分析的技术[20]。转录组分析不仅可以高通量地获得与基因表达RNA水平的有关信息, 还可以揭示基因表达与生命现象之间的内在联系, 表征生命体生理活动的规律, 确定其代谢特性[21]。药用植物的药效成分通常为次生代谢产物, 转录组测序技术可以挖掘与药用植物药效成分生物合成相关的转录本和关键酶序列, 探索其生物合成途径及调控机制, 从而提高药用植物有效成分的含量[22]。

目前对山楂的转录组报道较少, 项目组前期建立了一种适用于不同发育期山楂果实总RNA提取的方法[23], 同时对山楂的2个鲨烯合酶进行了克隆和原核表达研究[24], 在此基础上本研究利用Illumina Hiseq 2000高通量测序技术对不同发育时期的山楂果实进行转录组测序, 以期筛选出三萜和柠檬酸生物合成途径相关的基因, 并获得在山楂果实成熟期时相关基因的表达变化规律, 为初步阐明山楂三萜和柠檬酸生物合成途径奠定基础。

材料与方法

材料与试剂 本研究样品采自北京怀柔山区, 选定3株山楂树作为研究对象, 根据山楂果实的不同发育时期, 按照8月上旬 (S1)、8月下旬 (S2) 和10月上旬 (S3) 取样, 经中国中医科学院中药资源中心彭华胜研究员

鉴定。所有样品经鉴定后保存于−80 ℃冰箱, 凭证标本保存于中国中医科学院中药资源中心。TransStart FastPfu Fly DNA Polymerase高保真酶、凝胶回收试剂盒、感受态细胞Trans-T1购自于北京全式金生物技术有限公司; pMD18-T载体购自TaKaRa公司。所需引物由上海生工生物有限公司合成; 其他试剂均为国产分析纯。

总RNA提取及cDNA的合成 称取100 mg左右新鲜的山楂果实, 加液氮置于球磨机上研磨, 采用改良CTAB法从新鲜的山楂果实中提取总RNA[23], 1%的琼脂糖凝胶电泳用于评价RNA的质量。ND2000测定总RNA的A_{260}和A_{280}值, 选择A_{260}/A_{280}为1.8～2.0的总RNA进行反转录。用所提取的总RNA经TaKaRa反转录试剂盒 (Takara Prime Script™ 1st Strand cDNA Synthesis Kit) 反转录成cDNA。

cDNA文库构建与序列组装 将采自同一时期的3份山楂RNA样品等量混合, 由深圳华大基因科技服务有限公司构建测序文库, 使用Illumina HiSeq 2000测序平台进行转录组双末端测序。

Unigenes的功能注释 去除原始序列数据中含有接头及低质量的序列后, 得到Clean reads。使用Trinity软件将Clean reads组装成Contig, 然后使用序列聚类软件TGICL做进一步序列拼接和去冗余处理, 得到Unigenes。不同发育时期山楂样品的Unigenes进一步序列拼接和去冗余处理, 得到尽可能长的非冗余的Unigenes。将Unigenes序列与蛋白数据库Nr、Swiss-Prot、KEGG、COG和GO做BLASTX比对 (e值 < 0.000 01), 得到Unigenes的蛋白功能注释信息。

山楂转录组三萜和柠檬酸代谢途径基因筛选及热图分析 参考文献中柠檬酸化合物[10]和萜类[13]的生物合成代谢通路, 结合7个数据库注释结果, 筛选出本转录组中与三萜和柠檬酸合成相关的Unigenes, 以常用的基因表达水平估算方法中FPKM (fragments per kilobase million) 值进行表达量统计。运用TBtools软件对基因表达进行热图分析, 分析不同合成途径筛选出的基因在山楂不同发育时期的表达模式。

***CpSQE1*和*CpSQE2*基因cDNA的克隆** 山楂鲨烯环氧酶*CpSQE1*和*CpSQE2*的基因来源于山楂转录组数据库, 用Primer Premier 5.0软件设计特异性引物。上游引物: CpSQE1-F: 5'-ATGGATTACCAGTATGTTCTTGGTG-3', CpSQE2-F: 5'-ATGGCGGCAACTGTGGTTGTTCATC-3'; 下游引物: CpSQE1-R: 5'-TCACTGAAGAGGAGGAGCTCTGTAA-3', CpSQE2-R: 5'-TTACTCGGCAGGAGGAGCTCTATGA-3', 以反转录后的cDNA为模板进行PCR扩增, 用1%的琼脂糖凝胶电泳初步判断目的基因是否存在, 取阳性PCR的样品与pMD18-T载体进行连接, 并在连接产物中加入Trans-T1感受态细胞进行转化, 取转化后的沉淀物均匀涂在含氨苄抗性的LB培养板上, 12～16 h后挑菌并进行菌液PCR检测, 取阳性PCR样品进行测序, 将测序结果的序列及其翻译的氨基酸序列与山楂转录组数据进行比对, 判断目的片段是否与pMD18-T载体连接成功。

***CpSQE1*和*CpSQE2*基因的生物信息学分析** 将测序获得的序列结果使用ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/) 查找开放阅读框 (ORF)。利用在线工具ExPASy (https://web.expasy.org/protparam/) 预测*CpSQE*基因编码蛋白的分子质量、氨基酸数目、等电点、不稳定系数、脂肪指数和编码区全长等理化性质; 利用在线工具PDBsum (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/ Generate.html) 进行蛋白质二级结构分析; 利用Swiss Model (http://swissmodel.expasy.org/) 程序和PyMOL软件, 根据SQE氨基酸序列进行建模, 预测蛋白质的三级结构; 通过https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi进行蛋白质结构域分析; 利用在线工具TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/?tdsourcetag=s_pctim_aiomsg) 进行蛋白序列的跨膜结构域分析; 运用ClustalW软件及Bioedit软件与其他植物进行同源氨基酸序列比对; 用MEGA 6.06软件构建Neighbor-joining系统进化树, bootstrap重复次数为1 000次。

结果与分析

1 山楂转录组的测序和组装

山楂进行转录组测序后共获得54 475 600条raw reads, 过滤掉接头污染及低质量reads后, 得到49 370 732条clean reads, Q20比例为97.53%, GC含量为46.93%, 以上数据证明本次测序质量较好。通过Illumina Hiseq 2000平台测序, 总计产出14 518 598 800 nt数据。组装结果总Unigenes 78 496条, 总长49 538 667 nt, 平均长度941 nt, N50达到1 373 nt。Unigenes长度分布显示, 35 042条Unigenes长度大于500 nt, 占总数的60.2%; 17 949条Unigenes长度大于1 000 nt, 占总数的30.9%; 5 108条序列大于2 000 nt, 占总数的8.7%; 1 203条Unigenes长度大于3 000 nt, 占总数的2.1%。对山楂编码序列进行预测, 获取52 570条Unigenes作为CDS。30 605条CDS长度大于500 nt, 占总数的58.2%; 16 288条CDS长度大于1 000 nt, 占总数的30.9%; 3 510条序列大于2 000 nt, 占总数的6.7%; 703条CDS长度大于3 000 nt, 占总数的1.3% (表1)。

Table 1 Summary of length distribution from *Crataegus pinnatifida* Unigenes

Species		Number	≥500 nt	≥1 000 nt	≥2 000 nt	≥3 000 nt
C. pinnatifida	Unigenes	78 496	35 042	17 949	5 108	1 203
	CDS	52 570	30 605	16 288	3 501	703

2 Unigenes的注释结果统计

2.1 序列功能注释 将78 496条Unigenes序列分别与NR、NT、Swiss-Prot、KEGG、COG、GO库进行比对注释, 其中以NR数据库注释最多, 为53 972条, 占68.8%; 51 237条Unigenes (65.3%)在NT数据库比对成功得到注释, 在Swiss-Prot中注释34 557个 (44.0%), 在KEGG、GO等数据库获得注释的Unigenes数目依次为31 043 (39.5%)、35 051 (44.7%), COG数据库注释最少, 20 624条 (26.3%), 有58 395条Unigenes同时在所有数据库中注释 (表2)。

Table 2 The functional annotation statistics of *Crataegus pinnatifida* Unigenes in public databases

Database	78 496 Unigenes	
	Number annotated	Annotated unigenes ratio
NR	53 972	68.8%
NT	51 237	65.3%
Swiss-Prot	34 557	44.0%
KEGG	31 043	39.5%
COG	20 624	26.3%
GO	35 051	44.7%
ALL	58 395	74.4%

以NR数据库进行比对, 得到与注释信息相关的同源物种信息如图1所示, 在相似序列匹配度较高的物种中, 山楂与桃 *Amygdalus persica* 的同源性序列比例最高, 达71.7%; 其次为野草莓 *Fragaria vesca* (11.7%)、苹果 *Malus domestica* (3.6%)、葡萄 *Vitis vinifera* (1.9%)和大叶钻天杨 *Populus balsamifera* (1.1%)。这些物种为山楂的序列注释提供了参考。

2.2 COG分类 为预测Unigenes可能的功能, 将获得的Unigenes与COG数据库进行比对, 根据比对结果对Unigenes的功能分类并统计 (图2)。在COG 25个类别中, 一般功能基因的所占比例最大, 为16.67%; 其次为转录功能, 占9.76%, 复制、重组和修饰比例为7.66%, 翻译后修饰、蛋白质折叠和分子伴侣占7.29%, 翻译、核糖体结构与生物合成比例为7.07%。而核结构和细胞外结构的功能基因最少, 分别为17个 (0.04%)和5个 (0.01%)。此外有2 289个未确定其准确的生物学功能, 占所有功能注释信息的5.55%。

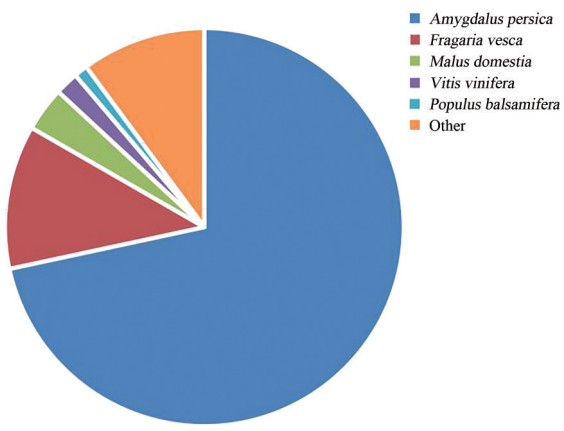

Figure 1 Species distribution of transcriptomic Unigenes against NR database

2.3 GO分类和KEGG代谢通路分析 通过对Unigenes的GO功能分析, 山楂转录组的Unigenes可分为3大类别, 包括生物过程、细胞组分和分子功能。生物过程类别中包括23个分组, 其中“细胞过程”和“发展过程”所占比例最高, 分别为20.39%、20.36%; 细胞组分中, 细胞、细胞部分所占的比例相同且最多, 为25.32%, 其次为细胞类脂质为19.03%; 分子功能中, 占比最多的为结合43.57%和催化活性39.78% (图3)。

山楂转录组数据中有42 881条Unigenes注释到了KEGG数据库中, 并被定位到128个代谢途径。其中, 细胞过程共注释到2 118条Unigenes (4.94%), 环境信息过程注释到2 090条Unigenes (4.87%), 遗传信息过程注释到9 286条Unigenes (21.66%), 新陈代谢注释到26 922条Unigenes (62.78%), 生物系统注释到2 465条Unigenes (5.75%)。

3 柠檬酸合成途径基因及关键酶分析

柠檬酸是山楂的有效成分之一, 本研究从山楂转录组中柠檬酸的生物合成通路中鉴定了15种酶52条基因, 分别为丙酮酸脱氢酶E1组分α亚单位 (AceE)、丙酮酸脱氢酶E2组分 (DLAT)、二氢硫辛酰胺脱氢酶 (DLD)、柠檬酸合酶 (CS)、柠檬酸ATP裂解酶 (ACLY)、乌头酸酶 (Aco)、异柠檬酸脱氢酶 (IDH1)、异柠檬酸脱氢酶 (IDH3)、酮戊二酸脱氢酶 (OGDH)、二氢硫辛酰琥珀酰转移酶 (DLST)、琥珀酰辅酶A合成酶α亚基 (LSC1)、琥珀酰辅酶A合成酶β亚基 (LSC2)、琥珀酸脱氢酶 (泛醌) 铁硫亚基 (SDHA)、富马酸水合酶 (fumA)、苹果酸脱氢酶 (MDH1) (图4, 表3)。在山楂3个不同发育时期, ACO3、IDH3、1个fumA和1个SDHA基因的表达量从8月初至8月底先上调表达, 在10月份呈明显的下调表达; 1个IDH3、2个IDH1、MDH1和2个DLST基因的表达量从8月至10月呈降低趋势; 1个SDHA

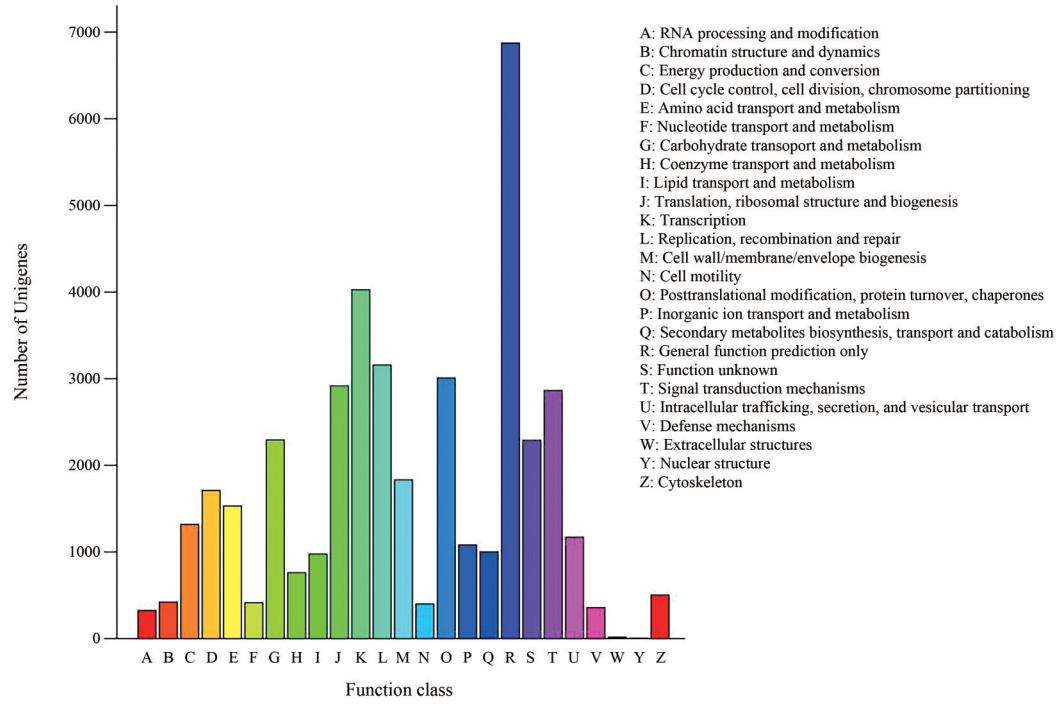

Figure 2 COG functional annotation distribution of Unigenes of transcriptome for *Crataegus pinnatifida*

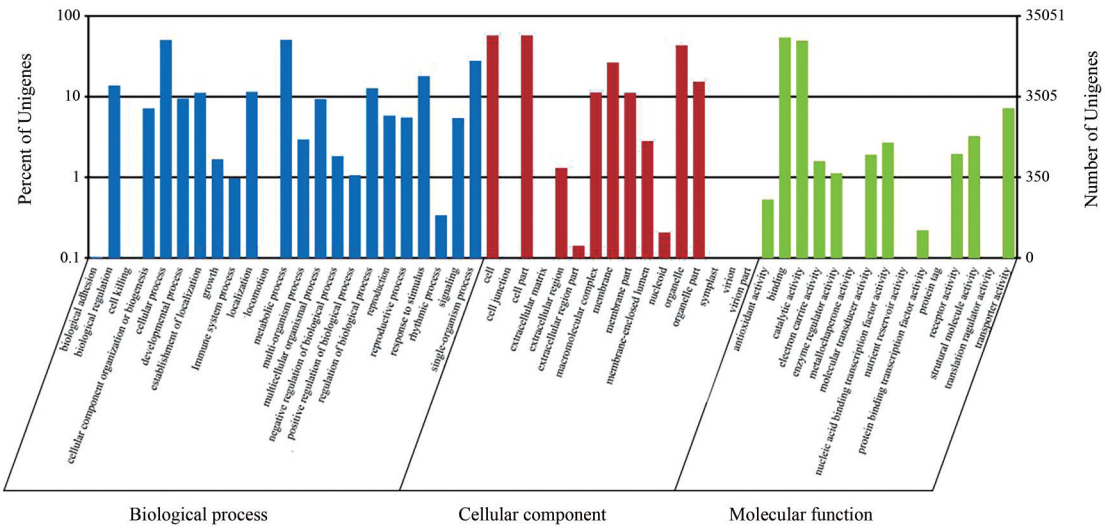

Figure 3 Functional classification of *Crataegus pinnatifida* using GO database

和1个DLST从8月初至8月底先下调表达, 从8月底至10月份呈上调表达; 柠檬酸合成途径中的AceE、DLAT等大部分基因表达量均从8月份至10月份呈上升趋势, 即在山楂果实成熟期的10月份达到最高。

4 三萜合成途径基因及关键酶分析

三萜化合物是一类重要的植物次生代谢物, 熊果酸、齐墩果酸和山楂酸是山楂的主要活性成分, 其生物合成过程主要包括甲羟戊酸途径 (MVA) 和2-*C*-甲基-*D*-赤藓醇-4-磷酸 (2-*C*-methyl-*D*-erythritol-4-phospate, MEP) 途径以及五环三萜碳环系统合成后的修饰过程。对山楂转录组数据库进行分析, 整条三萜代谢途径中各个环节的酶的62条编码基因均可被检测并注释, 分别为乙酰辅酶A酰基转移酶 (AACT)、HMG-CoA合成酶 (HMGS)、HMG-CoA还原酶 (HMGR)、甲羟戊酸激酶 (MK)、磷酸甲羟戊酸激酶 (PMK)、5-焦磷酸甲羟戊酸脱羧酶 (PMD)、DXP合成酶 (DXS)、DXP还原异构酶 (DXR)、CDP-ME合成酶 (MCT)、CDP-ME激酶 (CMK)、2-*C*-甲基-*D*-赤藓糖醇-2,4-环二磷酸合成酶 (MDS)、HMBPP合成酶 (HDS)、HMBPP还原酶 (HDR)、IPP异构酶 (IDI)、GPP合成酶 (GPPS)、FPP合成酶

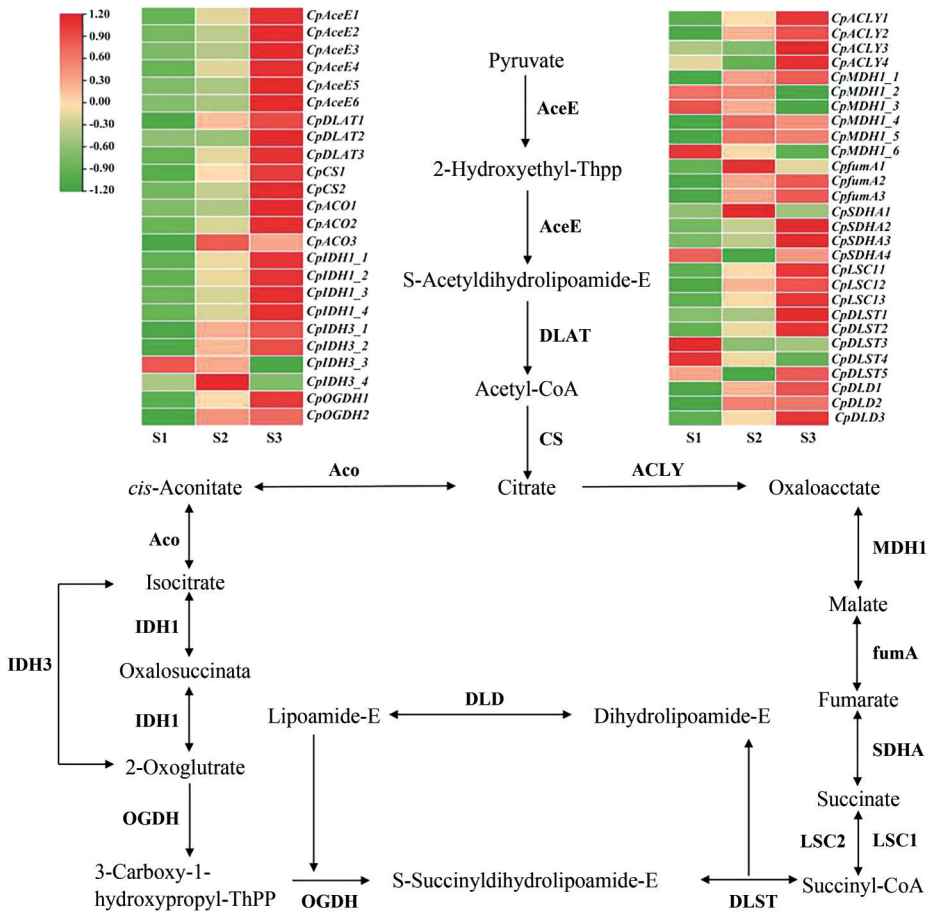

Figure 4 Citric acid biosynthesis pathway in *Crataegus pinnatifida*. S1, S2 and S3 correspond to early August, late August and October, respectively; red and green represent high and low expression levels, respectively

Table 3 Key enzyme genes involved in citric acid biosynthesis

Key enzyme	Abbreviation	EC number	Number of genes
Pyruvate dehydrogenase E1 component	AceE	EC:1.2.4.1	6
Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex	DLAT	EC:2.3.1.12	3
Dihydrolipoyl dehydrogenase	DLD	EC:1.8.1.4	3
Citrate synthase	CS	EC:2.3.3.1	2
ATP-citrate synthase	ACLY	EC:2.3.3.8	4
Aconitate hydratase	Aco	EC:4.2.1.3	3
Isocitrate dehydrogenase	IDH1	EC:1.1.1.42	4
Isocitrate dehydrogenase (NAD+)	IDH3	EC:1.1.1.41	4
2-Oxoglutarate dehydrogenase	OGDH	EC:1.2.4.2	2
Dihydrolipoamide succinyltransferase	DLST	EC:2.3.1.61	5
Succinate--CoA ligase [ADP-forming] subunit alpha	LSC1	EC:6.2.1.4	2
Succinate--CoA ligase [ADP-forming] subunit beta	LSC2	EC:6.2.1.5	1
Succinate dehydrogenase	SDHA	EC:1.3.5.1	4
Fumarate hydratase class I	fumA	EC:4.2.1.2	3
Malate dehydrogenase	MDH1	EC:1.1.1.37	6

(FPPS)、鲨烯合成酶 (SQS)、鲨烯环氧酶 (SQE)、β-香树脂醇 (β-amyrin) 和 α-香树脂醇 (α-amyrin) (表4)。为了更加清晰展示山楂三萜代谢途径关键酶基因在山楂不同生育时期的表达模式, 采用FPKM值对不同时期山楂关键酶基因表达量进行作图。从图5可以看出, 2个AACT、2个HMGS、1个HMGR、MK、PMK、PMD、CMK、GPPS、SQS、2个HDR、2个FPPS、1个SQE、1个α-AS、1个β-AS的表达从8月初至10月呈逐渐上升趋势; DXS、2个MCT、1个MDS、1个SQE的表达从8月至10月呈降低趋势, 即在10月份的表达量最低;

DXR、1 个 MCT、3 个 MDS、HDS、1 个 HDR、2 个 *α*-AS、1 个 *β*-AS 在 8 月初至 8 月底先上调表达, 在 10 月份呈明显的下调表达; 2 个 AACT、2 个 HMGS、3 个 HMGR、2 个 FPPS、1 个 *β*-AS 从 8 月初至 8 月底呈降低趋势, 8 月底表达量最低, 10 月份呈逐渐上升趋势。

5 *CpSQE1* 和 *CpSQE2* 的基因克隆和基因结构特征分析

5.1 *CpSQE* 的基因克隆 依据山楂转录组数据获得 *CpSQE1* 和 *CpSQE2* 的 cDNA, 利用 DNAMAN 软件结合 ORF Finder 在线软件对 *CpSQE1* 和 *CpSQE2* 基因 cDNA 序列进行分析, 预测 *CpSQE1* 含 1 594 bp 完整的开放阅读框, *CpSQE2* 含 1 597 bp 完整的开放阅读框, 通过设计特异性引物进行 RT-PCR 反应, 并将其重组入 pMD18-T 载体中, PCR 及测序结果与预测结果相一致 (图 6)。将 2 个基因提交到 GenBank 网站应用 BLASTX 进行序列比对, 比对结果显示 CpSQE 蛋白与 GenBank 中已注

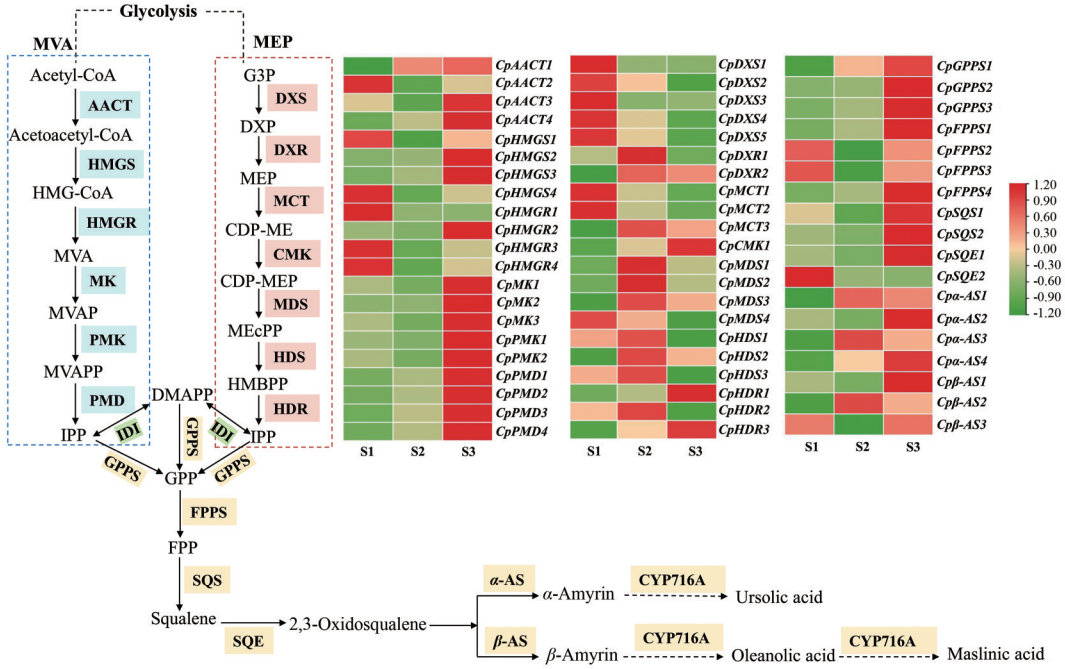

Figure 5 Triterpene biosynthesis pathway in *Crataegus pinnatifida*. S1, S2 and S3 correspond to early August, late August and early October, respectively; red and green represent high and low expression levels, respectively

Table 4 Key enzyme genes involved in triterpene biosynthesis

Key enzyme	Abbreviation	EC number	Number of genes
Acetyl-CoAc-acetyltransferase	AACT	EC:2.3.1.9	4
Hydroxymethylglutaryl-CoA synthase	HMGS	EC:2.3.3.10	4
3-Hydroxy-3-methylglutaryl-coenzyme A reductase	HMGR	EC:1.1.1.34	4
Mevalonate kinase	MK	EC:2.7.1.36	3
Phosphomevalonate kinase	PMK	EC:2.7.4.2	2
Mevalonate 5-diphosphosphate decarboxylase	MPD	EC:4.1.1.33	4
1-Deoxy-*D*-xylulose-5-phosphate synthase	DXS	EC:2.2.1.7	5
1-Deoxy-*D*-xylulose-5-phosphate reductoisomerase	DXR	EC:1.1.1.267	2
2-*C*-Methyl-*D*-erythritol 4-phosphate cytidylyltransferase	MCT	EC:2.7.7.60	3
2-*C*-Methyl-*D*-erythritol 2,4-cyclodiphosphate synthase	MDS	EC:4.6.1.12	4
4-(Cytidine-5-diphospho)-2-*C*-methylerythritol kinase	CMK	EC:2.7.1.148	1
4-Hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	HDS	EC:1.17.7.1	3
4-Hydroxy-3-methylbut-2-enyl diphosphate reductase	HDR	EC:1.17.1.2	3
Isopentenyl-diphosphate delta-isomerase	IDI	EC:5.3.3.2	1
Geranyl diphosphate synthase	GPPS	EC:2.5.1.1 2.5.1.10 2.5.1.29	3
Farnesyl diphosphate synthase	FPPS	EC:2.5.1.1 2.5.1.10	4
Squalene synthase	SQS	EC:2.5.1.21	3
Squalene epoxidase	SQE	EC:1.14.13.132	2
α-Amyrin synthase	*α*-AS	EC:5.4.99.40	4
β-Amyrin synthase	*β*-AS	EC:5.4.99.39	3

册植物 SQE 蛋白有较高的同源性, 其中 CpSQE1 和 CpSQE2 与山荆子 MbSQE (TQE12093.1) 相似度分别为 95.4%、69.30%, 与枇杷 EjSQE (AFI33133.2) 相似度分别为 91.54%、68.38%, 和桃 PpSQE (XP_007199788.1) 相似度分别为 87.13%、69.30%, 与苹果 MdSQE (XP_028948392.1) 相似度分别为 58.16%、94.77%。比对结果表明所获两条基因属于鲨烯环氧酶家族, 将该基因命名为 *CpSQE1* (GenBank 注册号为 MW915483)、*CpSQE2* (GenBank 注册号为 MW915484)。

5.2 蛋白理化性质分析 利用 ExPASy 在线软件对 *CpSQE1*、*CpSQE2* 基因编码蛋白质的理化性质进行分析, 结果显示 *CpSQE1*、*CpSQE2* 基因分别编码 530 和 531 个氨基酸; 蛋白分子质量分别为 57.6 kDa、57.5 kDa; 蛋白理论等电点 (PI) 分别为 8.71、8.88; CpSQE1 带负电残基 (Asp + Glu) 为 47, 带正电残基 (Arg + Lys) 为 54, 不稳定系数为 40.96, 属于不稳定蛋白; CpSQE2 带负电残基 (Asp + Glu) 为 45, 带正电残基 (Arg + Lys) 为 53, 不稳定系数为 42.45, 属于不稳定蛋白。

5.3 结构功能域与跨膜结构域分析 利用 NCBI 的 Conserved domains 在线工具测定分析 CpSQE1、CpSQE2 的功能结构域 (图 7), CpSQE1、CpSQE2 蛋白分别在 79～526 aa、42～527 aa 处保守区含有一个 PLN02985 superfamily 保守结构域, 均属于鲨烯单加氧酶超家族。在线软件 TMHMM 预测结果显示, CpSQE1、CpSQE2 均有 1 段跨膜结构域, 其中 CpSQE1 和 CpSQE2 蛋白 N 末端 4 aa 均位于膜内, 5～27 aa 均处于跨膜螺旋结构, 而 CpSQE1 的 28～530 aa 位于膜外, CpSQE2 的 28～531 aa 位于膜外。

5.4 二级结构和三级结构预测 PDBsum 预测表明, CpSQE1、CpSQE2 蛋白二级结构都由 α 螺旋 (alpha helix)、延伸链 (extended strand) 和无规则卷曲 (random coil) 组成。其中 CpSQE1 的 α-螺旋为 20.75%、延伸链 23.77%、无规则卷曲为 55.47%; CpSQE 2 的 α-螺旋为 26.37%、延伸链 20.90%、无规则卷曲为 52.73%。

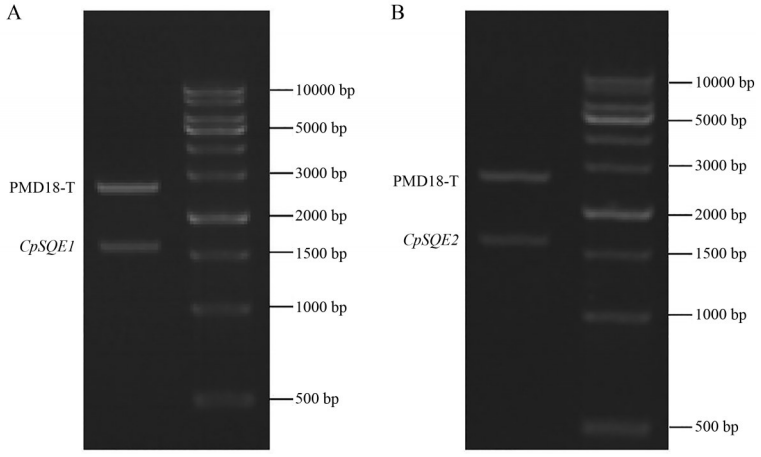

Figure 6 Cloning of *CpSQE1* and *CpSQE2*. A: Product of pMD-18T-*CpSQE1* digested by enzyme BamHI and SalI; B: Product of pMD-18T-*CpSQE2* digested by enzyme BamHI and SalI

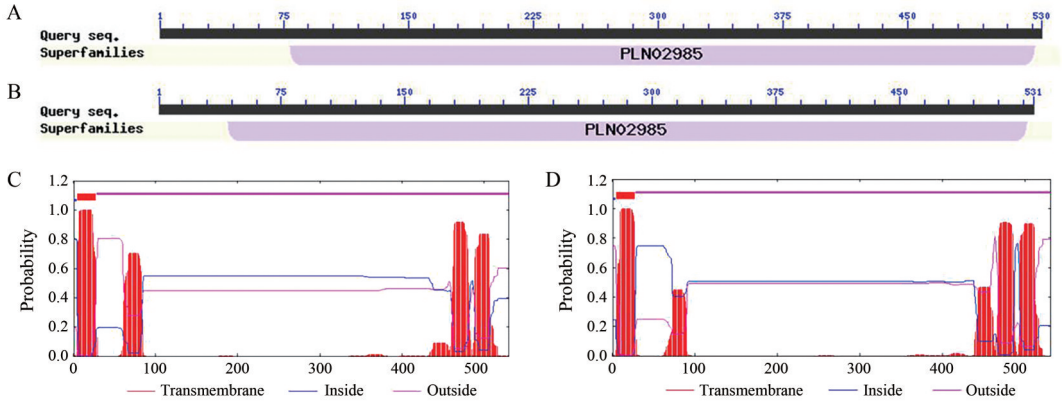

Figure 7 Structural functional domains and transmembrane domains of CpSQE1, CpSQE2. A: The predicted conserved domains of CpSQE1; B: The predicted conserved domains of CpSQE2; C: The predicted transmembrane regions of CpSQE1; D: The predicted transmembrane regions of CpSQE2

利用SWISS-MODEL和PyMOL对CpSQE进行同源建模, 得到了山楂鲨烯环氧酶的三维空间模型, 如图8所示。

5.5 氨基酸序列比对和系统进化树 *CpSQE1*和*CpSQE2*基因编码的蛋白序列的一致性为70.52%, 二者在前70 aa序列相似性较低, 具有较高的特异性。利用DNAMAN软件对山楂鲨烯环氧酶基因的氨基酸序列与其他植物的氨基酸序列进行比对, 与同属植物苹果的MdSQE的同源性为86.82%; CpSQE1和CpSQE2与芒柄花OsSQE1、OsSQE2同源性为81.40%; 与苜蓿MtSQE1、MtSQE2比对, 同源性为80.96%。比对结果显示, 山楂鲨烯环氧酶CpSQE1和CpSQE2均具有NAD(P)结合区和底物结合区。根据山楂SQE氨基酸序列与来源于不同物种SQE基因家族的氨基酸序列, 采用邻接法 (neighbor-joining) 构建系统进化树 (图9), 结果表明, 山楂的CpSQE1和CpSQE2与拟南芥的AtSQE1、AtSQE2、AtSQE3以及人参、雷公藤和芒柄花中的SQEs聚为一支, 拟南芥的AtSQE4, AtSQE5, AtSQE6聚成另外一支。

讨论

山楂是我国传统的药材, 位列于2017年国家卫生健康委员会公布的101种药食同源中药名单中, 其化学成分多样, 主要包括萜类、有机酸类、黄酮类、多糖类等。山楂中的有机酸主要有枸橼酸、苹果酸、亚麻酸等, 而三萜类化合物包括乌苏烷型、环阿屯烷型、齐墩果烷型、羊毛脂烷型和羽扇豆烷型等5种类型[25]。山楂果实的采收时间是影响山楂药用品质的重要因素之一, 采收时间过早导致产量低、山楂品质不达标, 采收时间过晚会影响山楂的储藏和加工[26]。山楂果实性状、药效成分在不同发育时期都呈现变化, 多位学者综合性状[27]和有机酸成分[28]的相关性研究确定山楂果实最佳采收期为10月份。目前对山楂不同发育时期中花色苷含量和相关合成基因也有相关研究[29], 如对不同发育期山楂果实中原花青素B2、金丝桃苷、异槲皮苷等成分进行了含量测定, 发现花色苷含量在果实成熟时达到最大值, 同时对花色苷生物合成途径28个基因的表达模式进行分析, 结果显示*Cp4CL2*、*CpCHI1*和*Cp3OGT*等15个基因与花色苷成分呈正相关, 推测以

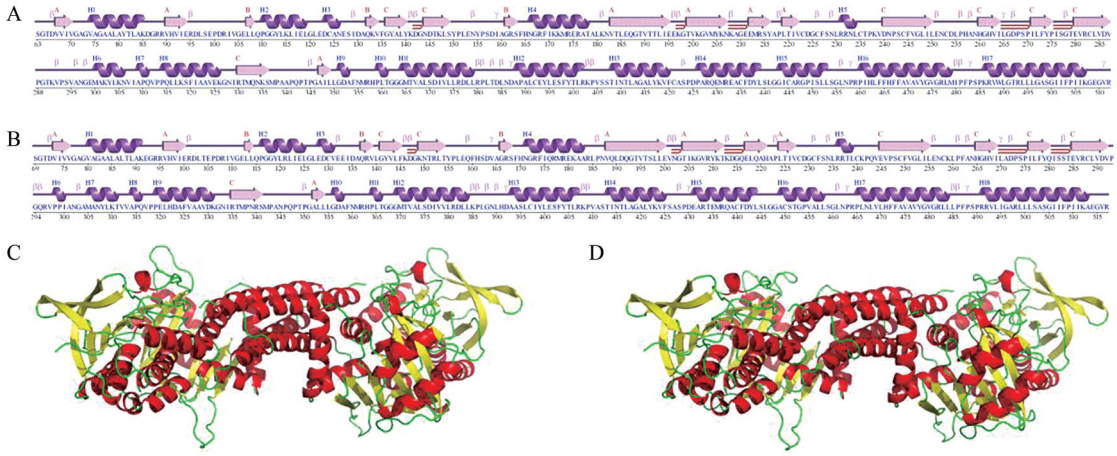

Figure 8 Secondary structure and tertiary structure model of CpSQE1 and CpSQE2. A: The predicted secondary structure of CpSQE1; B: The secondary tertiary structure of CpSQE2; C: The predicted tertiary structure of CpSQE1; D: The predicted tertiary structure of CpSQE2

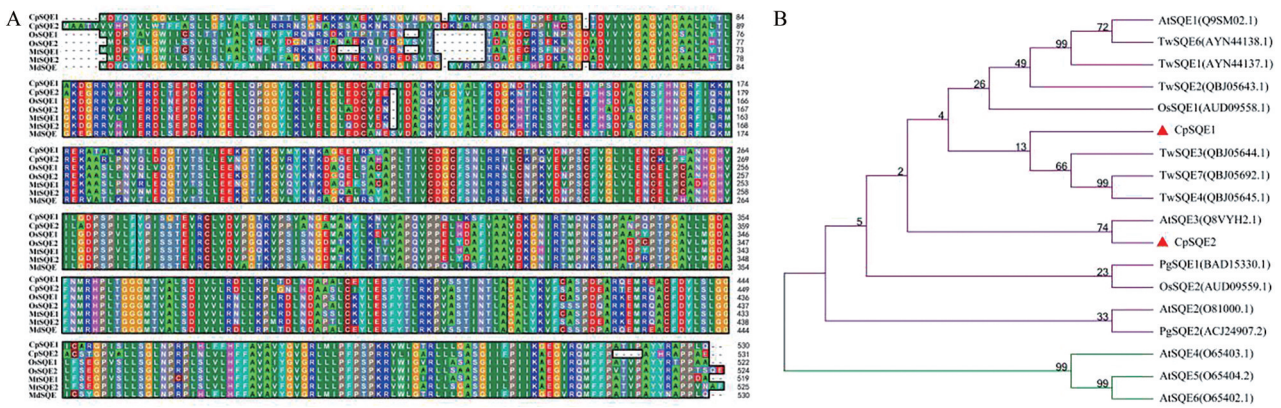

Figure 9 Alignment of amino acid sequences and phylogenetic tree analysis. A: Alignment of amino acid sequences of CpSQE1, CpSQE2 and other plant SQEs; B: Phylogenetic tree of SQE proteins

上基因可能是参与山楂果实花色苷生物合成的关键酶基因。目前对山楂的研究多集中于有效成分含量测定和花色形成相关基因、转录因子[30]等方面, 对于山楂次生代谢途径相关基因报道较少。因此本研究采用了Illumina Hiseq 2000高通量测序技术对同一产地不同发育时期的山楂果实进行转录组测序和生物信息学分析, 共获得78 496条Unigenes, 其中58 395条Unigenes能被公共数据库注释, KEGG数据库中42 881条Unigenes定位到了128个代谢途径中, 主要集中在遗传信息过程、新陈代谢等途径。同时本研究对药效成分有机酸合成途径和三萜合成途径相关基因进行了筛选和表达模式的分析, 为山楂次生代谢化合物的研究奠定基础。

柠檬酸 (citric acid), 又称枸橼酸, 药典中规定柠檬酸是山楂含量测定的指标成分。山楂中柠檬酸的积累与其生长发育阶段的关系十分密切, 山楂果实整个发育过程可分为3个时期: 幼果速长期、缓慢生长期和果实速长期。山楂处于幼果时含酸较少, 9月份有机酸开始迅速增加[31]。本研究对山楂果实8月上旬、8月下旬和10月上旬3个时期柠檬酸生物合成途径中15个代谢酶的52条Unigenes进行鉴定, 这些代谢酶参与柠檬酸生物合成的各个关键过程。柠檬酸生物合成过程中, MDH1、FumA和CS是参与柠檬酸生物合成的主要酶[32], 在山楂转录组数据中鉴定6条Unigenes编码MDH1, 3条在8月上旬高表达; 2条Unigenes编码FUM, 2条在10月上旬高表达; 2条Unigenes编码CS, 均在10月上旬高表达, 为研究这些酶调控有机酸类化合物的合成提供了基因信息。

三萜化合物是山楂中的另一类药效成分, 山楂中已发现的萜类化合物主要为四环三萜和五环三萜, 如山楂酸、科罗索酸、齐墩果酸、熊果酸等[33]。课题组前期研究发现, 熊果酸含量在山楂不同发育时期具有明显的差异, 其中S1时期最高, 在S2时期下降, S3时期最低[29], 这与李建中[34]对不同时期山楂测定的熊果酸含量变化趋势一致。本研究开展了山楂三萜生物合成途径中关键基因的挖掘与分析, 筛选出了62条20个关键酶基因, 其中DXS、2个MCT、1个MDS、1个SQE的表达从8月至10月呈降低趋势, 在10月份的表达量最低, 这与熊果酸含量变化一致。鲨烯环氧酶基因在三萜生物合成途径中起着重要作用, 法尼基焦磷酸在鲨烯合酶的催化下生成鲨烯, 鲨烯在鲨烯环氧化酶(SQE)催化下形成三萜和植物甾醇合成的共同前体2,3-氧化鲨烯[35]。本研究对山楂三萜代谢途径的2个关键酶SQE进行了克隆及生物信息学分析, 生物信息学分析表明CpSQE1和CpSQE2均属于鲨烯单加氧酶超家族。系统进化树中选取的序列为山楂*CpSQE1*和*CpSQE2*编码的氨基酸序列以及来源于其他植物中的SQE氨基酸序列, 结果分为了2支: 一支为已经验证没有鲨烯环氧酶功能的基因, 包括拟南芥中的*AtSQE4*、*AtSQE5*、*AtSQE6*, 另一支为已经鉴定出具有鲨烯环氧酶的功能的基因, 包括拟南芥[36,37]的*AtSQE1*、*AtSQE2*、*AtSQE3*和雷公藤[38]中的SQEs, 以及人参[39]和芒柄花[40]中的SQEs。从图9明显看出, 山楂CpSQE1和CpSQE2蛋白与有鲨烯环氧酶功能的基因聚在一起, 因此, 初步推测本研究中克隆的2个*CpSQE*基因很可能具有鲨烯环氧酶功能。

不同的*SQE*基因在植物中参与不同的次生代谢途径[41], 罗汉果中的2个*SQE*基因在果实中均有表达, 其中*SgSQE2*表达较为稳定, 而*SgSQE1*在15 d表达丰度最高, 与罗汉果中三萜合成途径关键酶*SgCS*基因具有完全一致的共表达模式, 推测*SgSQE1*更可能参与了甜苷的生物合成。通过对不同时期山楂关键酶基因表达量的FPKM热图分析发现, *CpSQE1*和*CpSQE2*在不同时期山楂果实中同样具有不同的表达模式, *CpSQE2*从S1期到S3期表达量呈现递减趋势, 这与山楂中的三萜化合物熊果酸含量变化一致, 推测*CpSQE2*很可能参与了熊果酸的生物合成; 而*CpSQE1*的基因表达模式与*CpSQE2*完全相反, 推测*CpSQE1*有可能参与非三萜类成分如胆固醇类、甾醇类的合成或参与三萜类的另一分支——齐墩果酸、山楂酸等化合物的形成。

根据本研究获得的不同时期山楂转录组数据的注释结果, 共有52条Unigenes编码15个关键酶参与柠檬酸循环, 62条Unigenes编码20个关键酶参与山楂的三萜合成通路, 并对2条SQE基因进行了克隆和表达分析, 为深入研究山楂生长发育和次生代谢等生物学过程功能基因的发掘提供了支持。

作者贡献: 查良平和彭华胜设计实验并提供资金; 吴君贤和徐睿参与实验、分析数据和撰写及修改论文; 尹旻臻、李景、余函纹、刘梦丽提供实验技术支持。

利益冲突: 所有作者均声明不存在利益冲突。

References

[1] Li SZ. Compendium of Materia Medica (本草纲目) [M]. Beijing: People's Medical Publishing House, 1996: 1773.

[2] Chinese Pharmacopoeia Commission. Pharmacopoeia of the People's Republic of China: Vol I (中华人民共和国药典: 第一部) [S]. Beijing: China Medical Science Press, 2020: 33.

[3] Lan SB. Research status and application prospect of the germplasm resources of *Crataegus pinnatifida* Bge. [J]. J Anhui Agric

Sci (安徽农业科学), 2016, 44: 182-184.

[4] Wu L, Zheng Q, Zhang KN, et al. Advances in safety evaluation of medicine and food homology of vegetable categories, cereal categories and others of Chinese materia medica [J]. Chin Tradit Herb Drugs (中草药), 2019, 50: 2505-2512.

[5] Wu JQ, Peng W, Qin RX, et al. *Crataegus pinnatifida*: chemical constituents, pharmacology, and potential applications [J]. Molecules, 2014, 19: 1685-1712.

[6] Zhang LL, Zhang LF, Xu JG. Chemical composition, antibacterial activity and action mechanism of different extracts from hawthorn (*Crataegus pinnatifida* Bge.) [J]. Sci Rep, 2020, 10: 8876.

[7] Dehghani S, Mehri S, Hosseinzadeh H. The effects of *Crataegus pinnatifida* (Chinese hawthorn) on metabolic syndrome: a review [J]. Iran J of Basic Med Sci, 2019, 22: 460-468.

[8] Niu CS, Chen CT, Chen LJ, et al. Decrease of blood lipids induced by Shan-Zha (fruit of *Crataegus pinnatifida*) is mainly related to an increase of PPAR*α* in liver of mice fed high-fat diet [J]. Horm Metab Res, 2011, 43: 625-630.

[9] Wen LR, Guo XB, Liu RH, et al. Phenolic contents and cellular antioxidant activity of Chinese hawthorn "*Crataegus pinnatifida*" [J]. Food Chem, 2015, 186: 54-62.

[10] Ou JM, Yang X, Shan CM, et al. Transcriptome analysis of "*Langmei*" fruits and key enzyme genes structure and function prediction involved in citric acid biosynthesis [J]. China J Chin Mater Med (中国中药杂志), 2020, 45: 4606-4616.

[11] Lu XP, Li FF, Xie SX. Citrate accumulation in citrus fruit: a molecular perspective [J]. J Fruit Sci, 2018, 35: 118-127.

[12] Li Y, Luo HM, Sun C, et al. EST analysis reveals putative genes involved in glycyrrhizin biosynthesis [J]. BMC Genomics, 2010, 11: 268.

[13] 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.

[14] Tang Q, Ma X, Mo C, et al. An efficient approach to finding *Siraitia grosvenorii* triterpene biosynthetic genes by RNA-seq and digital gene expression analysis [J]. BMC Genomics, 2011, 12: 343.

[15] Ni LH, Zhao ZL, Wu JR, et al. Analysis of transcriptomes to explore genes contributing to iridoid biosynthesis in *Gentiana waltonii* and *Gentiana robusta* (Gentianaceae) [J]. Acta Pharm Sin (药学学报), 2019, 54: 944-953.

[16] Niu YY, Zhu XX, Luo HM, et al. Development of the devices for synthetic biology of triterpene saponins at an early stage: cloning and expression profiling of squalene epoxidase genes in *Panax notoginsen* [J]. Acta Pharm Sin (药学学报), 2013, 48: 211-218.

[17] Guo HH, Li RF, Liu SB, et al. Molecular characterization, expression, and regulation of *Gynostemma pentaphyllum* squalene epoxidase gene 1 [J]. Plant Physiol Biochem, 2016, 109: 230-239.

[18] Li HH, Liu XJ, Xiao Y, et al. Cloning of squalene epoxidase genes (SQEs) and correlation analysis between their expression and ursolic acid (UA) content in suspension cells of loquat (*Eriobotrya japonica* L.) under temperature stress [J]. J Agric Biotechnol (农业生物技术学报), 2015, 23: 481-491.

[19] Wang D, Liu Y, Xu JY, et al. Construction of efficient yeast cell factories for production of ginsenosides precursor dammarenediol-II [J]. Acta Pharm Sin (药学学报), 2018, 53: 1233-1241.

[20] Cui K, Wu WW, Diao QY. Application and research progress on transcriptomics [J]. Biotechnol Bull (生物技术通报), 2019, 35: 1-9.

[21] Wang YL, Huang LQ, Yuan Y. Research advances on analysis of medicinal plants transcriptome [J]. China J Chin Mater Med (中国中药杂志), 2015, 40: 2055-2061.

[22] Liu HB, Shangguan YN, Pan YC. Applications of RNA-Seq technology on medicinal plants [J]. Chin Tradit Herb Drugs (中草药), 2019, 50: 5346-5354.

[23] Yan SM, Yuan Y, Yang B. Good method for isolating total RNA from Crataegi Fructus at different developmental stages [J]. Chin J Exp Tradit Med Form (中国实验方剂学杂志), 2016, 22: 39-43.

[24] Shan TY, Yu DQ, Han XJ, et al. Cloning and prokaryotic expression analysis of squalene synthase *CpSQS1* and *CpSQS2* from *Crataegus pinnatifida* [J]. China J Chin Mater Med (中国中药杂志), 2020, 45: 1334-1341.

[25] Dong JQ, Chen JP, Gong SX, et al. Research progress on chemical constituents and pharmacological effects of Crataegi Fructus and predictive analysis on Q-Marker [J]. Chin Tradit Herb Drugs (中草药), 2021, 52: 2801-2818.

[26] Li YZ, Huan LL. Study on fruit characteristics of Hawthorn [J]. Deciduous Fruits (落叶果树), 1995, 27: 16-18.

[27] Du RJ, Qu YJ, Zhang Y. Evaluation on optimum harvest time of hawthorn in Mudanjiang area [J]. Hort Seed (园艺与种苗), 2021, 41: 25-26.

[28] Cui J, Liu XY, Yang X, et al. Effects of different harvest time on the comprehensive quality of *Cralaegus pinnatifida*. Bge. var. *major* N. E. Br. fruits from different habitats [J]. Northwest Pharm J (西北药学杂志), 2020, 35: 633-638.

[29] Yan SM. Screening, Cloning and Characterizing of the Anthocyanidin-3-*O*-Galactosyl Transferase Gene in *Crataegus pinnatifida* (山楂矢车菊素-3-半乳糖基转移酶基因筛选、克隆及功能验证) [D]. Beijing: China Academy of Chinese Medical Sciences, 2016.

[30] Chen KQ, Lei YY, Guo YN, et al. Cloning and functional identification of a lignin regulation transcription factor MYB46 in hawthorn [J]. J Shenyang Agric Univ (沈阳农业大学学报), 2020, 51: 395-401.

[31] Qi XJ, Li ZX. Research progress on growth and development characteristics of *Crataegus pinnatifida* [J]. Northern Fruits (北方果树), 2004, 1: 4-7.

[32] Shangguan L, Sun X, Zhang C, et al. Genome identification and

analysis of genes encoding the key enzymes involved in organic acid biosynthesis pathway in apple, grape, and sweet orange [J]. Sci Hort, 2015, 185: 22-28.

[33] Liu RH, Shao F, Deng YQ, et al. Research progress on chemical constituents of *Crataegus pinnatifida* [J]. J Chin Mater Med (中药材), 2008, 31: 1100-1103.

[34] Li JZ. Analysis of Active Components in Chinese Hawthorn Fruit by Using HPLC (中国山楂果实中活性成分的HPLC分析研究) [D]. Baoding: Hebei Agricultural University, 2004.

[35] Xue Z, Tan Z, Huang A, et al. Identification of key amino acid residues determining product specificity of 2, 3-oxidosqualene cyclase in *Oryza* species [J]. New Phytol, 2018, 218: 1076-1088.

[36] Rasbery JM, Shan H, LeClair RJ, et al. *Arabidopsis thaliana* squalene epoxidase 1 is essential for root and seed development [J]. J Biol Chem, 2007, 282: 17002-17013.

[37] Laranjeira S, Amorim-Silva V, Esteban A, et al. *Arabidopsis* squalene epoxidase 3 (SQE3) complements SQE1 and is important for embryo development and bulk squalene epoxidase activity [J]. Mol Plant, 2015, 8: 1090-1102.

[38] Liu Y, Zhou JW, Hu TY, et al. Identification and functional characterization of squalene epoxidases and oxidosqualene cyclases from *Tripterygium wilfordii* [J]. Plant Cell Rep, 2020, 39: 409-418.

[39] Han JY, In JG, Kwon YS, et al. Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in *Panax ginseng* [J]. Phytochemistry, 2010, 71: 36-46.

[40] Almeida A, Dong L, Khakimov B, et al. A single oxidosqualene cyclase produces the seco-triterpenoid α-onocerin [J]. Plant Physiol, 2018, 176: 1469-1484.

[41] Zhang H, Guo J, Tang Q, et al. Cloning and expression analysis of squalene epoxidase genes from *Siraitia grosvenorii* [J]. China J Chin Mater Med (中国中药杂志), 2018, 43: 3255-3262.