DOI:10.13275/j.cnki.lykxyj.2017.04.001

滇东南濒危植物长梗杜鹃转录组微卫星特征分析

李太强，刘雄芳，万友名，李正红，李钰莹，刘秀贤，马 宏*

（中国林业科学研究院资源昆虫研究所，云南 昆明 650233）

摘要：[**目的**]全面了解滇东南特有濒危植物长梗杜鹃转录组 SSR 位点的分布及序列特征，为长梗杜鹃的保护和合理开发利用提供遗传学资料，为同属植物及近缘种 SSR 标记的开发及遗传研究提供便利。[**方法**]利用 Illumina Hiseq 4000 高通量测序平台对长梗杜鹃叶片进行转录组测序，再通过 MISA 软件对测序所得 Unigenes 进行 SSR 位点的发掘和分析。[**结果**]发现含 SSR 的序列 17 354 条，共得到 23 192 个 SSR，出现频率为 31.30%，平均每 3 kb 出现 1 个 SSR。二碱基和三碱基重复为长梗杜鹃 SSR 主要重复单元类型，分别占 SSR 总数的 69.25% 和 15.07%，187 种重复基元中，所占比例最高的是(AG/CT)n(62.01%)，其次是(A/T)n(12.34%)、(AC/GT)n(4.52%)和(AAG/CTT)n(4.23%)。在 SSR 和 CDS 的交集基因中，共发现 15 908 个 SSR 位点，其中 2 792 个位于编码区，出现频率为 0.076 SSR/kb，而非编码区为 0.344 SSR/kb，在基因编码区中出现频率最高的是三碱基重复(1 356，48.57%)。在不同长度重复单元中，二碱基重复 SSR 长度变异程度最高，其次是单碱基重复。长梗杜鹃 SSR 的频率和长度呈显著负相关($P<0.01$)，相关系数为 -0.566。[**结论**]长梗杜鹃转录组 SSR 位点的出现频率高、分布密度大、基元类型丰富、重复次数较高、长片段较多，具有较高的多态性潜能，用于遗传分析的潜力很大，能满足该物种的保护遗传学研究。

关键词：长梗杜鹃；转录组；微卫星特征；潜在多态性

中图分类号：S685.21 **文献标识码**：A **文章编号**：1001-1498(2017)04-0533-09

Characteristic Analysis of Microsatellites in the Transcriptome of *Rhododendron longipedicellatum*, an Endangered Species Endemic to Southeastern Yunnan, China

LI Tai-qiang, LIU Xiong-fang, WAN You-ming, LI Zheng-hong, LI Yu-ying, LIU Xiu-xian, MA Hong

(Research Institute of Resource Insects, Chinese Academy of Forestry, Kunming 650233, Yunnan, China)

Abstract: [**Objective**] To comprehensively understand the distribution and sequence characteristics of SSR loci in the *Rhododendron longipedicellatum* transcriptome, and to provide a theoretical basis for further development of high efficient SSR markers. [**Method**] Transcriptome sequencing was conducted on young leaves of *R. longipedicellatum* by using Illumina Hiseq 4000. Then the SSR loci were sought and analyzed using MISA software from the obtained unigenes. [**Result**] A total of 23,192 SSRs were identified in 17,354 unigenes, with an average density of one SSR per 3 kb. Dinucleotide and trinucleotide repeat were the main SSR types, accounting for 69.25% and 15.07% of all SSRs, respectively. Among all the 187 repeat motifs, (AG/CT)n was the most frequent repeat motif (62.01%), followed by (A/T)n (12.34%), (AC/GT)n (4.52%) and (AAG/CTT)n (4.23%). A total of 15,908 SSRs occurred in the intersection of SSR and CDS, only 2792 of which occurred in protein-coding regions of these sequences. The density of SSRs was 0.076 SSR/kb in coding regions which was significantly lower than that

收稿日期：2016-07-14

基金项目："云南省技术创新人才"培养对象项目(2016HB007)

作者简介：李太强(1993—)，男，云南凤庆人，硕士，主要从事杜鹃属植物保护生物学研究.

* 通讯作者：马宏，男，副研究员，主要从事西南特色野生花卉种质资源创新与遗传多样性研究. E-mail：hortscience@163.com.

in non-coding regions (0.344 SSR/kb). Moreover, trinucleotide repeat was the most abundant in coding regions (1356, 48.57%). In terms of different length repeat units, the variation of the length of dinucleotide repeat SSR was the most abundant, followed by the mononucleotide. There was a significant negative correlation ($P < 0.01$) between the frequency of SSR and the length, with the correlation coefficient of −0.566. [**Conclusion**] The SSR loci in the *R. longipedicellatum* transcriptome showed high frequency and density of distribution, rich repeat motifs, high repeat times, more long fragment and significant potential of polymorphism. The SSR loci could be applied in genetic analysis and conservation genetics of *R. longipedicellatum* in the future.

Keywords: *Rhododendron longipedicellatum*; transcriptome; microsatellites characteristics; potential of polymorphism

杜鹃花是杜鹃花科(Ericaceae)杜鹃属(*Rhododendron*)植物的总称,是“世界三大园艺植物”和“中国十大天然名花”之一。我国具有最丰富的资源蕴藏量,为世界杜鹃花育种做出了巨大贡献。近百年来,国外培育出了数以千计的杜鹃花新品种,既改变了国外园林的风貌,又使杜鹃花形成了一种世界性园艺产业[1]。而我国杜鹃花引种驯化工作起步较晚,育种工作断断续续,所育品种较少[2]。目前,国际上杜鹃花的花色育种趋势为纯色花,特别是纯正、明亮的黄色和恬静的蓝色等更显珍贵[3];同时,周年供应鲜花对于杜鹃花生产具有重要意义[4]。因此,选择观赏性高、抗逆性强、花期长等优良种质作为杂交育种的亲本材料尤为重要,其中长梗杜鹃(*Rhododendron longipedicellatum* Lei Cai & Y. P. Ma)就是众多野生资源中难能可贵的育种材料。

长梗杜鹃系杜鹃属、杜鹃亚属(Subg. *Rhododendron*)、越桔杜鹃组(Sect. *Vireya*)、类越桔杜鹃亚组(Subsect. *Pseudovireya*)常绿植物。花冠颜色为明亮的纯黄色,无任何斑点。更令人称奇的是,其花期11月下旬至翌年的2月上旬,时值春节且长达3个月之久[5]。由于人类活动使得生境破坏日益严重,该种分布范围已非常狭窄,仅分布于滇东南海拔1 183～1 316 m左右的石灰岩山上。为了保护以及合理开发利用这一珍稀杜鹃种类,本课题组目前正在开展针对该稀有濒危种的引种驯化及保护生物学研究。

遗传多样性是生物多样性最基本的组成部分,也是保护生物学研究的核心目标。近年来,基于微卫星(microsatellite or simple sequence repeat)标记的杜鹃属植物遗传多样性和遗传结构研究已有一些报道。吴富勤[6]利用14个SSR标记分析了极小种群野生植物大树杜鹃(*R. protistum var. giganteum* Forrest et Tagg chambeniain)2个残存居群的遗传结构、遗传多样性和历史动态;Wang等[7]利用8个SSR位点评估了当地居民采食花朵对大白花杜鹃(*R. decorum* Franch.)的遗传影响。但目前杜鹃花中可利用的SSR标记较少,限制了其在杜鹃花种质资源评价中的应用。鉴于此,本研究利用Illumina Hiseq 4000最新高通量测序平台,对长梗杜鹃叶片进行转录组测序和组装,从获得的Unigenes序列中检测SSR位点,并对其序列特征、组成和变异规律开展分析,以期为后续长梗杜鹃大批量EST-SSR标记开发,进而进行遗传多样性和遗传结构分析,以及长梗杜鹃的保护和合理开发利用提供遗传学资料。同时,也丰富了杜鹃属植物的EST数据库,为同属植物及近缘种SSR标记的开发及遗传研究提供便利。

1 材料与方法

1.1 供试材料

采自云南省麻栗坡县,海拔高度约1 270 m。于2016年10月采集长梗杜鹃植株的幼嫩叶片,立即置于液氮中,带回实验室于-80℃冰箱中保存备用。

1.2 转录组测序

用“试剂盒提取法”对所采集的材料进行RNA提取,送华大基因有限公司(BGI)进行高通量测序。测序完成后先对原始数据进行过滤,然后使用Trinity对过滤后的reads进行de novo组装,最后使用Tgicl进行聚类去冗余得到最终的Unigenes。

1.3 SSR位点的搜索与分析

利用Perl操作平台下的MISA软件(misa-microsatellite identification tool, MISA, http:// pgrc. ipk-gatersleben. de/misa/)搜索长梗杜鹃Unigenes中潜在的1～6 bp的SSR位点,参数设置为:单碱基、二碱基、三碱基、四碱基、五碱基、六碱基的最短重复分别为12、6、5、5、4、4,复合SSR两个位点间最大间隔碱基数为100。采用Excel软件统计长梗杜鹃SSR

位点的数量、出现频率、分布的平均距离、重复单元类型和比例、重复单元碱基组成以及序列长度变异,并结合 SSR 和 CDS 的位置信息判断 SSR 的落点,全面了解其转录组 SSR 的序列特征。

2　结果与分析

2.1　长梗杜鹃转录组测序组装结果及统计

测序获得 58.30 Mb 的 Raw Reads,过滤后得到 44.85 Mb 的 Clean Reads,总碱基数为 6.73 Gb,Q20(质量值大于 20 的碱基数目占总碱基数目的比例)为 98.22%,所得序列的数量及质量均较高。对 Clean Reads 进行组装得到 94 906 个转录本(Transcripts),其长度主要分布在 200 ～2 000 bp 之间,占转录本总数的 89.85%。将这些转录本进一步聚类去冗余得到 74 092 条 Unigenes,其中聚类(clusters)的 Unigenes 为 51 505 条,单独(singletons)的 Unigenes 为 22 587 条;GC(碱基)含量为 43.20%,长度在 1 kb 以上的有 23 879 条,占 Unigenes 总数的 32.23%(表 1)。

表 1　长梗杜鹃转录组组装测序结果

Table 1　Assembly sequencing results of transcriptome of *R. longipedicellatum*

长度范围 Length range /bp	转录本 Transcripts	比例 Proportion /%	无冗余的 Contig Unigenes	比例 Proportion /%
200 ～1000	67 851	71.49%	50 213	67.77%
1000 ～2000	17 423	18.36%	14 892	20.10%
2000 ～3000	6 445	6.79%	5 936	8.01%
≥3000	3 187	3.36%	3 051	4.12%
总数 Total/个	94 906		74 092	
总长 Total length/bp	82 078 113		69 505 225	
N50 长度 N50 length/bp	1 470		1 616	
平均长度 Mean length/bp	864		938	

2.2　长梗杜鹃转录组中 SSR 位点的分布丰度与距离

利用 Perl 操作平台下的 MISA 软件对长梗杜鹃转录组所得 74 092 条 Unigenes 中 1 ～6 bp 的 SSR 进行查找,共搜索到 23 192 个 SSR 位点,包含 2 826 个复合型 SSR,分布于 17 354 条 Unigenes 上,其中 4 402 条 Unigenes 含有 2 个或 2 个以上的 SSR,部分 SSR 信息见表 2。

表 2　长梗杜鹃转录组 SSR 数据库的部分结果

Table 2　Partial result of SSR data in transcriptome of *R. longipedicellatum*

Gene_ID	重复类型 Repeat type	重复单元 Repeat motif	SSR 长度 Length of SSR/ bp
]CL10. Contig4	单碱基 Mononucleotide	(A)13	13
CL2. Contig2	二碱基 Dinucleotide	(TC)10	20
CL4. Contig1	三碱基 Trinucleotide	(GAG)5	15
CL761. Contig4	四碱基 Tetranucleotide	(CAAA)5	20
CL7329. Contig1	五碱基 Pentanucleotide	(GGATA)5	25
CL3540. Contig3	六碱基 Hexanucleotide	(ATAATC)4	24
CL3646. Contig2	复合模式 Compound pattern	(GAC)5 (GAG)6	33

序列组装去冗余后总长度为 69 505 225 bp(表 1),SSR 的发生频率(含 SSR 位点的 Unigenes 数与总 Unigenes 之比)为 23.42%,包含 SSR 的一致序列出现频率(检出的 SSR 个数与总 Unigenes 序列数之比)为 31.30%。SSR 的分布密度为 0.334 SSR/kb,平均每 3 kb 出现 1 个 SSR 位点;搜索到的 SSR 序列总长度为 543.322 kb(0.78%),说明在长梗杜鹃转录组中 SSR 序列小于整个转录组序列的百分之一(表 3)。

2.3　长梗杜鹃转录组中 SSR 位点的重复单元类型

在长梗杜鹃转录组 SSR 数据库中,以二碱基为重复单元的 SSR 含量最多,占总数的 69.25%,其次是三碱基和单碱基,分别占 15.07% 和 12.45%。而四、五、六碱基重复单元所占比例均较低且依次递增

表 3　长梗杜鹃转录组 SSR 各重复类型的分布特征

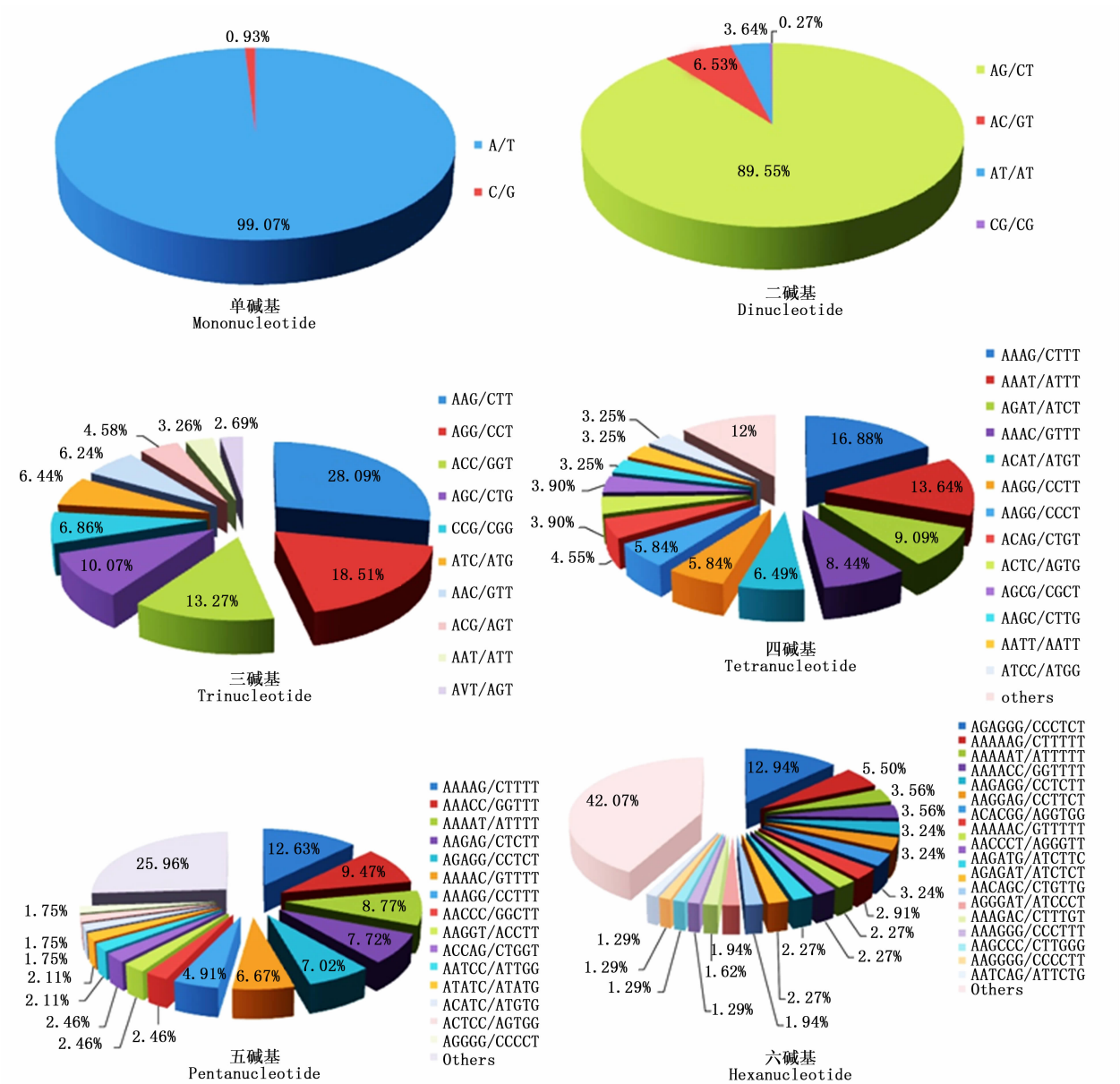
Table 3　The characteristic of various SSR repeat types in R. longipedicellatum transcriptome

重复类型 Repeat type	SSR 数量 Number of SSR	所占比例 Proportion/%	出现频率 Frequency/ %	平均距离 Mean distance/ kb	分布密度 Distribution density /(个. Mb^{-1})	平均长度 Mean length/ bp
单碱基	2 888	12.45	3.90	24.07	41.55	14.67
二碱基	16 060	69.25	21.68	4.33	230.95	22.99
三碱基	3 496	15.07	4.72	19.88	50.30	17.83
四碱基	154	0.66	0.21	451.33	2.22	21.63
五碱基	285	1.23	0.38	243.88	4.10	21.32
六碱基	309	1.33	0.42	224.94	4.45	28.89
小计 total	23 192	100	31.3	3.00	333.33	21.23

(表 3)。相应地不同重复单元的 SSR 含量、出现频率、分布密度以及分布的平均距离变化也很大。其中,SSR 含量、出现频率、分布密度的变化规律一致,依次为:二碱基 > 三碱基 > 单碱基 > 六碱基 > 五碱基 > 四碱基;与之对应的平均距离以四碱基最高,为 451.33 kb;以二碱基最低,为 4.33 kb,且二者的差异达 104 倍,即该转录组序列中每出现 104 个二碱基重复类型才出现 1 个四碱基重复类型的 SSR。

2.4 长梗杜鹃转录组中 SSR 重复基元碱基组成

考虑碱基互补作用,在长梗杜鹃转录组 23 192 个 SSR 中共发现 187 种重复基元,其中单、二、三、四、五、六碱基重复分别有 2、4、10、22、56 和 93 种,不同碱基的重复基元所占比例差异较大(图 1)。单碱基重复类型中以 A/T 为主要重复基元,占该类型的99.07%;二碱基重复类型中各基元所占比例依次为:AG/CT(89.55%) > AC/GT(6.53%) > AT/AT(3.64%) > CG/CG(0.27%);三碱基重复类型中 AAG/CTT 最多(28.09%),其次是 AGG/CCT(13.27%)、ACC/GGT(13.27%);AAAG/CTTT(16.88%)、AAAAG/CTTTT(12.63%)和 AGAGGG/CCCTCT(12.94%)分别为四、五、六碱基重复类型的优势重复基元,且分别有 5、20、41 种基元里只有 1 个 SSR。



注:others 表示未列出的其余基元的统称

Note: others: The rest of all repeat motifs not for being listed in the bar

图 1 长梗杜鹃转录组 SSR 不同重复类型各基元的比例

Fig. 1 Motif proportions of each types of repeat in *R. longipedicellatum* transcriptome

整体来看,在长梗杜鹃转录组中最丰富的 SSR 类型是二碱基重复,其次是三碱基重复,最主要的优势重复基元分别是(AG/CT)n、(A/T)n、(AC/GT)n 及(AAG/CTT)n,分别占总 SSR 数量的 62.01%、12.34%、4.52%和 4.23%。此外,还发现了 44 个在植物转录组中不常见的 CG/CG 基元,以及 240 个在双子叶植物中很少见的 CCG/CGG 基元。

2.5　长梗杜鹃转录组中 SSR 在编码区中的分布特征

对 SSR 和 CDS(编码区)的交集基因进行检测,共发现 15 908 个 SSR 位点,其中仅有 2 792 个位点存在于编码区,位于非编码区的位点达到 12 555 个,另有 561 个位点跨越了蛋白编码区和非编码区。编码区 SSR 的出现频率(编码区中检出的 SSR 个数与 CDS 总长度之比)为 0.076 SSR/kb,而在非编码区中为 0.344 SSR/kb,这说明非编码区 SSR 出现频率大约是编码区的 4.5 倍。在基因编码区 2 792 个位点中,所占比例最高的是三碱基重复(1 356, 48.57%),其次是二碱基重复(808, 28.94%)和单碱基重复(275, 9.85%),此外还发现(225, 8.06%)个复合型 SSR。非编码区则是二碱基重复最多(8 306, 66.16%),其次是单碱基重复(1 283, 10.22%)。

2.6　长梗杜鹃转录组中 SSR 基元重复次数

SSR 重复次数的不同会导致重复片段长度发生变异,进而影响其多态性。长梗杜鹃转录组中 SSR 各重复类型的重复次数分布范围较广,波动于 4～117 次,且多集中于 4～25 次(图 2)。

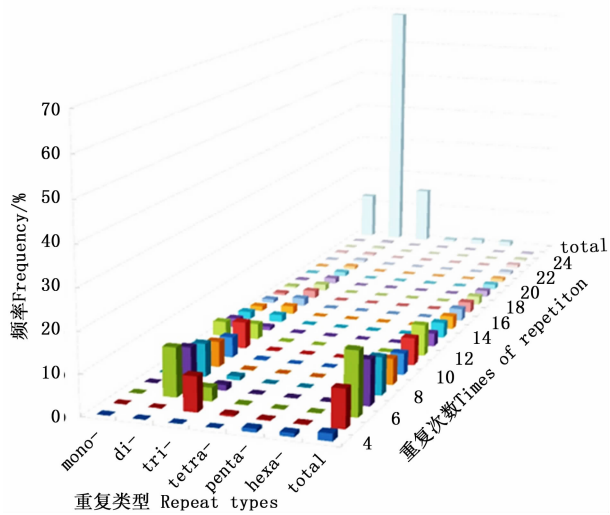


图 2　长梗杜鹃转录组 SSR 各重复类型不同重复次数分布频率

Fig. 2　Percentage of various repeat types with different number of repeats in *R. longipedicellatum* transcriptome

其中,单、二、三、四、五、六碱基分别重复 12～117、6～50、5～22、5～10、4～8 和 4～15 次,且表现为随着重复次数以及碱基数量的增加,SSR 出现的频率降低,仅当二碱基重复从 10 次增加到 11 次时,SSR 数量出现了较大增加的情况。重复基元以重复 6 次的频率最高,共有 SSR 3 630 个,占 SSR 总数的 15.65%,其次是 7 次(2 587,11.15%)、5 次(2 176,9.38%)、8 次(2 144,9.24%),25 次以上的 SSR 仅有 340 个,占总 SSR 的 1.47%。总体来看,SSR 的重复次数以 4～10 次较多,占 59.12%,11～20 次的占 35.97%,而重复次数在 20 次以上的不足 5%,表现为 SSR 数量随着重复次数的增加呈明显下降的趋势(图 3)。

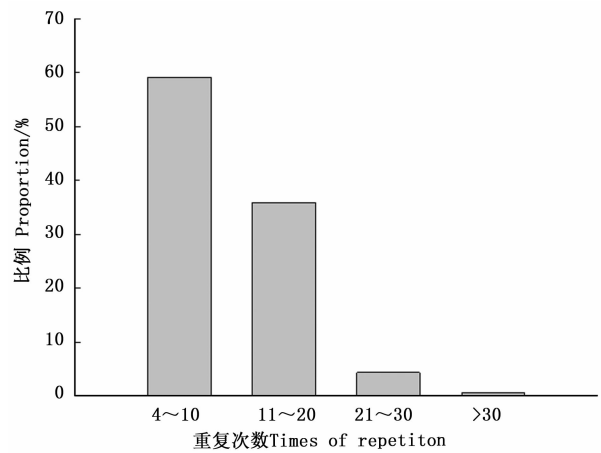


图 3　长梗杜鹃转录组 SSR 重复次数分布频率

Fig. 3　Frequency of repeat number of SSR in *R. longipedicellatum* transcriptome

2.7　长梗杜鹃转录组中 SSR 序列长度分布及变异情况

长梗杜鹃转录组中 SSR 序列的长度存在显著变异,长度由 12～117 bp 不等,平均长度为 21.23 bp,通过正态性检验,其偏度(Sk)和峰度(Ku)均大于零,不符合正态分布;单碱基重复长度变化范围最大(12～117 bp),其中以 A/T 基元长度变化范围最大(12～117 bp),其次是 AG/CT(12～100 bp)。单碱基、二碱基、三碱基、四碱基、五碱基和六碱基的平均长度分别为 14.67、22.99、17.83、21.63、21.32 和 28.89 bp(表 3),且各碱基重复类型均表现为随着重复片段长度的增加,SSR 出现的频率降低,即各碱基重复区段片段长度与其对应的 SSR 数量成相反的变化趋势。从全部碱基来看,12 bp 长的 SSR 在长梗杜鹃转录组中所占比例最高,为 14.46%,其次是 15 bp

(10.56%)、14 bp(10.48%)和 18 bp(9.53%),其中长度≥20 bp 的 SSR 位点有 7 698 个,占 SSR 总数的 42.90%(图 4)。

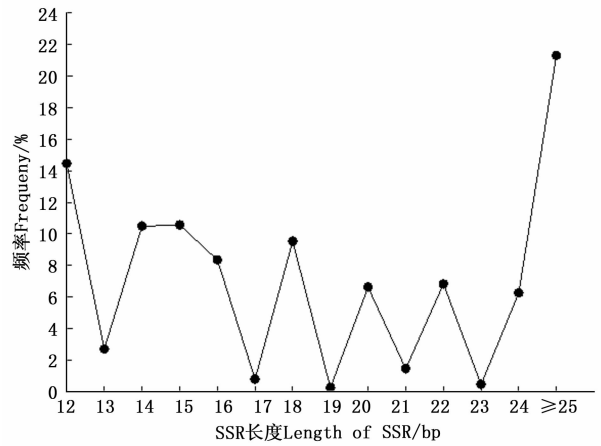


图 4 长梗杜鹃转录组中 SSR 的长度分布

Fig. 4 Length distribution of SSR in *R. longipedicellatum* transcriptome

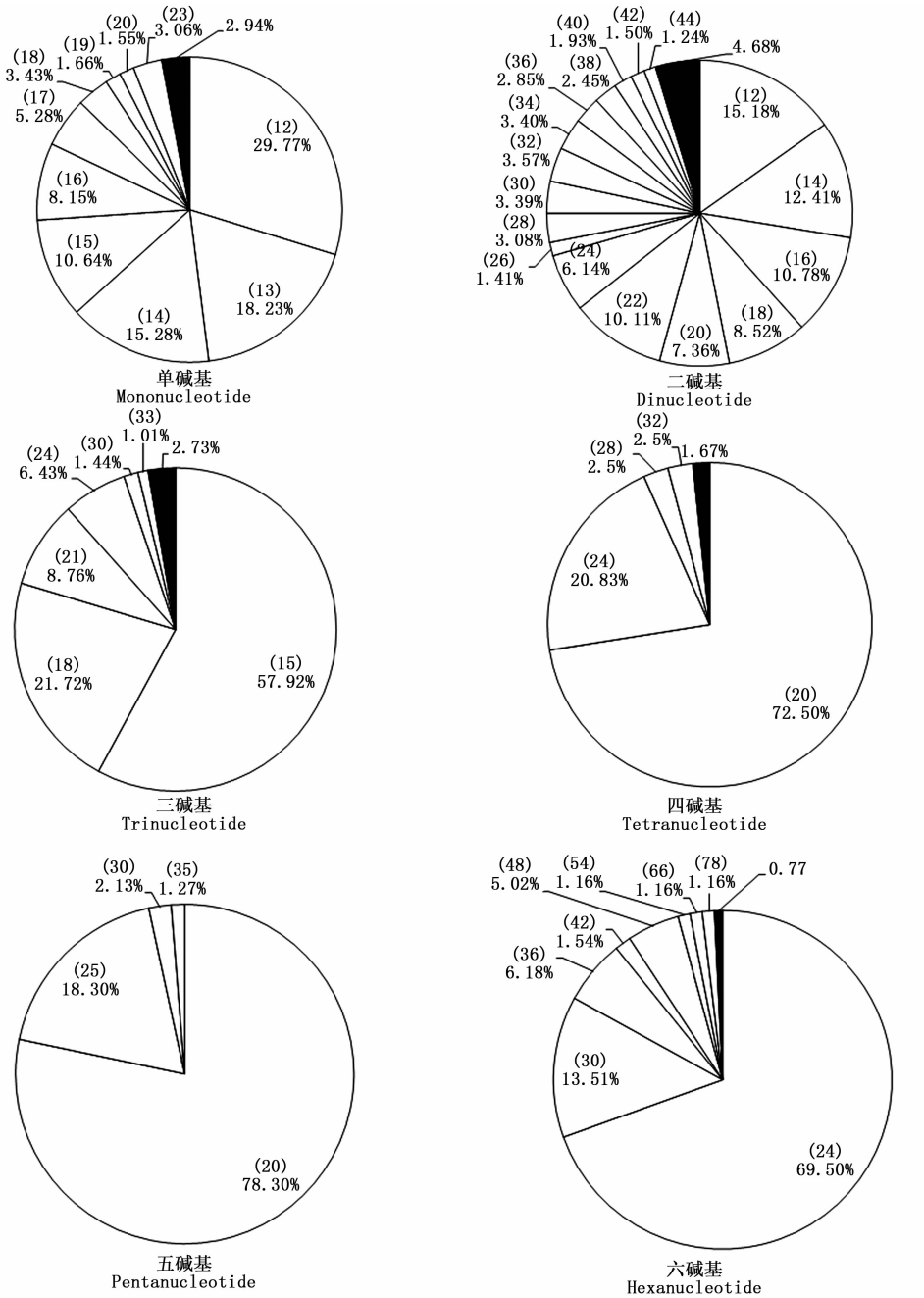
进一步对长梗杜鹃不同长度重复单元 SSR 的长度变异情况进行分析,分别描述了各碱基重复不同长度 SSR 在饼图中的占比,图中各扇区对应不同长度的 SSR,频率≤1% 的 SSR 合并在同一黑色扇区内(图 5)。从图中可知,二碱基重复 SSR 的长度变异程度最高,有 40 种不同 SSR 变化长度;其次是单碱基,有 28 种;三碱基、六碱基、四碱基重复 SSR 长度变异程度依次降低,五碱基最低,仅 4 种变化长度。长梗杜鹃转录组 SSR 的序列长度与其出现频率的 Pearson 相关性分析表明二者在 0.01 水平(双侧)上显著负相关,相关系数为 -0.566。

3 讨论

本研究通过长梗杜鹃叶片转录组测序,组装、聚类去冗余后获得 74 092 条 Unigenes,利用 Perl 操作平台下的 MISA 软件共搜索到 23 192 个 SSR 位点,包含 SSR 的一致序列出现频率为 31.30%,分布密度为 0.334 SSR/kb,平均每 3 kb 出现 1 个 SSR 位点。与大多数双子叶植物如杜仲(*Eucommia ulmoides* Oliver)[8](0.038 SSR/kb)、碧桃(*Prunus persica* cv. *duplex* Rehd.)[9](0.287 SSR/kb)和短丝木犀(*Osmanthus serrulatus* Rehd.)[10](0.183 SSR/kb)的 EST-SSR 相比,长梗杜鹃转录组中 SSR 的分布密度较高;但低于高粱(*Sorghum bicolor*(L.) Moench)(0.646 SSR/kb)、水稻(*Oryza sativa* L.)(0.739 SSR/kb)等单子叶植物[11],这可能是二者的进化因素不同使得双子叶植物的 SSR 分布偏低[12],另外出现这种差异也可能与物种间 SSR 的分布、含有 SSR 基因的表达丰度、搜索的序列来源、搜索软件的选择以及搜索的标准等不一致有关。总体而言,长梗杜鹃转录组中 SSR 数量比较丰富。

在获得的长梗杜鹃转录组所有 SSR 中,二碱基重复为最主要重复类型,占所有 SSR 的 69.52%,其次是三碱基重复,占 15.07%,这与许多物种以二、三碱基重复类型居多一致[13-15]。袁阳阳等[16]在荟菜(*Nymphoidespeltata*(Gmel.) O. Kuntze)转录组发现的 12 319 个 EST-SSR 位点中,二碱基和三碱基重复单元是主导类型,分别占总 SSR 的 57.31% 和 30.87%;李美芹等[17]从 NCBI 公共数据库现有杜鹃花相关 EST 中获得的 435 个 SSR 序列也以二、三碱基重复为主。一般认为,低级重复单元的大量存在暗示着该物种进化水平较高,而高级重复单元出现频率高的物种具有较短的进化时间或较低的变异频率[18,19]。长梗杜鹃中单、二和三碱基重复类型共占总 SSR 的 96.77%,可能预示着其具有较高的变异频率或较长的进化历史,这或许在一定程度上支持了方瑞征和闵天禄[20]所得结论,杜鹃属植物起源于距今约 6 700 万年至 13 700 万年中生代的白垩纪,具有悠久的进化历史。相比较而言,4 ~ 6 bp 重复类型较少,且随着重复单元碱基数的增加,SSR 出现频率、SSR 含量以及分布密度随之升高,即六碱基 SSR 类型较多。在云南松(*Pinus yunnanensis* Franch.)[21]转录组 SSR 分布特征研究中,也表现为六碱基较四、五碱基多。这可能与密码子以三碱基为一个单元有关,造成了三碱基位移[22]。

SSR 分布在不同物种间存在较大差异,且物种本身碱基组成也是选择的结果。在长梗杜鹃单碱基重复类型中,A/T 基元占绝大多数,四、五、六碱基中 AAAT/ATTT、AAAAT/ATTTT 和 AAAAAT/ATTTTT 基元含量也相对较高,表现出一定的 A/T 优势,这可能与碱基所含的能量有关[23]。但是主要重复类型二、三碱基的优势重复基元是 AG/CT 和 AAG/CTT,分别占 SSR 总数的 62.01% 和 4.23%,与蜡梅(*Chimonanthus praecox*(Linn.) Link)[24]、碧桃[9]、短丝木犀[10]等植物转录组 SSR 分布的研究结果一致。在三碱基重复中,AAG/CTT、AGG/CCT 和 ACC/GGT 基元所占比例最高,与王书珍等[25]报道的杜鹃花 EST-SSR 序列三碱基中的优势基元 AAG、



注:饼图每一扇区对应不同长度的 SSR 标注于所占比例上部括号内,若对应长度 SSR 频率≤1%,则一起合并在黑色扇区内。

Note: SSR in different lengths are demonstrated in separate slices. If the corresponding percentage≤1%, slices were combined for percentages(black slices).

图 5　长梗杜鹃转录组不同长度重复单元 SSR 长度变异情况

Fig. 5　Length diversification of SSR in *R. longipedicellatum* transcriptome

ACC、AGA 比较相似,许玉兰等[14]对多数物种的统计也表明三碱基中 AAG、AGC 和 AGG 较多,这些较多的重复基元可能在 EST 序列中较为普遍,也可能是优势的蛋白或 DNA 家族[26]。此外,长梗杜鹃中还发现了 44 个在植物转录组二碱基重复中比较罕见的 CG/CG 和 240 个在双子叶植物中分布较少的 CCG/CGG 重复基元,其含量远高于大多数植物,如甘蓝(*Brassica oleracea* L.)[27](1 个 CG)、蜡梅[24](6 个 CG)、杜仲[8](1 个 CG)和短丝木犀[12](13 个 CG、43 个 CCG)等,较多的 CG 和 CCG 重复基元可能与某些特定的功能相关,如抗逆性、转录调控和信号转导等[28]。也进一步证明所得长梗杜鹃转录组

SSR 具有较高的特异性。

许多研究表明三碱基重复 SSR 是目前为止基因编码区中发现最多的 SSR 类型[29,30]。长梗杜鹃也不例外,结合 SSR 和 CDS 的位置信息,对 SSR 的分布区间进行统计,发现长梗杜鹃转录组 SSR 序列主要分布在非编码区,编码区 SSR 出现频率仅为非编码区的 11/50,且编码区中三碱基 SSR 显著富集,占总检测量的 48.57%,而非编码区以二碱基重复较多。这可能是密码子选择作用的结果,由于三碱基重复单元重复次数的变化对基因读码框和表达产物的影响较小,从而使其在编码区的容受性优于其他类型。这一现象也说明三碱基重复 SSR 富集是基因编码区 SSR 在基因组中得以保存的重要机制[31]。Reddy 等[32]报道了人类基因组研究已经发现三碱基重复 SSR 与某些疾病的发生相关;将长梗杜鹃转录组测序所得全部 Unigenes 映射到 KEGG 代谢库,发现了 176 条与人类疾病相关的 Unigenes,这是否与基因编码区富集的三碱基重复有关,对长梗杜鹃的生长发育又有什么意义仍有待进一步研究。

SSR 位点多态性主要原因是基元重复数和碱基数不同而形成的序列长度多态性[33],一般重复次数越多,变异性越大,其多态性潜力越高。长梗杜鹃 SSR 重复次数波动于 4 ~ 117 次,以 4 ~ 10 次重复较多,其次是 11 ~ 20 次;其中单碱基因容易发生错配不考虑在内,其余的碱基重复类型重复次数也集中于 4 ~ 36 次,甚至有高达 50 次的。从片段长度来看,当 SSR 长度≥20 bp 时多态性较高,在 12 ~ 20 bp 之间多态性中等,<12 bp 时多态性极低[34],本研究在筛选过程中已经将 <12 bp 的低多态 SSR 过滤掉,最终发现长梗杜鹃 SSR 序列长度变化范围是 12 ~ 117 bp 之间,平均长度为 21.23 bp,其中≥20 bp 的高多态重复序列占 42.90%,其比例高于云南松[21](14.76%)、碧桃[9](12.13%)、短丝木犀[10](13.47%)等大多数植物,由此推测长梗杜鹃转录组挖掘出的 23 192 个 SSR 位点大部分具有高多态性潜能。通过 SPSS 软件对 SSR 发生频率与重复片段长度进行 Pearson 相关性分析,发现二者显著负相关,相关系数为 -0.566。在长梗杜鹃不同长度重复单元 SSR 长度变异分析中,二碱基重复 SSR 长度变异程度较高,有 40 种不同 SSR 变化长度,即二碱基类型获得或失去重复基元的活跃程度较高;其次是单碱基(28 种),而五碱基最低(仅 4 种),且各重复类型均表现为 SSR 长度越长,出现的频率越低。表明由短重复单元组成的 SSR 比由长重复单元组成的 SSR 可能具有更丰富的多态性。

4 结论

本研究通过 Perl 操作平台下的 MISA 软件对长梗杜鹃转录组中 SSR 序列进行查找,共搜索到 23 192 个 SSR 位点,对其分布频率、重复单元类型、重复基元碱基组成、在编码区中的分布特征、重复次数和序列长度分布及变异情况进行分析,得出大多数位点具有高多态性潜能,用于遗传分析的潜力很大,为长梗杜鹃 SSR 分子标记的大规模开发提供了重要的信息资源和数据保障。尤其是分布于编码区的序列,可能与某一特定功能相关联,有助于长梗杜鹃功能性 SSR 标记的开发,进而为该物种遗传多样性和遗传结构、遗传资源分类和进化以及分子标记辅助育种等方面的研究奠定基础。加之,EST-SSR 具有较高的转移性,进一步开发的 SSR 标记有望用于杜鹃属植物及其它亲缘关系较近物种的研究中。

参考文献:

[1] 张长芹. 云南杜鹃花[M]. 昆明: 云南科技出版社, 2008: 1.

[2] 张长芹,高连明,薛润光,等. 中国杜鹃花的保育现状和展望[J]. 广西科学, 2004, 11(4): 354 - 359.

[3] 程金水,刘青林. 园林植物遗传育种学(第2版)[M]. 北京: 中国林业出版社, 2010: 452.

[4] 兰 熙,张乐华,张金政,等. 杜鹃花属植物育种研究进展[J]. 园艺学报, 2012, 39(9): 1829 - 1838.

[5] Cai L, Neilsen J, Dao Z L, *et al*. *Rhododendron longipedicellatum* (Ericaceae), a new species from Southeastern Yunnan, China[J]. Phytotaxa, 2016, 282(4): 296 - 300.

[6] 吴富勤. 极小种群野生植物大树杜鹃的保护生物学研究[D]. 云南: 云南大学, 2015.

[7] Wang X, Huang Y, Long C. Assessing the genetic consequences of flower-harvesting in *Rhododendron decorum* Franchet (Ericaceae) using microsatellite markers[J]. Biochemical Systematics and Ecology, 2013, 50: 296 - 303.

[8] 黄海燕,杜红岩,乌云塔娜,等. 基于杜仲转录组序列的 SSR 分子标记的开发[J]. 林业科学, 2013, 5: 176 - 181.

[9] 马秋月,廖卓毅,张得芳,等. 碧桃花瓣转录组微卫星特征分析[J]. 南京林业大学学报:自然科学版, 2015, 3: 34 - 38.

[10] 陈 林,李龙娜,杨国栋,等. 特有植物短丝木犀(Osmanthus serrulatus)转录组微卫星特征分析[J]. 分子植物育种, 2016, 14(4): 959 - 965.

[11] Cavagnaro P F, Senalik D A, Yang L, *et al*. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.)[J]. BMC Genomics, 2010, 11(1): 569.

[12] 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: 429 – 437.

[13] Aggarwal R K, Hendre P S, Varshney R K, *et al*. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species[J]. Theoretical and Applied Genetics, 2007, 114(2): 359 – 372.

[14] 许玉兰,蔡年辉,康向阳,等. EST-SSR 标记的开发及其在木本植物中的分布特点[J]. 中国农学通报, 2012, 28(4): 1 – 7.

[15] 饶龙兵,杨汉波,郭洪英,等. 基于桤木属转录组测序的 SSR 分子标记的开发[J]. 林业科学研究, 2016, 29(6): 875 – 882.

[16] 袁阳阳,王青锋,陈进明. 基于转录组测序信息的水生植物莕菜 SSR 标记开发[J]. 植物科学学报, 2013, 31(5): 485 – 492.

[17] 李美芹,潘叶羽,钱萍仙,等. 杜鹃花 EST-SSR 标记的开发及遗传多样性分析[J]. 植物生理学报, 2016, 52(3): 356 – 364.

[18] Harr B, Schlotterer C. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide under representation[J]. Genetics, 2000, 155(3): 1213 – 1220.

[19] 阮桢媛,王兵益,欧阳志勤,等. 极度濒危植物巧家五针松基因组微卫星特征分析[J]. 植物研究, 2016, 36(5): 775 – 781.

[20] 方瑞征,闵天禄. 杜鹃属植物区系的研究[J]. 云南植物研究, 1995, 17(4): 359 – 379.

[21] 蔡年辉,许玉兰,徐　杨,等. 云南松转录组 SSR 的分布及其序列特征[J]. 云南大学学报:自然科学版, 2015, 37(5): 770 – 778.

[22] Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA[J]. Genome Research, 2000, 10(1): 72 – 80.

[23] 童治军,肖炳光. 3 种烟草基因组 SSR 位点信息分析和标记开发[J]. 西北植物学报, 2014, 34(8): 1549 – 1558.

[24] 李　响,杨　楠,赵凯歌,等. 蜡梅转录组 EST-SSR 标记开发与引物筛选[J]. 北京林业大学学报, 2013, 35(1): 25 – 32.

[25] 王书珍,张传进,程　华, 等. 杜鹃花表达序列标签资源中的微卫星信息分析[J]. 湖北林业科技, 2014, 43(2): 7 – 10.

[26] 张得芳,李淑娴,夏　涛. 蔷薇科 6 个属植物 EST-SSR 特征分析[J]. 植物研究, 2014, 34(6): 810 – 815.

[27] 陈　琛,庄　木,李康宁,等. 甘蓝 EST-SSR 标记的开发与应用[J]. 园艺学报, 2010, 37(2): 221 – 228.

[28] Li S X, Yin T M. Map and analysis of microsatellites in the genome of Populus: The first sequenced perennial plant[J]. Science in China Series C: Life Sciences, 2007, 50(5): 690 – 699.

[29] Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.[J]. Nature Genetics, 2002, 30(2): 194 – 200.

[30] 王丽鸳,韦　康,张成才,等. 茶树花转录组微卫星分布特征[J]. 作物学报, 2014, 40(1): 80 – 85.

[31] 刘菁菁,戴晓港,王　洁,等. 杨树微卫星序列对基因表达频率的影响及表达序列中微卫星特征的分析[J]. 南京林业大学学报:自然科学版, 2011, 35(1): 11 – 14.

[32] Reddy P S, Housman D E. The complex pathology of trinucleotide repeats[J]. Current Opinion in Cell Biology, 1997, 9(3): 364 – 372.

[33] 王　森,张　震,姜倪皓,等. 半夏转录组中的 SSR 位点信息分析[J]. 中药材, 2014, 37(9): 1566 – 1569.

[34] Temnykh S, Declerck G, Lukashova A. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.) frequency, length variation, transposon associations, and genetic marker potential[J]. Genome Research, 2001, 11(8): 1441 – 1452.

(责任编辑:张　玲)