罗勒花和叶的转录组数据组装及基因功能注释

刘　雷[1]，赵　欢[2*]，冉茂中[3]，杜保国[1]，邵镏钎[1]，伍希来[1]，贾白慧[1]，查　英[1]

1. 绵阳师范学院生命科学与技术学院，四川 绵阳　621000
2. 西华师范大学生命科学院，四川 南充　637009
3. 南充职业技术学院 农业科学技术系，四川，南充　637131

摘　要：目的　系统了解罗勒 *Ocimum basilicum* 花和叶的转录组特征，丰富罗勒转录组数据信息。**方法**　选取罗勒新鲜花朵和叶片为材料，利用 Illumina HiSeq™ 2500 进行高通量测序，构建罗勒花和叶转录组文库，并用测序评估、转录组数据组装和基因功能注释等生物信息学方法进行分析。**结果**　对原始数据进行除杂之后，共获得 86 331 137 个 reads 片段，包含了 6 455 365 309 个核苷酸序列信息；将质控后得到的高质量序列进行 *de novo* 拼接，得到了 90 341 个 Unigene 片段。将 Unigene 与 COG 数据库进行比对表明，罗勒的花和叶转录组中的 Unigene 根据功能可大致分为 25 类；根据 GO 功能分类可大致分为生物过程、细胞组分和分子功能 3 大类 43 分支；利用 KEGG 数据库作为参考，依据代谢通路可以将转录组中的数据分成 111 类，包括生化代谢通路、植物病原体互作、DNA 剪切、植物激素生物合成、苯丙氨酸生物合成、萜类化合物与类固醇类化合物合成、脂类代谢、RNA 降解等。MISA 软件成功设计 15 617 对 SSR 引物，检测到 10 254 个 SNP 位点。**结论**　研究结果为后期罗勒功能基因的挖掘、代谢途径及调控机制的研究奠定了理论基础。

关键词：罗勒；转录组；功能基因；代谢通路；萜类化合物

中图分类号：R282.12　　**文献标志码：**A　　**文章编号：**0253 - 2670(2017)17 - 3612 - 07

DOI: 10.7501/j.issn.0253-2670.2017.17.025

Transcriptome data assembly and gene function annotation of flowers and leaves of *Ocimum basilicum*

LIU Lei[1], ZHAO Huan[2], RAN Mao-zhong[3], DU Bao-guo[1], SHAO Qiang-qian[1], WU Xi-lai[1], JIA Bai-hui[1], ZHA Ying[1]

1. College of Life Science & Biotechnology, Mianyang Normol University, Mianyang 621000, China
2. College of Life Science, China Western Normal University, Nanchong 637009, China
3. Nanchong Vocational and Technical College, Department of Agriculture Science and Technology, Nanchong 637131, China

Abstract: Objective　To understand the transcriptome data of flowers and leaves of *Ocimum basilicum*, and analyze the transcriptome sequencing and bioinforamtics of *O. basilicum*. **Methods**　Selecting fresh flowers and leaves of *O. basilicum* as samples, the transcriptome libraries of *O. basilicum* were constructed using Illumina HiSeqTM 2500 sequencing technique and analyzed using the bioinformatics methods subsequently, such as sequencing assess, transcriptome data assembly, and gene function annotation. **Results**　After transcriptome sequencing and removing insignificant reads, 86 331 137 reads of *O. basilicum* were obtained. All of the reads contained 6 455 365 309 nucleotides. After *de novo* splicing, 90 341 Unigenes were obtained. The Unigenes were aligned in COG database, and searching result demonstrated that UniGenes were devided into 25 classes according to function. The Unigene GO functions could be broadly divided into biological processes, cellular components and molecular function categories of 43 branches. In KEGG database, the data in transcriptome could be divided into 111 classes according to the metabolic pathway which included the biochemical pathway in plants-Pathogens interaction, terpenoid and steroid compounds synthesis, lipid metabolism, RNA degradation and so on. Totally 15 617 pairs of SSR primers were successful designed by MISA software, and 10 254 SNP loci were detected. **Conclusion**　The results of this study can provide the further development of functional gene excavation, mentabolic pathways and

收稿日期：2017-02-05

基金项目：四川省科技厅科技支撑计划项目（2016NZ0058）；四川省教育厅重点科研项目（16ZA0323）；2016 年地方高校国家级大学生创新创业训练计划项目（201610639002，201610639014）

作者简介：刘　雷（1980—），男，四川彭州人，讲师，主要从事植物资源的评价与利用的研究。Tel: 13990177957　E-mail: 33020897@qq.com

*通信作者　赵　欢　Tel: (0817)2568353　E-mail: zhaohuan_2010@163.com

their regulatory mechanism for *O. basilicum* with theatrical basis.

Key words: *Ocimum basilicum* L.; transcriptome; functional genes; biochemical pathway; terpenoid

罗勒 *Ocimum basilicum* L. 为唇形科罗勒属一年生草本植物，俗称香佩兰、零陵香、九层塔、金不换和圣约瑟夫草，原产于以印度为中心的亚洲热带地区和非洲，在我国主要分布于新疆、吉林、河北、浙江、江苏、安徽、江西、湖北、湖南、广东、广西、福建、台湾、贵州、云南及四川等省[1]。罗勒具有疏风行气、化湿消食、活血、解毒之功能，主要用于治疗外感头痛、食胀气滞、脘痛、泄泻、月经不调、跌打损伤、蛇虫咬伤、皮肤湿疮等症[2-3]。此外，其还可作为提取精油、食品调味料、减肥代餐食品等原料，具有较大的经济效益和社会效益[4-5]。

罗勒化学成分的研究近年来主要集中于挥发油、酚酸、黄酮和甾体类。罗勒中含有大量的挥发油，通过水蒸气蒸馏法提取和 GC-MS 分析发现其含有多种萜烯类的含氧衍生物。近年来大量研究表明不同产地的罗勒挥发油成分和含量差异较大。兰瑞芳等[6]、帕丽达等[7]、Jose 等[8]、李建文等[9]分别对我国福建、新疆产和肯尼亚产及栽培罗勒挥发油进行研究，结果都表明芳樟醇相对量最高，达 50%左右，其他主要成分有茴香脑、对烯丙基苯甲醚、1,8-桉叶素、表姜烯酮、杜松烯醇和樟脑。卢汝梅等[10]对桂产罗勒挥发油的研究结果表明，相对量最高的为对烯丙基茴香醚（50.26%），其他主要成分有双环倍半水芹烯和 3,7,11-三甲基-(*Z*,*E*)-1,3,6,10-十二碳四烯。汪涛等[11]从河南产罗勒挥发油中鉴定出相对量最高的是 1,7-二甲基-1,6-辛二烯-3-醇（29.87%），其他主要成分有 1-己烯、3-己酮、环氧乙烷。胡西旦•格拉吉丁[12]从新疆产罗勒中鉴定出相对量最高的为 α-萜品油烯（30.97%），其他主要成分有香榧烯醇、α-萜品油、β-月桂烯、δ-愈创水烯和杜松烯。

关于罗勒中挥发油及生物碱的量及在不同栽培环境、措施及品种间的差异有一定的报道，但是上述研究均未能从本质上揭示活性物质的生物合成机制、代谢调控途径及调控水平。分子生物学是研究罗勒活性成分代谢调控途径的重要手段。随着后基因组时代的到来，转录组学、蛋白质组学、代谢组学等各种组学技术相继出现，其中转录组学是率先发展起来以及应用最广泛的技术。转录组是特定组织或细胞在某一功能状态下转录出来的所有 RNA 的总和，包括 mRNA 和非编码 RNA[13-14]。目前转录组测序及分析技术可以解决新基因的深度发掘、低丰度转录本的发现、转录图谱绘制、可变剪接的调控、代谢途径确定、基因家族鉴定及进化分析等各方面的问题[15-18]。为此，本课题组在前期对罗勒资源收集、评价，有效成分提取、分离及检测的研究基础上，采用转录组测序的方法，对罗勒花和叶片中功能基因进行功能注释和分类，为后期功能基因的挖掘、代谢途径及调控机制的研究奠定理论基础。

1 材料与方法

1.1 材料

供试植物于 2017 年采集于四川省绵阳市，经绵阳师范学院生命科学与技术学院罗明华教授鉴定为罗勒 *Ocimum basilicum* L.。

1.2 RNA 的提取与分离

采用 GENEOUTTM 植物 RNA 提取试剂盒（多糖多酚样本，成都兰博生物科技有限公司）提取罗勒花期花和叶的总 RNA，使用磁力架（厂商 Invitrogen）以磁珠法分离 mRNA[15]。分离到 mRNA 之后进行扩增、构建文库以及测序。

1.3 转录组数据的获得

将上述获得的罗勒总 RNA，以 5 μg 的起始量建库；采用磁珠法分离 mRNA，打断 mRNA（Truseq^{TM} RNA sample prep Kit）；双链 cDNA 合成、补平、3’端加 A、连接 index 接头（Truseq^{TM} RNA sample prep Kit）；文库富集，PCR 扩增 15 个循环；2%琼脂糖胶回收目的条带（Certified Low Range Ultra Agarose）；TBS380（Picogreen）定量，按数据比例混合上机；cBot 上进行桥式 PCR 扩增，生成 clusters；在 Hiseq2000 测序平台进行 2×100 bp 测序。

1.4 原始数据处理及生物信息学分析

采用 Base Calling 将测序得到的原始图像数据转化为序列数据，采用 FASTQ 文件格式来储存结果文件。将原始测序数据进行统计和评估，再根据接头信息去除有接头污染的序列。得到原始的 FASTQ 数据后，首先对其进行质控得到高质量的测序结果（clean data），然后再进行 *de novo* 拼接。

在 RNA-seq 分析过程中，将测序得到的 reads 与前面所得的拼接结果进行比对（mapping）。通过对定位到基因组区域的测序序列（clean reads）的数量来估计基因的表达水平，采用 Trinity 软件对拼接结果进行开放阅读框（ORF）预测。通过 GO（gene ontology）数据库和 COG 数据库对基因的功能进行分类；基于 KEGG 数据库，采用 BLAST 算法（blastx/blastp 2.2.24＋）将罗勒所有基因与 KEGG 的基因数据库（GENES）进行比对，再根据比对所得到的 KO 编号去查找具体的生物学通路，提供所分析基因可能参与的所有生物学通路。

2　结果

2.1　转录组数据组装

对罗勒花和叶进行了转录组测序之后，经原 reads 片段除杂，共获得 86 331 137 个高质量 reads 片段，包含了 6 455 365 309 个核苷酸序列信息。将质控后得到的高质量序列进行 *de novo* 拼接。结果显示，拼接得到的总 Unigene 片段达到 90 341 条，平均长度为 1 314.85 bp，最长的 Unigene 片段为 14 243 bp，最短的 Unigene 片段为 351 bp。总共得到 48 762 条基因。

在拼接得到的 90 341 条 Unigene 片段中，长度在 400～600 bp 的片段最多，达 20 765 条，其次是长度在 600～800 bp 和 800～1 000 bp 左右的片段，分别为 12 986 条和 9 506 条。

转录本的丰度体现基因的表达水平，转录本丰度越高，则基因表达水平越高。在分析中，将测序得到的 reads 与前面所得的拼接结果进行比对。结果显示，罗勒花与叶的转录本丰度均较高（≥70%），分别为 75%和 70%。

2.2　转录组数据拼接结果及基因预测

基于罗勒的花和叶的转录组测序，通过 Trinity 软件对拼接结果进行 ORF 预测，总共预测到具有 ORF 的序列有 60 476 条，另外 39 865 个序列未预测到 ORF。对具有 ORF 的序列进行蛋白质预测，共预测到 62 895 条蛋白质序列。

2.3　基因功能注释

2.3.1　GO 分类　GO 是基因本体论联合会建立的数据库。本研究将罗勒花和叶转录组获得的 Unigene 在 GO 功能数据库中比对分析发现，共有 219 789 条 Unigene 与数据库中的基因具有相似性，较多的单条 Unigene 能够与多种基因相对应，建立了 219 789 条对应关系，从而得到尽可能多的注释和分类。罗勒花和叶转录组中的 Unigene 根据 GO 功能大致可分为生物过程、细胞组分和分子功能 3 个大类 43 个分支（图 1）。通过对每一类的基因数量进行统计分析，结果表明，在生物过程这一大类中，代谢过程涉及的基因最多，有 97 468 条；在细胞组分这一大类中，细胞部分涉及的基因最多，达 18 072 条，其次是组成细胞器的基因，有 13 448 条。在分子功能这一大类中，具有催化活性功能涉及的基因最多，达 20 770 条，其次是具有结合功能的基因，有 19 243 条（表 1）。

2.3.2　KEGG 代谢途径分类　利用 KEGG 数据库作为参考，依据代谢通路将转录组中的数据分成 111

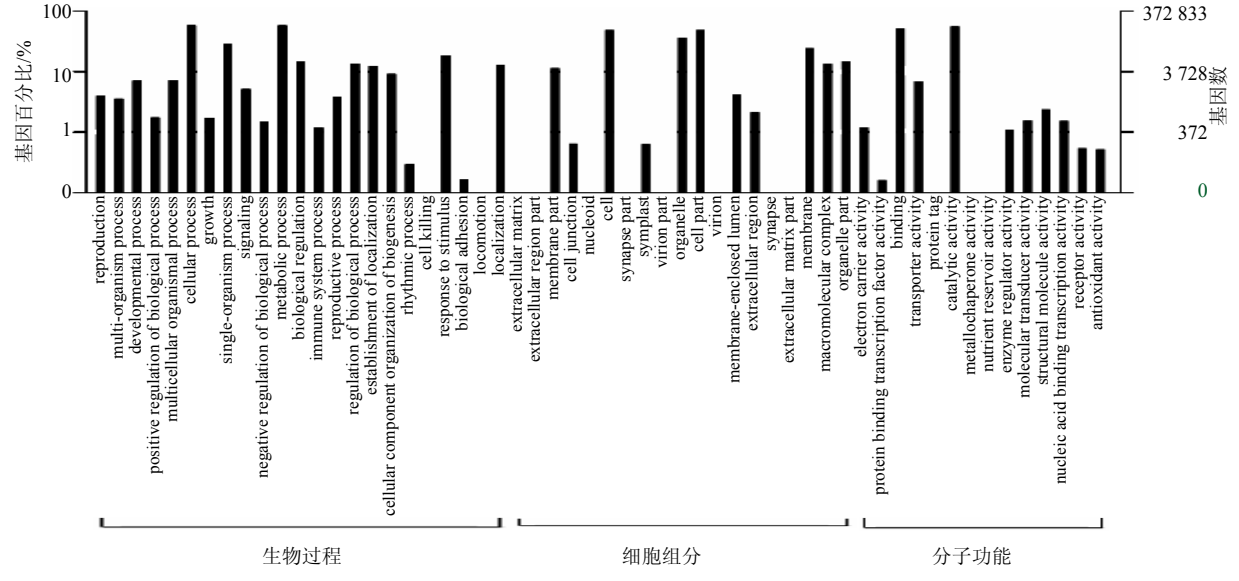

图 1　罗勒花和叶转录组的 Unigene GO 功能分类

Fig. 1　Unigene GO functional classification of transcriptome in flower and leaf of *O. basilicum*

表 1 罗勒转录组的 Unigene GO 功能分类

Table 1 Unigene GO functional classification of transcriptome of *O. basilicum*

本体功能类别	基因功能	基因数量	本体功能类别	基因功能	基因数量
生物过程	繁殖	1 482	细胞组分	突触部分	2
生物过程	多生物过程	1 321	细胞组分	神经元突触	2
生物过程	发展过程	2 643	细胞组分	细胞外基质成分	1
生物过程	生物过程的正调控	645	细胞组分	膜	9 103
生物过程	多细胞生物过程	2 688	细胞组分	大分子复合物	5 007
生物过程	细胞过程	21 483	细胞组分	细胞器部分	5 504
生物过程	生长	640	细胞组分	细胞外基质	8
生物过程	单生物过程	10 493	细胞组分	细胞外区域组成	16
生物过程	信号	1 904	细胞组分	分隔膜	4 254
生物过程	生物过程的负调控	555	细胞组分	细胞连接	240
生物过程	代谢过程	21 459	细胞组分	拟核	30
生物过程	生物调控	5 482	细胞组分	细胞	18 072
生物过程	免疫系统过程	441	细胞组分	突触部分	2
生物过程	生殖过程	1 405	分子功能	电子转运体活性	447
生物过程	生物过程调控	5 026	分子功能	蛋白结合转录因子	60
生物过程	细胞定位	4 597	分子功能	连接体	19 243
生物过程	细胞组织成分生物合成	3 391	分子功能	转运体活性	2 569
生物过程	节律过程	109	分子功能	蛋白质标签	4
生物过程	细胞杀伤	4	分子功能	催化活性	20 770
生物过程	对刺激的反应	6 896	分子功能	金属伴侣蛋白活性	5
生物过程	生物黏附	61	分子功能	营养盐储层活性	19
生物过程	运动	37	分子功能	酶调节活性	408
生物过程	定位	4 706	分子功能	分子转换器活性	580
细胞组分	共质体	236	分子功能	结构分子活性	885
细胞组分	病毒体部分	8	分子功能	核酸结合转录因子活性	572
细胞组分	细胞器	13 448	分子功能	受体活性	202
细胞组分	细胞部分	18 072	分子功能	抗氧化活性	193
细胞组分	病毒粒子	12	分子功能	电子转运体活性	447
细胞组分	膜封闭腔	1 554			

类，包括生化代谢通路、植物病原体互作、DNA 剪切、植物激素生物合成、苯丙氨酸生物合成、萜类化合物与类固醇类化合物合成、脂类代谢、RNA 降解等，共涉及基因 22 955 条（表 2）。其中，次生代谢物涉及基因 2 158 条，占整体的 9.4%。如黄酮类（包括了黄酮、黄酮醇、类黄酮）的基因有 44 条，占总体的 0.192%；萜类（包括了单萜、二萜、三萜、倍半萜、萜类化合物骨架）涉及的基因有 201 条，占整体的 0.88%；类胡萝卜素代谢途径中涉及基因有 81 条，占总体的 0.35%。

2.3.3 COG 功能分类 将罗勒花和叶的转录组 Unigene 片段与 COG 数据库进行比对，发现共有 25 596 条 Unigene 与 COG 数据库中的基因具有相似性，且较多的单条 Unigene 能够与多种基因相对应，建立了 25 596 条对应关系。罗勒花和叶转录组中的 Unigene 根据功能大致可分为 25 类，并对每一类的基因数量进行了统计。结果显示，Unigene 的 COG 功能种类比较全面，涉及到大多数的生命活动，仅作为一般功能预测的基因数量最多，有 5 299 条；其次是基因的复制、重组、修复等，涉及的基因为 2 721 条。其他种类基因的表达丰度也不尽相同，具体种类和数量见表 3。

2.3.4 转录组序列中 SSR 重复基元分析 从 http://pgrc.Lpk-gatersleben.de/misa 网站下载 est-trimmer.pl，

表 2 罗勒转录组 Unigene KEGG 的代谢途径分类

Table 2 KEGG classification of Unigene in transcriptome of *O. basilicum*

编号	代谢途径	注释度	代谢途径 ID	编号	代谢途径	注释度	代谢途径 ID
1	丁酸代谢	85	ko04664	56	次生代谢产物的生物合成	2	ko01110
2	错配修复	68	ko03430	57	磷脂酰肌醇信号系统	208	ko04070
3	光转导	72	ko04745	58	昼夜节律	20	ko04711
4	烟酸与烟酰胺代谢	45	ko00760	59	减数分裂	172	ko04113
5	蛋白质输出	125	ko03060	60	氨酰 tRNA 生物合成	124	ko00970
6	同源重组	103	ko03440	61	苯乙烯降解	25	ko00643
7	碱基切除修复	87	ko03410	62	黏多糖的降解	25	ko00531
8	嘧啶代谢	267	ko00240	63	钙信号途径	191	ko04020
9	萜类化合物骨架生物合成	103	ko00900	61	磷酸肌醇代谢	162	ko00562
10	倍半萜、三萜生物合成	59	ko00909	65	柠檬烯和蒎烯的降解	44	ko00903
11	二萜化合物生物合成	29	ko00904	66	细胞周期	228	ko04110
12	单萜化合物生物合成	10	ko00902	67	异喹啉类生物碱的合成	45	ko00950
13	嘌呤代谢	343	ko00230	68	核糖体	640	ko03010
14	多环芳烃的降解	18	ko00624	69	*N* 糖链的合成	179	ko00510
15	果糖和甘露糖代谢	127	ko00051	70	各种类型的 N 种糖链的合成	111	ko00513
16	泛素介导的蛋白质水解	269	ko04120	71	其他类型 *O* 他聚糖的生物合成	36	ko00514
17	氨基酸与核苷酸糖代谢	267	ko00520	72	维生素 B_6 代谢	20	ko00750
18	戊糖与葡糖糖醛酸转换	198	ko00040	73	卟啉与叶绿素代谢	84	ko00860
19	丙酮酸代谢	223	ko00620	74	真核生物核糖体合成	349	ko03008
20	赖氨酸降解	101	ko00310	75	不饱和脂肪酸的生物合成	127	ko01040
20	色氨酸代谢	110	ko00380	76	植物昼夜节律	110	ko04712
22	甘氨酸、丝氨酸、苏氨酸代谢	130	ko00260	77	趋化因子信号通路	76	ko04062
23	赖氨酸生物合成	77	ko00300	78	膦酸盐和膦酸盐的代谢	20	ko00440
24	组氨酸代谢	81	ko00340	79	非同源末端连接	40	ko03450
25	丙氨酸、天冬氨酸、谷氨酸代谢	129	ko00250	80	植物病原体相互作用	481	ko04626
26	β-丙氨酸代谢	131	ko00410	81	生物素代谢	30	ko00780
27	*D*-谷氨酰胺与 *D*-氨谷氨酸代谢	9	ko00471	82	内吞作用	369	ko04144
28	苯丙氨酸、酪氨酸、色氨酸生物合成	95	ko00400	83	核黄素代谢	17	ko00740
				84	玉米素的生物合成	35	ko00908
29	氰基氨基酸代谢	95	ko00460	85	硫胺素代谢	44	ko00730
30	缬氨酸、亮氨酸、异亮氨酸降解	185	ko00280	86	氮素代谢	98	ko00910
31	缬氨酸、亮氨酸、异亮氨酸生物合成	35	ko00290	87	脂肪酸延长	62	ko00062
		206	ko00330	88	其他多糖降解	24	ko00511
32	精氨酸与脯氨酸代谢	171	ko00270	89	内质网中的蛋白质处理	529	ko04141
33	半胱氨酸与甲硫氨酸代谢	103	ko00350	90	类胡萝卜素生物合成	81	ko00906
34	酪氨酸代谢	124	ko00360	91	谷胱甘肽代谢	183	ko00480
35	苯丙氨酸代谢	590	ko01230	92	泛酸和辅酶 A 合成	86	ko00770
36	氨基酸的生物合成	26	ko00362	93	光传导	61	ko04744
37	苯甲酸降解	127	ko00640	94	ABC 转运	73	ko02010
38	丙酸代谢	283	ko00010	95	植物激素信号转导	448	ko04075
39	糖酵解途径	153	ko00710	96	硫辛酸代谢	31	ko00785
40	光合生物碳固定	81	ko00195	97	过氧化物酶体	216	ko04146
41	光合作用	130	ko00030	98	核苷酸的切除修复	134	ko03420
42	磷酸戊糖途径	476	ko00190	99	甘油磷脂代谢	451	ko00564
43	氧化磷酸化	67	ko00061	100	淀粉和蔗糖的代谢	515	ko00500
44	脂肪酸生物合成	189	ko00020	101	黄酮和黄酮醇的生物合成	12	ko00944
45	三羧酸循环	299	ko03018	102	类黄酮的合成	32	ko00941
46	RNA 降解	507	ko03013	103	油菜素内酯的合成	13	ko00905
47	RNA 转运	82	ko03020	104	类固醇合成	94	ko00100
48	RNA 聚合酶	568	ko03040	105	吲哚生物碱合成	2	ko00901
49	剪接	408	ko03015	106	光合作用天线蛋白	24	ko00196
50	mRNA 监测途径	82	ko03030	107	乙醛酸和二羧酸的代谢	139	ko00630
51	DNA 复制	52	ko04623	108	咖啡因的代谢	2	ko00232
52	DNA 胞浆检测	42	ko00591	109	硫代谢	56	ko00920
53	亚油酸代谢	70	ko04210	110	矿物吸收	35	ko04978
54	细胞凋亡	68	ko00790	111	亚麻酸的代谢	69	ko00592
55	叶酸生物合成						

表 3 罗勒转录组的 Unigene COG 功能分类

Table 3 COG function classification of transcriptome in *O. basilicum*

编号	功能分类	基因数量
A	核糖核酸加工与修饰	268
B	染色质结构与动力学	234
C	能量生产和转换	844
D	细胞周期调控、细胞分裂、染色体分割	462
E	氨基酸转运与代谢	1 064
F	核苷酸转运与代谢	411
G	碳水化合物转运与代谢	1 340
H	辅酶转运与代谢	484
I	脂类转运与代谢	817
J	翻译、核糖体结构和生物合成	1 543
K	转录	2 538
L	复制、重组和修复	2 721
M	细胞壁、膜、膜的发生	482
N	细胞活力	9
O	翻译后修饰、蛋白质周转、伴侣	1 419
P	无机离子转运与代谢	758
Q	次生代谢产物的生物合成、运输和代谢	692
R	一般功能预测	5 299
S	功能未知	718
T	信号转导机制	2 332
U	细胞内转运、分泌、和小泡运输	386
V	防御机制	412
W	胞外结构	0
Y	核结构	5
Z	细胞骨架	358

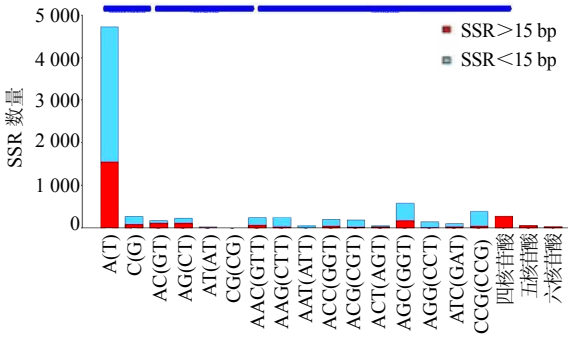

图 2 罗勒转录组序列中 SSR 统计情况

Fig. 2 SSR statistics of transcriptome in *O. basilicum*

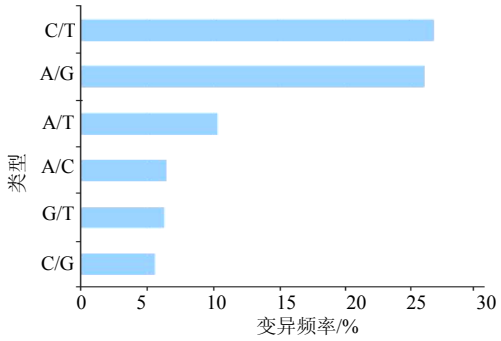

图 3 罗勒转录组中序列中 SNP 统计情况

Fig. 3 SNP statistics of transcriptome in *O. basilicum*

去除转录组序列中过短的序列和过长的序列；从 http://www.bioinformatics,org/cd-hit/中下载 CD-HIT 软件，去除冗余序列。从 http://pgrc.lpk-gatersleben.de/misa 网站下载使用 MISA 软件以识别和定位序列中 SSR，参数设置如下：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的重复次数至少为 10、6、5、3、3、3。使用 Primer3 批量设计 SSR 引物，网址：http://sourceforge.net/projects/primer3/files/primer3/1.1.4/primer3-1.1.4-WINXP.zip/download，引物设计参数为引物长度 18～22 bp，退火温度（T_m）55～65 ℃。其中前后引物 T_m 值相差 4 ℃，产物大小为 100～300 bp。应用本方法成功设计了 15 617 对 SSR 引物，SSR 密度分布出现频率最高是的单碱基微卫星，所占比例最高是 A/T，如图 2 所示。

2.3.5 转录组序列中 SNP 情况分析 本研究共检测到对照样品 SNP 10 254 个，注释到基因上的有 7 242 个。统计发现，6 种单核苷酸变异种 C/T 和 A/G 发生频率最高，大于 20%，其他 4 种单核苷酸变异 A/C、G/T、C/G、A/T 发生频率均在 15%以下（图 3）。6 种变异类型中 C/T 变异频率最高，可能是因为 CpG 二核苷酸上甲基化的胞嘧啶残基易自发脱去氨基而形成胸腺嘧啶的原因。

3 讨论

本实验首次建立了药用植物罗勒的花和叶的转录组数据库，获得了大量的转录本信息，结果共获得 86 331 137 个 reads 片段，包含 6 455 365 309 个核苷酸序列信息，对 reads 进行 *de novo* 拼接，共获得 90 341 个 Unigene 片段。对罗勒转录组所有 Unigene 的 ORF 进行 blast 分析，共获得 60 476 个 ORF 序列；在蛋白质数据库中对罗勒转录组所有的 Unigene 进行 blast 分析后，共获得了 62 895 个蛋白质序列；将 Unigene 和 COG 数据库进行比对发现共有 25 596 条 Unigene 与数据库中的基因具有相似性，共有 219 789 条 Unigene 可以与 GO 数据库中的基因具有相似性。

根据 KEGG 数据库作为参考，对上述 Unigene 进行代谢途径分析，可将其分为 111 类，参与到罗勒的生化代谢通路、植物激素生物合成、苯丙氨酸生物合成、萜类化合物与类固醇类化合物合成、脂类代谢、RNA 降解碳水化合物代谢、脂类代谢等过程中，其中，萜类和黄酮类涉及的基因共 245 条，该结果可为

发掘罗勒次生代谢产物合成途径上的关键基因克隆、功能验证等方面提供了数据依据。特别是萜类（包括了单萜、二萜、三萜、倍半萜、萜类化合物骨架）涉及的基因有 201 条，对进一步深入理解该物种挥发性成分生物合成关键酶基因奠定了基础。此外，以 COG 数据库为参考，将本实验测序所得的罗勒花和叶的转录组 Unigene 片段进行基因功能分类，可从基因组水平上找寻直系同源体，预测未知 ORF 的生物学功能，可以大大提高基因功能注释的准确性，本结果进一步支持了贾新平等[19]的观点。

罗勒属植物品种及变种繁多，Rastogi 等[20]对甜罗勒 *Ocimum basilicum* L. 和圣罗勒 *O. sanctum* L. 的叶片进行了转录组测序及综合比较分析，对其主要化学成分萜类化合物及苯丙素类化合物代谢途径进行了分析，建立了罗勒属植物的第一个转录组数据库。Zhan 等[21]对罗勒属一变种 *Ocimum americanum* var. *pilosum* (Wild.) Benth. 在低温胁迫下进行了转录组测序，构建了罗勒属植物首个冷响应转录组数据库。本课题组则对本地栽培罗勒品种花和叶混合库进行了转录组测序分析，并对数据进行了组装、基因功能注释和生物信息学分析，丰富了罗勒属植物转录组数据库信息。挖掘了其中的萜类等次生代谢成分生物合成途径关键酶候选 Unigene，值得注意的是 412 条防御进程相关基因的获得，对利用分子生物学手段提升该物种产量和品质奠定了初步基础。通过对罗勒转录组的统计分析及功能分类，可从次生代谢调控途径中挖掘相关的功能基因及其序列，并直观了解其在代谢通路中的调控位置。通过后期的基因序列验证，及其与活性成分量的相关性分析，以及对外源诱导因子的响应等研究，揭示罗勒活性成分代谢途径的调控机制，并对限制性步骤的关键酶基因进行转基因技术研究，通过调整其表达量的手段，从而提高活性成分的量。因此，本研究为罗勒属植物新品种（系）选育、分子标记开发及活性成分量的提高奠定了坚实的理论基础。

参考文献

[1] 汪荣斌, 王存琴, 秦亚东, 等. 罗勒的本草考证 [J]. 中药材, 2015, 38(5): 1081-1084.

[2] 董泽科, 徐先祥, 吴雅清, 等. 罗勒的化学成分和药理作用研究进展 [J]. 中国民族民间医药, 2013, 22(9): 46-48.

[3] 官玲亮, 吴丽芬, 庞玉新, 等. 芳香植物罗勒的研究进展 [J]. 热带农业科学, 2013, 33(8): 42-46.

[4] 胡尔西丹•伊麻木, 热娜•卡斯木, 阿吉艾克拜尔•艾萨. 罗勒子挥发油成分及抗氧化活性分析 [J]. 安徽农业科学, 2012, 40(2): 752-754.

[5] 张海弢, 付 娟, 杨世海, 等. 罗勒研究进展及开发利用 [J]. 人参研究, 2012, 24(3): 35-39.

[6] 兰瑞芳, 冯 珊. 闽产罗勒油化学成分的研究 [J]. 海峡药学, 2001, 13(1): 51-52.

[7] 帕丽达, 米仁沙, 丛媛媛, 等. 新疆罗勒芳香油的化学成分研究 [J]. 中草药, 2006, 37(3): 352-356.

[8] José S, Dambolena A, Maria P, *et al*. Essential oils composition of *Ocimum basilicum* L. and *Ocimum gratissimum* L. from Kenya and their inhibitory effects on growth and fumonisin in production by Fusarium verticillioides [J]. *Innov Food Sci Emerg Technol*, 2010, 11(6): 410-414.

[9] 李建文, 陈贵林, 何洪巨. GC-MS 法测定罗勒中芳香成分 [J]. 现代仪器与医疗, 2003(2): 19-20.

[10] 卢汝梅, 李耀华. 桂产罗勒芳香油化学成分的分析 [J]. 广西植物, 2006, 26(4): 456-458.

[11] 汪 涛, 崔书亚, 胡晓黎, 等. 罗勒挥发油成分研究 [J]. 中国中药杂志, 2003, 28(8): 740-742.

[12] 胡西旦•格拉吉丁. 气相色谱-质谱法分析罗勒中挥发油的化学成分 [J]. 光谱实验室, 2008, 25(2): 127-131.

[13] Verculescu V E, Zhang L, Zhou W, *et al*. Characterization of the yeast transcriptome [J]. *Cell*, 1997, 88(2): 243-251.

[14] 王晓玥, 宋经元, 谢彩香, 等. RNA-Seq 与道地药材研究 [J]. 药学学报, 2014, 49(12): 1650-1657.

[15] Chen J W, Hou K, Qin P, *et al*. RNA-Seq for gene identification and transcript profiling of three *Stevia rebaudiana* genotypes [J]. *BMC Genom*, 2014, 15: 571-582.

[16] 朱孝轩, 朱英杰, 宋经元, 等. 基于全基因组和转录组分析的赤芝密码使用偏好性比较研究 [J]. 药学学报, 2014, 49(9): 1340-1345.

[17] 李翠婷, 张广辉, 马春花, 等. 野三七转录组中 SSR 位点信息分析及其多态性研究 [J]. 中草药, 2014, 45(10): 1468-1572.

[18] Liu H C, Wu W, Hou Kai, *et al*. Deep sequencing reveals transcriptome re-programming of *Polygonum multiflorum* Thunb. roots to the elicitation with methyl jasmonate [J]. *Mol Gen Genom*, 2016, 291(1): 337-348.

[19] 贾新平, 叶晓青, 梁丽建, 等. 基于高通量测序的海滨雀稗转录组学研究 [J]. 草业学报, 2014, 23(6): 242-252.

[20] Rastogi S, Meena S, Bhattacharya A, *et al*. *De novo* sequencing and comparative analysis of holy and sweet basil transcriptomes [J]. *BMC Genom*, 2014, 15(1): 588-605.

[21] Zhan X Q, Yang L, Wang D, *et al*. *De novo* assembly and analysis of the transcriptome of *Ocimum americanum* var. *pilosum* under cold stress [J]. *BMC Genom*, 2016, 17(1): 1-12.