文章编号:1001-1498(2016)04-0500-08

蕨类植物芒萁幼孢子体转录组高通量测序及特征分析

刘丽婷[1,2], 温 强[2], 黄小春[2], 刘琪璟[1*]

(1. 北京林业大学 北京 100083; 2. 江西省林业科学院省植物生物技术重点实验室 江西 南昌 330013)

摘要:[**目的**]采用高通量测序技术 Illumina MiSeq250 获得蕨类植物芒萁的孢子体转录组数据,以期为芒萁的生长、发育、代谢调控、微进化机制分析等提供重要的分子信息。[**方法**]应用生物信息学方法对测序获得的大量单基因簇(Unigene)进行基因功能注释、代谢途径及微卫星分析等。[**结果**]本研究共获得 18 463 296 条序列读取片段(reads),总碱基数为 4.62Gbp 序列信息,经序列组装最终得到 63 169 个 Unigene,平均单条 Unigene 长度为 863 bp,N50 为 1 587 bp,其中分布在 200～500 bp 长度区间 Unigene 占总数的 55.4%。数据库中的序列同源性比较表明,26 826 个 Unigene 与其他物种的已知基因具有不同程度的同源性。芒萁转录组中的 Unigene 根据 GO 功能大致可分为细胞组成、分子功能和生物学过程 3 大类 47 个分支,其中有大量的 Unigene 与细胞进程、绑定活性、代谢过程和催化活性相关。将 Unigene 与 COG 数据库进行比对,根据其功能大致可分为 26 类。以 KEGG 数据库作为参考,依据代谢途径可将 Unigene 定位到 257 个代谢途径分支。此外,利用 MISA 软件检索 2～6 碱基微卫星,共找到 13 286 个 SSR。在不同长度微卫星中,三核苷重复数量最多,占总数的 40.41%。在各重复基序类型中出现频率最高的为 AG/CT(14.45%)与 AAG/CTT(12.39%)。利用重复基序开发的多态性 SSR 标记,可应用于芒萁不同个体的基因型分型鉴定。[**结论**]本研究获得了较高质量的芒萁转录组数据库,揭示了芒萁孢子体生长发育过程中表达基因的功能总体特征,可为芒萁进一步的功能基因挖掘和分子标记规模化开发奠定基础。

关键词:芒萁;转录组;功能注释;微卫星

中图分类号:S718.46 **文献标识码**:A

De novo Sequencing and Characterization of Juvenile Sporophyte Transcriptome of a Fern, *Dicranopteris dichotoma*

LIU Li-ting[1,2], *WEN Qiang*[2], *HUANG Xiao-chun*[2], *LIU Qi-jing*[1]

(1. Beijing Forestry University, Beijing 100083, China; 2. Jiangxi Provincial Biotech Key Lab for Plant, Jiangxi Academy of Forestry, Nanchang 330032, Jiangxi, China)

Abstract:[**Objective**] The sporophyte transcriptome of *Dicranopteris dichotoma* was sequenced by Illumina MiSeq 250 to provide molecular information of its growth, development, metabolism, and the micro evolutionary mechanism. [**Method**] The functional annotations, metabolic pathways and microsatellite analysis of some Unigenes were conducted using bioinformatics methods. [**Result**] A total of 18 463 296 reads containing 4.62 Gbp of sequence information were generated. A total of 63 169 unigenes were formed by initial sequence splicing, with an average read length of 863 bp and N50 value of 1 587 bp. 26 826 unigenes were annotated using BLASTX searches against the Nr, Nt and SwissProt databases. The unigenes of the transcriptome of *D. dichotoma* were roughly divided into cellular components, molecule function and biological processes categories of 47 branches by gene ontology, of which re-

收稿日期:2015-11-25

基金项目: 国家林业局 948 项目(213-4-62);江西省重大财政专项青年人才培养计划(芒萁根系微生物环境与成片发育研究)。

作者简介: 刘丽婷,博士研究生。主要从事森林生态与植物资源利用研究。Email:39191393@qq.com

* 通讯作者:刘琪璟

lated with cellular process cell, binding, metabolism processes and catalytic activities. Further annotated based on COG category, Unigenes could be grouped into 26 functional categories. KEGG pathway analysis showed that Unigenes could be divided into 276 classes based on their metabolic function. Meanwhile, 13 286 SSRs (simple sequence repeats) were mined with repeat motif of 2 to 6 bp by MISA. The trinucleotide repeats were most dominant, accounting for a total of 40.41%. AG/CT (14.45%) and AAG/CTT (12.39%) were the most common repeat motifs. Polymorphic SSR markers were developed from repeat motifs, which could be used for genotyping of different individuals of *D. dichotoma*. [**Conclusion**] A higher quality of transcriptome database was obtained in this study, which could reveal the general characteristics of gene expression in the process of growth and development, and lay the foundation for further gene function mining and the large-scale development of molecular markers of *D. dichotoma*.

Keywords: *Dicranopteris dichotoma*; transcriptome; gene annotation; simple sequence repeat

芒萁(*Dicranopteris dichotoma* (Thunb.) Bernh.)属水龙骨目(Polypodiales)里白科(Gleicheniaceae)芒萁属(*Dicranopteris*),是典型酸性土壤指示植物,也是亚热带丘陵山区马尾松(*Pinus massoniana* Lamb.)、杉木(*Cunninghamia lanceolat* (Lamb.) Hook.)暖性针叶林、疏灌草丛等次生植被的"识别种"及"标志种"[1]。芒萁除孢子繁殖外兼营克隆繁殖[2],其孢子体(2n = 78)[3]匍匐根茎发达,在林冠下层易形成稳定的片层结构[4]。作为丘陵红壤林区立地破坏后最先侵入的下层植被种类之一[5],芒萁具有重要的水土保持与植被恢复作用[6-7]。此外,芒萁孢子体对砷(As)[8]、铅(Pb)[9]、稀土[10-11]等重金属有较强的吸收富集作用,是典型的金属型植物(metallophytes)[12-13],已成为矿区废弃地植被重建的先锋植物。研究芒萁在困难立地种群扩散过程及对逆境应答机制,对于发挥其生态价值具有重要的现实意义。

、近年来,新一代高通量转录组测序被广泛应用于非模式植物[14-15],可以高通量地测定 cDNA 序列,揭示特定细胞或组织中表达的全部基因或表达序列标签(Expressed sequence tag, EST),获得大量 SSR 等遗传标记等。该技术已成为揭示植物优良特性及研究其环境互作等复杂分子机制的重要手段。蕨类植物由于拥有特殊的系统进化位置及独特的生活史,是研究陆生植物系统演化的代表性物种[16]。然而蕨类植物与种子植物相比具更复杂的染色体组成及较大的基因组[17-18],使得该类植物遗传信息资源非常有限,局限了其分子生物学研究[19]。早期有报道开展诸如江南卷柏(*Selaginella moellendorffii* Herb.)[20]、铁线蕨(*Adiantum capillus-veneris* Linn.)[21]等的转录组文库构建研究;而基于高通量测序的转录组学研究较少,仅见蕨(*Pteridium aquilinum* (Linn.) Kuhn.)[22]、水蕨(*Ceratopteris richardii* Linn.)[23]、鸟巢蕨(*Asplenium nidus* Linn.)[24]及海金沙(*Lygodium japonicum* (Thunb.) Sw.)[19]有研究报道。

目前有关芒萁的转录组学研究未见报道,该物种分子标记开发及抗逆机理等相关研究相对滞后。本研究旨在应用 Illumina Miseq250 高通量测序技术开展芒萁孢子体转录组学研究,采用生物信息学等方法对获得的大量 Unigene 进行基因功能注释、代谢途径分析等,从功能基因组水平上分析芒萁孢子体生长发育过程中重要基因的表达水平,为进一步功能基因挖掘和分子标记开发奠定基础。

1　方法

1.1　试验材料

采集当年新萌的芒萁孢子体幼叶,经液氮速冻后于 -70℃储存备用于 RNA 提取。用于检测 SSR 标记多态性的芒萁群体样本来自江西泰和县千烟洲(115°0.527′ E,25°22.445′ N),参考改进的 CTAB 高盐法[25]提取基因组总 DNA。

1.2　转录组测序与序列组装

RNA 提取试剂盒(TIANGEN)提取总 RNA。采用带有 Oligo(dT)的磁珠富集 mRNA,并将其随机打断成短片段作为模板,六碱基随机引物合成一链 cDNA,随后在 DNA polymerase Ⅰ作用下合成二链 cDNA。双链 cDNA 经纯化、加 poly(A)及连接测序接头后进行 PCR 扩增,得到测序用 cDNA 文库。采用 Illumina MiSeq 测序平台,利用双末端测序(Paired-end,PE)的方法,PE250 的测序策略进行高通量测序。测序得到的原始序列去除其中的接头及低质量序列,经 Trinity 软件拼接组装成一个转录组,同时取每条基因中最长的转录本(Transcripts)作为

单基因簇(Unigene)[26]。

1.3 Unigene 功能注释、GO 分类和代谢通路分析

将拼接得到的 Unigene 序列与 NR(NCBI non-redundant protein sequences),NT(NCBI nucleotide sequences),SwissProt(SwissProt protein database),COG(Cluster of orthologous groups)数据库进行 BLAST 比对获得注释(其中 NR、NT、SwissProt 数据库比对 E 值≤1e-5,COG 比对 E 值≤1e-3);通过 HMMER3 程序,搜索已建好的蛋白结构域的 HMM 模型,对 Unigene 进行蛋白家族(Protein family,Pfam)注释;另据 NR 和 Pfam 两部分蛋白注释结果,使用 Blast2GO 软件得到 Unigene 的 GO(Gene Ontology)条目,并用 WEGO 软件对所有的 Unigene 进行 GO 功能分类统计,最后进行 KEGG(Kyoto encyclopedia of genes and genomes)数据库代谢路径 KO(KEGG ORTHOLOG)注释分析。若前述各数据库之间的比对结果有出入,则按 NR、Swiss-Prot 的优先级确定 Unigene 的序列方向,比对不上的 Unigenes 则用软件 ESTScan 预测其编码区并确定序列方向。

1.4 微卫星分析与 SSR 标记应用

MISA 软件检索 Unigene 序列中的简单重复序列(Simple sequence repeats,SSR)。检索标准:单、二、三、四、五、六核苷酸基序(motif)至少重复次数分别为 10、6、4、3、3、3,包括精确型(perfect)及复合型(compound)重复基序(motif)[27],进而对微卫星基序开展统计分析。

随机选择微卫星重复基序长度大于等于 18 bp 的 Unigene 序列,利用 PRIMER3.0 软件进行 SSR 引物批量设计。本试验随机合成引物 20 对,编号 Dd_eSSR1-20。经优化确定芒萁孢子体 SSR 标记体系:10 μL 中含 10×PCR 缓冲液 1 μL,Mg^{2+} 2.5 mmol·L^{-1},dNTPs200 μmol·L^{-1},上下游引物各 0.2 μmol·L^{-1},Taq 聚合酶 0.5U,DNA30 ng 左右。PCR 反应程序:94℃预变性 3 min;94℃30 s,55℃30 s,72℃30 s,30 个循环;最后 72℃延伸 1 min,8℃保存。供试样本 PCR 产物采用 8% 聚丙烯酰胺凝胶,在 DYCZ-32 型垂直电泳槽(北京六一)中进行电泳分离,50bpMarker 作为标准分子量。银染检测电泳结果,同位点条带从大到小顺序以 A、B、C…编号,按照等位基因型进行判读。

2 结果与分析

2.1 转录组测序产出与基因表达分析

测序获得序列经过滤得到总的片段数(clean reads)为 9 231 648 条,总碱基数为 2.31 Gbp,GC 含量平均值为 47.76%。序列质量评估,碱基 Q20 为 98.24%,Q30 为 97.55%。原始数据经 Trinity 拼接后,共获得 110 051 个转录本,最短转录本长度为 201 bp,平均单条转录本长度为 1 238 bp,N50 为 2 182 bp。转录本经取舍获得 63 169 条 Unigene 序列,最短 Unigene 长度与转录本一致,平均单条 Unigene 长度为 863 bp,N50 为 1 587 bp。Unigene 序列在 200～500 bp 长度区间的数量占总数的 55.4%(34 982),组装效果符合 PE250 测序特点。利用 Blast 搜索预测了 23 064 个 CDS,其中 81.3%(18 740)的序列长度大于 300 nt,而长度大于 1 000 nt序列占 36.0%(8 301);其他未能用 Blast 比对上的 Unigene 序列采用 ESTScan 预测了 37 778个 CDS,其中 40.5%(15 308)的序列长度大于 300 nt。数据总体表明测序质量符合后续分析要求。

采用 FPKM(expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced)算法[28]估算芒萁 Unigene 的表达水平。该算法可消除基因长度差异和测序深度对基因表达水平估计的影响。芒萁 Unigene 的 RPKM 平均值为 17.42,最大值为 31 231.04;230 条 Unigene 的 FPKM 值大于 500,其中有 142 条 Unigene 序列在后续 NR 数据库中得到功能注释。各基因的 FPKM 表达量值集中在 3.16～36.12,显示本次测序低水平表达基因检测数量较多。

2.2 序列功能注释与功能分类

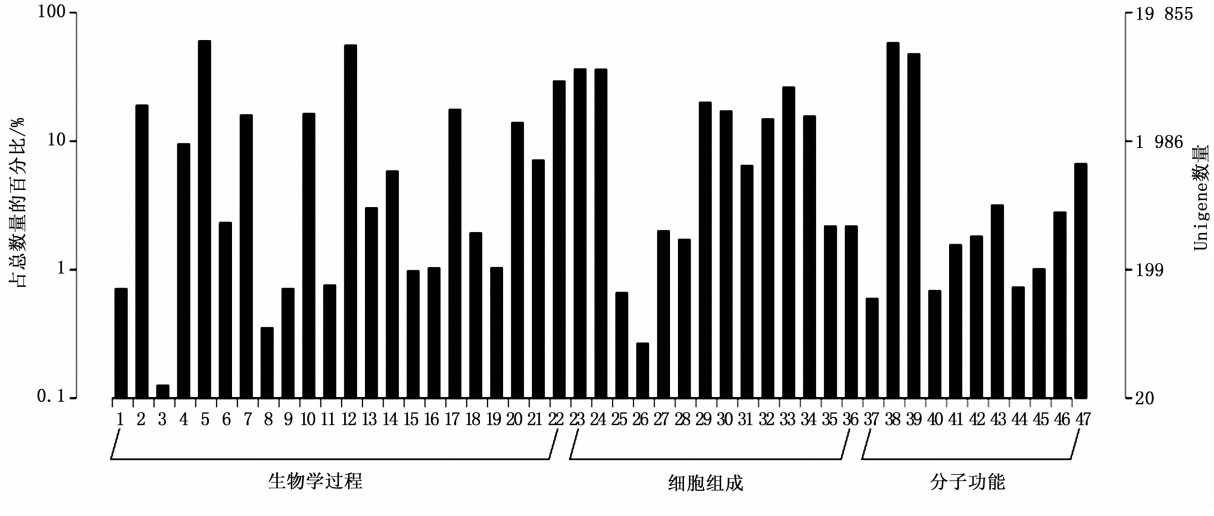
2.2.1 *功能注释序列比较* 将 Unigenes 序列与 Nr,Nt,SwissProt,Pfam,GO,COG,KEGG 数据库做比较,获得 Unigenes 的注释信息。统计最终获得注释信息的 Unigenes 序列共有 26 826 条,注释率为 42.46%,在各数据库中获得注释序列数量见表 1。经 NR 数据库比对,芒萁孢子体 Unigenes 序列与苔藓植物小立碗藓(*Physcomitrella patens*(Hedw.)Bruch & Schimp.)及同为蕨类植物的江南卷柏的 Unigenes 序列匹配相似数量最多,各占被注释总序列的 18.5% 与 18.4%,此外依次与北美云杉(*Picea sitchensis*(Bong.) Carr.)(17.0%)、葡萄(*Vitis vinifera* Linn.)(10.5%)及大豆(*Glycine max*(Linn.) Merr.)(3.8%)也能匹配到一定数量的相似序列。同时由于缺乏芒萁基因组信息,尚存一定数量 Unigenes 序列未能获得匹配。

表 1　Unigene 序列在各数据库的功能注释情况

数据库类型	被注释 Unigene 数量	占总数量的百分比/%
NR	22 897	36.24
NT	3 427	5.42
KO	6 240	9.87
SwissProt	18 016	28.52
PFAM	18 685	29.53
GO	19 855	31.43
COG	8 907	14.1
在 7 个数据库中均有注释数量	1 484	2.34
总的被注释 Unigene 数量	26 826	42.46

2.2.2　*GO 数据库注释分类*　本研究结合 GO 数据库,从宏观上对芒萁的 Unigene 进行功能分类,以了解其孢子体生长发育过程中表达基因的功能分布总体特征。19 855 个 Unigenes 可分成 3 个基本的功能本体,并区划成 47 个组别(图 1)。其中执行生物学过程功能 Unigenes 最多有 52 242 条,执行细胞组成有 36 032 条 Unigenes,涉及分子功能的 Unigenes 有 24 770 条。从各功能区划来看,各功能本体中 Unigenes 数量功能区划规律基本与鸟巢蕨孢子体发育进程的基因表达谱[24]一致。其中测得的芒萁孢子体 Unigenes 涉及较多的几个功能组为细胞过程(11 945条)、代谢过程(11 044 条)、细胞(7 196 条)、细胞要素(7 159 条)、绑定(11 538 条)及催化活性(9 457 条)等。

植物重金属硫结合蛋白(MT)关乎植物体重金属离子维持与毒害解除及调节运输等[29],作为典型的金属型植物,本研究重点关注了芒萁 MT 蛋白相关序列注释情况。本研究中共获得 23 个 GO 功能注释跟重金属结合蛋白密切相关的 Unigene 序列,其功能预测显示主要集中为铜、锌、镉等结合蛋白。在这 23 条有功能注释的 Unigene 序列中 FPKM 值大于 500 的有 2 条:comp28891_c0(GO:0046872//GO:0003950)、comp9834 _ c0 (GO:0008270//GO:0046872),其中 comp9834_c0 序列 FPKM 值最大为 6 424.29,GO 功能预测为锌结合蛋白,来自植物 PEC 金属硫蛋白家族。



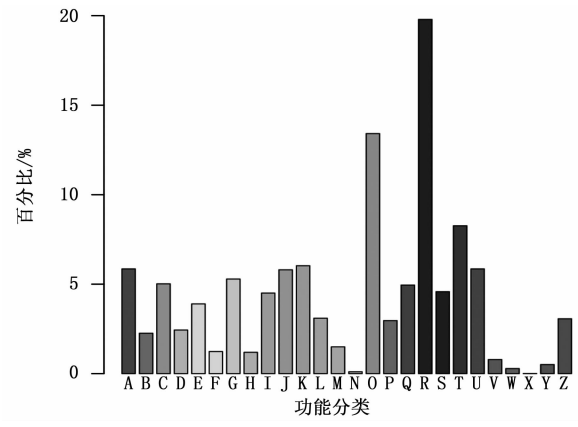
1. 生物附着;2. 生物调节;3. 细胞杀伤;4. 细胞组成或生物合成;5. 细胞过程;6. 发育过程;7. 定位活性;8. 生长;9. 免疫系统;10. 定位;11. 运动;12. 代谢过程;13. 多细胞进程;14. 多个有机体过程;15. 负调节;16. 正调控;17. 生物调节;18. 繁殖;19. 繁殖过程;20. 应激反应;21. 信号传导;22. 单一有机体过程;23. 细胞;24. 细胞要素;25. 细胞外基质;26. 细胞外基质要素;27. 胞外区;28. 胞外区要素;29. 大分子复合物;30. 膜;31. 膜封闭腔;32. 膜要素;33. 细胞器;34. 细胞器要素;35. 病毒体;36. 病毒体要素;37. 抗氧化活性;38. 绑定;39. 催化活性;40. 通道调节活性;41. 酶调节活性;42. 分子转导活性;43. 核酸结合转录因子活性;44. 蛋白结合转录因子活性;45. 受体活性;46. 结构分子活性;47. 转运活性。

图 1　GO 注释分类图

2.2.3　*COG 数据库注释分析*　将芒萁 Unigene 与 COG 蛋白质直系同源数据库进行比对,预测 Unigene 的功能并进行分类统计。研究结果见图 2,数据显示芒萁 Unigene 所涉及的 COG 功能类别较为全面,可将 10 035 个 Unigene 根据其功能大致分为 26 类。对每一类的 Unigene 进行统计分析显示,仅一般功能预测类基因最多(1 763 条),其次是翻译后修饰,蛋白折叠和分子伴侣类基因(1 195 条)、信号传导机制类基因(736 条)和转录类基因(537 条);而胞外结构类基因(25 条)和细胞运动类基因较少(9

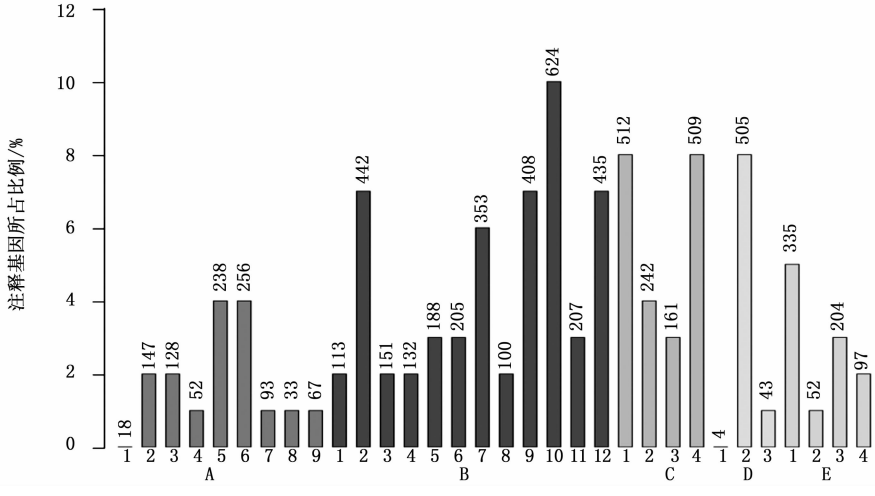
条),另外还存在 1 条未知蛋白基因;其他类别的基因表达丰度不尽相同。从总的 COG 功能类别来看本研究的 Unigene 基本涉及到芒萁大多数的生命活动。

2.2.4 生物学代谢 KEGG 分析 据 KEGG 数据库的注释信息进一步将芒萁 Unigene 进行 pathway 注释,其中 6 240 条 Unigene 获得对应的 KO 编号,这些 Unigene 参与或涉及相关代谢途径。5 个代谢通路大类中,主要包括碳水化合物代谢、翻译、信号转导、蛋白折叠,分类和降解、脂类物质代谢、运输和代谢等 32 类代谢途径(图 3)。32 类代谢途径下具体又可分为 257 个代谢分支,各分支中 Unigene 被注释到数量相对较多的路径(表 2)依次为碳代谢(248 条)、氨基酸生物合成(210 条)、内质网中蛋白质加工(192 条)等。研究表明重金属污染区域的芒萁自身可以通过控制光合活性来避免重金属离子对其光合系统的影响[11],而碳代谢是植物光合作用的重要内容,本研究中获得注释的 Unigene 数量可为今后开展芒萁相关研究提供序列基础。



A:RNA 的加工与修饰;B:染色体的结构域动力学;C:能源产生与转化;D: 细胞周期调控,细胞分裂,染色体分离 ;E:氨基酸转运与代谢;F: 核酸转运与代谢;G:碳水化合物转运与代谢;H: 辅酶转运与代谢;I:脂类转运与代谢;J: 翻译,核糖体结构和生物合成;K:转录;L:复制,重组和修饰;M: 细胞壁/细胞膜生物发生;N:细胞运动;O:翻译后修饰,蛋白折叠和分子伴侣;P:无机离子转运与代谢;Q:次生代谢物的生物合成,转运和代谢;R:仅一般功能预测;S:未知功能;T:信号传导机制;U:细胞内分泌和囊泡运输;V:防御机制;W:胞外结构;X:未知蛋白;Y:核结构;Z:细胞骨架

图 2 COG 注释分类图



A:有机系统;1. 感觉系统;2. 神经系统;3. 免疫系统;4. 排泄系统;5. 适应环境;6. 内分泌系统;7. 消化系统;8. 发展 9. 循环系统;B:代谢:1. 外来物质的降解和代谢;2. 总代谢;3. 核苷酸代谢;4. 萜类和酮类化合物;5. 其他氨基酸代谢;6. 代谢辅助因子和维生素;7. 脂质代谢;8. 糖链的生物合成与代谢;9. 能量代谢;10. 碳水化合物代谢;11. 其他次生代谢产物的生物合成; 12. 氨基酸代谢;C:遗传信息处理:1. 翻译;2. 转录;3. 复制和修复;4. 折叠,分类和降解;D:环境信息处理代谢:1. 信号分子的相互作用;2. 信号转导;3. 膜转运;E:细胞过程:1. 运输和代谢;2. 细胞运动;3. 细胞生长和死亡;4. 细胞通讯。

图 3 芒萁 Unigene 的 KEGG 分类

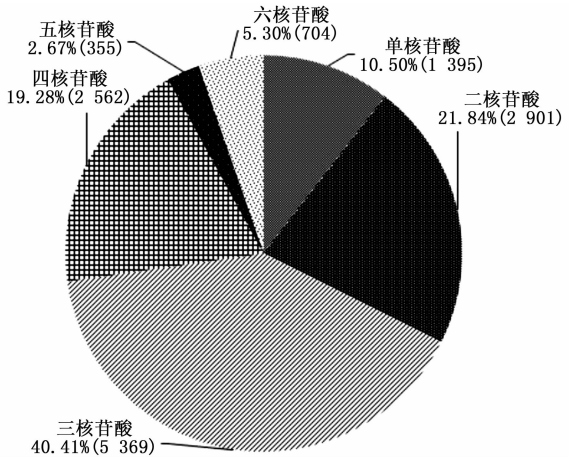
2.3 微卫星信息分析及 EST-SSR 有效性

在被检索的 63 169 个 Unigene 序列中,检测含有微卫星重复基序的序列总数为 10 120 个,包含微卫星总量为 13 286 个,其中复合型重复基序共 1 337 个,总的微卫星重复基序发生频率为 0.21,微卫星序列平均跨度为 4 100.97 bp。在被检索到的微卫

表 2　注释 Unigene 数量最多的 10 个代谢通路

编号	代谢通路	基因数目
Ko01200	碳代谢	248
Ko01230	氨基酸生物合成	210
Ko04141	内质网中的蛋白质加工	192
Ko03040	剪切体	187
Ko03010	核糖体	174
Ko03013	RNA 转运	162
Ko04075	植物激素信号转导	161
Ko00500	淀粉和糖代谢	151
Ko04626	植物与病原菌互作	144
Ko00190	氧化磷酸化	143

星基序中,单核苷酸到六核苷酸重复类型均存在。各类型重复基序分布数量及比例见图 4。在各重复基序中,以二、三短重复基序为主导,其中三核苷酸重复基序最多,占总数的 40.41%,而长的重复基序数量则相对较少,其中五核苷酸重复基序最少,仅占总数的 2.67%。在所有检测到的 290 种重复基序中,1 ～6 核苷酸重复基序中出现频率最高类型依次为,A/T(1 013 个,7.62%),AG/CT(1 920 个,14.45%),AAG/CTT(1 646 个,12.39%),AAAC/GTTT(463 个,3.48%),AAAAC/GTTTT(23 个,0.17%),AAAAAC/GTTTTT(10 个,0.08%),其中 AG/CT 与 AAG/CTT 同时为所有被检测重复基序中出现频率最高的两种。



注:括号内数值为对应重复基序的总值

图 4　芒萁微卫星重复基序分布比例图

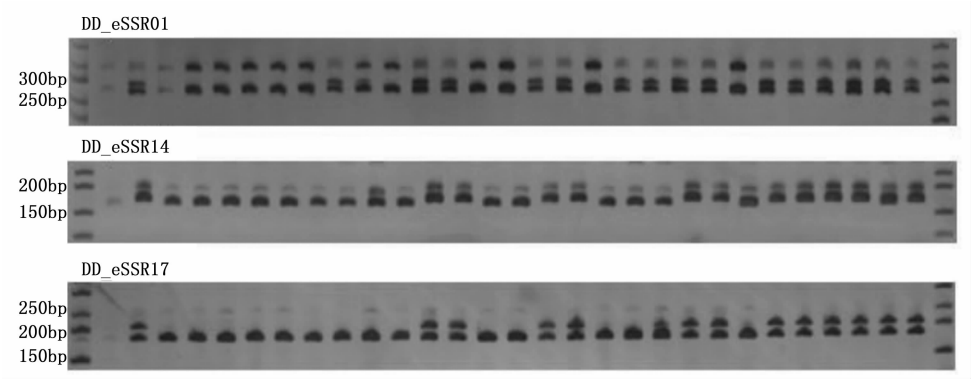
利用来自江西泰和县千烟洲的芒萁群体 29 个样本检测随机开发的 20 对 SSR 引物的扩增有效性与多态性,并初步尝试利用标记组合对各供试样本进行基因型分型鉴定。试验结果表明,共有 11 对引物具有良好的扩增,有效引物占 54.55%,而其中有 5 对引物在个体间存在多态性,多态引物信息见表 3。图 5 为 SSR 位点 DD_eSSR01、DD_eSSR14、DD_eSSR17扩增电泳图,组合 3 个位点的扩增结果,可初步判定 29 个样本包含 4 种基因型(分别为 BBBBBB、ABAAAB、ABBBBB 及 ABBBAB)。

表 3　多态性引物信息

位点	引物序列(5′～3′)	重复基序	退火温度/℃	扩增片段/ bp
DD_eSSR01	F:TCGTCCCCTTTACATTAGCCAC R:GCCAGTGTTGATACAGCTTGC	$(AC)_9$	56	280
DD_eSSR10	F: TCTGTCAGGCTTCGAACGAG R:TGGGGTTCTGAAAAATTGTAGCA	$(TC)_9$	55	273
DD_eSSR14	F: CCACTGGCACATTGTTCACA R TGAGACCCCTCTTTAGCAGGA	$(AC)_9$	55	170
DD_eSSR15	F: GCTTCTCCAGCCCTCCATTT R: CCTGTGCTTGGATTGGCAAC	$(GA)_{10}$	55	450
DD_eSSR17	F: CGAGGGTTCGGATTTCCCAA R: GGGCGGCTACAAGTGTGTAT	$(AATC)_4$	58	191

3　讨论

对于缺乏基因组信息的非模式物种而言,采用转录组测序技术可获得大量的转录本信息,对解决其基因进化、遗传育种以及生态等诸多方面的问题具有重要意义[30]。一般 Hiseq 策略的 Illumina 高通量测序通量大,但通量的增加注定会以牺牲序列片段长度为代价。由于非模式生物缺乏参考基因组信息,因而测序读长越长,越有利于测序片段的后续装配。这使得 454 技术(平均读长 400 bp)在非模式生物转录组研究中应用较为广泛,但该技术价格高昂且通量不高。本研究采用 MiSeq250 策略开展芒萁 Illumina 高通量测序,获得的 63 169 条 Unigene 序列,平均单条 Unigene 长度为 863 bp,N50 为 1 587 bp,序列质量评估 Q30 达到 97.55%,最短转录本长度为 201 bp,而获得的 Unigene 序列分布在 200 ～500 bp 长度区间的占总数的 55.4%。与同类研究相比[22,24],本次测序在降低成本的前提下,既保证



注:图中左右两侧为 50 bpMarker,自上而下依次对应为 SSR 位点 DD_eSSR01、DD_eSSR14、DD_eSSR17 PCR 扩增电泳检测结果。

图 5 芒萁群体 PCR 扩增电泳图

了测序的通量,又兼顾了单序列的长度与质量。

将芒萁 Unigene 序列与 NR 等 7 个数据库比对,共有 26 826 条获得注释,仅占总数的 42.46%。由于蕨类植物基因组学及转录组学研究尚处起步阶段,部分序列暂时无法获得相应的功能注释。有研究认为在目前转录组研究中,表达量很低的基因在 EST 数据库中很难找到,而表达量较高的数据过量存在[31]。本研究中芒萁 Unigene 的 FPKM 表达量值集中在 3.16 ~ 36.12,显示本次测序检测到的低表达水平基因的比例较高,可能原因在于研究对象的差异性。经 NR 数据库比对,芒萁与江南卷柏及苔藓植物小立碗藓的 Unigene 序列有较高的匹配,其中三者比较,芒萁与小立碗藓匹配率更高。Der 等[22]采用相似方法比较蕨、江南卷柏及小立碗藓的 Unigenes,显示蕨与小立碗藓具更多的相似性,推断原因可能在于与江南卷柏的异型孢子不同,蕨与小立碗藓同属同型孢子世代,三者生活史的差异影响了基因的进化。芒萁与蕨同为真蕨植物具有相似的生活史,结果的一致性进一步验证了前人的推测。同时,笔者注意到与本研究材料来源于孢子体不同,Der 的研究为世界第一个配子体转录组,初步分析蕨类植物不同世代功能基因表达存在一定的共性问题。

开发用于基因分型的 SSR 分子标记,可作为研究芒萁克隆生活史性状及对环境条件的响应机制的重要工具。本研究检索了芒萁 Unigenes 序列中的微卫星重复基序,显示重复基序进化趋向于较短的序列,诸如三核苷酸重复基序是该物种的主要微卫星基序,而五核苷酸重复基序最少。在不同类型核苷酸重复基序中出现频率最高类型为 AG/CT 与 AAG/CTT,这与同为蕨类植物的鸟巢蕨[24]相比有些不同,后者二核苷酸重复基序最多类型与前者一致,但三核苷酸重复基序以 AGG/CCT 为主;芒萁的这一微卫星分布规律与大青杨(*Populus cathayana* Rehd.)、油茶(*Camellia oleifera* Abel.)等[32-33]种子植物一致。随机开发 20 对 SSR 引物,有效引物占54.55%,其中有 5 对引物在芒萁群体检测存在多态性,表明引物开发效率较高。

4 结论

本研究应用 Illumina 高通量测序技术采用 MiSeq250 的策略开展芒萁孢子体转录组测序。通过生物信息学方法对测序获得的大量 Unigene 进行基因功能注释分类、代谢途径及微卫星特征等分析,从而初步揭示了芒萁孢子体生长发育过程中表达基因的功能总体特征。可为进一步开展芒萁生长、发育、代谢调控、微进化机制分析等研究提供重要的分子信息。此外作为世界蕨类植物基因组序列的重要补充,也可为其他蕨类植物功能基因挖掘及 SSR 标记的规模化开发提供丰富的基础数据。

参考文献:

[1] Xu X L, Li Q K, Wang J Y, *et al*. Inorganic and Organic Nitrogen Acquisition by a Fern *Dicranopteris dichotoma* in a Subtropical Forest in South China[J]. PLOS ONE, 2014, 9(5): e9005.

[2] 董 鸣. 克隆植物生态学[M]. 北京:科学出版社. 2011,5-6.

[3] 岩槻邦男. 日本野生植物-蕨类(日文)[M]. 东京: 平凡社. 1999,311.

[4] 张明如, 何 明, 温国胜, 等. 芒萁种群特征及其对森林更新影响评述[J]. 内蒙古农业大学学报, 2010, 31(4): 303-308.

[5] 李小飞, 陈志彪, 陈志强, 等. 南方红壤侵蚀区芒萁生长特征及其对环境因子的响应[J]. 水土保持通报, 2013, 33(3): 33-37.

[6] 刘迎春, 刘琪璟, 汪宏清, 等. 芒萁生物量分布特征[J]. 生态学

杂志, 2008, 27(5): 705 – 711.

[7] 侯晓龙, 刘明新, 蔡丽平, 等. 安溪崩岗侵蚀区不同植被配置模式与恢复效果研究[J]. 亚热带水土保持, 2010, 22(1): 5 – 10.

[8] Wei C Y, Wang C, Sun X, *et al*. Arsenic accumulation by ferns: a field survey in southern China[J]. Environmental Geochemistry and Health, 2007, 29(3): 169 – 177.

[9] 刘足根, 杨国华, 杨　帆, 等. 赣南钨矿区土壤重金属含量与植物富集特征[J]. 生态学杂志, 2008, 27(8): 1345 – 1350.

[10] 李小飞, 陈志彪, 陈志强. 南方稀土采矿恢复地土壤稀土元素含量及植物吸收特征[J]. 生态学杂志, 2013, 32(8): 2126 – 2132.

[11] 王立丰, 季红兵, 田维敏. 重稀土矿区芒萁稀土元素精细地位及光抑制对其光合活性的影响[J]. 中国稀土学报, 2010, 28(3): 379 – 386.

[12] 骆永明. 金属污染土壤的植物修复[J]. 土壤, 1999, 33(5): 261 – 265.

[13] 李交昆, 龚育龙, 唐璐璐, 等. 金属型植物的研究进展[J]. 生命科学研究, 2011, 15(6): 560 – 564.

[14] 邓敏捷, 董焱鹏, 赵振利, 等. 基于 Illumina 高通量测序的泡桐转录组研究[J]. 林业科学, 2013, 49(6): 30 – 36.

[15] Wang Z W, Jiang C, Wen Q, *et al*. Deep sequencing of the *Camellia chekiangoleosa* transcriptome revealed candidate genes for anthocyanin biosynthesis[J]. Gene, 2014, 538(1): 1 – 7.

[16] Barker M S, Wolf P G. Unfurling fern biology in the genomics age [J]. Bioscience, 2010, 60: 177 – 185.

[17] Barker M S. Evolutionary genomic analyses of ferns reveal that high chromosome numbers are a product of high retention and fewer rounds of polyploidy relative to angiosperms[J]. Amer Fern J, 2009, 99: 136 – 137.

[18] Nakazato T, Barker M S, Rieseberg L H, *et al*. Evolution of the nuclear genome of ferns and lycophytes [M]// Ranker T A, Haufler C H. Biology and Evolution of Ferns and Lycophytes. Cambridge University Press. 2008, 175 – 198.

[19] Aya K, Kobayashi M, Tanaka J, *et al*. De novo transcriptome assembly of a fern, *Lygodium japonicum*, and a web resource database, Ljtrans DB[J]. Plant & Cell Physiology, 2015, 56(1): e5.

[20] Weng J K, Tanurdzic M, Chapple C. Functional analysis and comparative genomics of expressed sequence tags from the lycophyte *Selaginella moellendorffii*[J]. BMC Genomics, 2005, 6: 85 – 97.

[21] Yamauchi D, Sutoh K, Kanegae H, *et al*. Analysis of expressed sequence tags in prothallia of *Adiantum capillus-veneris*[J]. Journal of Plant Research, 2005, 118: 223 – 227.

[22] Der J D, Barker M S, Wickett N J, *et al*. De novo characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*[J]. BMC Genomics, 2011, 12(1): 99 – 113.

[23] Bushart T J, Cannon A E, Haque U L A, *et al*. RNA-seq analysis identifies potential modulators of gravity response in spores of *Ceratopteris* (Parkeriaceae): evidence for modulation by calcium pumps and apyrase activity[J]. Amer J Bot, 2013, 100: 161 – 174.

[24] 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.

[25] 温　强, 叶金山, 雷小林, 等. 油茶 ISSR 反应体系建立及优化[J]. 中南林学院学报, 2006, 26(6): 22 – 26.

[26] 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, 644 – 652.

[27] Weber J L. Infomativeness of human (dC-dA)n-(dG-dT)n polymorphisms[J]. Genomics, 1990, 7: 524 – 530.

[28] Trapnell C, Williams B A, Pertea G, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation[J]. Nature Biotechnology, 2010, 28, 511 – 515.

[29] 全先庆, 张洪涛, 单　雷, 等. 植物金属硫蛋白及其重金属解毒机制研究进展[J]. 遗传, 2006, 28(3): 375 – 382.

[30] 刘洪亮, 郑丽明, 刘青青, 等. 非模式生物转录组研究[J]. 遗传, 2013, 35(8): 955 – 970.

[31] 梁　烨, 陈双燕, 刘公社. 新一代测序技术在植物转录组研究中的应用[J]. 遗传, 2011, 33(12): 1317 – 1326.

[32] 雷淑云, 张发起, Khan G, 等. 利用高通量测序分析青藏高原地区青杨的 SSR 和 SNP 特征[J]. 林业科学研究, 2015, 28(1): 37 – 43.

[33] 温　强, 徐林初, 江香梅, 等. 基于 454 测序的油茶 DNA 序列微卫星观测与分析[J]. 林业科学, 2013, 49(8): 43 – 50.

(责任编辑:彭南轩)