牡丹 EST 序列中的微卫星分析

贾小平,史国安,孔祥生,范丙友,张 龙

(河南科技大学 农学院,河南 洛阳 471003)

摘 要:NCBI 的 dbEST 数据库中 2 204 条牡丹 EST 序列经过预处理获得 2 048 个高质量的序列,经拼接,获得 318 个 Contigs 和 636 个 Singlets。其中 142 条拼接序列共检测出 167 个微卫星(简单序列重复,SSR),平均 1.18 个 SSR/Contig 或 Singlet。这些微卫星从二至四核苷酸重复都有出现,二、三核苷酸重复类型占多数(97.61%),其中二核苷酸重复为 46.71%,三核苷酸重复为 50.90%。二核苷酸重复类型中出现最多的是 AG/TC 和 GA/CT,分别占 SSR 总数的 20.36% 和 16.77%,三核苷酸重复类型以 AGA/TCT、GAT/ATC 最多,分别占总 SSR 的 5.40%。最后对含有微卫星的 Contigs 和 Singlets 进行基因注释、功能分类。

关键词:牡丹;EST 序列;微卫星分布;功能注释

中图分类号:S 685.11 **文献标识码**:A **文章编号**:1001－0009(2011)17－0131－03

微卫星也称简单序列重复(Simple Sequence Repeat,SSR),广泛存在于真核生物基因组中,因其具有含量丰富、多态性高、技术简单、共显性等优点,目前已广泛应用于植物群体遗传多样性研究、遗传图谱构建、亲缘关系鉴定及 QTL 定位。目前公共数据库存贮的大量的 EST 序列为开发植物微卫星标记提供了有利条件,相比传统构建基因组文库的开发方法,利用数据库开发微卫星标记不仅省去测序所需大量费用,还加快了开发速度。目前已经在许多植物中开发了大量的 EST-SSR 标记,并且成功用于遗传图谱构建、比较基因组学研究及遗传多样性研究[1-5]。

牡丹(*Paeonia suffruticosa* Andr.)属芍药科芍药属多年生木本落叶灌木。我国作为牡丹的起源地,具有丰富的遗传资源,被称为世界牡丹王国。牡丹具有很高的观赏价值,自古以来就被我国人民当作富贵、吉祥的象征而倍受推崇。此外牡丹的花还可以用来提取香精,根皮还可以加工成中药丹皮,有清热、活血等功能,可以说牡丹在我国是最受欢迎的花卉之一。近十多年来一些分子标记如 RAPD、AFLP 已经开始用于牡丹亲缘关系鉴定、遗传多样性研究、品种鉴定、杂交种的快速鉴定[6-11]。但是以上分子标记存在不同程度的局限性,RAPD 虽然操作简单,但是稳定性差,易受环境影响;AFLP 标记多态性水平高,但是操作要求高,需要酶切、接头连接等复杂步骤,应用范围受到限制。因此开发稳定、可靠、简单方便的微卫星标记,无论对牡丹基础研究的发展,还是对牡丹野生资源的多样性评估、核心种质库的构建与保护、牡丹 QTL 定位及标记辅助育种等应用研究方面的推动都具有重要意义。该研究从公共数据库中下载的 2 204 条牡丹 EST 序列中搜索微卫星,并且对其分布规律、所代表基因的功能进行初步分析和预测,为开发牡丹 EST-SSR 标记奠定基础。

1 材料与方法

1.1 试验材料

从 GenBank/dbEST(http://www.ncbi.nlm.nih.gov/entrez)下载的 2 204 条牡丹 EST 序列为研究材料。

1.2 试验方法

1.2.1 EST 序列的预处理 利用 SEQTRIM(http://www.scbi.uma.es/cgi-bin/seqtrim/seqtrim_form.cgi)在线分析软件对所有 EST 序列进行预处理,去除载体序列、接头序列、多聚腺苷酸、串联体及小于 100 bp 的低质量序列。

1.2.2 EST 序列的聚类与拼接 利用 CAP3(http://deepc2.psi.iastate.edu/cgi-bin/cap_pm.pl)在线软件对预处理获得的高质量 EST 进行序列聚类与拼接,以便获得更长的一致性序列。

1.2.3 微卫星序列搜索 用 SSRIT(http://acorn.cshl.org/db/searches/ssrtool)微卫星搜索工具搜索拼接序列中的微卫星,所设条件:二、三核苷酸重复不少于 5 次,四至六核苷酸重复不少于 4 次,对搜索到的微卫星统计各种重复类型出现的频率。

1.2.4 基因注释及功能分类 用 BLASTX 或 BLASTN(http://blast.ncbi.nlm.nih.gov/Blast.cgi)进行基因功能注释含微卫星的 Contigs 和 Singlets,若 E 值≤1E－10,则认为查询一致序列与已知基因具有同源性,用 GENEONTOLOGY(GO)(http://www.ge-neontology.org/)进行功能分类。

第一作者简介:贾小平(1973-),男,博士,副教授,现从事植物分子遗传学方面研究工作。E-mail:jiaxiaoping2007@163.com。

基金项目:河南科技大学青年科学基金资助项目(2009QN001)。

收稿日期:2011－06－10

2　结果与分析

2.1　牡丹 EST 序列的预处理及聚类拼接

首先对 2 204 条牡丹 EST 原始序列数据进行预处理，使用的 SEQTRIM 软件可以同时去除载体污染、接头序列、多聚腺苷酸、串联体及低质量序列。处理后获得了 2 048 条高质量的 EST 序列，接着对这些高质量的 EST 序列进行聚类与拼接，获得了 318 个 Contigs 和 636 个 Singlets。

2.2　微卫星序列的搜索

经过拼接后获得的 318 个 Contigs 和 636 个 Singlets 用微卫星搜索软件查询所包含的微卫星序列，共计有 142 个 Contigs 和 Singlets 含有微卫星，所占比例为 14.9%(142/954)。从 142 个一致性序列共发现 167 个 SSR，平均 1.18 个 SSR/Contig 或 Singlet。对这些 SSR 进行分析表明，二至四核苷酸重复单元都有出现，二、三核苷酸重复类型占多数(97.61%)，其中二核苷酸重复为 46.71%，三核苷酸重复为 50.90%，四核苷酸重复类型所占比例较小，为 2.39%，没有发现五、六核苷酸重复类型(表 1)。

二核苷酸重复类型中出现最多的是 GA/TC 和 AG/CT，分别占 SSR 总数的 20.36%和 16.77%，其余的二核苷酸类型依次是 AT/TA(4.2%)、CA/TG(3.6%)、AC/GT(1.8%)，没有发现 GC/CG 重复类型。三核苷酸重复类型共发现 22 种，以 AGA/TCT、GAT/ATC 最多，分别占总 SSR 的 5.4%，其次是 GGT/ACC，占 SSR 总数的 4.80%，再往后是 AAG/CTT、TGG/CCA，各占 SSR 总数的 4.2%，出现频率最低的包括 AAT/TAA、CAC/GTG、AGT/ACT、CAA/TTG、ACA/TGT、ACG/CGT，各占 SSR 总数的 0.6%(图 1-A，图 1-B)。四核苷酸重复类型只获得 4 个：ATGT、GTAC、GTAT、AAAG。

表 1　不同重复单元分布特点

重复单元	各类重复单元数目	占总 SSR 百分数/%	各种重复单元丰度
二核苷酸	78	46.71	1SSR/8.60kb
三核苷酸	85	50.90	1SSR/7.89kb
四核苷酸	4	2.39	1SSR/167.73kb
总数	167		

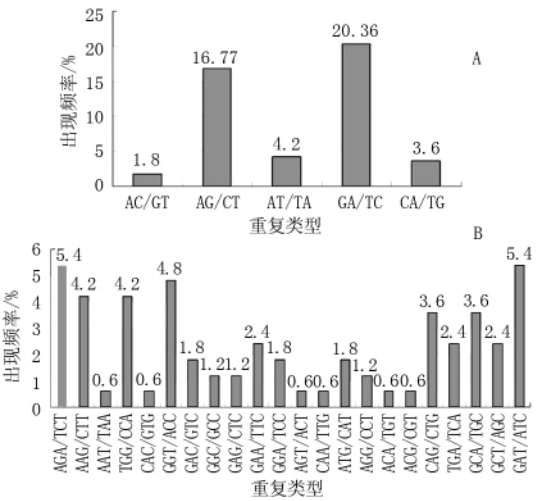

图 1　二、三核苷酸重复类型分布特点

注：A：二核苷酸重复类型分布，B：三核苷酸重复类型分布。

2.3　含有微卫星的拼接序列基因注释及功能分类

对含有微卫星的 142 条 Contigs 或 Singlets 用 NCBI 的 BLASTX、BLASTN 程序进行功能预测，有 19 个序列在数据库中找不到显著同源的基因，剩余 123 条序列都能找到相应的同源基因。这些同源基因多数来源于葡萄和蓖麻(56%)，且其中有相当比例同源基因功能尚未阐明(34%)，而功能确定的同源基因中以核糖体蛋白和核酸结合蛋白居多，占 34%(表 2)。

表 2　利用 BLASTX 和 BLASTN 推测的基因功能

牡丹 EST 拼接序列	推定基因	基因登陆号	来源生物	E 值
Contig2	金属硫蛋白类似蛋白	ACU14245.1	大豆	7e-19
Contig9	钙调蛋白	P93087.3	辣椒	6e-79
Contig17	功能未知蛋白	XP_002272660.1	葡萄	6e-128
Contig22	功能未知蛋白	XP_002263560.1	葡萄	4e-54
Contig50	水分胁迫诱导蛋白	AY571333.1	甘蓝	6e-27
Contig63	60s 核糖体蛋白	ADE42972.1	栀子	2e-94
Contig94	NADH 脱氢酶基因	XP_002512710.1	蓖麻	4e-37
Contig95	真核翻译起始因子	P24922.1	烟草	1e-80
gi|225903330	功能未知蛋白	XP_002315460.1	杨毛果	2e-52
gi|225903203	核酸结合蛋白	XP_002518811.1	蓖麻	4e-52
gi|225903194	新生多肽相关复合物	NP_177466.1	拟南芥	5e-63
gi|225903174	功能未知蛋白	XP_002516936.1	蓖麻	8e-17
gi|225903131	膜联蛋	XP_002527036.1	蓖麻	3e-42
gi|225903097	腺苷酸-活化蛋白激酶	XP_002511194.1	蓖麻	2e-82
gi|225903052	功能未知蛋白	XP_002272138.1	葡萄	3e-64

对 123 条拼接后的序列进行基因功能分类，根据基因参与的生物过程共分为 13 类：蛋白质合成、非生物胁迫应答、代谢、细胞生长分化、运输体、细胞结构、能量、胞内转运、转录、蛋白质定向储存、蛋白质降解、疾病/防卫、信号传导。123 条序列中有 68 条能够明确划分对应的类群，余下的 55 条序列则未能划分到 13 个类群中。在 13 个类群中包含基因最多的类群是蛋白质合成相关基因，有 19 个，其次是非生物胁迫应答基因，有 11 个，再往下是细胞生长分化相关基因，有 8 个，而蛋白质定向相关基因和能量产生相关基因及疾

病防卫相关基因则较少,每个类群只有 1~2 个。该研究中未能按功能进行分类的序列达到 55 个,将近一半,说明牡丹 EST 序列中有很多基因为新基因,对应的功能目前在已测序的拟南芥、水稻等植物中还没有被揭示出来,这些基因如果被克隆并弄清功能将拓宽人们对植物基因的认识水平(图 2)。

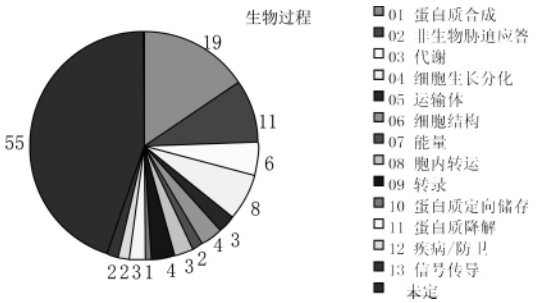

图 2 拼接序列所代表基因的功能分类

3 讨论与结论

该研究对牡丹 EST 序列微卫星分布规律进行了统计,从微卫星的分布密度来说,牡丹 EST 序列微卫星分布密度要低于油菜(ISSR/4.34kb)[2]、花生(ISSR/6.8 kb)[4]、橡胶(ISSR/3.93kb)[5]等植物。而从不同重复类型出现频率来说,三核苷酸重复出现频率略高于二核苷酸重复,与甘蔗、葡萄、柑橘、甜瓜分布特点一致[12-14];而在橡胶、茶树、猕猴桃、杏树中则是二核苷酸重复占了主导[15-17]。在牡丹中三核苷酸类型优势重复单元为 AGA/TCT、GAT/ATC 两类,而多数已报道的双子叶植物中以 AAG/TTC 重复为主;牡丹二核苷酸类型优势重复单元为 GA/TC 和 AG/CT 两类,与报道的多数植物情况相一致。该研究共发现牡丹二核苷酸基序 5 种,三核苷酸基序 13 种,低于橡胶(二核苷酸基序 6 种、三核苷酸基序 26 种),但高于花生(二核苷酸基序 3 种、三核苷酸基序 10 种),四核苷酸基序有 4 种,而橡胶、花生获得的四核苷酸基序有 5 种,该研究没有发现五、六核苷酸基序类型。由于不同研究收集的 EST 数量不同,因此分析结果可能存在一定差异,要弄清这种差异是物种之间本来存在的还是由于研究材料数目不同造成的,还需要在掌握更广泛的数据基础上做进一步研究。

参考文献

[1] 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.

[2] 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.

[3] 苏芳,郭绍贵,宫国义,等. 甜瓜基因组学研究进展[J]. 分子植物育种,2007,5(4):540-547.

[4] 梁炫强,彦彬,陈小平,等. 花生栽培种 EST-SSRs 分布特征及应用研究[J]. 作物学报,2009,35(2):246-254.

[5] 安泽伟,赵彦宏,程汉,等. 橡胶树 EST-SSR 标记的开发与应用[J]. 遗传,2009,31(3):311-319.

[6] 孟丽,郑国生. 部分野生与栽培牡丹种质资源亲缘关系的 RAPD 研究[J]. 林业科学,2004,40(5):110-115.

[7] 苏雪,张辉,董莉娜,等. 应用 RAPD 技术对甘肃栽培牡丹品种的分类鉴定研究[J]. 西北植物学报,2006(4):48-53.

[8] 侯小改,尹伟伦,李嘉珏,等. 部分牡丹品种遗传多样性的 AFLP 分析[J]. 中国农业科学,2006,39(8):1709-1715.

[9] 杨淑达,施苏华,龚洵,等. 滇牡丹遗传多样性的 ISSR 分析[J]. 生物多样性,2005(2):19-25.

[10] 周兴文,杨秋生,李永华. 牡丹基因组 AFLP 银染反应体系的建立和优化[J]. 河南农业大学学报,2006(6):34-38.

[11] 索志立. 牡丹品种鉴定用 ISSR 引物的筛选与开发[J]. 生物技术通报,2006(S1):342-346.

[12] Cordeiro G M,Casu R,McIntyre C L,et al. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum [J]. Plant Sci,2001,160:1115-1123.

[13] Chen C X,Zhou P,Choi Y A,et al. Mining and characterizing microsatellites from citrus ESTs[J]. Theor Appl Genet,2006,112:1248-1257.

[14] 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.

[15] 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,2006,26:17-23.

[16] 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.

[17] 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.

Analysis of Microsatellites in Tree Peony EST Sequences

JIA Xiao-ping,SHI Guo-an,KONG Xiang-sheng,FAN Bing-you,ZHANG Long
(College of Agriculture,Henan University of Science and Technoloy,Luoyang,Henan 471003)

Abstract: After precleaning 2 204 tree peony(*Paeonia suffruticosa* Andrews) expressed sequence tags(ESTs) from dbEST of NCBI,2 048 high-quality ESTs were obtained. 318 contigs and 636 singlets were obtained by assembling of these 2 048 ESTs. 167 microsatellites(simple sequence repeats,SSRs) were found in 142 of the 954 assembled sequences,on average 1.18 SSRs per assembled sequence. The repeat unit of SSRs included from dinucleotide repeat to tetramer repeat,most of the SSRs were dinucleotide repeat type and trinucleotide repeat type(97.61%),the two repeat types accounted for 46.71% and 50.90% respectively. Among dinucleotide repeat type,AG/TC and GA/CT were the most(20.36% and 16.77%),while in trinucleotide repeat type,AGA/TCT,GAT/ATC were the most (5.40% and 5.40%). Finally,bioinformatic analysis such as gene annotation,gene function group were done on contigs and singlets containing SSR.

Key words: tree peony; EST sequence; microsatellite distribution; function annotation