基于高通量测序技术的三叉苦幼苗转录组数据分析

廖小庭[1]，马新业[1*]，马 庆[2]，韩正洲[2]，詹若挺[1]

1. 广州中医药大学 中药资源科学与工程研究中心，岭南中药资源教育部重点实验室，国家中成药工程技术研究中心，广东 广州 510006
2. 华润三九医药股份有限公司，广东 深圳 518002

摘 要：目的 获得三叉苦 *Melicope pteleifolia* 转录组信息特征。**方法** 以三叉苦幼苗根、茎、叶混合样品为对象，采用二代高通量测序平台 Illumina HiSeq™ 2000 进行转录组测序并进行系统的生物信息学分析。**结果** 转录组测序分析共获得 47 045 040 条高质量序列（clean reads），Trinity de novo 组装获得 67 956 条 unigenes，平均长度 787 nt。BLAST 分析显示分别有 42 749（61.92%）、31 152（45.84%）、26 563（39.0 9%）、17 481（25.72%）条 unigenes 在 NR、Swiss-port、KOG、KEGG 数据库得到注释信息，参与生物过程、细胞组分和分子功能 3 个 GO 类别的 47 个小组，共 9 807 条 unigenes 注释到 130 个 KEGG 代谢通路中，筛选到 19 条次生代谢通路，KOG 功能分类分析获得 25 个不同的 KOG 功能类群。预测共有高等植物转录因子 56 个家族；借助 MISA 软件发现 7 748 个 SSRs，三碱基重复 SSRs 数量最丰富，有 4 117 个，出现频率为 53.1%，五碱基重复 SSRs 相对较少，占 2.2%。**结论** 利用高通量测序技术和生物信息分析获得三叉苦转录组信息特征，为后续三叉苦功能基因的挖掘、次生代谢途径解析及其调控机制研究奠定基础。

关键词：三叉苦；转录组；功能基因；代谢通路；简单重复序列

中图分类号：R282.12 **文献标志码：**A **文章编号：**0253 - 2670(2020)14 - 3777 - 08

DOI: 10.7501/j.issn.0253-2670.2020.14.023

Transcriptomic data analyses of *Melicope pteleifolia* via Illumina high-throughput sequencing technology

LIAO Xiao-ting[1], MA Xin-ye[1], MA Qing[2], HAN Zheng-zhou[2], ZHAN Ruo-ting[1]

1. Research Center of Chinese Herbal Resource Science and Engineering, Key Laboratory of Chinese Medicinal Resource from Lingnan of Ministry of Education, Laboratory of National Engineering Research Center for Pharmaceutics of Traditional Chinese Medicines, Guangzhou University of Chinese Medicine, Guangzhou 510006, China
2. China Resources Sanjiu Medical & Pharmaceutical Co., Ltd., Shenzhen 518002, China

Abstract: Objective To obtain the transcriptome sequence database of *Melicope pteleifolia*. **Methods** The transcriptome sequencing and systematic bioinformatics analysis were carried out using the second generation high-throughput sequencing platform Illumina HiSeq™ 2000 with mixed root, stem and leaf samples of *M. pteleifolia*. **Results** A total of 47 045 040 high quality sequences (clean reads) were obtained by transcriptome sequencing analysis. A total of 67 956 unigenes were assembled by Trinity *de novo*, with an average length of 787 nt. BLAST analysis showed that 42 749 (61.92%), 31 152 (45.84%), 26 563 (39.09%), and 17 481 (25.72%) unigenes were annotated in NR, Swiss port, KOG and KEGG databases respectively, and 47 groups were involved in three GO classification: biological process, cellular component and molecular function. A total of 9807 unigenes were annotated to 130 KEGG metabolic pathways, 19 secondary metabolic pathways were screened. Twenty-five different KOG functional groups were obtained by the analysis of KOG functional classification. It was predicted that there were 56 families of higher plant transcription factors. A total of 7 748 simple sequence repeats (SSRs) were found by MISA software. The number (4 117) of the tri-nucleotide SSRs was the richest,

收稿日期：2019-11-06

基金项目：国家自然科学基金委青年基金项目（81102764）；广东省教育厅重点提升平台建设项目——岭南中药资源教育部重点实验室（2014KTSPT016）；广东省教育厅创新团队项目——中药资源创新团队（2016KCXTD015）

作者简介：廖小庭（1994—），女，硕士研究生，研究方向为芳香药用植物资源学。Tel: 18897919485 E-mail: 1206459247@qq.com

*通信作者 马新业（1976—），男，博士，副研究员，主要从事中药资源学研究。Tel: 15817036306 E-mail: usermxy@163.com

with a frequency of 53.1%, and the number of the penta-nucleotide SSRs was relatively small, accounting for 2.2%. **Conclusion** The transcriptome information characteristics of root, stem, and leaf of *M. pteleifolia* can be obtained by high-throughput Illumina sequencing technology and bioinformatics analysis, which will lay a foundation for further research on functional gene mining, secondary metabolic pathway analysis and regulation mechanism of *M. pteleifolia*.

Key words: *Melicope pteleifolia* (Champion ex Bentham) T. G. Hartley; transcriptome; functional gene; metabolism pathway; simple sequence repeats

三叉苦 *Melicope pteleifolia* (Champion ex Bentham) T. G. Hartley 为芸香科蜜茱萸属植物，别名三丫苦、三桠苦、三部虎、三叉虎、三枝枪、斑鸠花、小黄散、鸡骨树等，主要分布在我国南部地区，如广东、广西、海南、福建、贵州和云南等地[1]。三叉苦入药部位有根、茎和叶等，性苦、寒，归心、肝经，有清热解毒、祛风除湿、消肿止痛等功效，属于岭南常用中药。主治发热、外感风热、黄疸型肝炎、咳嗽、喘促、咽喉肿痛、肺痈、疟疾寒热、风湿痹痛、湿疹、皮炎、胃脘疼痛和虫蛇咬伤等多种病症[2]。因其独特的功效，临床上应用广泛，为感冒灵颗粒、三九胃泰颗粒、消结安胶囊、双龙风湿跌打膏等多种中成药的组方药材。另外，三叉苦是一种药食同源的植物，是广东凉茶主成分之一[3]，食用价值较高。目前对三叉苦的研究主要集中在其化学成分[4-9]、药理作用[10-11]、质量标准[12]、种质资源遗传多样性[13]、本草源流考证[14]、生药学研究[15]等方面，整体上还处于基础阶段，主要有效成分并不明确，质量标准还未建立。利用现代转录组测序技术和生物信息学分析手段，充分挖掘三叉苦遗传信息，如解析某些特征活性成分生物合成途径及关键酶基因序列、探讨生长发育机制和预测分子标记等，将为更深入推进三叉苦基础和应用研究奠定良好数据基础。

转录组指细胞或组织内全部的 RNA 转录本，反映了生物体在不同生命阶段、不同生理状态、不同组织类型以及不同环境条件下全部基因的表达情况，是后基因组时代最活跃的研究领域之一[16]。转录组测序（RNA-seq）指将细胞或组织中全部或部分 RNA 进行测序分析。目前常见以高通量测序技术依托，以 Illumina 公司第二代测序平台为主流[17]。近年来，转录组高通量测序已在药用植物功能基因鉴定、次生代谢途径探索等方面广泛应用，比如已获得人参[18]、夏枯草[19]、冬凌草[20]、厚朴[21]和黄三七[22]等多种药用植物转录组数据，为中草药功能基因研究推进提供了丰富的信息资源。本研究利用二代高通量测序平台 Illumina HiSeq^{TM} 2000 对三叉苦幼苗的根、茎、叶混合样品进行转录组测序，以期获得三叉苦转录组整体数据特征，为有效发掘和鉴定次生代谢产物合成及其调控相关基因等研究提供数据基础。

1 材料与方法

1.1 材料

植物材料三叉苦幼苗采自广东省云浮市三九种植基地，经广州中医药大学中药资源科学与工程研究中心詹若挺教授鉴定为三叉苦 *Melicope pteleifolia* (Champion ex Bentham) T. G. Hartley。混合取新鲜小苗的根、茎、叶，无菌水清洗干净后用吸水纸吸干水分，迅速置于液氮中冷冻保持 5 min，随即置于−80 ℃冰箱保存备用。

1.2 总 RNA 提取、文库构建和转录组测序

总 RNA 使用北京艾德莱生物科技有限公司 EASYspin Plus 植物快速提取试剂盒提取。用 1%琼脂糖凝胶电泳检验其 RNA 完整性，用微量紫外分光光度计测定其 RNA 浓度和纯度。文库构建与测序工作委托广州基迪奥生物科技有限公司完成。

1.3 测序数据质量处理及组装

测序仪产生的原始图像数据经 base calling 转化为序列数据 raw reads，经过初步的过滤后得到 clean reads，再经过去除含 adaptor 的 reads、去除 N 的比例大于 10%的 reads 和去除低质量 reads（质量值 Q≤20 的碱基数占整个 read 的 40%以上）的处理过程后，最终获得高质量 clean reads。接着使用短 reads 组装软件 Trinity 做转录组从头组装，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。

1.4 转录组功能注释

通过 BLAST 将 unigenes 序列比对到蛋白数据库 NR、Swiss-Prot、KEGG 和 KOG（E 值＜1×10^{-5}），得到和给定 unigenes 具有最高相似性的蛋白序列，从而得到该 unigenes 的蛋白功能注释信息。使用 Blast2GO 软件，以 NR 注释信息为依据，获得 unigenes 的 GO 注释信息。在每个 unigenes 都得到

GO 注释后，将所有 unigenes 使用 WEGO 软件做 GO 功能分类统计，有助于从宏观上获得该物种的基因功能分布特征。

1.5 转录因子预测

按 NR、Swiss-Prot、KEGG 和 KOG 的优先级顺序将 unigenes 序列与以上蛋白库做 BLASTx 比对（E 值＜1×10^{-5}），如果某个 unigenes 序列可比对上高优先级数据库中蛋白信息，则不进入下一轮比对，否则自动和后续低优先级数据库做比对，如此循环直到和 4 个蛋白库比对完成。取 BLAST 比对结果中排名最高的蛋白确定该 unigenes 的编码区序列，然后根据标准密码子表将编码区序列翻译成氨基酸序列，从而得到该 unigenes 编码区的核酸序列（序列方向 5’→3’）和氨基酸序列。和以上蛋白库皆比对不上的 unigenes 用软件 ESTScan 预测其编码区，得到其编码区的核酸序列（序列方向 5’→3’）和氨基酸序列，将所有预测蛋白序列同植物转录因子数据库（plant TFDB）进行 hmmscan 比对搜索转录因子家族及其成员。

1.6 简单重复序列特征检测

以三叉苦转录组为研究对象，使用 MISA（Micro satellite）软件进行 SSRs 搜索。对 unigenes 进行 SSRs 检测，参数设置为“2-6、3-5、4-4、5-4、6-4”，“2-6”说明对于 2 个核苷酸重复单元，需要至少 6 个重复才会被认为是 SSRs，3 个核苷酸重复单元需要 5 个重复，4 个核苷酸重复单元需要 4 个重复，以此类推。

2 结果与分析

2.1 三叉苦转录组组装与质量分析

采用二代高通量测序平台 Illumina HiSeq™ 2000 对三叉苦小苗根茎叶进行转录组测序，共得到 47 728 770 条 raw reads，过滤产生了 47 045 040 条 clean reads，包含 6 937 000 904 个核苷酸信息，Q20（碱基≥20%）为 98.57%，Q30（碱基量≥30%）为 95.53%，GC 量为 45.04%，说明测序质控良好，clean reads 质量合格，数据可靠，可用于后面的转录组组装。Trinity de novo 组装获得 67 956 条 unigenes，平均长度 787 nt，最长达到 12 630 nt，最短序列为 201 nt，N50 为 1 351 nt。据统计，其中 16 386 条 unigenes 长度超过 1 000 nt，5 792 条序列长度在 2 000 nt 以上（图 1）。所含 reads 数量在 1～100 的 unigenes 数量最多，为 27 226 条；其次为 reads 数量在 11～50 的 unigenes，为 16 323 条；接着 reads 数量从大到小依次排列为在 1 001～10 000、501～1 000、51～100、101～200 的 unigenes 分别为 8 234、3 884、3 630、3 097 条；其余 reads 分布区域对应的 unigenes 数量均相对较少，均在 2 000 以下（图 2）。

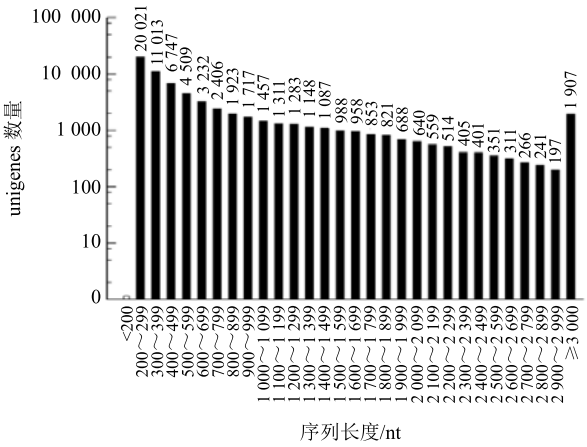

图 1 三叉苦转录组 unigenes 长度分布

Fig. 1 Length distribution of *M. pteleifoliar* transcriptomic unigenes

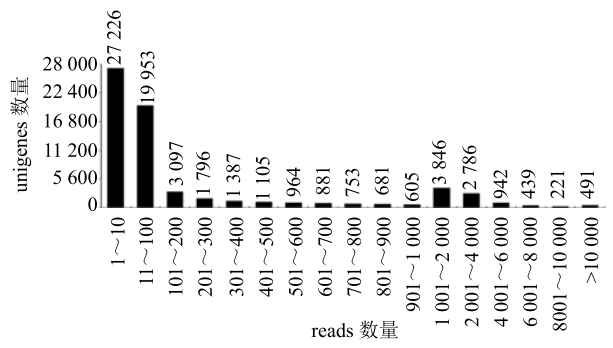

图 2 三叉苦转录组 reads 在 unigenes 上的覆盖统计

Fig. 2 Coverage statistics of *M. pteleifoliar* transcriptome reads on unigenes

2.2 三叉苦转录组 unigenes 功能注释

通过 BLASTx 将 unigenes 序列比对到蛋白数据库 NR、Swiss-Prot、KOG 和 KEGG 等数据库中，并统计注释到各个数据库的 unigenes 数量，进而获得三叉苦转录组 unigenes 的功能注释信息。结果表明有 42 749 条 unigenes 在 NR 数据库中比对成功获得注释，所占比例为 61.92%；31 152 条 unigenes 在 Swiss-port 数据库中比对成功获得注释，所占比例为 45.84%；26 523 条 unigenes 在 KOG 数据库中比对成功获得注释，所占比例为 39.09%；17 481 条 unigenes 在 KEGG 数据库中比对成功获得注释，所占比例为 25.72%。至少有一种数据库注释成功的

unigenes 共 43 635 条（64.21%），24 321 条 unigenes 未获得注释（表 1）。

表 1　三叉苦转录组 unigenes 注释情况

Table 1　Annotations of *M. pteleifoliar* transcriptome unigenes

公共数据库	注释基因数量/条	注释率/%
Nr	42 749	62.91
Swiss-Prot	31 152	45.84
KOG	26 563	39.09
KEGG	17 481	25.72
在以上数据库中至少一种数据库注释成功	43 635	64.21
总计	67 956	100.00

利用 BLASTx 将组装出来的 unigenes 序列与 NR 数据库进行比对后，取每个 unigenes 在 NR 库中比对结果最好（*E* 值最低）的一条序列为对应同源序列（如有并列，取第 1 条），确定同源序列所属物种，展示同源序列数量最多的前 10 个物种（图 3）。在相似序列匹配度较高的物种中，甜橙 *Citrus sinensis* (Linn.) Osbeck 占比最高，20 205 条，其次为羊膜花 *Anthurium amnicola* L.，5 254 条，土瓶草 *Cephalotus follicularis* Labill.，1 872 条，可可 *Theobroma cacao* (Cocoa) Seed Butter，1 536 条，后面 6 种物种分别为圆果种黄麻 *Corchorus capsularis* L.、棉花 *Gossypium arboretum* L.、木豆 *Cajanus cajan* (Linn.) Millsp.、拟南芥 *Arabidopsis thaliana* L.、油菜 *Brassica napus* L.、长果种黄麻 *Corchorus olitorius* L.，分别有 530、530、528、509、507、434 条。

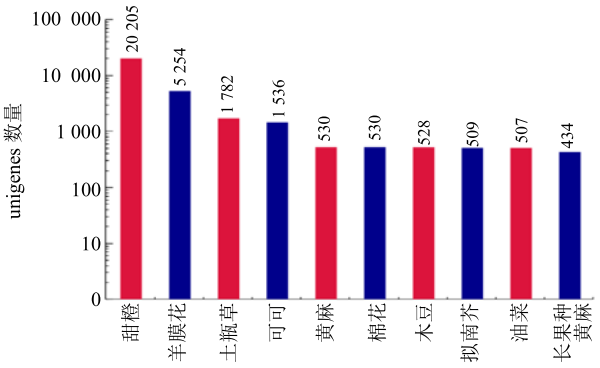

图 3　三叉苦转录组 unigenes 与 NR 数据库匹配前 10 个物种分布

Fig. 3　Top 10 species distribution of *M. pteleifoliar* transcriptome unigenes matched with NR database

三叉苦小苗根茎叶转录组数据中有 26 563 条 unigenes 获得 KOG 注释，共得到 25 个不同的 KOG 功能类群（图 4），种类齐全，包括大多数的生命活动。一般功能预测的注释结果最高，为 8 539 条，翻译后修饰、蛋白反转和伴侣注释结果次之，为 5 536 条，信号转导机制次之，为 4 979 条。此外，1 348 条 unigenes 被注释为“次生代谢合成、转运和代谢”，表明三叉苦根茎叶转录组中包含信号传导和次生代谢相关基因，为后续数据挖掘提供依据。

在三叉苦根茎叶转录组中，共有 16 566 条 unigenes 被注释到参与生物过程、细胞组分和分子功能 3 个 GO 类别 47 个小组（图 5）。参与的生物过程主要分布在代谢过程（metabolic process）、细胞过程（cellular process）、单一有机体（single-organism process），基因数分别为 9 631、8 893、7 068 条。细

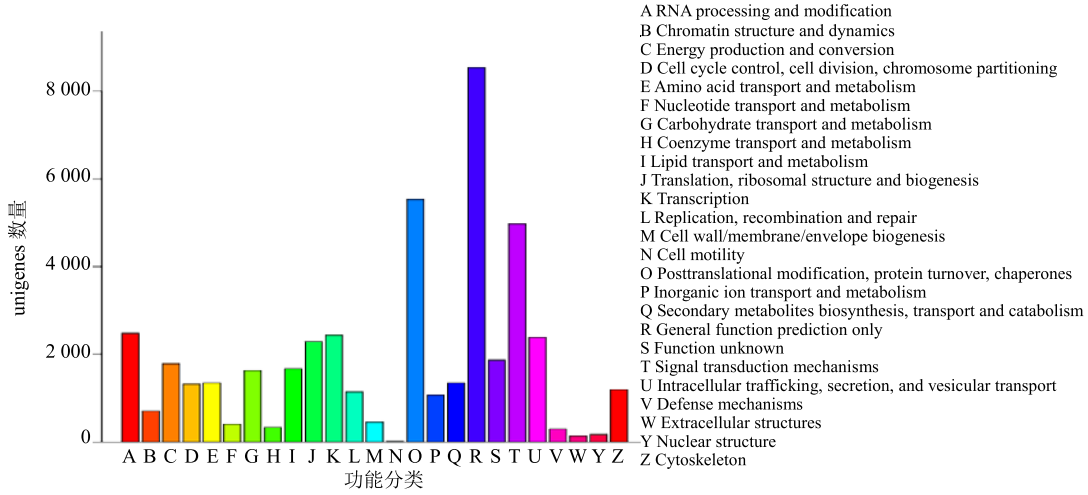

图 4　三叉苦转录组 unigenes KOG 功能分类图

Fig. 4　KOG functional classification of *M. pteleifoliar* transcriptomic unigenes

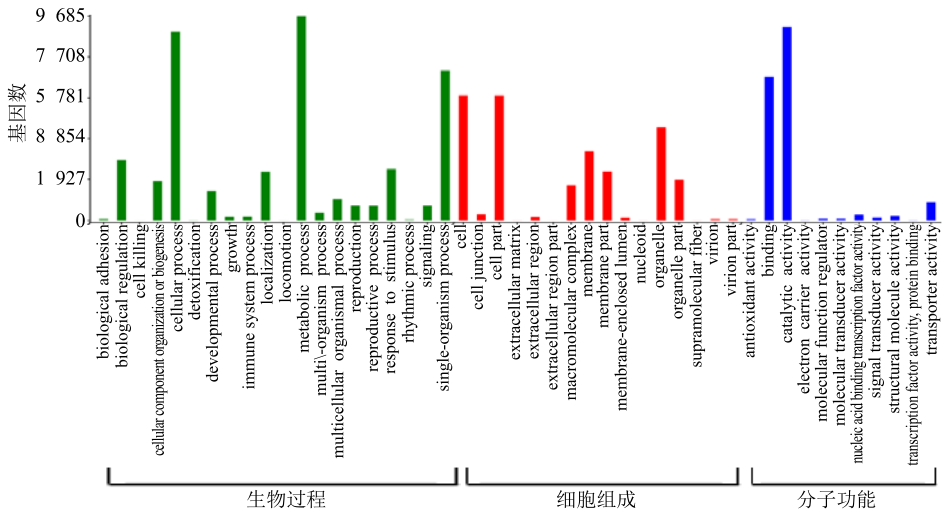

图 5 三叉苦转录组 unigenes GO 功能分类图

Fig. 5 GO functional classification of *M. pteleifoliar* transcriptomic unigenes

胞组分中分布则较均匀，基因数占前 3 的分别是细胞（cell）5 900 条、细胞部分（cell part）5 898 条、细胞（organelle）4 396 条。分子功能中具有催化活性（catalytic activity）的基因数量最高，为 9 127 条，结合功能（binding）次之，为 6 773 条。

三叉苦转录组中总共有 17 418 条 unigenes 成功注释到 KEGG 数据库中，三叉苦转录组 unigenes 注释到 KEGG 数据库的通路中共分为 5 类，分别是遗传信息处理、代谢、细胞过程、生物系统和环境信息处理，分别为 4 613、9 619、926、460、586 条。其中代谢涉及的通路最多，遗传信息处理次之，有机系统涉及的通路最少，根据 KEGG 数据库的注释信息能进一步得到 unigenes 的通路注释。根据注释结果显示，有 9 807 条 unigenes 注释到 130 条代谢通路中，以选取通路包含 unigenes 数量大于 300 条为标准，将基因注释数量从大到小依次排列，本文获得前 14 个代谢通路信息（表 2）。KEGG 代谢通路分析发现有 830 条 unigenes 参与三叉苦苯丙素、萜类、黄酮类、生物碱类等生物合成相关的 19 个次生代谢标准通路（表 3）。其中，苯丙素生物合成代谢通路基因数量最多，为 224 条，萜类骨架生物合成代谢通路基因数量次之，为 105 条，类黄酮生物合成代谢通路基因数量居第 3 为 66 条，另外分别有 10、7 条基因与黄酮和黄酮醇、异黄酮的生物合成代谢通路相关，对苯二甲酸、哌啶和吡啶生物碱的生物合成代谢通路基因数量有 49 条，异喹啉、吖啶酮生物碱合成相关的基因数分别有 39 和 1 条，有 41 条 unigenes 参与二萜生物合成代谢通路，36 条 unigenes 参与倍半萜和三萜生物合成代谢通路。

表 2 三叉苦转录组 unigenes KEGG 通路分析情况

Table 2 KEGG pathway analysis of *M. pteleifoliar* transcriptome unigenes

编号	代谢通路	unigenes 数量/条	占比/%	通路 ID
1	核糖体	770	7.85	ko03010
2	碳代谢	675	6.88	ko01200
3	内质网蛋白质加工	547	5.58	ko04141
4	剪接体	526	5.36	ko03040
5	氨基酸生物合成	519	5.29	ko01230
6	RNA 转运	426	4.34	ko03013
7	内吞作用	405	4.13	ko04144
8	植物病原相互作用	388	3.96	ko04626
9	植物激素信号转导	355	3.62	ko04075
10	氧化磷酸化	354	3.61	ko00190
11	嘌呤代谢	352	3.59	ko00230
12	泛素介导的蛋白质水解	331	3.38	ko04120
13	RNA 降解	307	3.13	ko03018
14	淀粉和蔗糖代谢	303	3.09	ko00500

2.3 转录因子分析

对三叉苦转录组所有 unigenes 进行转录因子分析，预测共有 1 741 条 unigenes 分属于 56 个家族。其中最多的转录因子类型是 ERF 类有 142 条，约占 8.15%；其次是 BHLH 类有 135 条，约占 7.75%；C2H2 类有 126 条，约占 7.24%（图 6）。

2.4 SSRs 特征分析

利用 MISA 脚本对拼接得到 unigenes 进行 SSRs

表 3　三叉苦转录组 unigenes 次生代谢 KEGG 通路注释统计

Table 3　Secondary metabolism KEGG pathway annotation analysis of *M. pteleifoliar* transcriptome unigenes

编号	代谢通路	unigenes 数量/条	占比/%	通路 ID
1	苯丙素生物合成	224	2.28	ko00940
2	萜类骨架生物合成	105	1.07	ko00900
3	类黄酮生物合成	66	0.67	ko00941
4	二苯乙烯、二芳基庚烷和姜辣素生物合成	62	0.63	ko00945
5	类胡萝卜素生物合成	55	0.56	ko00906
6	对苯二甲酸、哌啶和吡啶生物碱生物合成	49	0.50	ko00960
7	二萜生物合成	41	0.42	ko00904
8	异喹啉生物碱生物合成	39	0.40	ko00950
9	倍半萜和三萜生物合成	36	0.37	ko00909
10	柠檬烯和松烯降解	30	0.31	ko00903
11	油菜素类固醇生物合成	28	0.29	ko00905
12	单环 β-内酰胺合成	28	0.29	ko00261
13	玉米素生物合成	24	0.24	ko00908
14	单萜类生物合成	15	0.15	ko00902
15	黄酮和黄酮醇生物合成	10	0.10	ko00944
16	异黄酮生物合成	7	0.07	ko00943
17	咖啡因代谢	7	0.07	ko00232
18	花青素生物合成	3	0.03	ko00942
19	吖啶酮生物碱生物合成	1	0.01	ko01058

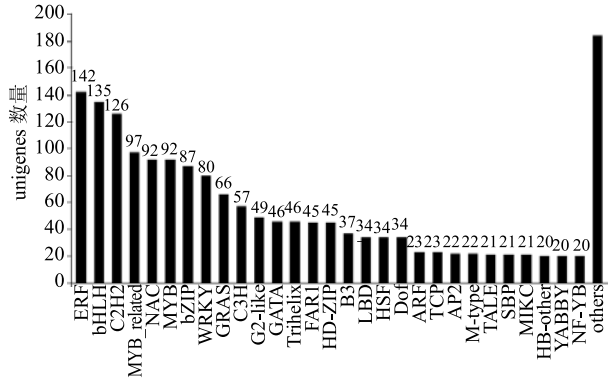

图 6　三叉苦转录组 unigenes 的 TF 家族分布图

Fig. 6　Transcription factor family distribution of *M. pteleifoliar* transcriptome unigenes

位点检测。结果显示共检测到 7 748 个 SSRs 位点，6 281 条 unigenes 含有 SSRs 位点（表 4）。其中三核苷酸重复 SSRs 数量最丰富，有 4 117 个，占比 53.1%；其次是二核苷酸重复 SSRs 数量为 2 635 个，占比 34.0%；四核苷酸和六核苷酸重复分别为 562、261 个，各占 7.3%、3.4%；五核苷酸重复相对较少，仅占 2.2%。此外，还发现 SSRs 重复单元数量也存在一定变化，其中重复 5、6 次的比例最高，重复 4、7、8 次的次之。将检测到的 SSRs 按照短串联重复单元的类型分类（图 7），三核苷酸重复中，AAG/CTT 类型的比例最高，在二核苷酸重复中，AG/CT 重复类型数量最多。

表 4　三叉苦转录组 unigenes SSRs 分析

Table 4　SSRs analysis of *M. pteleifoliar* transcriptome unigenes

重复单元数量/个	重复数					
	二核苷酸	三核苷酸	四核苷酸	五核苷酸	六核苷酸	合计/个
4	0	0	399	134	202	735
5	0	2 291	106	29	42	2 468
6	833	1 053	31	8	7	1 932
7	516	484	13	1	5	1 019
8	365	97	7	0	3	472
9	296	23	1	1	1	322
10	217	74	2	0	0	293
11	121	26	0	0	0	147
12	25	5	1	0	1	32
13	7	14	0	0	0	21
14	7	7	2	0	0	16
≥15	248	43	0	0	0	291
合计/个	2 635	4 117	562	173	261	7 748
占比/%	34.0	53.1	7.3	2.2	3.4	100

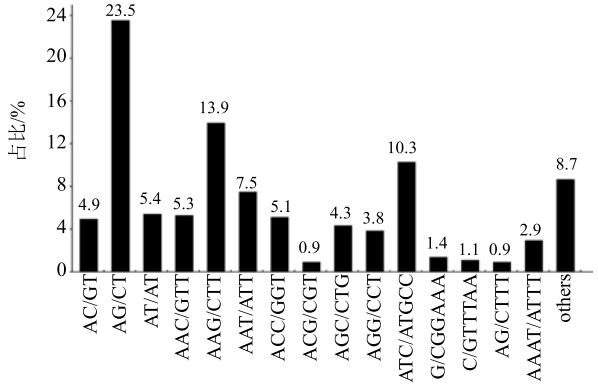

图 7　三叉苦转录组 unigenes 不同串联重复单元类型的 SSRs 在总 SSRs 中所占比例统计

Fig. 7　Proportion of SSRs with different types of tandem repetitive units in total SSRs of *M. pteleifoliar* transcriptome unigenes

3　讨论

三叉苦是众多中成药的重要原料药材，但目前整体上研究，其有效成分尚未明确，质量标准建立不完全。本实验尝试利用组学技术挖掘遗传信息，

为推动三叉苦分子生物学研究奠定基础。将三叉苦小苗根茎叶的混合样品，采用二代高通量测序平台 Illumina HiSeq™ 2000，进行转录组测序，总共获得约 4.7 G 转录组数据，经过拼接组装获得 67 956 条序列 unigenes，Q20、Q30 均大于 90%，GC 量接近 45%，表明测序和组装质量较好，可用于后续所有序列数据库的建立和基因功能注释、分类。混合样品转录组 unigenes 信息量大，基本涵盖全转录组信息，能够清晰反映传统中药三叉苦基因表达特征，为深入研究生长发育、次生代谢、转录调控等生物学过程中的基因功能提供数据资源。

三叉苦转录组 unigenes 通过 BLASTx 注释到 NR、Swiss-Prot、KOG 和 KEGG 等 4 大数据库中的 unigenes 的条数分别是 42 749、31 152、26 563、17 481、43 635，总体注释成功的 unigenes 共 43 635 条，占所有序列的 64.21%，表明已有一半以上的 unigenes 研究将有参考信息依据，其余 24 321 条 unigenes 未获得注释，占比 35.79%，说明三叉苦转录组中存在着大量未知的序列，这种情况可能是由于部分序列长度太短、相关数据库信息不够全面的原因。

利用 KOG 数据库对三叉苦根茎叶转录组 unigenes 进行基因功能分类，可从基因组水平上寻找直系同源体，预测未知 ORF 的生物学功能，能够在很大程度上提高基因功能注释的准确性。三叉苦转录组 KOG 种类比较全面，KOG 注释共得到 25 个不同的 KOG 功能类群。三叉苦转录组在 GO 数据库中共有 16 566 条 unigenes 被注释，细分为 47 个小组，归属分子功能、细胞组分和生物过程 3 大类，揭示了三叉苦转录组基因表达功能分布。KEGG 是系统分析基因产物在细胞中的代谢途径以及这些基因产物的功能的数据库。在三叉苦转录组数据中，发现 KEGG 通路注释中有 130 个 KEGG 标准代谢通路，这些代谢通路的基因可能参与三叉苦的水分吸收、矿质营养、光合作用和水解作用等生命代谢活动；发现 830 条 unigenes 参与苯丙素、萜类、黄酮类、生物碱类等与生物合成相关的 19 个次生代谢标准通路，有利于三叉苦次生代谢物生物合成途径的分子生物学研究。

基因表达的转录调控在植物的生长发育及环境适应方面发挥重要作用。转录因子又称反式作用因子，是能够与真核基因启动子区的顺式作用元件进行特异相互作用的结合蛋白，通过转录因子间及与其他相关蛋白间的相互作用激活或抑制转录[23]。本研究获得的三叉苦 unigenes 通过转录因子预测发现有 56 个家族，说明三叉苦生命活动代谢涉及非常复杂的转录调控机制。

SSRs 也被称为微卫星 DNA，这是一种以一个或几个碱基可变次数的串联重复序列，这些序列广泛存在于真核细胞的基因组内，被认为是一种理想的分子标记，由若干个串联重复单元组成，每单元含 1～10 个核苷酸，广泛分布在真核生物基因组中，是一种广泛应用的分子标记技术[24]。本研究使用软件 MISA，发现三叉苦根茎叶转录组中 6 281 个 unigenes 中存 7 748 个 SSRs 位点，从双核苷酸类型到六核苷酸类型均具备，表明三叉苦基因组内该形式分子标记位点丰度较高。重复类型以三核苷酸为主，双核苷酸所占比例次之，去掉单核苷酸重复后，三核苷酸重复占 53.1%，三核苷酸重复占 34.0%，与大多数植物 SSRs 与二、三核苷酸重复为主要类型的统计数据较一致，如番红花[25]、红景天[26]、大黄[27]等。三叉苦双核苷酸重复 SSRs 中以 AG/CT 类型为主，三核苷酸重复中 AAG/CTT 类型最多，ATC/GAT 次之。这与黄三七[22]、川芎[28]、人参[29]等的研究结果一致，说明 SSRs 重复类型可能存在一定的普遍性。

本研究基于二代高通量测序技术对三叉苦小苗混合样品进行转录组测序并分析获得了大量三叉苦遗传信息和基因表达特征。其中参与苯丙素、萜类、黄酮类、生物碱类等生物合成相关的 19 个次生代谢标准通路信息，为确定三叉苦有效成分和挖掘其次生代谢物的生物合成途径关键基因提供了基础资料；丰富的 SSRs 信息可启发后续三叉苦种质资源奠定、遗传多样性分析和本草源流考证技术和手段。

参考文献

[1] 广东省药品监督管理局. 广东省中药材标准（第 3 册）[M]. 广州: 广州科技出版社, 2018.

[2] 全国中草药汇编编写组. 全国中草药汇编（上册）[M]. 北京: 人民卫生出版社, 1975.

[3] 杨增明, 马志伟, 袁玲玲. 傣医药研究 [M]. 昆明: 云南科学技术出版社, 2012.

[4] 李斯达. 三桠苦化学成分研究 [D]. 广州: 广州中医药大学, 2017.

[5] Tang Y Q, Ti Y Q, Xie Y B, *et al*. Evodialones A and B: Polyprenylated acylcyclopentanone racemates with a 3-ethyl-1,1-diisopentyl-4-methylcyclopentane skeleton

from *Evodia lepta* [J]. *J Nat Prod*, 2018, 81(6): 1483-1487.

[6] Sichaem J, Jirasirichote A, Sapasuntikul K, *et al*. New furoquinoline alkaloids from the leaves of *Evodia lepta* [J]. *Fitoterapia*, 2014, 92: 270.

[7] 魏荷琳, 周思祥, 姜 勇, 等. 三叉苦叶的化学成分研究 (英文) [J]. 中国中药杂志, 2013, 38(8): 1193-1197.

[8] Li G L, Zhu D Y. Two dichromenes from *Evodia lepta* [J]. *J Asian Nat Prod Res*, 1999, 1(4): 337.

[9] Li G L, Zhu D Y. Two new dichromenes from *Evodia lepta* [J]. *J Nat Prod*, 1998, 61(3): 390.

[10] 梁 粤. 三叉苦抗炎镇痛作用及脂溶性化学成分研究 [D]. 广州: 广东药学院, 2010.

[11] 林紫微, 赵智萍, 林志军, 等. 海南三叉苦抗炎作用及机制研究 [J]. 海南医学, 2016, 27(13): 2079-2081.

[12] 甘杰华, 严 萍, 马 庆, 等. 三叉苦质量标准的研究 [J]. 中成药, 2019, 41(3): 511-515.

[13] 罗登花. 三叉苦种质资源 ISSR 遗传多样性分析及其品质评价 [D]. 广州: 广州中医药大学, 2018.

[14] 陈彩英, 黄永秋, 王小平, 等. 三叉苦本草源流考证 [J]. 中药新药与临床药理, 2017, 28(4): 543-548.

[15] 杨卫丽, 毛彩霓, 刘明生, 等. 三叉苦生药学研究 [J]. 中国热带医学, 2008, 8(5): 851-853.

[16] 王尧龙, 黄璐琦, 袁 媛, 等. 药用植物转录组研究进展 [J]. 中国中药杂志, 2015, 40(11): 2055-2061.

[17] 崔 凯, 吴伟伟, 刁其玉. 转录组测序技术的研究和应用进展 [J]. 生物技术通报, 2019, 35(7): 1-9.

[18] 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(9): 1593-1601.

[19] 朱昀昊, 张梦佳, 李 璐, 等. 夏枯草的转录组测序与次生代谢产物生物合成相关基因的挖掘 [J]. 中草药, 2019, 50(5): 1220-1226.

[20] 陈延清, 胡志刚, 刘大会, 等. 药用植物冬凌草高通量转录组测序与分析 [J]. 中国现代中药, 2018, 20(12): 1476-1482.

[21] 时小东, 顾雨熹, 代 娇, 等. 基于转录组的厚朴次级代谢产物途径基因挖掘及分析 [J]. 时珍国医国药, 2018, 29(1): 247-250.

[22] 李依民, 彭 亮, 杨冰月, 等. 基于高通量测序技术的黄三七根茎转录组数据分析 [J]. 中草药, 2018, 49(21): 4983-4990.

[23] 郭光艳, 柏 峰, 刘 伟, 等. 转录因子对木质素生物合成调控的研究进展 [J]. 中国农业科学, 2015, 48(7): 1277-1287.

[24] 陆 丹, 牛 楠, 李玥莹, 等. SSR 标记技术在植物基因组研究上的应用 [J]. 沈阳师范大学学报: 自然科学版, 2010, 28(1): 83-85.

[25] 陈国庆. 番红花 EST 资源的 SSR 信息分析 [J]. 广西植物, 2011, 31(1): 43-46.

[26] 雷淑芸, 高庆波, 付鹏程, 等. 基于 Solexa 高通量测序的唐古特红景天(Rhodiola algida)微卫星信息分析 [J]. 植物研究, 2014, 34(6): 829-834.

[27] 黑小斌, 李 欢, 李依民, 等. 药用大黄幼苗转录组高通量测序及蒽醌类生物合成基因筛选 [J]. 中国药学杂志, 2019, 54(7): 526-535.

[28] 袁 灿, 彭 芳, 杨泽茂, 等. 川芎转录组 SSR 分析与 EST-SSR 标记的开发 [J]. 中国中药杂志, 2017, 42(17): 3332-3340.

[29] 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(1): 245.