DOI:10.13275/j.cnki.lykxyj.2018.05.001

云南干热河谷地区余甘子转录组分析

刘雄芳，李太强，李正红，万友名，刘秀贤，张 序，安 静，马 宏*

（中国林业科学研究院资源昆虫研究所，云南 昆明 650224）

摘要：[**目的**]对云南干热河谷地区余甘子转录组特征进行描述，旨在为余甘子微卫星标记的开发和功能基因的挖掘提供较全面的背景信息。[**方法**]采用 Illumina Hiseq 4000 测序平台对余甘子叶片进行转录组测序，对原始数据进行过滤、de novo 组装及聚类去冗余等处理后，再与公共数据库进行比对，对 Unigenes 进行基本功能注释、CDS 预测、TF 编码能力预测及 R-Gene 预测等分析。[**结果**]本研究共获得 10.52 Gb 的 Clean reads，Q20、Q30 分别为 98.47%、95.28%。组装并去冗余后获得 76 881 条 Unigenes，平均长度、N50 分别为 713、1 257 nt。通过与 NR、COG、KEGG 和 SwissProt 数据库进行比对，44 768 条 Unigenes 获得功能注释。余甘子转录组 Unigenes 根据 COG 功能注释信息大致分为 25 类；按 GO 功能注释信息划分为生物学过程、细胞组分和分子功能 3 大类 47 亚类；参考 KEGG 注释信息，可归为 6 大代谢通路、21 类代谢途径，其中约 3/5 为代谢相关通路。根据以上注释结果共检测出 42 953 个 CDS，其余未比对上的 Unigenes 用 ESTScan 预测后得到 2 058 个 CDS。同时，预测到 56 个 TF 家族以及 18 种 R-Gene。[**结论**]本研究获得的余甘子转录组 Unigenes 序列的组装质量较高、完整性较好、基因丰富、功能多样，极大地扩充了余甘子基因信息库，为今后余甘子乃至叶下珠属植物功能基因挖掘、抗性机理分析、分子标记开发、分子辅助育种等研究提供了重要的基础数据。

关键词：余甘子；转录组；Unigene；功能注释；编码序列；转录因子；抗性基因

中图分类号：S722 **文献标识码**：A **文章编号**：1001-1498(2018)05-0001-08

Transcriptome Analysis for *Phyllanthus emblica* Distributed in Dry-hot Valleys in Yunnan, China

LIU Xiong-fang, LI Tai-qiang, LI Zheng-hong, WAN You-ming, LIU Xiu-xian, ZHANG Xu, AN Jing, MA Hong

(Research Institute of Resource Insects, Chinese Academy of Forestry, Kunming 650224, Yunnan, China)

Abstract: [**Objective**] To provide comprehensive genetic information for the development of microsatellite markers and the mining of functional genes in *Phyllanthus emblica* by characterizing the transcriptome of P. emblica in dry-hot valleys in Yunnan. [**Method**] Transcriptome sequencing was conducted on young leaves of *Ph. emblica* using Illumina Hiseq 4000, followed by filtering, de novo assembly and clustering. Sequence similarity analysis and annotation of the obtained Unigenes were performed based on databases like NCBI-non-redundant (NR) protein database, Gene Ontology (GO), Clusters of Orthologous Groups (COG), KEGG database, SwissProt, PlantTFDB, and PRGdb. [**Result**] In total, 10.52 Gb Clean reads with Q20 of 98.47% and Q30 of 95.28% were generated. A total of 76 881 Unigenes with an average length of 713 nt and N50 of 1 257 nt were obtained by de novo assembly and clustering with Clean reads. Out of them, 44 768 Unigenes were functionally annotated against four protein databases. The Unigenes were roughly divided into 25 categories according to COG function, and were grouped into three functional categories (including biological processes, cellular components and molecular function) and 47 sub-cate-

收稿日期：2018-02-26

基金项目：中央级公益性科研院所基本科研业务费专项(CAFYBB2016ZX003-2)；“云南省技术创新人才”培养对象项目(2016HB007)

作者简介：刘雄芳(1992—)，女，硕士研究生，主要从事林木种质资源及遗传多样性研究. E-mail:liuxiongfang16@163.com

* 通讯作者：马 宏，男，副研究员，主要从事野生花卉及多功能植物种质资源创新与遗传多样性研究. E-mail:hortscience@163.com

gories based on GO functional annotation. KEGG analysis showed that the Unigenes could be fallen into six categories and 21 metabolic pathways, of which about 3/5 were Metabolism. A total of 42 953 CDS were detected based on the results of functional annotation, and 2 058 CDS were predicted using ESTScan with the remaining Unigenes. And 56 Transcription Factor families and 18 resistance genes were predicted. [**Conclusion**] The Unigenes of transcriptome in *Ph. emblica* show high quality, good integrality, abundant genes and various functions, which could lay an important foundation for further study of functional gene excavation, resistance mechanism analysis, molecular marker development and molecular assisted breeding of *Ph. emblica* and other congeneric species.

Keywords: *Phyllanthus emblica*; transcriptome; unigene; functional annotation; CDS; transcription factor; R-Gene

滇川干热河谷地处我国西南生物多样性中心横断山脉的南缘,主要包括元江、怒江、金沙江和澜沧江四大干热河谷区,因地理位置、地形地貌及气候等因素,表现出独具特色的干、热生态特点[1]。由于气候和人类活动等因素的影响,干热河谷区域植被的原生类型几乎不复存在,现存的主要是与非洲萨王纳植被(即热带草原植被)具有一定相似性的半稀树草原植被[1],有的地方已变成光秃裸露的荒山,生物多样性和生态系统的稳定性遭到破坏,水土流失严重。生长在该地区的植物类群对这种严酷的干热生境形成了很好的适应性[2],余甘子就是其中典型的代表。

余甘子(*Phyllanthus emblica* L.)隶属叶下珠科(Phyllanthaceae)叶下珠属(*Phyllanthus*),是一种分布于热带和亚热带地区的重要食药同源经济树种,被世界卫生组织列为在世界范围内推广种植的三种保健植物之一[3]。其叶片、根和果实均含有维生素C、类黄酮、超氧化物歧化酶等有益于健康的成分,多种治疗功效在现代医药学研究中已被证实[4-5]。其中,余甘子果实是维生素C最丰富的天然来源之一,具有极高的商业开发潜力[6]。在生态学方面,余甘子因极耐干旱贫瘠环境而常被作为我国西南干热河谷荒山绿化的先锋树种,对水土流失严重的干热河谷地带有明显的保水、固土作用[3]。尽管兼具药用、经济和生态价值,但目前余甘子遗传背景仍然不清楚,特别是该植物的分子生物学背景研究较为薄弱。

随着高通量测序技术的发展,测序时间和成本显著减少。对于无参考基因组的物种,采用转录组高通量测序技术可获取大量的数据信息,构建cDNA文库,挖掘重要功能基因,同时也为大量分子标记的开发奠定基础,是开展植物优良性状研究的重要手段[7]。鉴于此,本研究以云南宾川干热河谷地区的余甘子为研究材料,采用Illumina Hiseq 4000平台对余甘子进行高通量转录组测序,并对测序的原始数据进行过滤和de novo组装,之后通过生物信息学的方法对获得的Unigenes进行功能注释、CDS预测、TF及抗性基因预测等分析,以期了解余甘子在一定生长发育时期基因表达情况及功能分布特征,为余甘子转录组水平上的研究以及微卫星标记的开发和抗旱相关基因的挖掘奠定基础,亦为干热河谷地区的生态恢复和经济发展提供参考。

1 材料与方法

1.1 试验材料

试验材料采自云南省宾川县(25°45′59″ N,100°26′29″ E,属金沙江干热河谷区),选择5株相距15 m以上的健康余甘子植株,每株各采集嫩叶5 g左右分别放在锡箔纸中包好,迅速放入液氮中,然后保存于-80℃冰箱中备用。

1.2 cDNA文库构建和转录组测序

各单株分别称取0.1 g叶片等量混合均匀后,按照Kumar等[8]的方法提取总RNA。RNA样品的浓度、纯度和完整性分别用Qubit 2.0、Nanodrop和Agilent 2100检测。以片段化的mRNA为模板,合成双链cDNA,纯化后进行末端修复、加A尾和接头,经片段大小筛选后进行PCR扩增富集得到cDNA文库。构建好的文库用Agilent 2100 Bioanalyzer质检合格后,使用Illumina HiSeq 4000进行测序。

1.3 数据过滤和de novo组装

为确保所获得reads的精确性,首先对原始数据进行一系列过滤(去除包含接头的reads、未知核苷酸比例>10%的reads以及质量值Q<10的碱基数占整个read 40%以上的reads),过滤后得到Clean reads。然后使用Trinity软件[9]对过滤后的Clean reads进行de novo组装获得转录本。最后使用CD-HIT[10]对转录本进行聚类得到Unigenes。

1.4　Unigene 注释、CDS、TF 编码能力及 R-Gene 预测

通过 BLASTx 将获得的 Unigenes 序列比对到 NR、COG、KEGG 以及 SwissProt 四大蛋白数据库,获得 Unigenes 序列的蛋白功能注释信息(*E*-value≤1e-5)。通过 BLASTp 将 Unigenes 比对到 PlantTFDB 和 R-Gene 数据库(PRGdb)分别预测 TF 家族和抗性基因。

2　结果与分析

2.1　余甘子转录组测序数据过滤及 de novo 组装

测序共产生 10.95 Gb 的 Raw reads,经过滤处理后获得 10.52 Gb 的 Clean reads,占原始读长的 98.03%。其中,中间未知碱基(N)含量为 0,Q20 为 98.47%,Q30 为 95.28%,GC 含量为 43.29%。说明余甘子转录组测序获得序列的数量、质量和精确性均较高,为下一步 de novo 组装提供了较好的原始序列数据。

通过组装共获得 97 628 条转录本,平均长度、GC 含量分别为 648 nt 和 39.69%,N50 长度较长,为1 116 nt,表明组装质量较好。进一步聚类去冗余得到 76 881条 Unigenes,平均长度、GC 含量及 N50 长度分别为 713 nt、39.74% 和 1 257 nt。对组装的转录本长度进一步分析可知(图 1A):22.34% 的转录本长度在 1 000 nt 以上,仅 1.79% 的转录本≥3 000 nt。对聚类去冗余后 Unigenes 的长度分布进行统计(图 1B),有 23.57% 的序列长度大于 1 000 nt,2.10% 的转录本序列大于 3 000 nt。由此可以看出:经 Trinity 软件组装后 Unigenes 的平均长度及 N50 长度有所增加,1 000 nt 及以上序列所占比例有所提高,表明组装及聚类去冗余效果较好,可进行后续分析。

2.2　余甘子转录组 Unigene 的 NR 功能注释

将 76 881 条 Unigenes 比对到 NR 等四大蛋白数据库(*E*-value≤1e-5),共有 44 768 条 Unigenes 获得注释,注释率为 58.23%。其中,在 NR 数据库中获得注释的 Unigenes 数量最多,为 44 270 条(57.58%),且与其它物种同源序列具有不同程度的匹配度。由图 2 可见:注释序列分布较多的 3 个物种分别是麻疯树(*Jatropha curcas* L.)、蓖麻(*Ricinus communis* L.)、胡杨(*Populus euphratica* Oliv.),分别占 NR 数据库所注释 Unigenes 总数的 14.09%、12.81%、10.56%,其余近一半的被注释 Unigenes 分布于其它 542 个物种中。由于缺乏余甘子基因组和

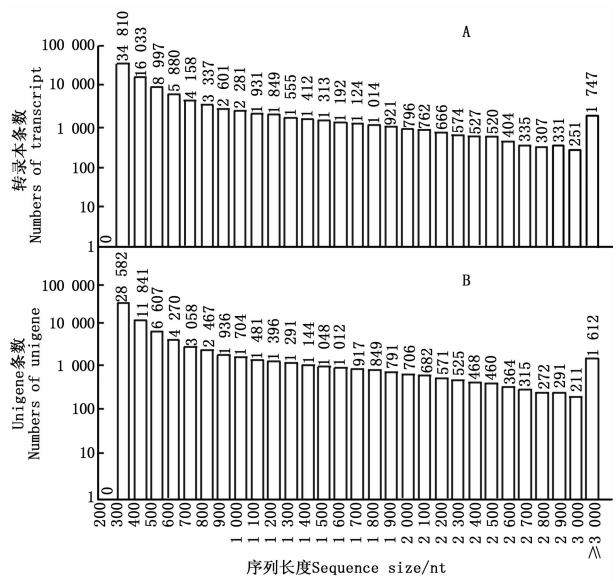


图 1　余甘子转录本和 Unigenes 的长度分布图

Fig. 1　Transcripts and Unigenes length distribution for P. emblica

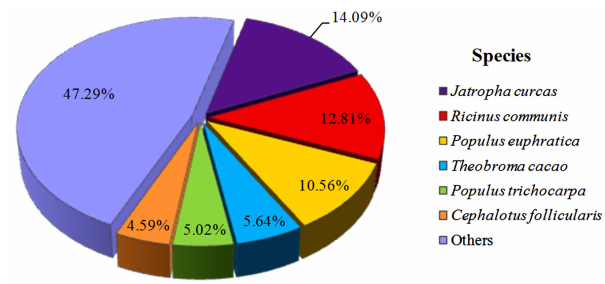


图 2　余甘子转录组 Unigenes 的 NR 注释物种分布

Fig. 2　NR annotated species distribution of Unigenes of transcriptome in *P. emblica*

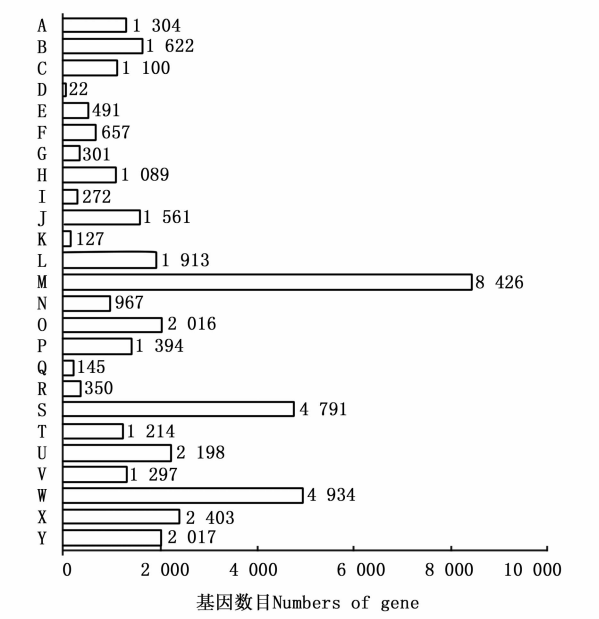
转录组信息,仍有一部分 Unigenes 未能在 NR 库中匹配到同源序列。

2.3　余甘子转录组 Unigene 的 COG 注释及其分类

COG 数据库是识别直系同源基因并对基因产物进行同源分类的数据库。共有 27 008 条 Unigenes(35.13%)分别被注释到 25 个 COG 功能分类中,共包含 42 611 个功能注释信息,基本涉及到余甘子大部分的生命活动(图 3)。各类功能的基因表达丰度各不相同,以一般功能基因为最大类别,占 COG 数据库总功能注释信息的 19.77%;其次是信号转导机制(11.58%)与翻译后修饰、蛋白质折叠和分子伴侣(11.24%);而细胞能动性基因较少,仅有 22(0.05%)个。

2.4　余甘子转录组 Unigene 的 GO 注释及其分类

根据 NR 注释信息对余甘子所有 Unigenes 进行 GO 功能分类统计,从宏观上认识该物种叶片中相关



注：A：氨基酸运输及代谢；B：辅糖类运输及代谢；C：细胞周期调控，细胞分裂，染色体分配；D：细胞运动；E：细胞壁/细胞膜/囊膜生源；F：染色质结构与动力；G：辅酶运输及代谢；H：细胞骨架；I：防御机制；J：能量生成与转化；K：胞外结构；L：未知功能；M：一般功能(预测)；N：无机离子的运输及代谢；O：胞内运输，分泌及小泡运输；P：脂类运输及代谢；Q：细胞核结构；R：核苷酸运输及代谢；S：翻译后修饰，蛋白质转换与分子伴侣；T：复制，重组与修复；U：RNA 加工与修饰；V：次生代谢产物生物合成，运输及代谢；W：信号转导机理；X：转录；Y：翻译，核糖体结构和生物合成。

Note：A：Amino acid transport and metabolism；B：Carbohydrate transport and metabolism；C：Cell cycle control，cell division，chromosome partitioning；D：Cell motility；E：Cell wall/membrane/envelope biogenesis；F：Chromatin structure and dynamics；G：Coenzyme transport and metabolism；H：Cytoskeleton；I：Defense mechanisms；J：Energy production and conversion；K：Extracellular structures；L：Function unknown；M：General function prediction only；N：Inorganic ion transport and metabolism；O：Intracellular trafficking，secretion，and vesicular transport；P：Lipid transport and metabolism；Q：Nuclear structure；R：Nucleotide transport and metabolism；S：Posttranslational modification，protein turnover，chaperones；T：Replication，recombination and repair；U：RNA processing and modification；V：Secondary metabolites biosynthesis，transport and catabolism；W：Signal transduction mechanisms；X：Transcription；Y：Translation，ribosomal structure and biogenesis.

图3 余甘子转录组 Unigenes 的 COG 功能注释分布统计图

Fig. 3 COG functional annotation distribution of Unigenes of transcriptome in *P. emblica*

基因功能分布特征。分析图 4 可知：有 19 749(25.69%)条 Unigenes 得到 97 102 个 GO 注释信息，平均每条 4.9 个；近 3/4 Unigenes 在 GO 库中未匹配到同源序列，进一步说明余甘子基因信息不足。

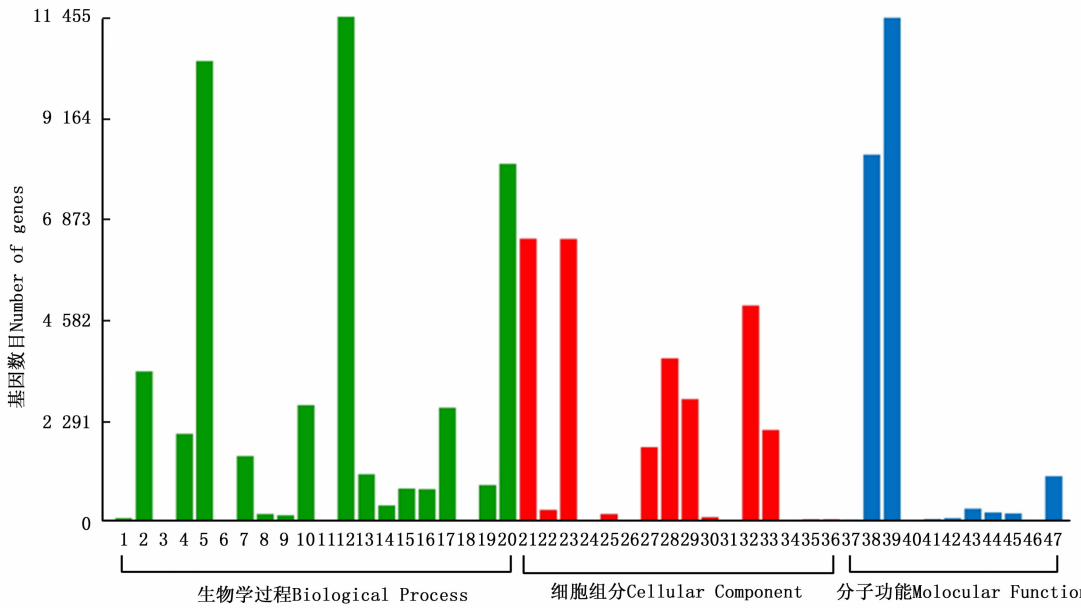
获得的 GO 注释可划分为生物学过程、细胞组分和分子功能三大类，其中执行生物学过程的注释最多(占 47.91%)，涉及分子功能的最少(占 22.45%)。这三大功能分类又被细分为 47 个功能亚类。参与生物学过程的46 519 个 GO 条目被划分为20 个亚类，其中，以代谢过程最多，占该类型的 24.62%。构成细胞组分的 28 786 个 GO 条目被划分为 16 个亚类，以细胞、细胞部分以及细胞器较多，分别占该类型的 22.32%、22.31% 和 17.06%。分子功能包含的 11 个亚类中，以催化活性和结合居多，分别占该类别的 52.48% 和 38.23%。总体来说，余甘子在细胞活动、代谢活动的基因表达丰度较高，表明余甘子自身具有较强的代谢能力。

2.5 余甘子转录组 Unigene 的 KEGG 代谢通路分析

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功能的数据库，利用 KEGG 可以进一步研究基因在生物学上的复杂行为。余甘子 KEGG 比对结果表明，获得注释的 18 175(23.64%)条 Unigenes 参与的代谢通路可归为 6 大类别、21 个亚类。6 大类代谢通路中，代谢和遗传信息处理相关的通路所占比例最大，分别为 58.42% 和26.66%；比例最小的是人类疾病相关的通路，仅占 1.26%；细胞过程、环境信息处理以及生物系统相关的通路分别占 5.65%、4.21% 和 3.82%。进一步细分为 21 个亚类，其中，代谢包含的 11 个亚类中，碳水化合物代谢所占比例最高，为 13.79%，而糖类生物合成与代谢所占比例最低，为 1.71%。遗传信息处理包含 4 个亚类，以翻译(10.89%)与折叠、分类和降解(8.41%)为主要代谢途径，而复制和修复(3.71%)以及转录(3.65%)在此时期代谢较弱。总的来说，该时期余甘子代谢活动和遗传信息处理能力较强。

2.6 余甘子转录组 Unigene 的 CDS 预测及与近缘模式生物同源基因 CDS 的比较

对 CDS 进行预测能为余甘子功能基因的挖掘利用和分子标记的开发提供重要参考。按照 NR、SwissProt、KEGG 和 COG 数据库的优先级顺序，通过 BLAST 比对共获得 42 953 条 Unigenes 的 CDS($E \leqslant$ 1e－5)，与以上蛋白数据库皆比对不上的 Unigenes 用软件 ESTScan[11] 预测后获得 2 058 个 CDS。从图 5A 可看出：通过 BLAST 获得的 CDS 序列长度较长，≥1 000 nt 的占 21.99%；而通过 ESTScan 预测后获得的 CDS 序列较短，主要集中在 200 ～ 600 nt(占 96.36%)(图 5B)。



注:1:生物粘附;2:生物调节;3:细胞杀伤;4:细胞组分的组织与生物合成;5:细胞过程;6:解毒作用;7:发育过程;8:生长;9:免疫系统过程;10:定位;11:移动;12:代谢过程;13:多细胞生物过程;14:多组织过程;15:繁殖;16:繁殖过程;17:应激反应;18:节律性过程;19:信号;20:单一的生物过程;21:细胞;22:细胞连接;23:细胞部分;24:胞外基质;25:胞外区;26:胞外区域部分;27:大分子复合物;28:膜;29:膜部分;30:膜封闭内腔;31:拟核;32:细胞器;33:细胞器部分;34:超分子纤维;35:病毒体;36:病毒体部分;37:抗氧化活性;38:结合;39:催化活性;40:电子载体活性;41:分子功能调节器;42:分子传感器活性;43:核苷酸结合转录因子活性;44:信号传感器活性;45:结构分子活性;46:蛋白质结合转录因子活性;47:转运载体活性。

Note:1:biological adhesion;2:biological regulation;3:cell killing;4:cellular component organization or biogenesis;5:cellular process;6:detoxification;7:developmental process;8:growth;9:immune system process;10:localization;11:locomotion;12:metabolic process;13:multicellular organismal process;14:multi - organism process;15:reproduction;16:reproductive process;17:response to stimulus;18:rhythmic process;19:signaling;20:single - organism process;21:cell;22:cell junction;23:cell part;24:extracellular matrix;25:extracellular region;26:extracellular region part;27:macromolecular complex;28:membrane;29:membrane part;30:membrane - enclosed lumen;31:nucleoid;32:organelle;33:organelle part;34:supramolecular fiber;35:virion;36:virion part;37:antioxidant activity;38:binding;39:catalytic activity;40:electron carrier activity;41:molecular function regulator;42:molecular transducer activity;43:nucleic acid binding transcription factor activity;44:signal transducer activity;45:structural molecule activity;46:protein binding transcription factor activity;47:transporter activity.

图 4　余甘子转录组 Unigenes 的 GO 功能分类统计图

Fig. 4　GO functional classification of Unigenes of transcriptome in *P. emblica*

对余甘子 Unigenes 的 CDS 长度与近缘模式生物麻疯树同源基因 CDS 长度的比较分析可知(图 6A):余甘子 45 011 条 Unigenes 的 CDS 中有一半以上(62.96%)可以映射到麻疯树基因组编码区,但仅有 25.22% Unigenes 的 CDS 与麻疯树同源基因 CDS 长度之比接近 1(0.9≤Ratio<1.1);覆盖深度≤300 时,20.89% Unigenes 的编码区覆盖接近 1(0.9≤Ratio<1.1);覆盖深度>300 时,仅有 4.33% Unigenes 的编码区覆盖接近 1。进一步分析可知(图 6B):有近 1/6 Unigenes 的 CDS 与近缘物种麻疯树同源基因 CDS 的覆盖度≥80%,仍有大部分未能覆盖到基因的完整编码区。

2.7　余甘子转录组 Unigene 的 TF 编码能力以及 R-Gene 的预测

通过 PlantTFDB 数据库共预测出 56 个 TF 家族共 1 374 个 TF 的 Unigenes(图 7),其中,bHLH、MYB-related、ERF、C2H2、NAC 以及 MYB 类 TF 为多基因家族,分别占总预测量的 8.01%、7.72%、7.21%、6.33%、5.82% 和 4.88%,而 SAP 和 HRT-like 所占比例较少,均为 0.073%。对这些 TF 家族特征进行统计分析可知:共有 203 条参与植物非生物胁迫应答反应的 TF,包括植物中特有的 ERF(99 条)、NAC(80 条)和 HSF(24 条);WRKY 超家族(66 条)和 bZIP 类 TF(60 条)则与植物生物胁迫的抗性密切相关。这些 TF 的发现表明余甘子在生长、发育以及适应外界环境过程中基因转录调控的复杂性,也为后续研究余甘子抵抗高温、干旱、盐碱以及病虫害等非生物、生物胁迫提供了新思路。

抗性基因(Resistance gene,R-Gene)分布在植物基因组中,其产物介导植物对特定病原体的抗性。

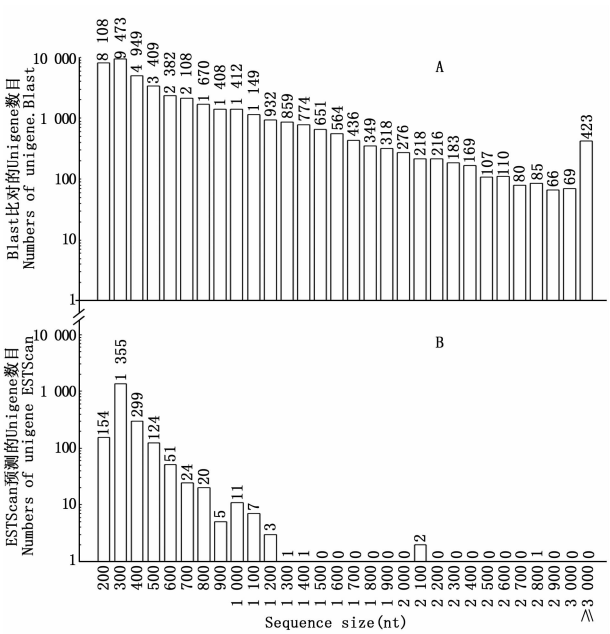
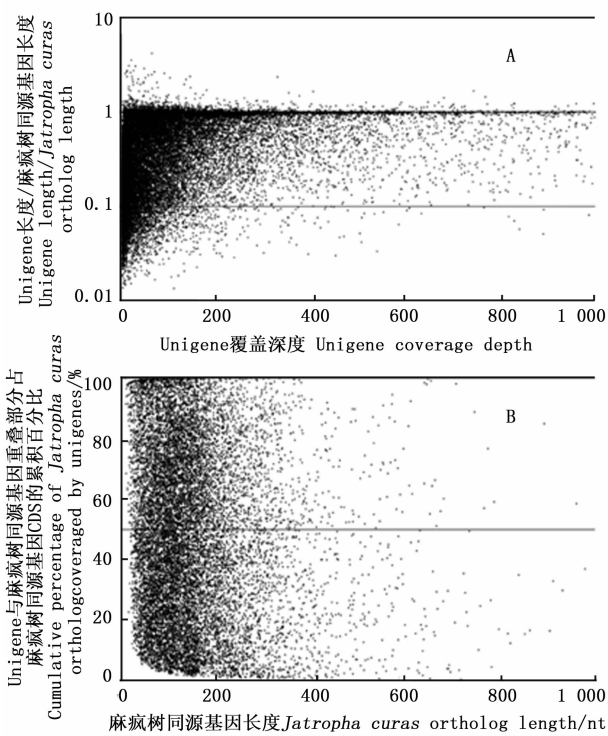


图5 余甘子转录组 Unigenes 的 CDS 长度分布图

Fig. 5 CDS length distribution of Unigenes of transcriptome in *P. emblica*



注:图中每个点代表1个基因。Note: Each point indicates one gene.

图6 余甘子 Unigenes 的 CDS 与近缘物种麻疯树同源基因 CDS 的比较分布图

Fig. 6 Comparison distribution of CDS between Unigenes of *P. emblica* and orthologous genes of Jatropha curcas

本研究预测出18种共3 806条R-Gene的Unigenes(图8),其中,RLP、N、NL及TNL较多,分别占总预

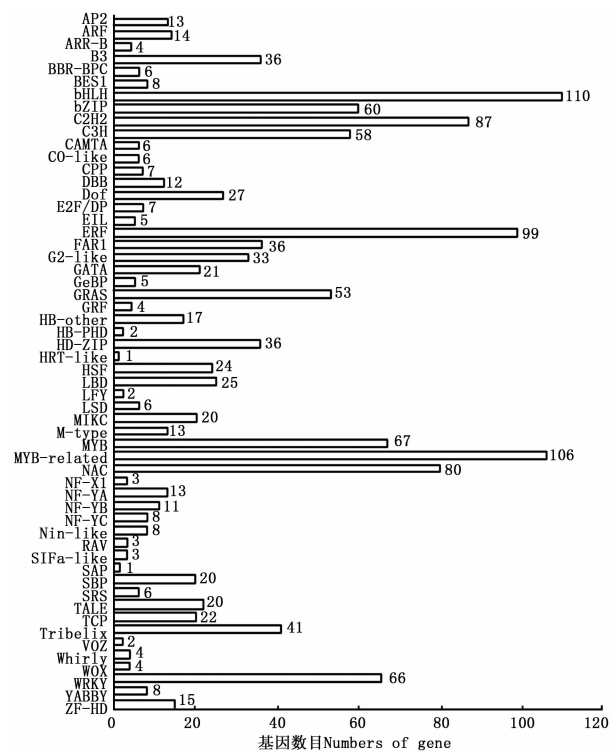


图7 余甘子转录组 Unigenes 的转录因子家族分类图

Fig. 7 Transcription Factor family classification of Unigenes of transcriptome in *P. emblica*

测量的29.51%、16.79%、13.77%和12.98%。这些R-Gene中大多数属于编码包含核苷酸结合位点(nucleotide-binding site,NBS)和/或亮氨酸重复结构(leucine-rich repeat,LRR)保守基序的抗病蛋白,如N基因、RLP基因以及RLK系列基因等[12-13]。

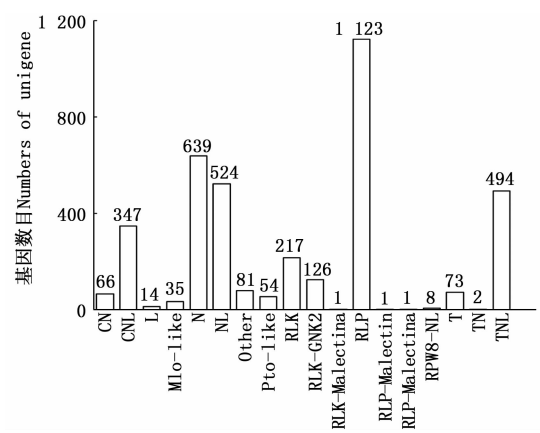


图8 余甘子转录组 Unigenes 的 R-Gene 分类图

Fig. 8 R-Gene classification of Unigenes of transcriptome in *P. emblica*

3 讨论

第二代高通量转录组测序技术(RNA-seq)因测

序用时短、成本低、数据量大而常被用于转录组学研究中。测序所获得的转录组数据不但能扩充某一物种的基因信息库,在揭示其特定代谢途径的分子机理中也具有重要作用。本研究通过 Illumina HiSeq 4000 平台对余甘子叶片进行转录组测序,共产生 10.95 Gb 的数据量,与苦味叶下珠(*Phyllanthus amarus* Schum. & Thonn.)叶片转录组相比[14],本研究获得的数据信息量较大。一般认为,当 Q30 值在 80% 以上时测序质量就很可靠[15],余甘子叶片转录组 Clean reads 碱基 Q20 和 Q30 分别为 98.47% 和 95.28%,测序质量和准确度比较高。经 de novo 组装获得 76 881 条 Unigenes,组装的 N50 值较大,为 1 257 nt,组装质量较好;所得 Unigenes 中长片段较多,且多数集中在 200 ~2 000 nt,近 1/4 的 Unigenes 长度在 1 000 nt 以上,组装序列比较长,为进一步的功能注释、分类、CDS 预测和 R-Gene 预测提供了较好的基础数据。

利用 NR、COG、KEGG 和 SwissProt 四大蛋白数据库,将获得的 76 881 条余甘子 Unigenes 序列进行比对分析,共有 44 768 条获得注释,占 Unigenes 总数的 58.23%,而且近一半 Unigenes 总体表达量 FPKM 值小于 1,表明本研究检测到余甘子低丰度表达的基因比例较高。四大数据库获得注释最多的 NR 数据库的 Unigenes 中,14.09% 的 Unigenes 注释为与麻疯树同源的基因,比例高于其它被注释的 544 个物种,可能是由于麻疯树已完成基因组测序且与余甘子在进化上亲缘关系较近的缘故[16];而仅有 33(0.074%)条 Unigenes 比对到 8 个同科物种(分属 5 个属)中,其中,与叶下珠属植物比对上的 Unigenes 最多,有 23(0.052%)条,这可能是现有叶下珠科植物基因组信息较为缺乏或是所得 Unigenes 多为余甘子特有的新基因而未能比对上相应的基因序列。总的未被注释的 Unigenes 有 32 113 条,占总体的 41.77%,这些序列可能是因为片段较短,或本身为非编码序列,也有可能是由于这四大蛋白数据库相关数据信息缺乏,或余甘子中存在新基因而未能比对到同源序列[17-19]。

通过 COG 数据库进行余甘子 Unigenes 直系同源基因功能注释及分类,获得了大量余甘子在一定时期表达的基因信息,27 008 条 Unigenes 共 42 611 个功能注释信息涉及到 25 类基因功能,包含余甘子大部分的生命活动,其中一般功能、信号转导机制、翻译后修饰、蛋白质折叠和分子伴侣的基因表达水平较高,而 4.5%(1 913 条)的 Unigenes 生物学功能未知,可能是由于注释信息不足,在很多树种转录组分析中也出现了类似的情况,如长梗杜鹃(*Rhododendron longipedicellatum* Lei Cai & Y. P. Ma)[18]、云南松(*Pinus yunnanensis* Franch.)[19]、锥栗(*Castanea henryi* (Skan) Rehd. et Wils.)[20]等。使用 GO 基因功能分类体系,将余甘子 Unigenes 参与的生物学过程、构成的细胞组分和执行的分子功能划分为 47 个功能亚类,以代谢过程、催化活性和细胞过程的基因表达丰度较高,表明余甘子在这一时期代谢能力较强。利用 KEGG 进一步研究余甘子 Unigenes 在生物学上的复杂行为,分析基因产物在细胞中可能的代谢途径,18 175 条 Unigenes 注释到 6 大代谢通路共 21 条代谢途径中,其中映射到代谢相关通路的基因数占多数,约为总注释量的 3/5,进一步说明余甘子在此时期具有较强的代谢活动。在余甘子 KEGG 代谢通路中还发现 203 条与人类疾病相关的 Unigenes,包括内分泌和代谢疾病(201 条)以及抗药性(2 条)。另外,还涉及到氨基酸代谢、碳水化合物代谢、脂类物质代谢、次生物质合成及环境适应等代谢途径,这些数据为探索余甘子抗性机理、挖掘余甘子特殊代谢途径相关基因以及与环境适应性相关基因等研究提供了基础数据。从总体的功能注释信息来看,余甘子基因含量丰富,在分子层面上解释了其具有较强适应性的可能原因。

TF 及其结构的多样性决定了其所介导的基因表达调控网络的多样性和复杂性,本研究预测出 56 个 TF 家族共 1 347 个编码 TF 的 Unigenes,这些 TF 在余甘子生长发育过程中发挥着重要的作用。与拟南芥、棉花、番茄以及杜鹃花等大多数植物不同[18,21],在余甘子中,数量最多的 TF 不是 MYB 家族而是 bHLH 家族,bHLH 家族是植物第二大类 TF,可能与余甘子的各种信号转导、合成代谢以及抗逆性等有关[22]。植物体免疫系统能够对抗诸如细菌、真菌、病毒、线虫和昆虫等各类病原体[23],本研究预测到 3 806 条编码 R-Gene 的 Unigenes,其中,NB-LRR 蛋白由植物中最大且最重要的基因家族之一编码,这些 NBS-LRR 蛋白可以直接或间接地识别病原体分泌的效应物,进而激活下游的信号通路,从而激活植物防御反应,对抗各种类型的病原体[13,24-25],余甘子 R-Gene 的预测及分析对了解植物抗病蛋白在植物抗病信号转导途径中的作用机制具有重要的生物学意义,并为余甘子抗病育种提供新的理论

指导。

4 结论

本研究使用 Illumina HiSeq 4000 测序平台构建了余甘子转录组数据库,获得的 76 881 条 Unigenes 与四大公共数据库进行比对,有 58.23% 的 Unigenes 获得功能注释信息,同时对其进行了 CDS 预测、TF 编码能力预测以及 R-Gene 预测等分析。研究结果揭示了余甘子丰富的基因信息,在转录组水平上支持了余甘子适应性强的特点,COG、GO、KEGG 的功能注释结果均表明余甘子在这一时期进行着较强的代谢活动。所得转录组信息丰富了余甘子基因信息库,为今后余甘子乃至同属物种微卫星标记开发、功能基因挖掘、抗性机理分析、遗传资源分类与进化以及分子辅助育种等研究奠定了重要基础。

参考文献:

[1] 金振洲, 欧晓昆. 元江、怒江、金沙江、澜沧江干热河谷植被[M]. 昆明: 云南大学出版社, 云南科技出版社, 2000.

[2] Zhou Z, Ma H, Lin K, *et al*. RNA-seq reveals complicated transcriptomic responses to drought stress in a nonmodel tropic plant, *Bombax ceiba* L.[J]. Evolutionary Bioinformatics, 2015, 11(S1): 27 – 37.

[3] 李巧明, 赵建立. 云南干热河谷地区余甘子居群的遗传多样性研究[J]. 生物多样性, 2007, 15(1): 84 – 91.

[4] Variya B C, Bakrania A K, Patel S S. *Emblica officinalis* (Amla): A review for its phytochemistry, ethnomedicinal uses and medicinal potentials with respect to molecular mechanisms [J]. Pharmacological Research, 2016, 111: 180 – 200.

[5] Chaphalkar R, Apte K G, Talekar Y, *et al*. Antioxidants of *Phyllanthus emblica* L. bark extract provide hepatoprotection against ethanol-induced hepatic damage: a comparison with Silymarin [J]. Oxidative Medicine and Cellular Longevity, 2017, 2017: 1 – 10.

[6] Srinivasan M. Vitamin C in plants: Indian Gooseberry (*Phyllanthus emblica*) [J]. Nature, 1944, 153(3892): 684.

[7] Alvarez M, Schrey A W, Richards C L. Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? [J]. Molecular Ecology, 2015, 24(4): 710 – 725.

[8] Kumar A, Singh K. Isolation of high quality RNA from *Phyllanthus emblica* and its evaluation by downstream applications [J]. Molecular Biotechnology, 2012, 52(3): 269 – 275.

[9] Grabherr M G, Haas B J, Yassour M, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome [J]. Nature Biotechnology, 2011, 29(7): 644 – 652.

[10] Fu L M, Niu B F, Zhu Z W, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data [J]. Bioinformatics, 2012, 28(23): 3150 – 3152.

[11] Iseli C, Jongeneel C V, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences [C]// Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. Menlo Park: AAAI Press, 1999: 138 – 158.

[12] 房卫平, 谢德意, 李志芳, 等. NBS-LRR 类抗病蛋白介导的植物抗病应答分子机制[J]. 分子植物育种, 2015, 13(2): 469 – 474.

[13] Chisholm S T, Coaker G, Day B, *et al*. Host-microbe interactions: shaping the evolution of the plant immune response [J]. Cell, 2006, 124(4): 803 – 814.

[14] Bose Mazumdar A, Chattopadhyay S. Sequencing, *de novo* assembly, functional annotation and analysis of *Phyllanthus amarus* leaf transcriptome using the Illumina platform [J]. Frontiers in Plant Science, 2016, 6(340): 1199.

[15] 贾新平, 孙晓波, 邓衍明, 等. 鸟巢蕨转录组高通量测序及分析[J]. 园艺学报, 2014, 41(11): 2329 – 2341.

[16] Sato S, Hirakawa H, Isobe S, *et al*. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. [J]. DNA Research, 2011, 18(1): 65 – 76.

[17] Bai T D, Xu L A, Xu M, *et al*. Characterization of masson pine (*Pinus massoniana* Lamb.) microsatellite DNA by 454 genome shotgun sequencing [J]. Tree Genetics & Genomes, 2014, 10(2): 429 – 437.

[18] 李太强, 刘雄芳, 万友名, 等. 基于高通量测序的极小种群野生植物长梗杜鹃的转录组分析[J]. 植物研究, 2017, 37(6): 825 – 834.

[19] 蔡年辉, 邓丽丽, 许玉兰, 等. 基于高通量测序的云南松转录组分析[J]. 植物研究, 2016, 36(1): 75 – 83.

[20] 张 琳, 范晓明, 林 青, 等. 锥栗种仁转录组及淀粉和蔗糖代谢相关酶基因的表达分析[J]. 植物遗传资源学报, 2015, 16(3): 603 – 611.

[21] 牛义岭, 姜秀明, 许向阳. 植物转录因子 MYB 基因家族的研究进展[J]. 分子植物育种, 2016, 14(8): 2050 – 2059.

[22] 王 翠, 兰海燕. 植物 bHLH 转录因子在非生物胁迫中的功能研究进展[J]. 生命科学研究, 2016, 20(4): 358 – 364.

[23] Jones J D G, Dangl J L. The plant immune system [J]. Nature, 2006, 444: 323 – 329.

[24] Gururani M A, Venkatesh J, Upadhyaya C P, *et al*. Plant disease resistance genes: current status and future directions [J]. Physiological and Molecular Plant Pathology, 2012, 78(51): 51 – 65.

[25] Dubey N, Singh K. Role of NBS-LRR proteins in plant defense [M]// Singh A, Singh I. Molecular Aspects of Plant-Pathogen Interaction. Singapore: Springer Singapore, 2018: 115 – 138.

(责任编辑:张 玲)