基于高通量测序 454 GS FLX 的丹参转录组学研究

李 滢[1], 孙 超[1], 罗红梅[1], 李西文[1,2], 牛云云[1], 陈士林[1*]

(1. 中国医学科学院、北京协和医学院药用植物研究所, 北京 100193; 2. 清华大学化学系, 北京 100084)

摘要: 本研究应用新一代高通量测序技术 454 GS FLX Titanium 对 2 年生丹参根的转录组进行测序, 研究其基因表达谱, 挖掘其功能基因。获得 46 722 表达序列标签 (express sequence tags, EST), 序列平均长度 414 bp, 与 Sanger 测序的长度相当。所得序列与 GenBank 丹参 EST 合并拼接, 获得 18 235 条 unigene, 其中, 454 高通量测序发现了 13 980 条新的 unigene。数据库中的序列同源性比较表明, 其中 73.0% (13 308 条) 与其他生物的已知基因具有不同程度的同源性。通过 BLAST 与 Gene Ontology 分析获得了可能参与丹参酮合成的序列 27 条 (编码 15 个关键酶), 参与丹酚酸合成的序列 29 条 (编码 11 个关键酶), 细胞色素 P450 序列 70 条, 转录因子序列 577 条。454 高通量测序技术作为药用植物功能基因组研究的重要手段可在丹参功能基因的发现中发挥重要作用, 这些基因的发现为丹参酮和丹酚酸类化合物生物合成研究奠定了基础, 同时也为丹参的转录组研究提供了基础数据。

关键词: 丹参; 454 GS FLX; 表达序列标签; 转录组

中图分类号: R931 **文献标识码**: A **文章编号**: 0513-4870 (2010) 04-0524-06

Transcriptome characterization for *Salvia miltiorrhiza* using 454 GS FLX

LI Ying[1], SUN Chao[1], LUO Hong-mei[1], LI Xi-wen[1,2], NIU Yun-yun[1], CHEN Shi-lin[1*]

(*1. Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China; 2. Department of Chemistry, Tsinghua University, Beijing 100084, China*)

Abstract: To investigate the profile of gene expression in *Salvia miltiorrhiza* and elucidate its functional gene, 454 GS FLX platform and Titanium regent were used to produce a substantial expressed sequence tags (ESTs) dataset from the root of *S. miltiorrhiza*. A total of 46 722 ESTs with an average read length of 414 bp were generated. 454 ESTs were combined with the *S. miltiorrhiza* ESTs from GenBank. These ESTs were assembled into 18 235 unigenes. Of these unigenes, 454 sequencing identified 13 980 novel unigenes. 73% of these unigenes (13 308) were annotated using BLAST searches (E-value ≤ 1e-5) against the SwissProt, KEGG, TAIR, Nr and Nt databases. Twenty-seven unigenes (encoding 15 enzymes) were found to be involved in tanshinones biosynthesis, and 29 unigenes (encoding 11 enzymes) involved in phenolic acids biosynthesis. Seventy putative genes were found to encode cytochromes P450 and 577 putative transcription factor genes. Data presented in this study will constitute an important resource for the scientific community that is interested in the molecular genetics and functional genomics of *S. miltiorrhiza*.

Key words: *Salvia miltiorrhiza*; 454 GS FLX; expressed sequence tags; transcriptome

丹参 (*Salvia miltiorrhiza* Bge.) 为唇形科 (Labiatae) 鼠尾草属常用中药, 以根和根茎入药[1]。目前丹参类药物年销售额逾 10 亿元, 至 2009 年 9 月, 国家食品药品监督管理局颁布的丹参制剂生产批文已近千份。现代化学及药理学研究表明丹参含有两类生理活性物质: 脂溶性的丹参酮类化合物和水溶性的丹酚酸类化合物。虽然丹参在化学和药理学研究方面已有了广泛的研究, 但其活性成分代谢途径的研究较少。丹

收稿日期: 2010-01-25.
基金项目: 国家自然科学基金资助项目 (30772735).
*通讯作者 Tel: 86-10-62899700, Fax: 86-10-62896313, E-mail: slchen@implad.ac.cn

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参酮类生物合成途径的研究, 仅限于萜类共同生物合成途径研究, 特别是GGPP下游合成途径报道更少, 仅柯巴基焦磷酸合酶 (CPS) 和类贝壳烯醇合酶 (KSL)[2]有报道。丹酚酸类化合物中的迷迭香酸生物合成途径已经得到阐明, 且已成功克隆途径中5个关键酶基因[3−7]。但丹酚酸类其他成分, 如丹酚酸A、B、C和E的生物合成途径还未见文献报道。随着化学创新药物研制难度的加大, 研究者把目光转向了药用植物有效成分生物合成的研究。

丹参基因组及转录组数据的缺乏给丹参酮类和丹酚酸类化合物的次生代谢途径的研究带来了困难。基因表达序列标签 (expressed sequence tags, EST) 技术被认为是一种研究转录组的有效方法, 广泛应用于新基因发现、基因表达分析和蛋白质组学[8−11]。传统 EST 技术也被应用到药用植物次生代谢基因的发掘中, 已有研究对一些重要的药用植物建立了 EST 文库, 如西洋参[12, 13]和蛇足石杉[14]。GenBank 已有10 288条丹参EST, 但现有数据尚不足以挖掘丹参的功能基因。这些 EST 数据都是通过传统 Sanger 测序方法获得的。新一代高通量测序技术 454 GS FLX (454 Life of Science, Roche)[15]是对传统测序方法的一次革命性变革。相对于传统的96道毛细管测序, 454 GS FLX 技术一次测序可产生100万条序列, 序列平均长度约 400 bp, 数据总量约 500 M。应用 454 GS FLX 高通量测序技术进行转录组研究, 大大降低了测序所需时间和成本, 使我们能够进行药用植物转录组和次生代谢相关基因的研究。

本实验将 454 GS FLX Titanium 高通量测序技术应用到药用植物转录组的研究中, 对丹参根转录组进行测序, 并应用生物信息学方法对所得 EST 序列与 GenBank 丹参 EST 序列进行分析, 试图从功能基因组水平上研究丹参重要基因的表达。这些重要基因包括丹参酮和丹酚酸类化合物生物合成相关的关键酶基因、参与植物次生代谢的基因、转录因子等。这些基因的发现为进一步克隆其全长、研究其功能提供了基础数据, 同时也为丹参酮和丹酚酸类化合物的生物合成研究奠定了基础。本实验为应用生物技术方法获取丹参有效成分或其中间体提供了一定的科学依据。

材料与方法

材料 2年生栽培丹参采自山东省平邑县, 流水洗净其根, 用吸水纸吸干表面水分, 迅速将根切成约2 mm厚的薄片, 立即用液氮冷冻, 存于−80 ℃备用。

RNA 提取和反转录 采用通用植物总 RNA 提取试剂盒 (百泰克公司) 提取丹参根总 RNA, Oligotex® mRNA kit (Qiagen) 分离纯化 mRNA。以 2 μg mRNA 为模板, SMART™ PCR cDNA Synthesis kit (Clontech) 反转录合成 cDNA。采用 PCR Advantage II polymerase (Clontech) 对 cDNA 进行扩增, 扩增条件为 95 ℃, 1 min; 94 ℃, 15 s; 65 ℃, 30 s; 68 ℃, 6 min, 13 个循环。采用 PureLink™ PCR Purification kit (Invitrogen) 去除体系中小于 300 bp 的片段。

454 文库构建和测序 应用新一代高通量测序平台 454 GS FLX Titanium 对 cDNA 样品测序。5 μg 双链 cDNA 打断为 300～800 bp 的片段后, 两端添加特异性衔接子 A 和 B, 变性为单链连接到磁珠上, 经 emPCR 富集后, 置于PicoTiterPlate板上, 上机测序[15]。

序列拼接 采用 GS-FLX Software 去除衔接子区域和低质量序列, 屏蔽 SMART PCR 引物。将测序序列与 GenBank 丹参 EST 序列合并, 经 GS De Novo Assembler Software 进行拼接。所有分析使用默认参数。

功能注释、分类和代谢途径分析 使用 BLAST 程序将拼接所得 unigene 与核酸、蛋白质序列数据库比对 (E 值<1e-5), 并选取最佳注释。蛋白质数据库包括 SwissProt、KEGG、拟南芥蛋白质组数据库 TAIR9 和 GenBank 非冗余蛋白数据库 Nr; 核酸数据库为 GanBank 非冗余核酸数据库 Nt。

根据 TAIR9 注释所含 Gene Ontology (GO) 信息, 对序列 (按分子功能、细胞组分、生物学过程) 进行分类[16]。根据 KEGG 注释的基因功能信息, 对参与次生代谢的序列 (按次生代谢物种类) 进行分类。

对所有注释信息整理, 搜索丹参酮和丹酚酸类化合物生物合成途径中的关键酶基因、细胞色素 P450 基因及可能参与或调控次生代谢的和转录因子等基因。

简单重复序列分析 (SSR) 搜索及分析 在 unigene 中搜索 SSR 位点, 设置参数如下: 总重复序列长度不低于 20 bp; 二核苷酸、三核苷酸、四核苷酸、五核苷酸和六核苷酸至少重复次数分别为 10、7、5、4 和 4[17]。

结果

1 454 测序和 EST 序列拼接

采用 454 GS FLX Titanium 高通量测序技术对 2 年生丹参根的转录组进行测序, 1/8 的测序反应即获得 46 722 条 EST 序列, 平均长度为 414 bp。所得高

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

通量测序数据提交至 GenBank Sequence Read Archive (SRA), 登录号 SRX017265。搜索 GenBank dbEST 数据库, 获得 10 288 条 Sanger 测序的丹参 EST 序列, 平均长度为 447 bp。合并两个 EST 文库, 经过软件拼接, 获得 18 235 条 unigene, 包括 6 620 个序列重叠群 (contig) 和 11 615 条 singleton, unigene 总长 7.89 Mb。18 235 条 unigene 中, 由 454 测序新鉴定的 unigene 为 13 980 条 (图 1)。

2 序列功能注释

通过 BLAST 搜索比对, 共有 13 308 条 unigene 获得了基因注释 (表 1)。根据拟南芥蛋白质组数据库注释结果, 被注释序列大约包含 7 800 个转录本。另有 4 927 条 unigene (27.0%) 未被注释, 认为是可能的新基因。

3 EST 文库中的高表达转录本

Unigene 所包含的 EST 数目代表了其表达丰度, 丹参根中表达丰度最高的前 10 个转录本见表 2。表达丰度最高的前两个转录本可能编码凝集素相关蛋白 (SMLII, related to legume lectin protein, gb|EF593952.1|), 表达丰度第三的转录本编码衰老相关的蛋白 (senescence-associated gene 21, AT4G02380.1)。其他高表达转录本涉及糖基水解酶家族蛋白、天冬氨酸蛋白酶、金属硫蛋白和病原相关蛋白。表达丰度最高的前 10 个转录本中有 6 个注释到已知的丹参的序列, 但这些序列的功能多数是通过生物信息学方法推测得到的。

4 功能分类研究

通过与拟南芥的蛋白质组序列比对, 获得丹参 unigene 的 GO 分类信息。11 068 条 unigene 归入分子功能 (molecular function), 10 852 条 unigene 归入生物学过程 (biological process), 10 629 条 unigene 归入细胞组分 (cellular component)。在 GO 分类体系中分子功能、生物学过程和细胞组分 3 个大的类别被划分为更详细的 45 个小的类别, 这一分类结果显示了丹参根基因表达谱的总体情况 (图 2)。

5 代谢途径分析

在丹参根 EST 文库中发现与次生代谢相关的 unigene 共 116 条。根据 KEGG 注释结果, 可将次生代谢途径按代谢物分为 15 类。这 15 类次生代谢物包括: 生物碱 I (alkaloid I)、生物碱 II (alkaloid II)、油菜素内酯 (brassinosteroid)、咖啡因 (caffeine)、类胡萝卜素 (carotenoid)、二萜 (diterpenoid)、黄酮和黄酮醇 (flavone and flavonol)、类黄酮 (flavonoid)、柠檬烯和蒎烯 (limonene and pinene)、新生霉素 (novobiocin)、苯丙烷 (phenylpropanoid)、链霉素 (streptomycin)、萜类 (terpenoid)、四环素 (tetracycline) 和玉米素 (zeatin)。参与各类次生代谢途径的 unigene 的数目如图 3 所示。丹参酮的生物合成涉及其中的萜

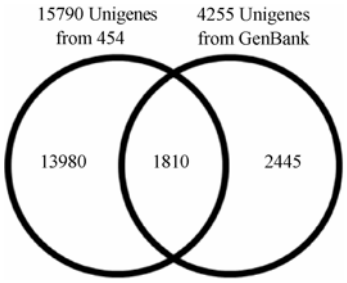

Figure 1 Comparison of *S. miltiorrhiza* unigenes from 454 EST and GenBank EST

Table 1 BLAST analysis results against important public databases

Database	18235 Unigenes in EST library	
	Annotated (n)	Percentage (%)
SwissProt	6 602	36.2
KEGG	10 502	57.6
TAIR	12 117	66.5
Nr	12 721	69.8
Nt	9 811	53.8
Total	13 308	73.0

Table 2 High expressed transcripts in *S. miltiorrhiza* EST library

Unigene ID	No. of ESTs	Subject ID	BLAST annotation
Contig06383	587	gb|EF593952.1|	SMLII [*Salvia miltiorrhiza*]
Contig00159	419	gb|EF593952.1|	SMLII [*Salvia miltiorrhiza*]
Contig06376	411	AT4G02380.1	Senescence-associated gene 21 [*Arabidopsis thaliana*]
Contig00350	358	gb|AAU86897.1|	Glycosyl hydrolase family-like protein [*Salvia miltiorrhiza*]
Contig00249	341	sp|P42211|ASPRX_ORYSJ	Aspartic proteinase [*Oryza sativa*]
Contig06480	323	gb|EF666996.1|	Putative metallothionin 2a [*Salvia miltiorrhiza*]
Contig00141	244	gb|EU182720.1|	Translationally controlled tumor protein (TCTP) [*Salvia miltiorrhiza*]
Contig00248	218	AT1G62290.2	Aspartyl protease family protein [*Arabidopsis thaliana*]
Contig00206	209	gb|EF621486.1|	Pathogen-related protein STH-2 gene [*Salvia miltiorrhiza*]
Contig06531	196	sp|O50001|PRU1_PRUAR	Major allergen [*Prunus armeniaca*]

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

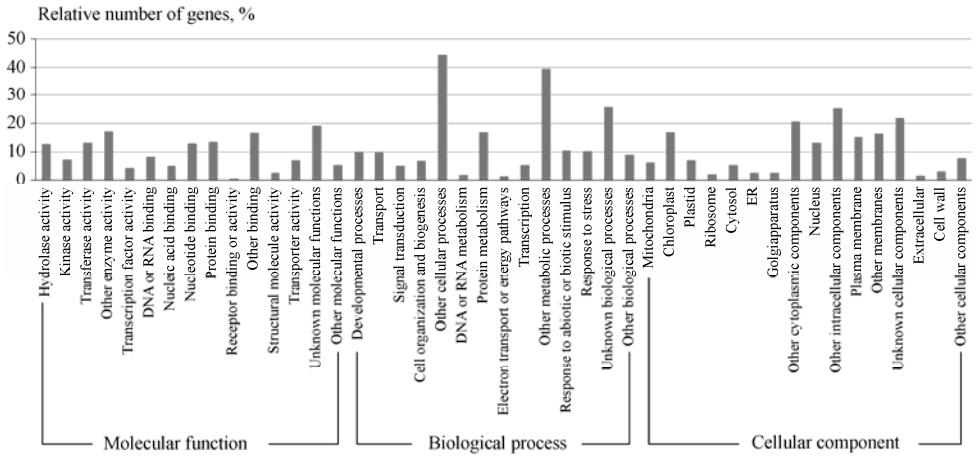

Figure 2　The unigenes were functionally categorized into 45 GO categories

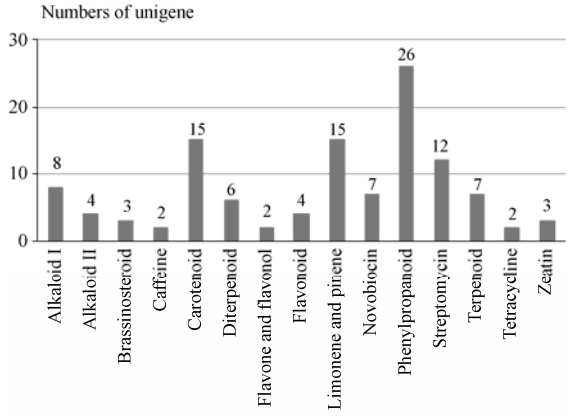

Figure 3　The unigenes related to secondary metabolites

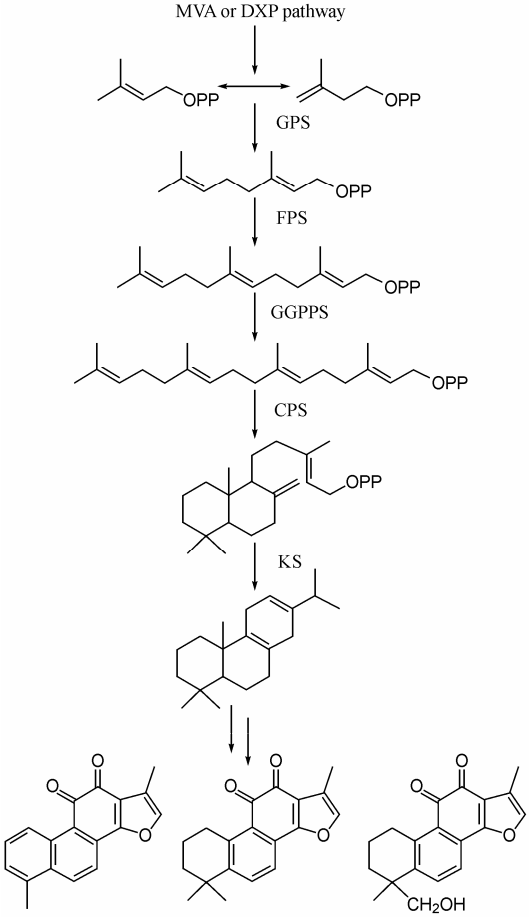

Figure 4　Tanshinome biosynthesis pathway[2, 18]. MVA: Mevalonate pathway; DXP: 1-Deoxy-*D*-xylulose 5-phosphate pathway; GPS: Geranyl diphosphate synthase; FPS: Farnesyl diphosphate synthase; GGPPS: Geranylgeranyl diphosphate synthase; CPS: Copalyl diphosphate synthase; KS: Kaurene synthase

类和二萜代谢两个途径。丹参酚酸的生物合成除涉及苯丙烷代谢途径外, 还涉及酪氨酸和苯丙氨酸代谢途径。

丹参两种主要药理活性成分丹参酮、丹酚酸类化合物的生物合成途径如图 4、图 5 所示。通过同源性搜索, 找到 27 条 unigene 可能编码丹参酮合成途径的 15 个关键酶, 包括甲羟戊酸途径 (MVA) 的酶 5 个、DXP 途径的酶 3 个和二萜骨架合成的酶 7 个。29 条 unigene 可能编码丹酚酸合成途径的 11 个关键酶, 包括迷迭香酸合成途径的酶 2 个和苯丙酮代谢途径的酶 9 个。此外, 细胞色素 P450 家族在丹参酮和丹酚酸的生物合成途径中都有非常重要的作用, 从丹参 EST 文库中找到 70 条可能编码细胞色素 P450 的 unigene。

6　SSR 分析

在丹参 unigene 中搜索到 223 个 SSR 位点, 占 unigene 序列的 1.22%, 平均每 100 kb 出现 2.8 个 SSR。SSR 种类丰富, 二至六核苷酸重复类型均存在, 重复次数也有很大的变化 (表 3)。在检出 SSR 中, 共发现 83 种基序 (motif) 类型, 出现频率最高的 5 类基序为: AG/GA/TC/CT (76 个)、AC/CA/TG/GT (25 个)、AT/TA (21 个)、AAG/GAA/AGA/CTT/TTC/TCT (12 个)、AAT/TAA/ATA/ATT/ TTA/TAT (11 个)。对这些 SSR 的鉴定,

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

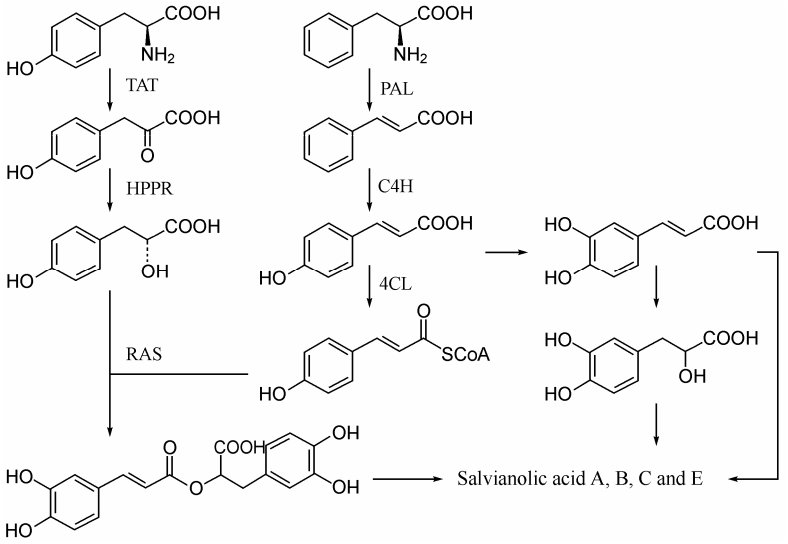

Figure 5 Salvianolic acid biosynthesis pathway[19]. TAT: Tyrosine aminotransferase; HPPR: Hydroxyphenylpyruvate reductase; RAS: Rosmarinic acid synthase; PAL: Phenylalanine ammonialyase; C4H: Cinnamate 4-hydroxylase; 4CL: 4-Coumarate: CoA ligase

将为进行丹参及其同属物种的基因组差异性分析、遗传图谱构建等研究提供帮助。

Table 3 SSR distribution in *S. miltiorrhiza* unigene

	Repeat				Total
	< 6	6−10	11−20	> 20	
Di-nuleotide		39	74	9	122
Tri-nucleotide		40	6	2	48
Tetra-nucleotide	8	2	1		11
Penta-nucleotide	14				14
Hexa-nucleotide	26	2			28
Total	48	83	81	11	223

讨论

本文首次采用 454 GS FLX Titanium 高通量测序技术对丹参 2 年生根的转录组进行测序和功能分析, 挖掘其次生代谢物合成相关基因。结合 GenBank 已有 Sanger 测序丹参 EST, 共拼接得到 18 235 条 unigene, 其中有 454 系统测序获得鉴定的新 unigene 13 980 条, 占总数的 76.7%。表明高通量测序技术是批量发现丹参功能基因的更为有效手段。与传统测序相比, 454 高通量测序的长度已经和 Sanger 测序的读长相当, 完全可以满足转录组测序的要求, 且 454 测序还具有速度快、通量高、成本低的优点。

通过同源性搜索, 获得丹参酮合成相关 unigene 共 27 条, 丹酚酸合成相关 unigene 共 29 条。编码细胞色素 P450 的 unigene 共 70 个, 这为筛选参与次生代谢的细胞色素 P450 提供了足够的候选序列。获得 577 个可能编码转录因子的 unigene, 参与丹参根的基因表达调控。丹参有效成分生物合成途径关键酶基因的发掘为克隆基因、研究基因功能提供了基础数据, 为研究有效成分的生物合成途径和调控机制奠定了基础, 同时为应用生物技术方法提高丹参有效成分含量、或直接生产有效成分和其中间体提供了可行性。

致谢: 中国医学科学院药用植物研究所李先恩研究员提供丹参实验材料。

References

[1] Zhou L, Zuo Z, Chow MS. Danshen: an overview of its chemistry, pharmacology, pharmacokinetics, and clinical use [J]. J Clin Pharmacol, 2005, 45: 1345−1359.

[2] Gao W, Hillwig ML, Huang L, et al. A functional genomics approach to tanshinone biosynthesis provides stereochemical insights [J]. Org Lett, 2009, 11: 5170−5173.

[3] Zhao SJ, Hu ZB, Liu D. Two divergent members of 4-coumarate: coenzyme A ligase from *Salvia miltiorrhiza* Bunge: cDNA cloning and functional study [J]. J Integr Plant Biol, 2006, 48: 1355−1364.

[4] Huang B, Yi B, Duan Y, et al. Characterization and expression profiling of tyrosine aminotransferase gene from *Salvia miltiorrhiza* (Dan-shen) in rosmarinic acid biosynthesis pathway [J]. Mol Biol Rep, 2008, 35: 601−612.

[5] Song J, Wang Z. Molecular cloning, expression and characterization of a phenylalanine ammonialyase gene (SmPAL1) from *Salvia miltiorrhiza* [J]. Mol Biol Rep, 2009, 36: 939−952.

[6] Xiao Y, Di P, Chen J, et al. Characterization and expression profiling of 4-hydroxyphenylpyruvate dioxygenase gene (Smhppd) from *Salvia miltiorrhiza* hairy root cultures [J].

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Mol Biol Rep, 2009, 36: 2019−2029.

[7] Huang B, Duan Y, Yi B, et al. Characterization and expression profiling of cinnamate 4-hydroxylase gene from *Salvia miltiorrhiza* in rosmarinic acid biosynthesis pathway [J]. Russ J Plant Physiol, 2008, 55: 390−399.

[8] Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project [J]. Science, 1991, 252: 1651−1656.

[9] Boguski MS, Tolstoshev CM, Bassett DE, et al. Gene discovery in dbEST [J]. Science, 1994, 265: 1993−1994.

[10] Bouchez D, Hofte H. Functional genomics in plants [J]. Plant Physiol, 1998, 118: 725−732.

[11] Ewing RM, Ben Kahla A, Poirot O, et al. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression [J]. Genome Res, 1999, 9: 950−959.

[12] Chen SL, Sun YQ, Song JY, et al. Analysis of expressed sequence tags (EST) from *Panax quinquefolium* root [J]. Acta Pharm Sin (药学学报), 2008, 43: 657−663.

[13] Wu Q, Song JY, Sun YQ, et al. Transcript profiles of *Panax quinquefolius* from flower, leaf and root bring new insights into genes related to ginsenosides biosynthesis and transcriptional regulation [J]. Physiol Plant, 2010, 138: 124−149.

[14] Luo HM, Sun C, Li Y, et al. Analysis of expressed sequence tags from the *Huperzia serrata* leaf for gene discovery in the areas of secondary metabolite biosynthesis and development regulation [J]. Physiol Plant, 2010, (in press, DOI: 10.1111/j.1399-3054.2009.01339.x).

[15] Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors [J]. Nature, 2005, 437: 376−380.

[16] Berardini TZ, Mundodi S, Reiser L, et al. Functional annotation of the *Arabidopsis* genome using controlled vocabularies [J]. Plant Physiol, 2004, 135: 745−755.

[17] Wang QH, Zhang BL. TRAP analysis of *Salvia miltiorrhiza* Bge from different places of production [J]. Acta Pharm Sin (药学学报), 2009, 44: 927−930.

[18] Dewick PM. Medicinal Natural Products: A Biosynthetic Approach [M]. 3rd edition. Chichester: John Wiley & Sons, 2009: 223.

[19] Duan YB. Molecular Cloning and Characterization of Phenylalanine Branch's Genes Involved in the Biosynthetic Pathways of Rosmarinic Acid From *Salvia miltiorrhiza* Bung (丹参中迷迭香酸生物合成途径的苯丙氨酸支路基因的克隆及研究) [D]. Second Military Medicial University, 2006.

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net