地黄全长转录组测序及苯乙醇苷合成途径催化酶基因鉴定

王丰青[1*], 杨 旭[1], 左 鑫[1], 苗春妍[1], 张重义[2]

(1. 河南农业大学农学院, 河南 郑州 450002; 2. 福建农林大学农学院, 福建 福州 350002)

摘要: 地黄为玄参科药用植物, 具有重要的药用价值。为了有效发掘地黄的转录组信息, 鉴定参与苯乙醇苷(PhGs) 类成分生物合成的催化酶基因, 本研究以地黄的叶、茎和块根为材料利用Pacific Biosiences RS II平台进行测序。共获得非冗余的转录本27 773条, 平均长度2 380 bp, 预测出27 236个蛋白编码序列 (CDS: coding sequence)。利用BLAST等软件在NR、NT、GO、COG、KEGG、Swissprot和Interpro等数据库共预测到27 399个注释的基因。NR注释表明, 与地黄转录本匹配数量最多的是芝麻 (*Sesamum indicum*), 有81.44%, 这与它们进化上的亲缘关系一致。推测了参与异毛蕊花糖苷、松果菊苷、肉苁蓉苷A、肉苁蓉苷F、2′-乙酰毛蕊花糖苷和leonoside F生物合成的催化酶, 并鉴定出143个参与苯乙醇苷类成分生物合成的转录本。19个催化酶基因在地黄12个组织中与毛蕊花糖苷的含量呈正相关, 其中多数基因在叶和花中具有较高的表达量。研究结果为地黄功能基因的挖掘提供了可靠的转录组数据, 为苯乙醇苷类成分生物合成的分子机制研究提供了依据。

关键词: 地黄; 苯乙醇苷; 全长转录组; 毛蕊花糖苷; 表达特性

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870(2022)03-0831-08

Full-length transcriptome sequence and identification of genes involved in phenylethanol glycoside biosynthesis in *Rehmannia glutinosa*

WANG Feng-qing[1*], YANG Xu[1], ZUO Xin[1], MIAO Chun-yan[1], ZHANG Zhong-yi[2]

(*1. College of Agronomy, Henan Agricultural University, Zhengzhou 450002, China; 2. College of Agriculture, Fujian Agriculture and Forestry University, Fuzhou 350002, China*)

Abstract: *Rehmannia glutinosa* belongs to the Scrophulariaceae family with important medicinal value. In order to effectively explore the transcriptome information of *R. glutinosa* and identify the genes encoding enzymes involved in phenylethanol glycoside (PhGs) biosynthesis, the leaves, stems and tuberous roots of *R. glutinosa* were used for transcriptome sequencing using Pacific Biosiences RS II platform. A total of 27 773 transcripts were generated with an average length of 2 380 bp, and 27 236 coding sequences (CDS) were predicted. Using BLAST software, non-redundant transcript sequences were annotated with NR, NT, GO, COG, KEGG, SwissProt and Interpro databases and a total of 27 399 annotated genes were obtained. Among them, the number of genes related to *Sesamum indicum* in the NR database was the highest (81.44%), which is consistent with their evolutionary relationship. Enzymes likely involved in the biosynthesis of isoacteoside, echinacoside, cistanosides A, cistanosides F, 2′-acetylacteoside and leonoside F were identified, and 143 genes were identified in *R. glutinosa* full-length transcriptome. The expression levels of 19 genes correlated with acteoside content in twelve tissues of *R. glutinosa*, and most showed higher expression levels in leaf tissues and floral organs. This study provides more reliable transcriptome data for screening *R. glutinosa* for functional genes and provides a foundation for the study of the molecular mechanisms of PhGs biosynthesis.

收稿日期: 2021-10-05; 修回日期: 2021-11-11.

基金项目: 国家自然科学基金资助项目 (81872950, 81473299, 82073952).

*通讯作者 Tel / Fax: 86-371-56990188, E-mail: fqwang@henau.edu.cn

DOI: 10.16438/j.0513-4870.2021-1440

Key words: *Rehmannia glutinosa*; phenylethanol glycoside; full-length transcriptom; acteoside; expression characteristics

地黄 (*Rehmannia glutinosa* L.) 为玄参科植物, 以块根入药, 为我国著名的“四大怀药”之一。地黄药材用量大、疗效确切, 在2012年版《国家基本药物目录》所列203种中成药组方中使用频率名列第二[1], 是我国中药材及中药饮片出口前十大品种之一[2]。地黄富含苯乙醇苷 (phenylethanoid glycosides, PhGs) 类成分, 其具有抗氧化、免疫调节、增强记忆、神经保护、抗炎、抗肿瘤、保肝等药理活性[3]。其中含量较高的毛蕊花糖苷 (acteoside) 曾作为《中国药典》(2015版) 中地黄质量控制的指标性成分。虽然近年来关于地黄毛蕊花糖苷的生物合成途径及其关键酶基因已有较多研究[4-6], 然而, 其他苯乙醇类成分的生物合成机制目前还未见报道。苯乙醇苷类化合物是一类由苯乙醇苷元经糖苷键与糖基结合而成的苷类化合物, 因多数化合物糖上连有咖啡酰基或阿魏酰基, 又称其为苯丙素类化合物 (phenylpropanoid glycosides, PPGs)。目前从地黄中分离出的苯乙醇苷成分至少有26种[7,8], 含量较高的有毛蕊花糖苷、肉苁蓉苷A、松果菊苷、2′-乙酰毛蕊花糖苷和异毛蕊花糖苷等。解析苯乙醇苷类生物合成的途径及调控机制对于保障地黄药材质量的稳定性具有重要意义。

转录组测序是获得植物基因转录水平序列信息及表达丰度的有效手段。基于Illumina平台和BGI平台的二代测序具有高通量、检测阈值宽、重复性好等特点, 是筛选差异表达基因的理想工具, 已经广泛用于地黄生长发育、逆境胁迫和品质形成等分子机制研究[5,9,10]。然而, 二代测序获得的转录本片段较小, 大量转录本没有完整的开放阅读框 (ORF), 获取的遗传信息量有限。基于PacBio RS II测序平台的第三代单分子测序技术具有长读、长测序的优势, 文库构建时不再需要将转录本打断, 不经过组装直接检测转录本的完整结构, 成功地获取高可信度的剪切位点和转录本模型, 对于无参考基因组物种的转录组研究具有非常突出的优势[11-13]。然而, 地黄的全长转录组测序目前还未见报道。

本研究基于地黄全长转录组测序获得地黄的全长转录本数据库, 推导了参与6个苯乙醇苷成分生物合成的催化酶基因, 在地黄全长转录本数据库中鉴定出参与苯乙醇苷类成分合成的催化酶基因的序列, 分析了催化酶基因的表达特性及其与毛蕊花糖苷生物合成的相关性, 为揭示地黄苯乙醇苷类成分生物合成的分子机制奠定了基础。

材料与方法

实验材料 转录组测序选择地黄主栽品种温85-5, 经河南农业大学王丰青副教授鉴定为*Rehmannia glutinosa* L.。地黄种植在河南省焦作市温县亢村 (35º3′31″N, 113º7′17″E) 怀药种植基地, 4月中下旬播种, 9月20日随机选取3株长势良好的地黄植株, 分离膨大的块根、茎和叶片液氮速冻后放入−80 ℃冰箱保存备用。

RNA提取、建库和测序 混合等量的冷冻叶、茎和块根组织, 以TRIzol法提取地黄的总RNA。利用Clontech SMART cDNA第一链合成试剂盒合成全长cDNA, 利用Blue Pippin™进行片段大小筛选和PCR扩增, 将得到的cDNA加上SMRT bell接头, 构建单分子实时 (SMRT) cDNA文库。测序平台为Pacific Biosiences RS II, 测序委托华大基因科技公司进行。利用生物信息软件对全长转录本进行聚类, 得到一致性序列, 获得地黄的全长转录组数据。

功能注释 利用BLAST软件将获得的全长转录本与NR (NCBI非冗余蛋白序列数据库)、NT (NCBI非冗余核苷酸序列数据库)、COG (同源蛋白数据库)、KEGG (京都基因与基因组数据库)、Swissprot (蛋白质序列数据库) 等数据库进行比对, 获得同源基因、蛋白的注释信息。使用Blast2GO及NR的注释结果进行GO (基因功能分类数据库) 注释。利用InterProScan5进行Interpro (蛋白结构域数据库) 注释。

CDS预测 使用TransDecoder软件识别全长转录本中的候选编码区域, 首先提取最长的开放阅读框, 然后通过Blast比对SwissProt数据库和Hmmscan搜索Pfam蛋白同源序列, 从而预测编码区域。同时, 未能注释的转录本用ESTScan进行编码区预测。

基因−代谢共表达网络构建 基于课题组利用二代转录组测序获得的地黄12个组织的转录组数据[14], 获得与全长转录本一致的序列, 根据基因表达量的FPKM值, 利用DPS软件[15]与不同组织的毛蕊花糖苷含量进行Pearson相关分析。利用Cytoscape[16]软件构建催化酶基因与毛蕊花糖苷的共表达网络。分别以不同颜色的实线代表统计的相关性, 红线代表极显著正相关 ($P < 0.01$), 黄线代表显著正相关 ($P < 0.05$), 蓝线代表显著负相关 ($P < 0.05$)。

基因表达谱分析 为了分析催化酶基因在地黄不同组织中的时空表达模式, 获得共表达网络中与毛蕊花糖苷含量显著相关的候选基因在幼嫩叶 (L1)、展开叶 (L3)、衰老叶 (L5)、上部茎 (S1)、中部茎 (S2)、下部茎 (S3)、种栽 (SS)、上部块根 (HTR)、中部块根 (MTR)、幼嫩的花蕾 (YB)、成熟的花蕾 (MaY) 和完全开放的花 (MF) (图 1) 中的 FPKM 值, 分别计算其 log2 值, 利用 MeV 4.9.0[17]绘制表达量热图。

结果与分析

1 地黄的 Pacific Biosience 全长转录组测序

通过提取地黄根、茎、叶样品的总 RNA, 构建了 2 个 ISO-Seq 文库, 利用 Pacific Biosience RS II 测序平台测序 2 个 SMRT cell (测序模块), 共获得 1 053 569 121 原始数据 (311 633 reads), 插入的序列 (ROI) 数分别为 172 573 条和 139 060 条, 平均长度分别为 3 316 bp 和 3 461 bp, 序列质量均超过 0.90。2 个模块测序共获得 311 633 个 ROI, 根据序列是否包含正确的 5′端引物和 3′端引物以及 poly-A 尾, 共获得非嵌合序列 115 241 条, 长度分别为 1 440 bp 和 1 825 bp (表 1)。通过序列聚类, 进一步获得一致性序列分别为 28 766 条和 44 426 条, 其中高质量序列分别为 19 764 条和 28 652 条, 平均长度分别为 2 512 bp 和 3 382 bp。2 个模块的高质量一致性序列去冗余后获得 27 773 条最终转录本序列, 碱基数为 66 103 978 bp, 平均长度 2 380 bp, N50 长度 2 694 bp, 用于后续的分析。

2 地黄基因功能注释和功能分类

应用 Blast、Blast2GO 和 InterProScan5 软件, 把所有的转录本与 NR、NT、GO、COG、KEGG、Swissprot 和 Interpro 等数据库进行比对, 结果共有 27 399 个基因 (98.65%) 得到注释 (表 2)。地黄转录组在 NR 数据库中共注释得到 26 552 个基因, 占比 95.6%, 其中与芝麻 (*Sesamum indicum*) 比对上的序列最多, 有 21 623 个基因, 占比达到 81.44% (图 2A)。其他比对率较高的物种分别为中粒咖啡 (*Coffea canephora*)、林烟草 (*Nicotiana sylvestris*) 和葡萄 (*Vitis vinifera*) 等, 分别有 597、333 和 315 个基因 (图 2A)。在 GO 数据库中共有 8 423 个基因被注释, 根据功能可分为生物进程 (biological_process)、细胞组分 (cellular_component) 和分子功能 (molecular_function) 3 个大类 (图 2B)。地黄转录组在与 NR、COG、KEGG、Swissprot 和 Interpro 5 个数据库比对之后均注释到的基因有 12 893 个, 分别单独注释到的基因分别有 347、0、13、10 和 561 个 (图 2C)。

3 CDS 预测

为了预测基因的蛋白编码序列 (CDS), 首先利用 BLSAT 把能够最好匹配到功能数据库的转录本作为 CDS, 共预测到 26 560 个 CDS, 平均长度为 1 019 bp, GC 含量为 44.42% (表 3)。此外, 未能注释的转录本用 ESTScan 进行 CDS 预测, 共鉴定出 676 个 CDS, 平均长度为 3 142 bp, GC 含量为 44.07%。利用两种方法共鉴

Figure 1 Features of the 12 tissues of *R. glutinosa*. L1: Tender leaf; L3: Fully expanded leaf; L5: Old leaf; S1: Top of stem; S2: Middle piece of stem; S3: Lower stem; SS: Seed stock; HTR: Head of tuberous root; MTR: Middle of tuberous root; YB: Young flower bud; MaB: Mature flower bud; MF: Fully opened flower

Table 1 Results of PacBio sequencing of *R. glutinosa*

Library	Reads of insert	5′-Prime reads	3′-Prime reads	Poly-A reads	Full-length non-chimeric reads	Full-length non-chimeric read length/bp
1	172 573	104 609 (60.62%)	110 860 (64.24%)	96 684 (56.02%)	45 959 (26.63%)	1 440
2	139 060	94 476 (67.94%)	95 698 (68.82%)	90 506 (65.08%)	69 282 (49.82%)	1 825

Table 2 Summary of functional annotation results for *R. glutinosa* transcripts. Nr: NCBI non-redundant protein sequences; Nt: NCBI nucleotide sequences; KEGG: Kyoto encyclopedia of genes and gonomes; COG: Clusters of orthologous groups; GO: Gene ontology

Values	Nr	Nt	Swissprot	KEGG	COG	Interpro	GO	Overall
Number	26 552	26 521	21 057	22 724	14 735	25 374	8 423	27 399
Percentage	95.60%	95.49%	75.82%	81.82%	53.06%	91.36%	30.33%	98.65%

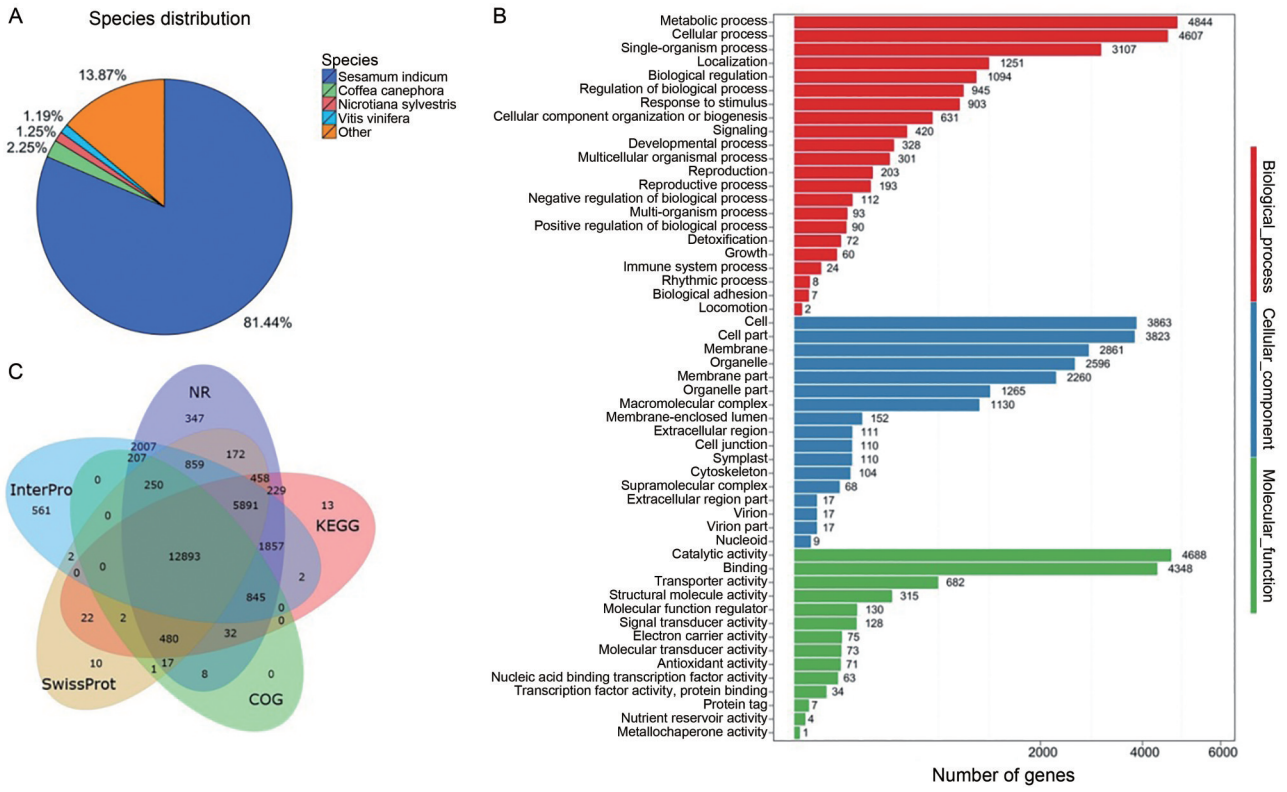

Figure 2 Functional annotation of full-length transcripts of *R. glutinosa*. A: Species distribution of annotated transcripts in NR database; B: Classification diagram of GO annotation; C: The venn diagram between NR, COG, KEGG, Swissprot and Interpro

Table 3 Quality metrics of predicted CDS. N50: A weighted median statistic that 50% of the total length is contained in CDS great than or equal to this value. GC/%: The percentage of G and C bases in all CDS

Software	Total number	Total length	Mean length	N50	N70	N90	GC/%
Blast	26 560	27 077 331	1 019	1 329	921	543	44.42
ESTScan	676	2 124 561	3 142	4 302	2 949	1 887	44.07
Overall	27 236	29 201 892	1 072	1 413	969	561	44.40

定出27 236个CDS, 平均长度为1 072 bp。

4 地黄苯乙醇苷类成分生物合成途径推导和催化酶基因鉴定

毛蕊花糖苷的生物合成途径目前已经比较清楚, 分别由苯丙氨酸 (*L*-phenylalanine) 经苯丙氨酸解氨酶 (phenylalanine ammonia-lyase, PAL)、肉桂酸-4-羟化酶 (cinnamate-4-hydroxylase, C4H)、香豆酸-3-羟化酶 (coumarate-3-hydroxylase, C3H) 和4-香豆酸辅酶A连接酶 (4-coumarate-CoA ligase, 4CL) 催化下形成的咖啡酰-CoA (caffeoyl CoA), 由酪氨酸 (tyrosine) 在多酚氧化酶 (polyphenol oxidase, PPO)、酪氨酸脱羧酶 (tyrosine decarboxylase, TyDC)/多巴脱羧酶 (DOPA decarboxylase, DODC)、铜胺氧化酶 (copper-containing amine oxidase, CuAO)、乙醇脱氢酶 (alcohol dehydrogenase, ALDH) 和糖苷转移酶 (UDP-glucose lucosyltransferase, UGT) 形成羟基酪醇苷 (hydroxytyrosol glucoside), 咖啡酰CoA和羟基酪醇苷在莽草酸-*O*-羟基肉桂酰转移酶 (shikimate O-hydroxycinnamoyltransferase, HCT) 和UGT催化下经缩合、糖苷化形成毛蕊花糖苷 (acteoside)[5]。然而, 有关其他苯乙醇苷类成分的生物合成途径仍不清楚。

本研究推测了在植物中含量比较高的苯乙醇苷类成分异毛蕊花糖苷 (isoacteoside)、松果菊苷 (echinacoside)、肉苁蓉苷A (cistanosides A)、肉苁蓉苷F (cistanosides F)、2′-乙酰毛蕊花糖苷 (2′-acetylacteoside) 和leonoside F生物合成中参与的催化酶 (图3)。肉苁蓉苷F由咖啡酸 (caffeic acid) 经UGT糖苷化而来, leonoside F则由羟基酪醇苷经UGT糖苷化和甲基化酶 (O-methyltransferase, OMT) 催化生成。异毛蕊花糖苷可能由羟基酪醇苷和咖啡酰CoA经HCT缩合及糖苷化形成, 也有可能是由毛蕊花糖苷转化而来。松果菊苷是毛蕊花糖苷在UGT催化下糖苷化合成, 而肉苁蓉苷A则由松果菊

苷甲基化而来。2′-乙酰毛蕊花糖苷是毛蕊花糖苷和乙酰CoA在乙酰转移酶作用下形成的, 与半乳糖苷乙酰转移酶 (EC: 2.3.1.18) 最为类似, 命名为2′-乙酰毛蕊花糖苷合酶 (2′-acetylacteoside synthase, AAS)。

根据转录组注释的结果, 共鉴定出143个转录本, 编码11个催化酶 (表4)。其中编码4CL的转录本最多, 有24个, 其次为HCT (19个), C4H和C3H的转录本比较少, 分别有5个和4个。由于目前半乳糖苷乙酰转移酶 (LacA) 的报道主要是在细菌和真菌中, 植物中未见报道, 在已经注释的基因里没有找到地黄的同源蛋白AAS编码基因。143个转录本平均片段大小为1 765.19 bp, 不同催化酶基因中*PAL*和*CuAO*的转录本片段平均值较大, 分别为2 045.57 bp和2 027.6 bp, *C3H*的转录本平均值最小, 仅有1 047.75 bp。

5 地黄苯乙醇苷生物合成途径催化酶基因与毛蕊花糖苷含量的相关性分析

利用获得的三代转录本数据, 以鉴定出的苯乙醇苷生物合成途径催化酶的转录本为查询序列, 在本实验室已有的数据库中进行同源比对, 共鉴定到78条高度一致的序列, 发现编码同一催化酶的多条全长转录本同时比对到同一条Unigene。课题组前期对地黄叶、茎、块根和花器官等12个组织的毛蕊花糖苷含量和转录组进行了分析[14], 通过分析78个催化酶基因与毛蕊花糖苷的相关性, 构建了毛蕊花糖苷与催化酶基因的共表达调控网络 (图4)。有14个催化酶基因与毛蕊花糖苷含量呈极显著正相关 ($P < 0.01$), 包括3个*PPO*基因 (Full_16519、Full_16855和Full_17485)、3个*PAL*基因 (Full_9693、Full_14537和Full_10312)、2个*HCT*基

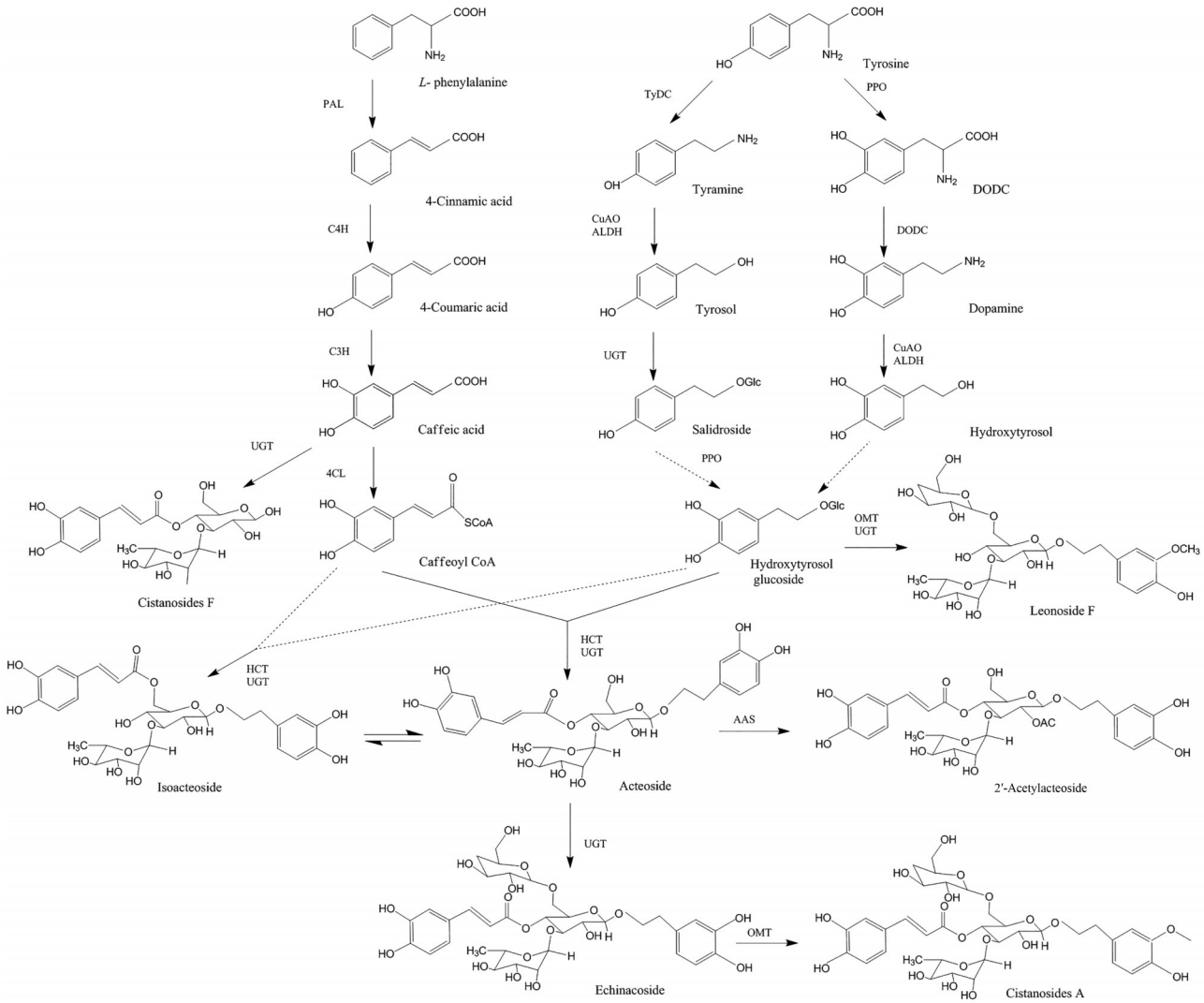

Figure 3 Biosynthetic pathway of phenylethanoid glycosides. PAL: Phenylalanine ammonia-lyase; C4H: Cinnamate-4-hydroxylase; C3H: Coumarate-3-hydroxylase; TyDC: Tyrosine decarboxylase; PPO: Polyphenol oxidase; CuAO: Copper-containing amine oxidase; ALDH: Alcohol dehydrogenase; UGT: UDP-glucose lucosyltransferase; 4CL: 4-Coumarate-CoA ligase; HCT: Shikimate O-hydroxycinnamoyltransferase; AAS: 2′-Acetylacteoside synthase; OMT: O-Methyltransferase

Table 4 Identified enzyme genes involved in phenylethanoid glycosides biosynthesis

Gene	No. of trancript	Average length/bp	Range of lengthe/bp	Enzyme	Enzyme code
PAL	14	2 045.57	549−2 533	Phenylalanine ammonia-lyase	4.3.1.24
C4H	5	1 376.80	674−1 776	*trans*-Cinnamate 4-monooxygenase	1.14.13.11
C3H	4	1 047.75	462−1 857	Coumaroylquinate (coumaroylshikimate) 3′-monooxygenase	1.14.13.36
4CL	24	1 867.88	476−3 890	4-Coumarate-coa ligase	6.2.1.12
CuAO	18	2 027.06	560−2 711	Copper amine oxidase	1.4.3.21
ALDH	9	1 574.56	1 256−2 138	Cinnamyl alcohol dehydrogenase	1.1.1.195
TyDC/DODC	8	1 700.13	996−2 353	Tyrosine decarboxylase	4.1.1.25
PPO	10	1 800.2	1 019−2 059	Polyphenol oxidase	1.10.3.1
UGT	17	1 748.76	1 495−2 432	UDP-glycosyltransferase	2.4.1.-
HCT	19	1 730.63	852−3 294	Shikimate O-hydroxycinnamoyltransferase	2.3.1.133
AAS	0	−	−	Galactoside O-acetyltransferase	2.3.1.18
OMT	15	1 487.6	709−3 495	O-Methyltransferase	2.1.1.-
Total	143	1 765.19	462−3 890		

因 (Full_19075 和 Full_19442)、2 个 *C4H* 基因 (Full_19380 和 Full_21774), *4CL*、*TyDC*、*UGT* 和 *C3H* 基因各 1 个。1 个 C4H 编码基因 Full_21774 的表达量与毛蕊花糖苷含量的相关性最强, 相关系数达到 0.92。此外, 还有 3 个 *UGT* 基因 (Full_19832、Full_15803 和 Full_20225) 和 2 个 *CuAO* 基因 (Full_9276 和 Full_18410) 的表达量与毛蕊花糖苷含量呈显著正相关 ($P < 0.05$)。另外, 还有 6 个基因的表达量与毛蕊花糖苷含量呈显著负相关 ($P < 0.05$), 是否参与其他苯乙醇苷类成分的生物合成有待于进一步研究。

6 地黄苯乙醇苷生物合成途径催化酶基因在地黄不同组织的时空表达模式

分析根据基因−代谢调控网络筛选到的与毛蕊花糖苷合成呈正相关或负相关的催化酶基因的表达特性, 结果 (图 5) 表明, 25 个催化酶基因被分为 4 个类群。类群 I 包含 6 个基因, 其主要特征是在叶片中具有

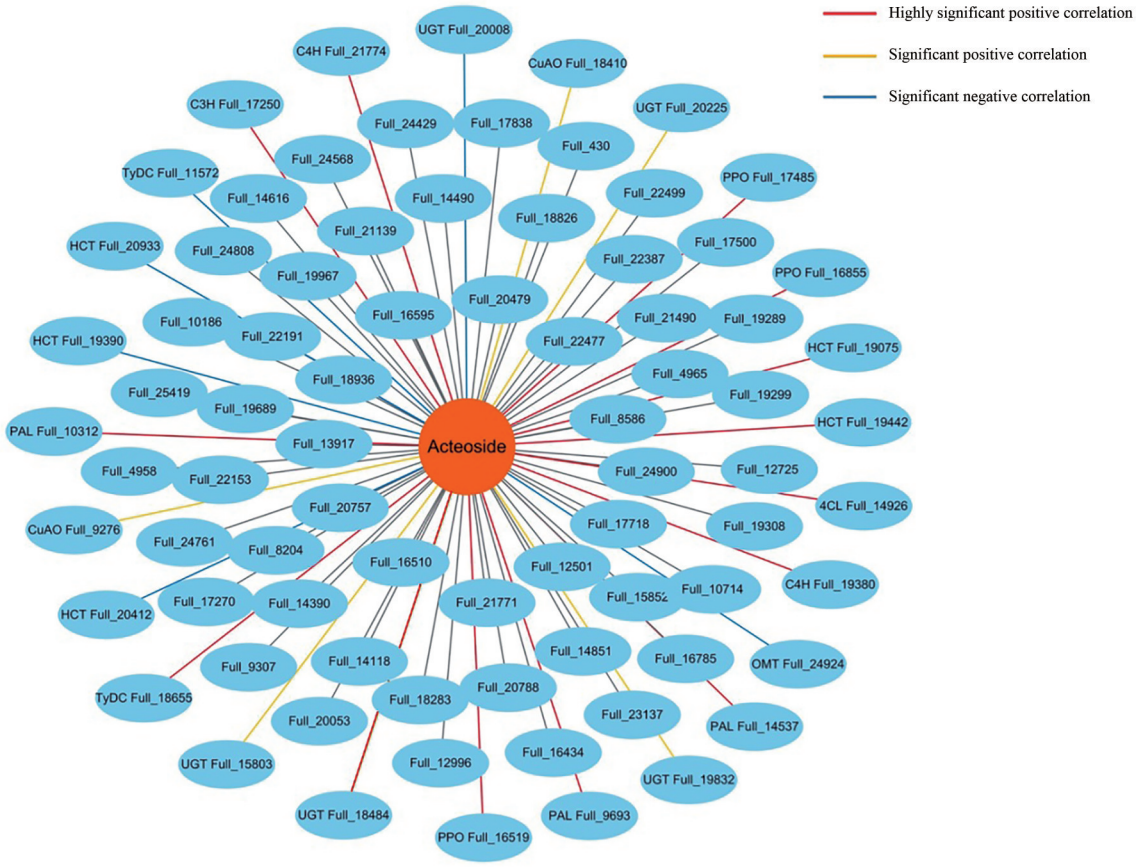

Figure 4 Gene-metabolite correlation network representing the enzyme genes and acteoside involved in phenylethanoid glycosides biosynthesis

较高的表达量, 尤其是衰老的叶片 (L5) 中的表达量高, 同时在花器官中也有较高的表达量, 尤其是幼嫩的花蕾 (YB), 这与地黄毛蕊花糖苷的积累模式相似。类群 II 包含 8 个基因, 它们的表达模式与类群 I 的类似, 主要差异表现为这些基因在茎的不同部位 (S1～S3) 和块根不同部位 (SS、HTR 和 MTR) 中的表达量较低, 尤其是块根中的表达量, 这与毛蕊花糖苷的含量变化一致性很高。类群 III 中包含 5 个基因, 其主要特征是在叶中的表达量较高, 尤其是衰老的叶片中表达量最高, 与毛蕊花糖苷的含量也呈正相关。类群 IV 包含 6 个基因, 这些基因的表达均与毛蕊花糖苷的含量呈负相关, 其主要特征是在茎和块根的不同部位中表达量较高, 如编码甲基化酶的 OMT 基因 Full_24924, 可能参与特定苯乙醇苷成分的生物合成。

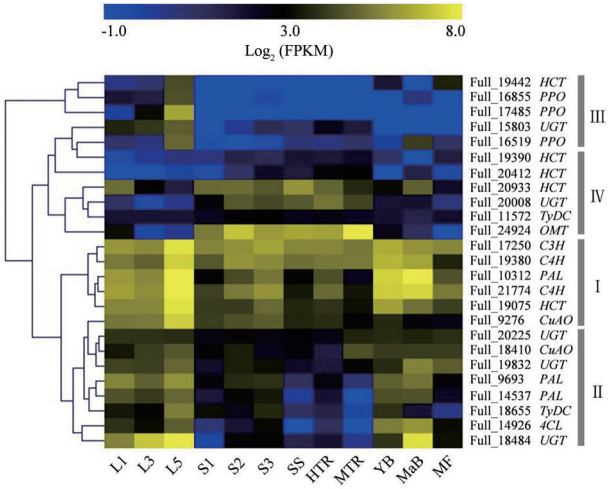

Figure 5 Heat map representing expression dynamics of enzyme genes involved in phenylethanol glycoside biosynthesis pathway in different tissues. L1: Tender leaf; L3: Fully expanded leaf; L5: Old leaf; S1: Top of stem; S2: Middle piece of stem; S3: Lower stem; SS: Seed stock; HTR: Head of tuberous root; MTR: Middle of tuberous root; YB: Young flower bud; MaB: Mature flower bud; MF: Fully opened flower

讨论

目前虽然地黄的基因组已被报道[18], 但由于地黄可能是同源四倍体, 且基因组庞大, 杂合度高, 完整解析地黄的遗传信息仍需要大量的转录组数据。基于第二代高通量测序获得的转录组数据库虽然能够为解析地黄发育、胁迫响应和次生代谢产物合成机制提供宝贵的基因信息[4,14,19], 但由于测序技术方面的不足, 冗余片段较多, 拼接片段短, 大量基因没有全长的编码序列, 限制了应用转录组技术对地黄基因的功能研究。如 Zhou 等[4]应用 Illumina HiSeq 2500 测序平台获得地黄块根和叶混合样品的 96 961 条 Unigene, 平均长度 678.3 bp。Zhi 等[19]利用 Illumina HiSeq 2500 测序平台测序 4 个地黄品种块根菊花心和非菊花心共 24 个样品的转录组, 组装获得 150 405 条 Unigene, 平均长度 1 244 bp。本研究利用第三代单分子测序共获得 27 773 条最终转录本序列, 平均长度达 2 380 bp, 测序质量远高于第二代测序技术。

苯乙醇苷类成分具有重要的药理活性, 在植物中广泛分布。毛蕊花糖苷是含量较高的一种苯乙醇苷成分, 据报道至少在 150 种植物中发现了毛蕊花糖苷[20], 其生物合成途径目前研究较多。最早是利用前体饲喂橄榄 (*Olea europaea*) 细胞实验初步构建了毛蕊花糖苷的生物合成途径, 认为来源于酪氨酸途径的羟基酪醇和来源于苯丙氨酸途径的咖啡酸共同合成毛蕊花糖苷[21]。课题组进一步完善了毛蕊花糖苷的生物合成途径, 推测咖啡酸在 4CL 催化下生成咖啡酰基辅酶 A, 红景天苷在 PPO 催化下生成羟基酪醇苷或羟基酪醇在 UGT 催化下生成羟基酪醇苷, 咖啡酰基辅酶 A 和羟基酪醇苷在 HCT 和 UGT 作用下经缩合、糖苷化修饰产生毛蕊花糖苷[5]。本研究通过分析几个在地黄中含量较高的苯乙醇苷类成分的分子结构, 推测了参与异毛蕊花糖苷、松果菊苷、肉苁蓉苷 A、肉苁蓉苷 F、2-乙酰毛蕊花糖苷和 leonoside F 的催化酶, 并在地黄中鉴定出除 AAS 以外的其他苯乙醇苷成分合成的催化酶基因, 为进一步研究这些成分在地黄中的合成机制提供了依据。

催化酶基因鉴定和功能研究是解析药效成分生物合成机制的关键。课题组前期利用二代转录组测序数据在地黄中鉴定出毛蕊花糖苷生物合成途径的 219 个催化酶基因, 其中 54 个基因在水杨酸诱导后上调表达[5]。本研究共鉴定出 143 个参与苯乙醇苷成分生物合成途径的催化酶基因转录本, 检测到的数量少于之前利用二代数据获得转录本, 但转录本的长度明显较二代转录本更长。可能的原因是三代测序获得的全长转录组中转录本数量少于二代测序, 虽然二代测序获得的转录本数量多, 但长度短, 且与测序时的样本量有关[22]。利用全长转录组获得的催化酶基因转录本的数量虽然少, 但质量更高, 进行表达谱分析及功能研究更加可靠。结合基因组信息是去除冗余转录本或假转录本的有效手段, 然而, 由于已经报道的地黄基因组测序结果仍需完善 (染色体挂载率约为 56.21%), 用于分析本研究鉴定的全长转录本存在困难。主要苯乙醇苷成分毛蕊花糖苷在地黄的叶和花器官中含量较高, 其次为茎, 在块根中含量最低[14]。基因-代谢调控网络分析发现 19 个催化酶基因的表达量与毛蕊花糖苷的含量呈正相关, 催化酶基因的表达量也是在叶和花中比较高, 在块根中表达量较低, 可能参与毛蕊花糖苷的生物合成。

6个催化酶基因与毛蕊花糖苷含量呈负相关, 其中包括1个OMT基因Full_24924, 说明此Full_24924可能参与leonoside F或肉苁蓉苷A的生物合成。

本研究首次利用Pacific Biosience RS II测序平台获得了地黄的三代全长转录组, 推导了苯乙醇苷类成分的生物合成途径并对催化酶基因进行了鉴定, 分析了催化酶基因与毛蕊花糖苷的调控关系, 为地黄苯乙醇苷类成分生物合成的分子机理奠定了基础。

作者贡献: 王丰青是本文的第一作者和通讯作者, 负责研究工作的实验设计, 生物合成途径推导、催化酶基因鉴定、表达谱分析和论文的撰写; 杨旭负责样品采集工作, 参与论文的撰写; 左鑫参与数据分析工作; 苗春妍参与实验材料的种植和管理; 张重义参与研究内容的设计和稿件修改。

利益冲突: 本文的研究内容无任何利益冲突。

References

[1] Yang CR, Xu M, Song H, et al. National essential drugs and traditional Chinese medicine resources [J]. Mod Chin Med (中国现代中药), 2016, 18: 1513-1520.

[2] Huo W, Jiang L. Analysis of export trend of Chinese medicinal herbs and decoction pieces in 2015 [J]. Mod Chin Med (中国现代中药), 2016, 18: 512-514.

[3] Xue ZZ, Yang B. Phenylethanoid glycosides: research advances in their phytochemistry, pharmacological activity and pharmacokinetics [J]. Molecules, 2016, 21: 991.

[4] Zhou YQ, Wang XN, Wang WS, et al. *De novo* transcriptome sequencing-based discovery and expression analyses of verbascoside biosynthesis-associated genes in *Rehmannia glutinosa* tuberous roots [J]. Mol Breed, 2016, 36: 139.

[5] Wang FQ, Zhi JY, Zhang ZY, et al. Transcriptome analysis of salicylic acid treatment in *Rehmannia glutinosa* hairy roots using RNA-seq technique for identification of genes involved in acteoside biosynthesis [J]. Front Plant Sci, 2017, 8: 787.

[6] Li XR, Zhi JY, Yang CF, et al. Cloning, subcellular location and expression analysis of an acteoside synthase gene from *Rehmannia glutinosa* [J]. Chin Tradit Herb Drugs (中草药), 2020, 51: 4739-4746.

[7] Li HW, Meng XL. Research progress on chemical constituents and pharmacological activities of *Rehmannia glutinosa* [J]. Drug Eval Res (药物评价研究), 2015, 38: 218-228.

[8] Gao Y, Peng CY, Chen XY, et al. Studies on the phenylethanoid glycosides from the fresh roots of *Rehmannia glutinosa* [J]. J Chin Med Mater (中药材), 2017, 40: 2073-2076.

[9] Yang YH, Li MJ, Che XJ, et al. *De novo* characterization of the *Rehmannia glutinosa* leaf transcriptome and analysis of gene expression associated with replanting disease [J]. Mol Breed, 2014, 34: 905-915.

[10] Wang FQ, Li XR, Yang CF, et al. Effects of shading on tuberous root traits, photosynthetic characteristics and gene transcription of *Rehmannia glutinosa* [J]. Chin Tradit Herb Drugs (中草药), 2019, 50: 4419-4429.

[11] Chen XZ, Li JR, Wang XB, et al. Full-length transcriptome sequencing and methyl jasmonate-induced expression profile analysis of genes related to patchoulol biosynthesis and regulation in *Pogostemon cablin* [J]. BMC Plant Biol, 2019, 19: 266.

[12] Chao YH, Yuan JB, Guo T, et al. Analysis of transcripts and splice isoforms in *Medicago sativa* L. by single-molecule long-read sequencing [J]. Plant Mol Biol, 2019, 99: 219-235.

[13] Zhang H, Jin JJ, Xu GY, et al. Reconstruction of the full-length transcriptome of cigar tobacco without a reference genome and characterization of anion channel/transporter transcripts [J]. BMC Plant Biol, 2021, 21: 299.

[14] Wang FQ, Li XR, Zuo X, et al. Transcriptome-wide identification of WRKY transcription factor and functional characterization of RgWRKY37 involved in acteoside biosynthesis in *Rehmannia glutinosa* [J]. Front Plant Sci, 2021, 12: 739853.

[15] Tang QY, Zhang CX. Data Processing System (DPS) software with experimental design, statistical analysis and data mining developed for use in entomological research [J]. Insect Sci, 2013, 20: 254-260.

[16] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks [J]. Genome Res, 2003, 13: 2498-2504.

[17] Howe EA, Sinha R, Schlauch D, et al. RNA-Seq analysis in MeV [J]. Bioinformatics, 2011, 27: 3209-3210.

[18] Ma LG, Dong CM, Song C, et al. *De novo* genome assembly of the potent medicinal plant *Rehmannia glutinosa* using nanopore technology [J]. Comput Struct Biotechnol J, 2021, 19: 3954-3963.

[19] Zhi JY, Li YJ, Zhang ZY, et al. Molecular regulation of catalpol and acteoside accumulation in radial striation and non-radial striation of *Rehmannia glutinosa* tuberous root [J]. Int J Mol Sci, 2018, 19: 3751.

[20] He J, Hu XP, Zeng Y, et al. Advanced research on acteoside for chemistry and bioactivities [J]. J Asian Nat Prod Res, 2011, 13: 449-464.

[21] Saimaru H, Orihara Y. Biosynthesis of acteoside in cultured cells of *Olea europaea* [J]. J Nat Med, 2010, 64: 139-145.

[22] Kang H, Zhao ZL, Ni LH, et al. Transcriptome analysis and exploration of genes involved in the biosynthesis of iridoids in *Gentiana crassicaulis* (Gentianaceae) [J]. Acta Pharm Sin (药学学报), 2021, 56: 2005-2014.