火炬松热胁迫 cDNA 文库的 EST-SSR 预测

林元震[1]，郭　海[2]，刘纯鑫[1]，黄少伟[1]，陈晓阳[1]

（1 热带亚热带林业生物技术实验室，华南农业大学 林学院，广东 广州 510642；
2 水利部水土保持植物中心，北京 100038）

摘要：以火炬松热胁迫 cDNA 文库的 EST 序列为材料，对 EST 序列进行聚类、拼接等处理后，再进行 EST-SSR 标记的预测. 结果表明：火炬松热胁迫 cDNA 文库 4 283 条 EST 序列经 CAP3 拼接后，获得 2 062 个 UniGene，其中 934 个 Contig，1 128 个 Singletons. 对 UniGene 利用 SSRIT 在线软件分析得到 110 条 EST-SSR. 拼接后的 UniGene 含有 SSR 位点的频率为 4.32%，SSR 在火炬松 EST 上的分布密度为每 14.6 kb 出现 1 个 SSR. 这些 EST 重复基元中，二核苷酸重复和三核苷酸重复最多，分别占 60.90% 和 36.36%；四、六核苷酸重复分别占 0.91%、2.73%；没有五核苷酸的重复基序. 所有的核苷酸重复基元中，二核苷酸 AT 所占比例最高，约占 42.73%；三核苷酸重复中，比例最高的是 AAG 和 AGG，均占 7.33%. 上述研究结果对于开发火炬松新分子标记与开展分子辅助育种的研究具有一定的指导意义.

关键词：火炬松；热胁迫；cDNA 文库；EST-SSR
中图分类号：Q523　　**文献标识码**：A　　**文章编号**：1001-411X(2009)03-0041-04

EST-SSR Prediction of Heat Induced cDNA Library of *Pinus teada*

LIN Yuan-zhen[1], GUO Hai[2], LIU Chun-xin[1], HUANG Shao-wei[1], CHEN Xiao-yang[1]

(1 Laboratory for Biotechnology in Tropical and Subtropical Forest Science, College of forestry, South China Agricultural University, Guangzhou 510642, China;
2 Plant Materials for Soil and Water Conservation, Ministry of Water Resources, Beijing 100038, China)

Abstract: EST sequences from heat induced cDNA library of loblolly pine (*Pinus teada*) were assembled with CAP3 and predicted for EST-SSR. The results showed that 4 283 ESTs from heat induced cDNA library were asembled into a total of 2 062 non-redundant UniGenes with 934 Contigs and 1 128 Singletons by the program CAP3. These non-redundant UniGenes were used for searching SSR through the software SSRIT. A total of 110 SSR loci for loblolly pine with the frequency of 4.32% in UniGene and one SSR per 14.6 kb in EST was identified. Dinueleotides and trinueleotides were observed at the highest frequencies, 60.90% and 36.36%, respectively. And then tetra-, hexa-, was 0.91%, 2.73%, respectively, while pentanuleotides was not found. AT (42.73%) was the most frequent repeat. AAG and AGG (7.33% each) was the most motif in trinucleotides. Findings from this study may provide important information for the new molecular marker discovery and molecular assisted selection in loblolly pine.

Key words: *Pinus teada*; heat stress; cDNA library; EST-SSR

林木生长周期长，遗传杂合性高，许多重要性状属于多基因控制的数量性状，遗传机理不明，利用常规育种手段往往难以满足不同目的定向培育树木新品种的要求，因此人们期望借现代分子生物学技术，尤其是分子辅助选择育种技术，进行林木分子遗传育种，该方法具有高效性和针对性，可弥补常规育种

收稿日期：2008-11-17
作者简介：林元震(1979—)，男，讲师，博士；通讯作者：陈晓阳(1958—)，男，教授，博士，E-mail：xychen@scau.edu.cn
基金项目：广东省自然科学基金(7118123)；国家“十一五”科技支撑计划专题(2006BAD07A04)

技术的不足,缩短育种周期,加速优质、高抗、速生、丰产林木新品种的选育进程.

了解生物个体间详尽的遗传关系,特别是DNA序列差异,有利于推动种质创新研究. DNA标记是评价材料间遗传差异的很好的工具. 简单序列重复(Simple sequence repeat,SSR),也称微卫星(Microsatellite),具有重复性好、多态性高、呈共显性遗传、数量丰富和遍布整个基因组等优点,这些优点使得SSR标记成为广泛使用的分子标记之一,然而,按照传统的方法开发SSR标记,不仅费时费力,而且效率低[1-2]. 随着国际公共数据库中的EST序列呈指数增长趋势,从表达序列标签(Expressed sequence taqs,ESTs)中开发SSR标记正在成为新标记开发的焦点. 自2000年起,人们相继展开了从植物(如葡萄、甘蔗、小麦、黑麦、和大麦)中发掘EST-SSR标记的研究[3]. 截止2008年11月7日,美国国立生物技术信息中心(NCBI)的dbEST公共数据库中共有58 281 007条EST. 其中,火炬松 *Pinus teada* 328 628条,海岸松 *P. pinaster* 27 288条,长白松 *P. sylvestris* 2 349条,意大利石松 *P. pinea* 289条,辐射松 *P. radiata* 151条,湿地松 *P. elliottii* 150条. 如此大量的EST序列,不仅可以发现、分离新基因,也可制备DNA芯片用于基因表达和比较基因组研究,而且还可开发分子标记用于遗传连锁图谱构建、基因定位等.

目前,国内外火炬松的研究,主要集中在常规育种与一些分子标记方面[4-7],关于火炬松EST-SSR方面的报道还比较少,尤其是EST-SSR标记的预测与开发. 在我国,火炬松主要分布在亚热带及其北部地区,其主要原因之一是火炬松耐热性差[8]. 本研究以火炬松热胁迫cDNA文库的EST序列为材料,进行聚类、拼接等处理,再进行EST-SSR标记的预测,为今后开发EST-SSR标记及其运用于火炬松耐热分子辅助育种的研究奠定基础.

1 材料与方法

1.1 材料

火炬松热胁迫cDNA文库所有原始ESTs序列均来自ForestTreeDB数据库[9].

1.2 ESTs序列组装

利用CAP3软件对EST序列进行组装[10]. CAP3的参数设置为默认值,其中,重叠一致百分比域值(Overlap percent identity cutoff) $N>80$,重叠长度域值(Overlap length cutoff) $N>40$,经过组装就可得到相应的UniGenes(Contig和Singleton).

1.3 EST-SSR位点的搜索

利用SSRIT对UniGenes搜寻SSR位点[11]. 搜索标准为:重复单位长度为2~6 bp,最少重复次数为5次. 得到SSR位点后,进行统计分析.

2 结果与分析

2.1 ESTs的聚类分析

利用CAP3软件对火炬松热胁迫cDNA文库的4 283条EST序列进行组装. 组装后产生了2 062个UniGene,其中包括934个Contig和1 128个Singleton. 用总序列数量减去总UniGene数量,再除以总序列数量的方式来估算EST的冗余率,火炬松EST的冗余率为51.86%. 火炬松UniGene总的覆盖长度约为1.6 Mb,平均长度为777 bp.

2.2 EST-SSR的分布频率

以重复基序长度2~6 bp和最少5次重复为标准,在火炬松UniGene上共搜寻到110个SSR位点,表现为36种不同的重复类型. 这些SSR位点分布在89个UniGene上,其中43个UniGene含有2个以上的SSR位点,单个UniGene最多含有3个SSR位点. 拼接后的UniGene和原始EST含有SSR位点的频率分别为4.32%和2.08%. 就UniGene总长度1.6 Mb而言,SSR在火炬松EST中的分布频率为每14.6 kb出现1个SSR.

2.3 EST-SSR重复基序长度分布

在火炬松EST-SSR中,二核苷酸长度的重复基序是最多的,有67个,占整个SSR位点总数的60.90%;其次是三核苷酸重复基序,有40个,占36.36%;其余的是四核苷酸和六核苷酸重复基序,占的比例比较小,分别是0.91%(1个)和2.73%(3个);没有五核苷酸的重复基序(图1).

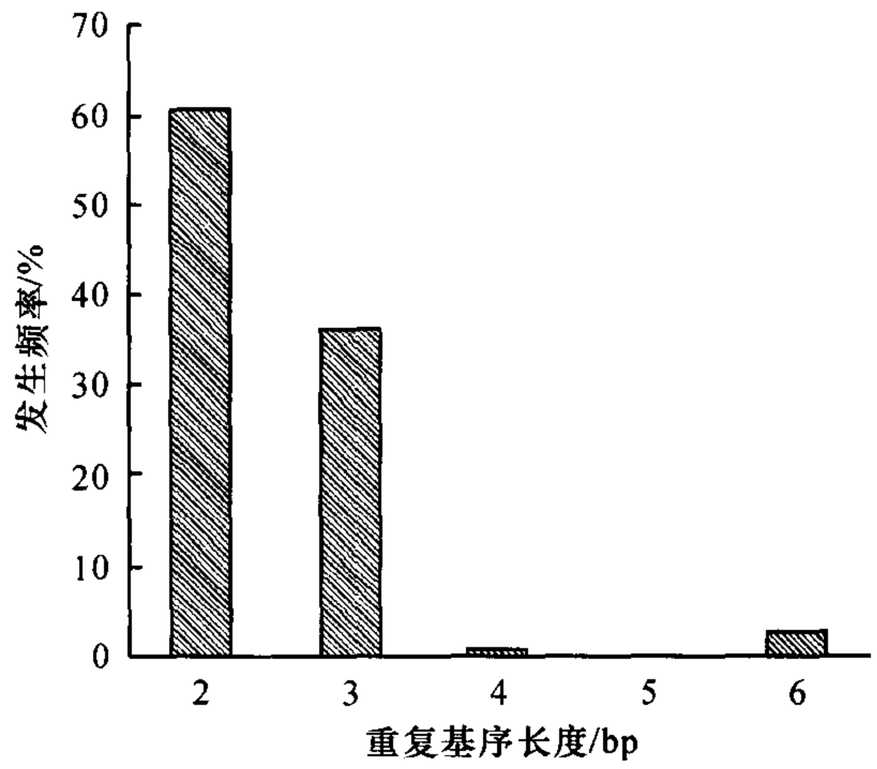

图1 火炬松EST-SSR重复基序长度的分布

Fig. 1 The motif length distribution of EST-SSR

2.4　EST-SSR 不同重复基序类型的分布

由于二核苷酸和三核苷酸重复基序占了 SSR 总数的 96% 左右,其他重复基序出现的次数很少,因此只统计二核苷酸和三核苷酸的基序类型,并比较它们的分布. 如图 2 所示,二核苷酸和三核苷酸重复基序类型出现次数的分布并不均匀. 对于二核苷酸,出现最多的基序类型是 AT,达到了 47 次,占火炬松 EST-SSR 总数的 42.73%;其次是 AG 基序类型,占总数的 14.55%. 对于三核苷酸重复基序,次数最多的是 AAG 和 AGG,均占 7.33%;其次是 ACC、CGC 和 ACG,分别占 5.45%、4.55% 和 3.64%.

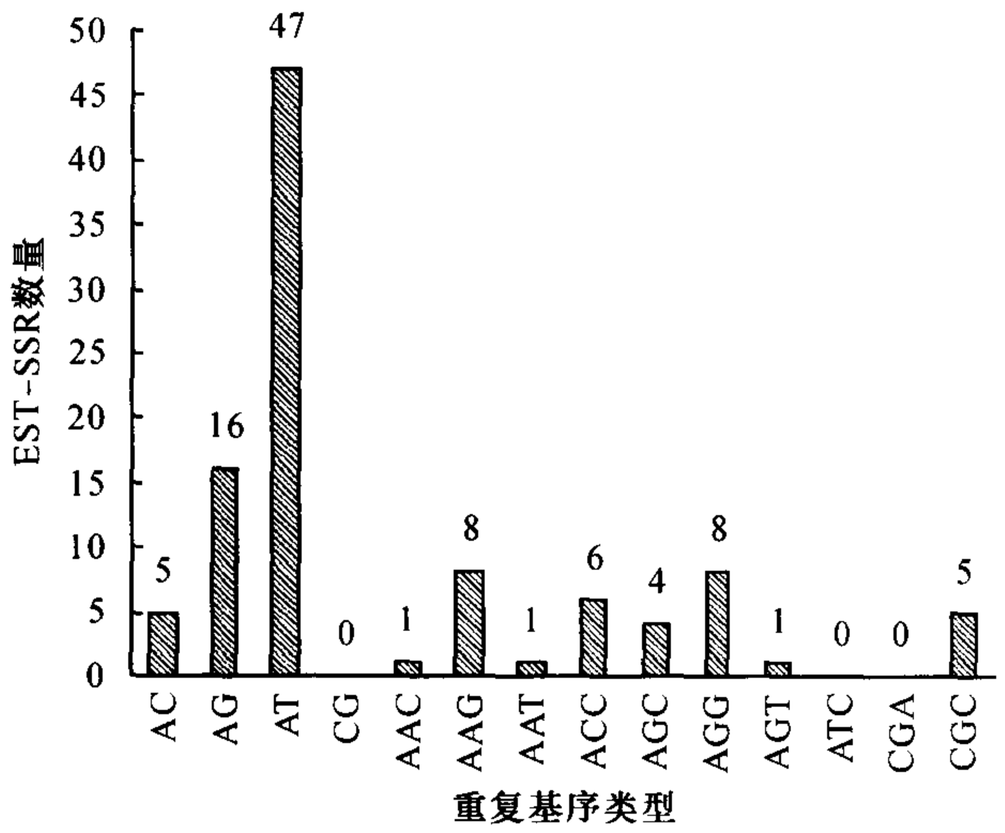

图 2　火炬松 EST - SSR 重复基序类型的分布

Fig. 2　The motif type distribution of EST - SSR

2.5　EST-SSR 位点序列长度的分布

火炬松 EST-SSR 的数量随着重复序列长度的增加而减少,在所检测到的 SSR 里,长度都在 100 bp 以内,而且 EST-SSR 位点的长度主要集中在 10 ~ 20 bp 范围内,超过 20 bp 的仅占很小比例(图 3).

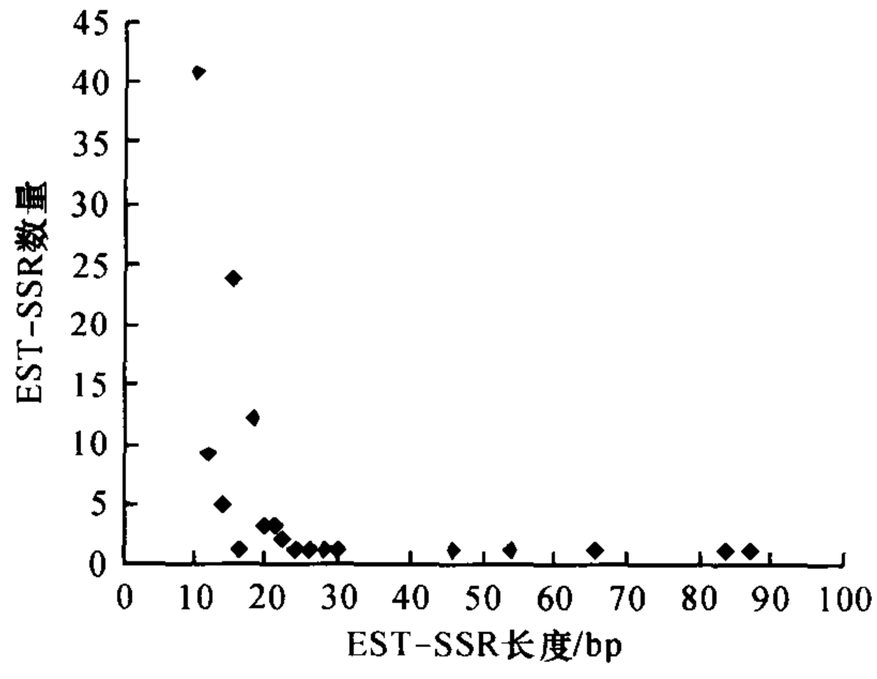

图 3　火炬松 EST - SSR 位点长度的分布

Fig. 3　The length distribution of EST - SSR loci

2.6　EST-SSR 基序重复次数的分布

将火炬松的 EST-SSR 位点按重复次数的不同进行统计后,得到图 4. 从图 4 中可以看出,随着重复次数的增加,火炬松的 EST-SSR 数量表现出逐渐递减的趋势. 其中,EST-SSR 数量最多的基序重复次数是 5 次,占总量的 62.73%;其次是 6 次重复,占 16.36%. 重复次数在 5 ~ 7 次的 EST-SSR 占到了总量的 86%. 此外,重复次数最多可达 29 次,但没有 15 ~ 21次以及 24 ~ 27 次的重复.

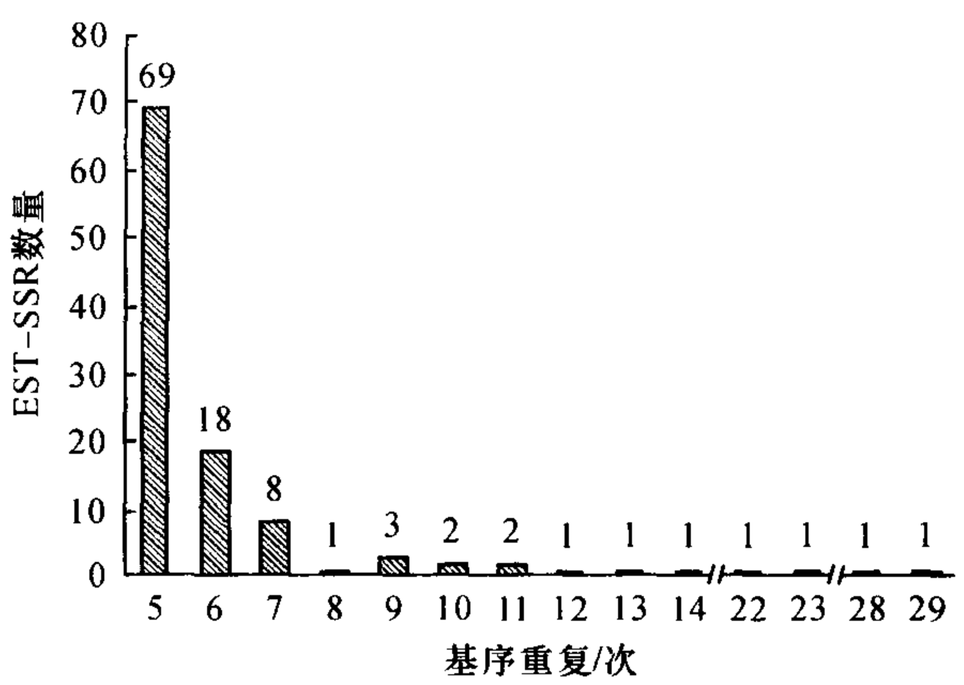

图 4　火炬松 EST - SSR 基序重复次数分布

Fig. 4　The repeating number distribution of EST - SSR motif

3　讨论

过去,开发 SSR 标记的费用比较高[3],往往只限在少数物种上;而现在可从 EST 序列中直接筛选 SSR,成本低,且简便、有效. 现已在小麦、大麦、甘蔗、玉米、高粱、水稻、番茄、葡萄等多种植物中建立了 EST-SSR 标记[3]. 基于 EST 序列开发的 EST-SSR 标记与传统的 SSR 标记相比有很多优点,如:EST-SSR 标记在相关物种之间具有高转移性;EST-SSR 标记往往也代表着某种基因或蛋白功能,可通过序列同源性分析获知;EST-SSR 还能够反映出转录区的差异,因此在种质遗传多样性分析、分子标记辅助选择和比较作图等研究中具有更高的使用价值[1-3]. 随着功能基因组学和比较基因组学的发展,EST 因为可以用于新基因的发现、基因表达调控研究和基因芯片的底物而被大量测序,并保存在公共序列数据库中,这些 EST 为 SSR、SNP 等新分子标记的开发提供了丰富的序列资源.

不同植物中 SSR 分布的频率差异很大,但 2% 以上 EST 序列都含有 SSR. 本文在 2.08% 的火炬松 EST 中搜索到 110 个 SSR,平均分布频率是每 14.6 kb出现 1 个 SSR,与小麦、番茄、棉花、拟南芥和杨树等植物相似[3],但明显小于水稻[3]、油菜[3]、猕猴桃[12]等植物,它们的平均分布频率大约在每 4 kb 出现 1 个 SSR. 不同植物 EST-SSR 的分布频率不同,

甚至差异显著,这与不同物种的基因组大小和重复DNA序列所占的比例以及基因组中转录部分的比例、低拷贝序列出现的频率有关[13]. 火炬松EST-SSR以二核苷酸重复为主,与大多数植物的EST-SSR都以三核苷酸重复为主不同[3,14],而与猕猴桃[12]、杏树和桃树[15]、茶树[16]的研究结果相一致. 这可能是不同物种基因组的组成特征不同造成的,也可能是EST-SSR位点搜寻标准不同所造成的[3].

火炬松EST-SSR中二核苷酸重复以AT、AG重复为主,这与已报道的多数植物相同[3]. 火炬松EST-SSR的三核苷酸重复中,以AAG、AGG重复最多,与大豆[17]、拟南芥[18]以及茶树[15]一致,进一步验证了Gao等[17]认为在双子叶植物中AAG重复丰度很高的推测. 这些占优势的重复基元可能与其编码相应蛋白质时使用频率较高有关,如在拟南芥中就存在此情形[19].

基序重复次数的变异引起位点序列长度的变化是EST-SSR产生多态性的主要原因,同时基序重复次数的分布也是SSR进化的表现和潜在动力[3]. SSR位点的产生是DNA复制过程中DNA聚合酶滑移的结果,而火炬松EST-SSR的重复次数86%以上都集中在5~7次重复,总体上低于基因组SSR和其他物种EST-SSR的重复次数[3],这可能是松属植物在转录区具有更严格的碱基错配修复机制的缘故.

参考文献:

[1] 忻雅,崔海瑞. 植物表达序列标签EST标记及其应用研究进展[J]. 生物学通报,2004,39(8): 4-6.

[2] 李虹,卢孟柱,蒋湘宁. 表达序列标签(EST)分析及其在林木研究中的应用[J]. 林业科学研究,2004,17(6): 804-809.

[3] 李永强,李宏伟,高丽锋,等. 基于表达序列标签的微卫星标记(EST-SSRs)研究进展[J]. 植物遗传资源学报,2004,5 (1): 91-95.

[4] TEMESGEN B, BROWN G R, HARRY D E, et al. Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.) [J]. Theor Appl Genet, 2001, 102: 664-675.

[5] KOMULAINEN P, BROWN G R, MIKKONEN M, et al. Comparing EST-based genetic maps between *Pinus sylvestris* and *Pinus taeda* [J]. Theor Appl Genet, 2003, 107: 667-678.

[6] 税珺,黄少伟,陈炳铨. 火炬松原生种源和引种群体RAPD遗传多样性[J]. 华南农业大学学报:自然科学版,2005,26(3): 74-81.

[7] 徐有明,林汉,班龙海,等. 不同环境下火炬松种源造纸材材性遗传差异与遗传稳定性分析[J]. 林业科学,2008,44(6): 157-163.

[8] 潘志刚. 我国三种主要国外松(湿地松、火炬松、加勒比松)树种及种源选择的研究初报[J]. 热带林业,1998,26(1): 12-14.

[9] PAVY N, JOHNSON J J, CROW J A, et al. ForestTreeDB: A database dedicated to the mining of tree transcriptomes [J/OL]. Nucleic Acids Research, 2007, 35: 888-894 [2008-08-20]. http://foresttree.org/ftdb.

[10] HUANG X, MADAN A. CAP3: A DNA sequence assembly program [J/OL]. Genome Research, 1999, 9: 868-877 [2008-09-15]. http://bioweb.pasteur.fr/seqanal/interfaces/cap3.html.

[11] TEMNYKH S, DECLERCK G, LUKASHOVA A, et al. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential [J/OL]. Genome Research, 2001, 11: 1441-1452 [2008-10-10]. http://www.gramene.org/db/searches/ssrtool.

[12] FRASER L G, HARVEY C F, CROWHURST R N, et al. EST-derived microsatellites from *Actinidia* species and their potential for mapping[J]. Theor Appl Genet, 2004, 108: 1010-1016.

[13] MORGANTE M, HANAFEY M, POWELL W. Microsatellites are prefertially associated with nonrepetitive DNA in plant genomes[J]. Nat Genet, 2000, 30: 194-200.

[14] VARSHNEY R K, GRANER A, SORRELLS M E. Genic microsatellite markers in plants: Features and applications [J]. Trends in Biotechnology, 2005, 23(1): 48-55.

[15] JUNG S, ABBOTT A, JESUDURAI C, et al. Frequency type distribution and annotation of simple sequence repeats in Rosaceae ESTs[J]. Funct Integr Genomics, 2005, 5: 136-143.

[16] 金基强,崔海瑞,陈文岳,等. 茶树EST-SSR的信息分析与标记建立[J]. 茶叶科学,2006,26(1): 17-23.

[17] GAO Li-fang, TANG Ji-feng, LI Hong-wei, et al. Analysis of microsatellites in major crops assessed by computational and experimental approaches [J]. Mol Breed, 2003, 12: 245-261.

[18] CARDLE L, RAMSAY L, MILBOURNE D, et al. Computational and experimental characterization of physically clustered simple sequence repeats in plants [J]. Genetics, 2000, 156: 847-854.

[19] 范三红,郭蔼光,单丽伟,等. 拟南芥基因密码子偏爱性分析[J]. 生物化学与生物物理进展,2003,30(2): 221-222.

【责任编辑 李晓卉】