

## 牡丹 EST 资源的 SSR 信息分析

侯小改<sup>1a</sup>, 王娟<sup>1a</sup>, 郭大龙<sup>1b\*</sup>, 刘改秀<sup>2</sup>, 马慧丽<sup>1b</sup>, 王小燕<sup>1b</sup>

(1.河南科技大学 a.农学院; b.林学院, 河南 洛阳 471003; 2.中国洛阳国家牡丹基因库, 河南 洛阳 471006)

**摘要:** 首先对 NCBI 数据库下载获得的 2 204 条牡丹 EST 序列进行比对, 去冗余后得到 1 658 条牡丹 EST 序列, 然后利用 MISA 软件对这些序列进行筛查, 结果在其中 901 条 EST 序列中发掘出 1 111 个 SSR, 出现频率为 67.00%, 平均每 1 004 bp 出现 1 个 SSR。在牡丹 EST-SSR 中, 单核苷酸重复是最主要的重复类型(89.38%), 其次是二核苷酸重复(6.67%)和三核苷酸重复(3.78%), 四核苷酸重复和六核苷酸重复分布极少, 没有五核苷酸重复。A/T 是优势重复基元, 占微卫星总数的 87.76%。牡丹 EST-SSR 基元类型的重复次数主要集中在 6~30 次, 其基元长度主要集中在 26~31 bp。

**关键词:** 牡丹; 表达序列标签; 简单重复序列; 频率; 特征

中图分类号: S685.11 文献标志码: A 文章编号: 1007-1032(2011)02-0172-05

## Analysis of SSR motifs in tree peony (*Paeonia suffruticosa* Andrews.) EST sequences

HOU Xiao-gai<sup>1a</sup>, WANG Juan<sup>1a</sup>, GUO Da-long<sup>1b\*</sup>, LIU Gai-xiu<sup>2</sup>, MA Hui-li<sup>1b</sup>, WANG Xiao-yan<sup>1b</sup>

(1.a.College of Agriculture; b.College of Forestry, Henan University of Science and Technology, Luoyang, Henan 471003, China; 2.Luoyang Chinese National GenBank of Tree Peony, Luoyang, Henan 471006, China)

**Abstract:** The development of genomic SSR markers is a difficult and time-cost work. EST database has become a rich source for *in silico* identification of SSRs and provided a cost-effective, labor-efficient approach for SSR marker development. Firstly, 2 204 ESTs of tree peony (*Paeonia suffruticosa* Andrews.) were downloaded from the database of NCBI and some redundant sequences were removed and 1 658 non-redundant ESTs were obtained. Subsequently, MISA, a web tool for SSR hunting, was used to analyze the SSR motifs in the EST sequences. The results showed that a total of 1 111 SSRs were detected which were distributed in 901 ESTs, with the frequency for the non-redundant ESTs being 67.00%, the average distribution distance of the EST-SSRs were about 1 004 bp. Among the different SSRs motifs, mononucleotide repeats were most abundant (89.38%), followed by dinucleotide (6.67%) and trinucleotide repeats (3.78%). Distribution of hexanucleotide repeat and tetranucleotide repeat were very few, and there is no existence of pentanucleotide. A/T was the most frequent repeat motif and accounted for 87.76% in all SSRs. Among all SSRs, The repeated times of motifs mainly concentrated in 6 to 30, and the motifs length mainly concentrated in 26 to 31 bp.

**Key words:** tree peony; expressed sequence tag(EST); simple sequence repeat(SSR); frequency; characteristics

牡丹(*Paeonia suffruticosa* Andrews.)花大色艳, 花香袭人, 有“国色天香”之称, 深受人们喜爱。

收稿日期: 2010-08-21

基金项目: 国家自然科学基金项目(NSFC31070620); 河南省高校科技创新人才支持计划项目(2010HASTTT002); 河南省重点科技攻关项目(092102110024); 河南省高等学校青年骨干教师资助计划(2010GGJS-072)

作者简介: 侯小改(1966—), 女, 河南焦作人, 博士, 教授, 主要从事园艺植物栽培生理、生物技术等方面的研究, hxg382@126.com;

\*通信作者, guodalong@mail.haust.edu.cn

牡丹不仅具有很高的观赏价值,而且还有极高的利用价值,其根可入药,花可食、酿酒、提取香精等。种质资源的保存和评价是利用牡丹的重要前提。目前用于牡丹种质鉴定的分子标记方法有 RAPD (random amplified polymorphic DNA)、AFLP (amplified fragment length polymorphisms)、基因组 SSR (simple sequence repeat) 等。这些分子标记或是扩增非编码区域,或是扩增基因组的随机片段,得到的位点一般与目标性状基因的距离较远,不利于快速应用于分子育种的实践。最近,公共数据库中数量增加的表达序列标签 (expressed sequence tag, EST) 极大地增强了对基于 EST 的 SSR 标记开发能力<sup>[1-6]</sup>。EST-SSR 是通过 EST 序列进行分析,找到含有重复单元的位置,在两侧设计引物而开发得到的,因此,EST-SSR 反映基因的编码部分,可直接获得基因的表达信息,并可对一些重要性状进行直接鉴定<sup>[5]</sup>。

迄今,在多种作物中,大量的 EST-SSR 标记已经得到了开发,并广泛用于遗传图谱的构建、基因发掘及比较基因组学的研究<sup>[7-8]</sup>中。截至 2009 年 9 月,在 NCBI 数据库中已登录 2 204 条牡丹 EST,但目前国内外还没有用这些 EST 大规模开发牡丹 SSR 的报道。笔者对现有牡丹 EST 中的 SSR 信息进行分析,以探寻牡丹 EST-SSR 的发生频率和特点。

## 1 材料与方 法

### 1.1 牡丹 EST 序列来源

登录网站 <http://www.ncbi.nlm.nih.gov/dbEST/index.html> 的 NCBI dbEST 数据库,以“peony”为关键词搜索 EST 序列。在“display”下拉框中,选择输出格式为 FASTA,在“send to”下拉框中,选择输出方式为 file,共得到牡丹(*Paeonia suffruticosa* Andrews.) EST 序列 2 204 条(截至 2010 年 1 月 31 日),保存至文本文档中备用。

### 1.2 EST 序列的前期处理

用网站 <http://www.ebi.ac.uk/clustal> 的 Clustal W

软件进行 EST 比对,剔除重复的 EST 序列。

### 1.3 SSR 位点的筛选

用网站 <http://pgrc.ipk-gatersleben.de/misa> 的 MISA 软件对优化后的 EST 序列进行 SSR 搜索,并结合手工查询。筛选标准为单核苷酸重复的次数在 10 次或 10 次以上,二核苷酸重复的次数在 6 次或 6 次以上,三至六核苷酸重复的次数在 5 次或 5 次以上,同时筛选被小于或等于 100 个碱基打断 (interrupted) 的复合型 SSR (compound microsatellite)。

### 1.4 SSR 引物的设计和评价

用网站 <http://frodo.wi.mit.edu/primer3> 的 Primer3 软件对所获得的包含有 SSR 位点的 EST 序列进行引物设计。设计引物的主要参数:引物长度为 15~25 bp; PCR 产物大小为 100~350 bp; 引物退火温度为 50~60 °C,上下引物间退火温差不超过 2 °C; (C+G) 含量为 30%~40%,最适为 50%,末端不能有 4 个连续相同的碱基,3'端结尾不能是 A 或 T 等。在 Oligo 6 中对候选引物上下游位置进行定位,锁定后用“Analyze”菜单的“PCR”选项对这对引物进行评价。出现频率=SSR 总数/无冗余 EST 总数;平均距离=无冗余 EST 总长度/SSR 总数。

## 2 结果与分析

### 2.1 牡丹 EST-SSR 的发掘

用 Clustal W 对 2 204 条牡丹 EST 序列进行冗余性查找,得到非冗余序列 1 658 条。再用 MISA 软件对优化后的牡丹序列进行 SSR 搜索,发现这些序列中有 1 111 个微卫星简单重复序列 (SSR),分布于 901 条 EST 序列中,占被调查 EST 的 67.00%。1 658 条非冗余 EST 序列拼接总长度为 1 115 499 bp,平均每 1 004 bp 出现 1 个 SSR。

### 2.2 牡丹 EST-SSR 的频率和分布密度

牡丹 EST 中微卫星含量较丰富。在 901 条 EST 中,含有单个 SSR 的 EST 有 729 条,含有 2 个或 2 个以上 SSR 的 EST 有 172 条,其中还有 102 条序

列出现 2 个 SSR 串联。本研究共检测出 1 111 个精确重复的 SSR, 占无冗余 EST 的 67.00%, 即为牡丹基因组中 EST-SSR 的出现频率。牡丹 EST-SSR 的优势重复基元为单、二和三核苷酸, 三者共占 EST-SSR 总数的 99.82%, 其中以单核苷酸所占的比例最大(89.38%), 二核苷酸重复次之(6.67%), 三核苷酸重复最少(3.78%), 四核苷酸重复和六核苷酸重复各 1 个, 没有五核苷酸重复(表 1)。

表 1 牡丹中 EST-SSR 的数量、比例和出现频率

Table 1 Number, percentage and frequency of EST-SSRs in tree peony EST sequences

重复类型	数量/个	所占比例/ %	出现频率/ %
单核苷酸	993	89.38	59.89
二核苷酸	74	6.67	4.46
三核苷酸	42	3.78	2.53
四核苷酸	1	0.09	0.06
六核苷酸	1	0.09	0.06
总计	1 111	100.00	67.00

### 2.3 牡丹 EST 中 SSR 基元类型及比例

在挖掘的牡丹 EST-SSR 中, 共观察到 16 种重复基元, 单、二、三、四和六核苷酸重复基元分别有 2、3、9、1、1 种。单核苷酸重复基元是最丰富的重复基元, 其次是二核苷酸重复基元和三核苷酸重复基元, 四核苷酸重复基元和六核苷酸重复基元出现的频率极低, 没有五核苷酸重复基元。不同类型基元的出现频率不一致, 存在明显的偏倚性。从出现的频率看, 在单碱基中, *poly A/poly T* 重复是最显著的, 单核苷酸重复类型所占的比例为 98.19%, C/G 较少, 仅有 18 个, 占 1.81%。

二碱基重复基元共有 3 种, AG/CT 丰度最高, 占该重复基元总数的 83.78%, 其次是 AC/GT 和 AT/AT, 各占 6.76% 和 9.46%; 在三核苷酸重复基元中, 最丰富的是 AAG/CTT, 共有 15 条, 占该重复基元的 35.71%。在此核苷酸重复中, 还包括 ACC/GGT、ACT/ATG、ACG/CTG、AGC/CGT、AAC/GTT、CCG/CGG、AAT/ATT、AGG/CCT 等 8 种重复基元 27 个, 共占该重复基元的 64.29%。除 AGG/CCT (仅占 2.38%) 外, 其他重复基元数均在 1

个以上。四核苷酸重复和六核苷酸重复较少, 分别只有 1 个重复基元, 且未发现五核苷酸重复(表 2)。

表 2 牡丹 EST 中主要重复基元

Table 2 Main repeat motifs in tree peony ESTs

重复类型	重复基元	数量/个	比例/ %	出现频率/ %
单核苷酸	A/T	975	87.76	58.81
	C/G	18	1.62	1.06
二核苷酸	AC/GT	5	0.45	0.30
	AG/CT	62	5.58	3.74
	AT/AT	7	0.36	0.42
三核苷酸	AAC/GTT	3	0.27	0.18
	AAG/CTT	15	1.35	0.90
	AAT/ATT	2	0.18	0.12
	ACC/GGT	6	0.54	0.36
	ACG/CTG	4	0.36	0.24
	ACT/ATG	6	0.54	0.36
	AGC/CGT	3	0.27	0.18
四核苷酸	AGG/CCT	1	0.09	0.06
	CCG/CGG	2	0.18	0.12
	ACAT/ATGT	1	0.09	0.06
六核苷酸	ACAGGGG/CCCTCT	1	0.09	0.06

### 2.4 牡丹 EST-SSR 重复次数和重复基元长度

SSR 重复次数的变异引起位点长度的变化是产生 SSR 多态性的主要原因。牡丹 EST-SSR 按重复次数可分为 3 个区间, 即 1~5 次重复为第一个区间, 6~30 次重复为第二区间, 30 次以上为第三区间。由统计结果可知, 在第一、三区间有少量核苷酸重复的分布, 牡丹 EST-SSR 主要分布在第二区间。除了五核苷酸重复没有出现外, 其余一至六核苷酸重复基元均有分布, 主要是单核苷酸重复, 其次是二核苷酸和三核苷酸重复, 最少的是四核苷酸和六核苷酸重复。

重复基元长度的变化是 EST-SSR 位点多态性的主要表现形式。从总体上看, 牡丹 EST-SSR 基元长度主要集中在 26~31 bp, 此范围几乎全是单核苷酸重复; 其次是 10~25 bp, 此范围仍是单核苷酸重复所占比例最大, 但二核苷酸重复和三核苷酸重复次数比例增多(图 1)。

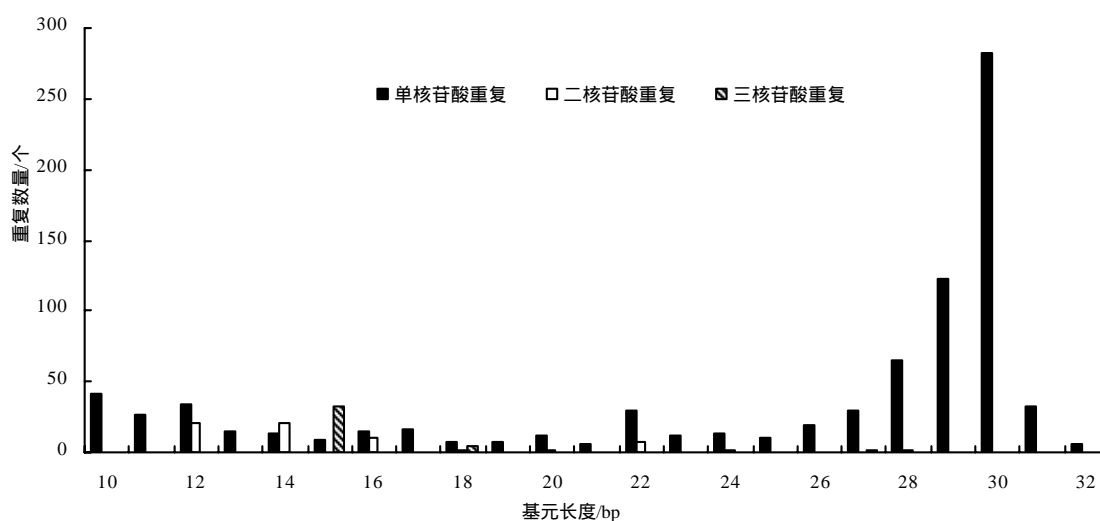


图 1 牡丹 EST-SSR 基元的重复数量

Fig.1 Distribution of EST-SSR motifs length in peony

### 3 结论与讨论

本试验中,牡丹 EST-SSR 中的主要重复类型为单、二和三核苷酸(约占总数的 99.82%),四、六核苷酸重复极少,没有五核苷酸重复,其中单核苷酸重复比例最大,占总 EST-SSR 的 89.38%。Gao 等<sup>[9]</sup>报道小麦、水稻、玉米和大豆都以三核苷酸重复为主,没有以单核苷酸重复为主的特征。这可能与牡丹的遗传特性有关,或是 EST 序列的来源不同所致。牡丹二核苷酸重复有 3 种,其中,AG/CT 占该重复类型的 83.78%,AT/AT 和 AC/GT 所占比例均不到 10.00%。在牡丹三核苷酸重复中, AAG/CTT 为优势重复基元,这与油菜<sup>[10]</sup>、木薯<sup>[11]</sup>和 水稻<sup>[12]</sup>等作物情况相同。范三红等<sup>[13]</sup>认为这些占优势的重复基元可能与其编码相应蛋白质的使用频率较高有关。无论从重复基元类型的变化还是从重复次数的变化来看,牡丹 EST 序列的 SSR 分布都不是均衡的,具有明显的偏倚性。哪种重复基元类型占主导地位,主要和该类型基元可翻译成哪几种氨基酸有关,并且这几种氨基酸在相应的物种中所含的蛋白质应该占较大的比例<sup>[13]</sup>。

Temnykh 等<sup>[14]</sup>报道,当 SSR 长度在 20 bp 以上时,在不同品种间显示出较高的多态性;当长度在

12~20 bp 时,SSR 多态性较低;当长度小于 12 bp 时,SSR 多态性很低。本研究结果显示,12 bp 以下的有 123 个,12~20 bp 的 SSR 有 168 个,20 bp 以上的有 820 个,即多态潜能高的 SSR 占 73.81%,次高的 SSR 占 15.12%,最低的占 11.07%,因此,按照 Temnykh 的观点,笔者所发掘的牡丹 EST-SSR 位点大部分都具有多态性潜能。基因组开发 SSR 标记难度大,开发成本高,费时费力,采用公共数据库登录的表达序列标签开发 EST-SSR 是一种相对简便、经济的途径。近年来,功能基因组学的发展促进了 EST 测序工作的开展,使得公共数据库中的 EST 数量迅速增长。这些大量且可以共享的 EST 序列为分子标记的开发和研究提供了丰富的资源。EST-SSR 标记相继在亚麻<sup>[15]</sup>、大麦<sup>[16]</sup>、小麦<sup>[17-18]</sup>和人参<sup>[19]</sup>等多种植物上得到开发。由于 EST-SSR 是基因的一部分,牡丹 EST-SSR 在某一牡丹群体的遗传多样性可直接反映相关基因功能的遗传多样性,这对于种质资源的收集、保存、评价及重要新基因的发掘均有重要意义,且物种间重要基因的高度保守性使 EST-SSR 具有较好的通用性,这可以有效地弥补物种分子标记的不足,丰富标记的数量,从而有利于构建高密度遗传图谱<sup>[20]</sup>。

## 参考文献:

- [1] 陈向明, 郑国生, 张圣旺. 牡丹栽培品种的 RAPD 分析[J]. 园艺学报, 2001, 28(4): 370-372.
- [2] Morgante M, Olivieri A M. PCR-amplified microsatellite as markers in plant genetics[J]. Plant J, 1993(3): 175-182.
- [3] Powell W, Machray G C, Provan J. Polymorphism revealed by simple sequence repeats[J]. Trends Plant Sci, 1993(1): 215-222.
- [4] 王显生, 贾钰莹, 顾汉艳, 等. 鹰嘴豆 EST 资源中 SSR 的信息分析[J]. 湖南农业大学学报: 自然科学版, 2010, 36(2): 137-141.
- [5] Kumpatal S P, Mukhopadhyay S. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species[J]. Genome Res, 2005, 48: 985-998.
- [6] 胡建斌, 刘颖, 王兰菊, 等. 甜瓜 EST 序列中微卫星的分布特征[J]. 植物生理学通讯, 2009, 45(3): 258-262.
- [7] 李永强, 李宏伟, 高丽峰, 等. 基于表达序列标签的微卫星标记(EST-SSR)研究进展[J]. 植物遗传资源学报, 2004, 5(1): 91-95.
- [8] 李小白, 崔海瑞, 张明龙. EST 分子标记开发及在比较基因组学中的应用[J]. 生物多样性, 2006, 4(6): 541-547.
- [9] Gao L F, Tang J F, Li H W, et al. Analysis of microsatellites in major crops assessed by computational and experimental approaches[J]. Mol Breed, 2003, 12(3): 245-261.
- [10] 李小白, 张明龙, 崔海瑞. 油菜 EST 资源的 SSR 信息分析[J]. 中国油料作物学报, 2007, 29(1): 20-25.
- [11] 彭丁文, 郑柳城, 朱宏波. 木薯 EST 资源的 SSR 信息分析[J]. 中国农学通报, 2008, 24(2): 433-436.
- [12] Kantety R V, Rota M L, Matthews D E, et al. Datamining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat[J]. Plant Molecular Biology, 2009, 48: 501-510.
- [13] 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): 221-225.
- [14] Temnykh S, DeClerck G, Lukashova A, et al. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.)[J]. Genome Research, 2001(8): 456-466.
- [15] 张建平, 王斌, 赵丽娟, 等. 亚麻 EST 序列中 SSR 标记的筛选[J]. 西北植物学报, 2009, 29(5): 910-915.
- [16] Thiel T, Michalek W, Varshney R K, et al. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.)[J]. Theor Appl Genet, 2003, 106: 411-422.
- [17] Peng J H, Nore L, Lapitan V. Characterization of EST-derived microsatellites in the wheat genome and development of SSR markers[J]. Funct Integr Genomics, 2005, 5: 80-96.
- [18] Rota L R, Kantety R V, Yu J K, et al. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley[J]. BMC Genomics, 2005, 6: 23-34.
- [19] 杨成君, 王军. 人参 EST 资源的 SSR 信息分析[J]. 植物生理学通讯, 2008, 44(1): 69-73.
- [20] 郭大龙. 牡丹种质资源遗传多样性研究进展[J]. 北方园艺, 2007(9): 61-65.

责任编辑: 王赛群

英文编辑: 罗文翠