

丹参基因组 SSR 位点的特征分析及引物开发^{*}

□徐海滨 钱 俊 汪 波 孙 超

(中国医学科学院 药用植物研究所 北京 100193)
(北京协和医学院)

谢丽芳(中国医学科学院高性能计算平台 北京 100730)

宋经元^{**}(中国医学科学院 药用植物研究所 北京 100193)
(北京协和医学院)

摘 要 :丹参为唇形科鼠尾草属多年生落叶开花植物 ,以根与根茎部位入药 ,广泛用于多种疾病的治疗。本研究采用生物信息学方法 ,从丹参基因组框架图序列中发掘出 4 832 个长度超过 40 bp 的基因组 SSR 位点。结果表明 ,二核苷酸重复基元和三核苷酸重复基元是构成丹参基因组 SSR 位点的主要类型 ,分别占全部 SSR 位点总数的 37.3%和 61.3%。此外 ,富含 A/T 碱基的 SSR 位点 ,包括 AT/TA、AAT/ATT、ATA/TAT、TAA/TTA ,分别占全部 SSR 位点总数的 30.5% ,21.6% ,17.1% ,20.4% ,在数量上明显高于其他类型。以基因组 SSR 位点为基础 ,共设计基因组 SSR 引物序列 1 079 对。这些引物可用于基因组多样性分析、遗传图谱构建、分子标记筛选等工作 ,为后期丹参群体遗传学与基因组学研究奠定基础。

关键词 :丹参 简单重复序列 引物设计

doi: 10.11842/wst.2013.03.006 中图分类号 :R282 文献标识码 :A

丹参(*Salvia miltorrhiza* Bge.)是唇形科鼠尾草属多年生落叶开花植物 ,以干燥根及新鲜根茎入药。丹参广泛分布于我国大部分地区 ,具有活血化瘀、通经止痛的功效 ,应用于心脑血管疾病及高脂血症等疾病的治疗 ,具有较高的经济和药用价值^[1-3]。目前 ,我国栽培丹参品种的产量和品质及不稳定 ,成

为制约丹参产业发展的严重问题^[4]。通过育种手段培育产量稳定且药用成分含量高的丹参品种 ,对丹参生产具有重要意义。

分子标记辅助育种的方法已广泛应用到大田作物的品种改良实践中 ,如水稻、大豆、小麦、玉米等。筛选与重要农艺性状紧密连锁的分子标记是分子标记辅助育种的重要工作。目前 ,常用的分子标记类型包括限制性片段长度多态性(RFLP)、随

收稿日期 :2013-05-24

修回日期 :2013-06-04

^{*} 科学技术部国家科技支撑计划(2012BAI29B01) :基于本草基因组的中药材活性成分合成及育种示范研究 ,负责人 :宋经元 ,教育部长江学者与创新团队发展计划(IRT1150) :中药资源学 ,负责人 :陈士林。

^{**} 通讯作者 :宋经元 ,研究员 ,博士生导师 ,主要研究方向 :药用植物资源学。

机扩增多态性 DNA(RAPD)、酶切扩增多态性序列 (CAPS)、单核苷酸多态性 (SNP)、简单重复序列 (SSR)等。其中,SSR 序列又称为微卫星序列,是由 1~6 个核苷酸为重复基元重复多次组成的 DNA 序列^[5]。研究表明,相同位点的 SSR 重复基元的重复次数在不同个体间变异较大,使得其可作为筛选品种间遗传多样性的重要分子标记来源。相对与其他类型的分子标记,SSR 标记为共显性标记,其实验操作简单且在基因组中分布广泛,目前已成为一类广泛应用的分子标记。

依 SSR 的序列来源不同可将其分为两类:基因组 SSR 和表达序列标签 (EST)SSR。目前已有报道利用美国 NCBI (National Center for Biotechnology Information) 公共数据库中的丹参及近缘种的 EST 序列开发 EST-SSR 标记,并将其应用于遗传多样性分析^[6,7]。相对与转录组 SSR,基因组 SSR 的序列重复次数更多、分布平均且多态率更高,使其可作为基因组多样性分析、遗传图谱构建、标记筛选等工作的主要分子标记来源^[8],但应用传统手段筛选基因组 SSR 序列需要构建小片段基因组文库,所需时间较长且费用较高。相比之下,通过生物信息学手段可快速获取具有全基因组框架图序列的基因组 SSR 位点。

目前,丹参基因组框架图序列和丹参叶绿体基因组完成图序列已完成^[9,10]。本研究通过分析目前已有的丹参基因组框架图序列,筛选出了其中 4 832 个基因组 SSR 序列位点,并分析了其分布特征。以此为基础,开发了 1 079 对基因组 SSR 引物,为丹参分子标记辅助育种工作提供基础。

一、材料和方法

1. 丹参基因组 SSR 的筛选及统计分析

以丹参基因组序列框架图为序列基础,应用 MISA 工具*搜索其中的 SSR 位点,搜索参数为:二核苷酸、三核苷酸、四核苷酸、五核苷酸和六核苷酸重复基元的最

少重复次数分别为 20 次,15 次,10 次,10 次和 10 次。因单核苷酸重复基元的 SSR 位点的实际应用较少,本研究未就其开展分析。利用 Mysql**和 Microsoft Excel 统计上述筛选结果。

2. SSR 引物设计

以筛选得到的 SSR 位点为基础,利用 EMBOSS 软件[△]提取 SSR 位点上、下游各 300 bp 序列。利用 primer3 软件^{△△}在默认参数下对每一个 SSR 位点设计引物。

二、结果与分析

1. 丹参基因组 SSR 重复基元特征分析

在丹参基因组框架图中,我们筛选出 62 种 2~6 核苷酸为重复基元的 SSR 位点共 4832 个。我们发现不同核苷酸重复基元的 SSR 数量间存在明显差异(表 1)。其中,二核苷酸重复基元的 SSR 位点 1 804 个,占 SSR 位点总数的 37.3%;三核苷酸重复基元的 SSR 位点 2 965 个,占 SSR 位点总数的 61.3%;四、五、六核苷酸重复基元的 SSR 位点分别出现 34、14、15 次,共仅占 SSR 位点总数的 1.4%。不同重复基元的 SSR 位点的平均长度在 45~76 bp 间(表 2)。

我们注意到,在占 SSR 位点总数 98.6%的二、三核苷酸重复基元的 SSR 位点中,不同重复基元 SSR 位点的数量间差异仍然较大(表 2),且存在较高 GC 含量偏差。所有全部由 A/T 碱基构成的 SSR 位点,包括 AT/TA、AAT/ATT、ATA/TAT、TAA/TTA,分别占全部 SSR 位点总数的 30.5%、21.6%、17.1%、20.4%,在数量上明显高于其他类型。相对应,我们并没有筛选出全部由 G/C 碱基构成的 SSR 位点。

表 1 丹参基因组 SSR 位点信息统计表

重复类型	基元类型数量	SSR 位点数量	占全部 SSR 位点的比例
二核苷酸	5	1 804	37.3%
三核苷酸	15	2 965	61.3%
四核苷酸	18	34	0.7%
五核苷酸	9	14	0.2%
六核苷酸	15	15	0.3%

* <http://pgsc.ipk-gatersleben.de/misa/>

** <http://www.mysql.com/>

△ <http://emboss.sourceforge.net/>

△△ <http://primer3.sourceforge.net/>

表 2 丹参基因组二、三核苷酸重复基元统计表

	基元类型	最小重复次数	最大重复次数	SSR 总数	重复总数	SSR 平均长度	占全部 SSR 位点比例
二核苷酸重复	AC/GT	20	27	12	268	45	0.2%
	AG/CT	20	60	159	3 955	50	3.3%
	AT/TA	20	76	1 474	32 602	44	30.5%
	CA/TG	20	30	43	959	45	0.9%
	GA/TC	20	66	116	2 936	51	2.4%
三核苷酸重复	AAC/GTT	15	25	20	379	57	0.4%
	AAG/CTT	15	22	8	138	52	0.2%
	AAT/ATT	15	986	1 042	26 505	76	21.6%
	ACA/TGT	17	20	5	90	54	0.1%
	ACT/AGT	17	19	3	54	54	0.1%
	AGA/TCT	15	28	11	196	53	0.2%
	ATA/TAT	15	357	827	20 006	73	17.1%
	ATC/GTA	15	20	12	203	51	0.2%
	ATG/CAT	15	22	10	183	55	0.2%
	CAA/TTG	15	25	17	317	56	0.4%
	CTA/TAG	21	21	1	21	63	0.0%
	GAA/TTC	15	21	12	204	51	0.2%
	GAT/ATC	15	28	7	138	59	0.1%
	TAA/TTA	15	156	984	24 191	74	20.4%
	TCA/TGA	15	27	6	112	56	0.1%

表 3 丹参基因组 SSR 引物序列示例表

基元类型	SSR 长度	左引物序列(5'-3')	右引物序列(5'-3')
GAT/ATC	45	AGGGTGAATCTTGGGGTGTT	AATATTCATGGCTCGGCTTG
TAA/TTA	66	TGAATTGGCTGAGAAGAGTGAG	AACGAATGGCATAGGTCAC TG
AG/CT	40	ACCCACCGTTCTCTTCT	CTTGGAATGGATGTTACGAA
TAA/TTA	60	CGGATTGAGTTGGATTTCGTT	ATGCTGTCAAGGGTGCTTTT
ATA/TAT	99	CGAAAGTCTAGCTCTGCCACT	ATGGTGGGGGAATGTGACTA
AAT/ATT	72	CATGTATTCTGGCCTTCACG	TGGAGCTTCAGCTATTTGAGTG
AG/CT	42	CGAAATTGCCAATGTGACAG	CCCCAATCAAAACCTTACC
AAT/ATT	45	ATCCATCCCCAGTCTTTGCT	GGATTGGGATGGAGGGAATA
AT/TA	40	ATCGATATGCGTGAGGAAGG	GCTGAACAGACGAAGGAAGC

2. 基因组 SSR 引物设计

为进一步方便将筛选出的基因组 SSR 位点应用于基因组多样性分析、遗传图谱构建、紧密连锁标记筛选等分子育种相关的研究中,我们利用基因组 SSR 位点及其上下游各 300 bp 基因组序列设计引物以扩增基因组 SSR 序列。结果表明,共有 1 079

个基因组 SSR 位点成功设计出引物,占基因组 SSR 位点总数的 22%。表 3 展示了设计成功的基因组 SSR 引物示例。

三、讨 论

本研究通过对丹参基因组框架图数据分析,筛

选出 4 832 个丹参基因组 SSR 位点。这些基因组 SSR 位点在位点数量和序列长度上均明显优于前人利用 EST 序列筛选 EST-SSR 位点的分析结果。由于长片段 SSR 位点在进化中有更高的概率产生长度变异,这些基因组 SSR 位点的筛选更加有利于丹参群体基因组多态性的发掘,为丹参分子育种工作提供重要资源。

在我们的筛选结果中,不同重复基元 SSR 位点的数量间存在较大差异且存在 GC 含量偏差。原因除基因组本身含有较少的高 GC 基因组 SSR 位点外,还有可能是由于基因组测序过程中可能存在较大的 GC 含量偏差,高 GC 含量区域并未包含于所用的基因组框架图序列中。

SSR 引物在品种间的通用性较高,因此针对测序品种获得的 SSR 位点可有效用于其他丹参品种的多态性筛选,甚至可用于丹参近缘种的研究。本研究成功针对 1 079 个基因组 SSR 位点设计出引物,为进一步将高通量测序的结果应用于丹参分子育种研究中奠定了基础。

参考文献

1 Zhong G X, Li P, Zeng L J, *et al.* Chemical characteristics of

Salvia miltiorrhiza (Danshen) collected from different locations in China. *J Agric Food Chem*, 2009, 57(15):6879~6887.

2 Zhou L, Zuo Z, MS C. Danshen: an overview of its chemistry, pharmacology, pharmacokinetics, and clinical use. *J Clin Pharmacol*, 2005, 45(12):1345~1359.

3 Cheng T. Cardiovascular effects of Danshen. *Int J Cardiol*, 2007, 121(1):9~22.

4 郝岗平, 孙立彦, 史仁玖, 等. 山东产丹参遗传多样性的扩增片段长度多态性指纹分析. *时珍国医国药*, 2007, 18(1): 51~53.

5 Litt M, Luty J A. A Hypervariable Microsatellite Revealed by *In Vitro* Amplification of a Dinucleotide Repeat within the Cardiac Muscle Actin Gene. *Am J Hum Genet*, 1989, 44(3):397.

6 宋杰, 严铸云, 马云桐, 等. 丹参 EST 序列中 SSR 信息的分析及分子标记的建立. *世界科学技术-中医药现代化*, 2009, 11(1):58~63.

7 邓科君, 张勇, 熊丙全, 等. 药用植物丹参 EST-SSR 标记的鉴定. *药学报*, 2009, 44(10):1165~1172.

8 Piquemal J, Cinquin E, Couton F, *et al.* Construction of an oilseed rape (*Brassica napus* L.) genetic map with SSR markers. *Theor Appl Genet*, 2005, 111(8):1514~1523.

9 Qian J, Song J, Gao H, *et al.* The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS One*, 2013, 8(2):e57607.

10 Ma Y, Yuan L, Wu B, *et al.* Genome-wide identification and characterization of novel genes involved in terpenoid biosynthesis in *Salvia miltiorrhiza*. *J Exp Bot*, 2012, 63(7):2809~2823.

Characterization and Primer Design for Simple Sequence Repeat Loci in *Salvia miltiorrhiza* Genome

Xu Haibin¹, Qianjun¹, Wang Bo¹, Sun Chao¹, Xie Lifang², Song Jingyuan¹

(1. Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100193, China;

2. High Performance Computing Platform of Chinese Academy of Medical Sciences, Beijing 100730, China)

Abstract: *Salvia miltiorrhiza* Bge. is a perennial deciduous flowering plant. Its medicinal root and rhizomes part is widely used in the treatment of various diseases. In this study, bioinformatics analysis was performed to identify 4832 genome SSR loci with length longer than or equal to 40 bp from the draft genome assembly of *S. miltiorrhiza*. The results showed that the dinucleotide repeat motifs and trinucleotide repeat motifs constitute the main types of genome SSR loci, accounting for 37.3% and 61.3% respectively. SSR types enriched with A/T bases showed significantly higher abundance than other types, including AT/TA AAT/ATT, ATA/TAT, TAA/TTA, accounting for 30.5%, 21.6%, 17.1%, 20.4% of the total number of SSR loci, respectively. 1079 primer pairs were designed for these genome SSR loci. These primers can be used for genomic diversity analysis, genetic map construction, genetic marker screening. These data could lay the foundation for population genetics and genomics research of *S. miltiorrhiza*.

Keywords: *Salvia miltiorrhiza*, simple sequence repeat, primer design

(责任编辑 李沙沙 张志华, 责任译审 陈士林)