党参转录组中 SSR 位点信息分析

王 东，曹玲亚，高建平*
山西医科大学药学院，山西 太原 030001

摘 要：目的 采用生物信息学方法分析党参转录组文库 EST 序列简单重复序列（SSR）位点，快速、大规模鉴定党参功能性 SSR。**方法** 使用 MicroSAtellite（MISA）软件分析党参高通量转录组 SSR 的分布频率和重复基元的类型特征，利用软件 Primer3 设计引物，并通过 SSRFinder 校验 SSR，筛选 SSR 引物。**结果** 从 45 511 条 Unigenes 中共搜到 7 327 个 SSR 位点，分布在 6 017 条 Unigenes 序列中，发生频率为 12.22%，共有 415 种重复基元，平均每 4 520 bp 含 1 个 SSR 位点，二核苷酸重复占主要地位，发生频率为 58.67%，在所有重复基元中，AG/CT 出现频率最高。共获得 4 329 条 SSR 引物。**结论** 大规模的 SSR 分子标记开发将有助于党参遗传多样性与分子育种研究。

关键词：党参；转录组；SSR；MicroSAtellite 软件；引物
中图分类号：R282.12 **文献标志码：**A **文章编号：**0253 - 2670(2014)16 - 2390 - 05
DOI: 10.7501/j.issn.0253-2670.2014.16.021

Data mining of simple sequence repeats in *Codonopsis pilosula* transcriptome

WANG Dong, CAO Ling-ya, GAO Jian-ping
College of Pharmacy, Shanxi Medical University, Taiyuan 030001, China

Abstract: Objective To rapidly and largely identify the functional simple sequence repeat (SSR) in transcriptome dataset of *Codonopsis pilosula* based on bioinformatics analysis. **Methods** MicroSAtellite (MISA) sofeware was used to analyze the distributing frequency of high-flux transcriptome SSR and the basic characteristics of repeat motifs. Primers were designed using Primer3, and SSRFinder was used to check SSR and screen the SSR primers. **Results** SSR loci (7 327) were obtained from 45 511 unigenes, distributed in 6 017 unigenes (12.22%). There are 415 kinds of repeat motifs existing in *C. pilosula* transcriptome. On average, SSRs occurred every 4 520 bp in length. Dinucleotid repeats predominated with an occurrence frequency of 58.67%, and AG/CT was the most frequent one among all the repeat types. A total of 4 329 primer pairs were designed for marker development. **Conclusion** The large number of SSR genetic markers developed in the present study should contribute greatly to research into genetic diversity and germplasm characterization in *C. pilosula*.

Key words: *Codonopsis pilosula* (Franch.) Nannf; transcriptome; simple sequence repeat; MicroSAtellite; primer

党参为桔梗科党参属党参 *Codonopsis pilosula* (Franch.) Nannf、素花党参 *C. pilosula* Nannf. var. *modesta* (Nannf.) L. T. Shen 或川党参 *C. tangshen* Oliv. 的干燥根，性平、味甘，具有补中益气、健脾益肺之功效[1]。党参产地众多，种类各异，主产于山西、甘肃、陕西、四川、湖北等省，主要为栽培品，少量为野生。目前我国党参资源破坏严重，野生资源逐渐变得稀少，急需对药源进行保护[2]。利用分子标记对其遗传学背景进行研究，研究结果可对制定合理有效的保护策略提供科学依据。

目前尚缺乏党参遗传和基因组信息研究，遗传多样性研究和分子标记研究也只有张建清等[3]利用 RAPD 分析甘肃栽培的党参和素花党参在居群水平上的遗传多样性；李忠虎等[4]利用 g-SSR 技术分析发现党参野外居群具有较丰富的遗传多样性水平，并且筛选出 5 对 g-SSR 引物能够在党参 4 个近缘物种成功扩增；赵莎等[5]曾采用党参药材中提取的 ITS/ITS2 序列对党参药材及其混淆品进行鉴定，但不能成功鉴别党参 3 个基原物种；He 等[6]利用 ITS/ITS2 序列分析从甘肃和重庆搜集的 65 个党参样

收稿日期：2014-03-24
基金项目：国家自然科学基金资助项目（81072987）；国家“十二五”科技支撑计划（2011BAI07B07）
作者简介：王 东（1989—），男，山西阳曲人，在读硕士，研究方向为中药资源及其活性成分研究。
***通信作者** 高建平（1963—），女，山西阳泉盂县人，教授，硕士研究生导师，研究方向为中药资源及其活性成分研究、中药鉴定新技术与中药质量评价。Tel: (0351)4690345 E-mail: jpgao123@sina.com

本，成功鉴别党参 3 个基原物种；苏强等[7]初步建立党参 AFLP 分析体系。目前还缺乏能够大量应用于党参种质鉴定、遗传图谱构建、功能基因定位等研究的简便、高效、稳定且具有种属特异性的分子标记体系。

简单重复序列（simple sequence repeat，SSR）又称微卫星 DNA，SSR 标记是较为理想的遗传标记，具有数量丰富、分布广泛、多态性强和共显性遗传易于分析等特点[8]。SSR 按来源分，有基因组 SSR（g-SSR）和转录组来源的 SSR（EST-SSR），相对于 g-SSR，EST-SSR 标记无需构建基因组文库、杂交、测序，避免了大量人力、物力和时间的投入，同时 EST-SSR 多态性可能与基因功能直接相关，在相近植物间具有良好通用性[9]，而且转化率也高。目前药用植物中已有人参[10]、丹参[11]、西洋参[12]和金银花[13]等借助现有测序数据开发了 EST-SSR。研究主要集中于遗传多样性评价、种质鉴定、标记通用性等方面。为此，本研究基于党参转录组文库高通量测序结果，利用 MicroSAtellite（MISA）[14]软件搜索党参 SSR 位点，分析其分布、组成特征，并进行初步可用性评价，以期为党参遗传图谱构建、遗传多样性分析、功能基因定位和比较作图等研究提供理论基础。

1 材料和方法

1.1 党参转录组数据来源

转录组测序样本采自于山西省晋城市陵川县党参 GAP 基地，原植物由山西医科大学药学院高建平教授鉴定为党参 *Codonopsis pilosula* (Franch.) Nannf。党参转录组数据来自于课题组前期利用 Illumina 高通量测序平台对党参进行的全转录组测序，共有 45 511 条 Unigenes。

1.2 党参 EST-SSR 的筛选

为了检测党参中的 SSR 位点，使用软件 MISA[14]对组装得到的 Unigene 序列进行 SSR 分析。所检测 SSR 位点包括单核苷酸重复、二核苷酸重复、三核苷酸重复、四核苷酸重复、五核苷酸重复和六核苷酸重复 6 类。判断标准为单核苷酸重复至少 12 次；二核苷酸重复至少 6 次；三核苷酸重复、四核苷酸重复至少 5 次；五核苷酸重复和六核苷酸重复至少 4 次。

1.3 党参 EST-SSR 引物设计

使用软件 Primer3 对 SSR 重复单元前后的序列设计引物，每条 SSR 产生 5 条引物。主要的引物设置参数如下：扩增产物长度 80～300 nt，引物序列长度 18～28 nt（最适宜长度为 23 nt），退火温度为 55～65 ℃（最佳退火温度为 60 ℃），GC 量为 40%～65%（最适宜为 50%）。

1.4 党参 EST-SSR 引物筛选

将设计出的引物通过以下方式筛选：（1）引物不能存在 SSR；（2）将获得的引物比对到 Unigene 序列，引物的 5’端允许有 3 个碱基的错配，3’端允许有 1 个碱基的错配；（3）去掉比对到不同 Unigene 上的引物，筛选唯一匹配的引物；（4）使用 SSRFinder 校验 SSR，使用产物序列来寻找 SSR，检验结果是否与 MISA 结果相同，并筛选出相同的 SSR 产物。

2 结果与分析

2.1 党参转录组中的 SSR 位点的数量与分布

利用软件 MISA 对党参转录组中的 45 511 条 Unigenes 的 cDNA 序列进行搜索。结果显示，7 327 个 SSR 分布于 6 017 条 Unigenes 序列中，SSR 发生频率（含有 SSR 的 Unigene 数目与总 Unigene 的数目之比）为 12.22%，出现频率（检出 SSR 个数与总 Unigene 数目之比）为 16.10%。平均每 4 520 bp 就含有 1 个 SSR 位点（表 1）。其中 5 006 条 Unigenes 只包含单个 SSR 位点，1 011 条 Unigenes 包含 2～4 个 SSR 位点。

党参 EST-SSR 重复类型丰富，单核苷酸至六核苷酸重复都可以发现。从 SSR 位点数量上看，出现最多的为一至三核苷酸重复，占到总 SSR 位点数量的 94.1%。其中二核苷酸重复比例最高，可以占到 58.7%；其次为三核苷酸和单核苷酸重复，分别为 21.6%和 13.8%。四、五、六核苷酸重复类型的数量很少，总计不足 6%（表 1）。

表 1 党参 EST-SSR 不同重复基元分布情况

Table 1 Distribution of different repeat motifs in *C. pilosula* transcriptome

重复类型	SSR 数	占总 SSR 比例 / %	发生频率 / %	平均长度 / bp	总长度 / bp	平均距离 / kb
单核苷酸	1 012	13.81	2.22	15.42	15 601	32.74
二核苷酸	4 299	58.67	9.45	16.03	68 896	7.71
三核苷酸	1 585	21.63	3.48	16.75	26 547	20.91
四核苷酸	71	0.97	0.16	20.85	1 480	466.70
五核苷酸	109	1.49	0.24	20.64	2 250	304.00
六核苷酸	251	3.43	0.55	24.00	6 024	132.01
总计	7 327	100.00	16.10	16.49	120 798	4.52

党参转录组中 SSR 位点的序列总长度达到 120 798 bp，SSR 位点平均长度为 16.49 bp，各类型 SSR 位点的平均长度分别为 15.42、16.03、16.75、20.85、20.64、24.00 bp。党参转录组中 SSR 位点重复次数以 6 次（1 466）最多，占总 SSR 的 20%；其次为重复次数 7、5、10、9 和 8 次，SSR 位点个数在 706～1 006。统计 4～10 次重复的 SSR 位点共有 6 101 个，占 83.2%；11～24 次重复的 SSR 位点共有 1 226 个，占 16.7%（图 1）。

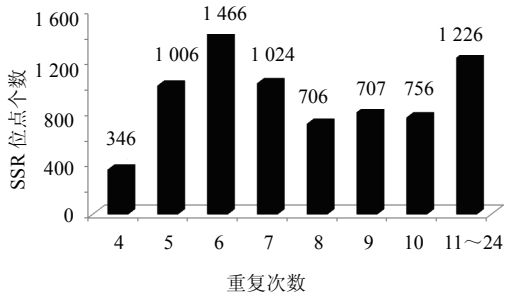

图 1 党参 EST-SSR 重复次数分布图

Fig. 1 Distribution of EST-SSR repeat frequency in *C. pilosula* transcriptome

2.2 党参 EST-SSR 的特性

在党参转录组 7 327 个 SSR 位点中，共发现到 415 种重复基元，其中一至六核苷酸重复基元分别有 2、4、10、18、48 和 122 种（表 2）。以二核苷酸重复基元 AG/CT 最多，总 SSR 的 43.78%；其次为二核苷酸重复基元 AC/GT 和单核苷酸重复基元 A/T，分别占 11.83%和 10.65%；二核苷酸重复基元中，以 AG/CT、AC/GT 和 AT/AT 出现的数量最多，三者共占二核苷酸 SSR 的 99.88%，CG/CG 出现次数最少，仅占到二核苷酸 SSR 的 0.12%。在三核苷酸重复基元中，

表 2 党参 EST-SSR 重复基元的类型

Table 2 EST-SSR repeat motifs in *C. pilosula* transcriptome

重复类型	重复基元	重复个数	发生频率 / %	所占比例 / %
单核苷酸	A/T	780	1.71	10.65
	C/G	232	0.51	3.17
二核苷酸	AG/CT	3 208	7.05	43.78
	AC/GT	867	1.91	11.83
	AT/AT	219	0.48	2.99
	CG/CG	5	0.01	0.07
三核苷酸	AAG/CTT	439	0.96	5.99
	ACC/GGT	311	0.68	4.24
	ATC/ATG	226	0.50	3.08
	AGG/CCT	180	0.40	2.46
	AGC/CTG	122	0.27	1.67
	AAC/GTT	112	0.25	1.53
	CCG/CGG	72	0.16	0.98
	AAT/ATT	59	0.13	0.81
	ACG/CGT	33	0.07	0.45
	ACT/AGT	31	0.07	0.42
四核苷酸	AAAG/CTTT	13	0.03	0.18
	ACAT/ATGT	12	0.03	0.16
	AAAT/ATTT	7	0.02	0.10
	AGCC/CTGG	7	0.02	0.10
	AAAC/GTTT	5	0.01	0.07
	剩余四核苷酸	27	0.06	0.37
五核苷酸	AAAAG/CTTTT	10	0.02	0.14
	AAAAT/ATTTT	10	0.02	0.14
	AAAAC/GTTTT	8	0.02	0.11
	AAGAG/CTCTT	8	0.02	0.11
	AAATC/ATTTG	6	0.01	0.08
	剩余五核苷酸	67	0.15	0.91
六核苷酸	AAAGAG/CTCTTT	11	0.02	0.15
	AAAAAG/CTTTTT	7	0.02	0.10
	AAGAGG/CCTCTT	7	0.02	0.10
	AAGCCC/CTTGGG	7	0.02	0.10
	AAGGAG/CCTTCT	7	0.02	0.10
	剩余六核苷酸	212	0.47	2.89

AAG/CTT 出现次数最多占三核苷酸 SSR 的 27.7%；其次是 ACC/GGT、ATC/ATG 和 AGG/CCT，分别为 19.62%、14.26%和 11.36%。其他四核苷酸、五核苷酸和六核苷酸重复基元类型较多，数量较少，出现频率均较低。

2.3 党参 EST-SSR 的可用性评价

SSR 分子标记的多态性是判断其可用性的重要依据[15]。SSR 的长度是影响其多态性高低的重要因素，当 SSR 长度大于或等于 20 bp 时多态性较高，长度在 12～20 bp 的 SSR 多态性中等，而长度在 12 bp 以下时多态性极低[16]。党参 EST-SSRs 的长度在 12～25 bp，其中长度在 12～20 bp 的 SSR 有 6 483 条，占 SSR 总数的 88.48%，这类 SSR 位点具有中等多态性；而长度在 20 bp 以上的 SSR 达到 844 条，占 SSR 总数的 11.52%，这类 SSR 具有较高多态性。此外，根据 Dreisigacker 等[17]的发现，高级基元 SSR 多态性普遍比低级基元的低。经统计发现，长度在 20 bp 以上的党参 EST-SSR 中，包含低级基元一、二、三核苷酸 SSR 共 564 条。可以预计这部分多态性潜能高的 SSR 在党参上应具有较高的利用价值。

2.4 党参 EST-SSR 引物设计与筛选

为进一步在实验中利用筛选出的党参 EST-SSR，本课题组使用 Primer3 软件对上下游序列均不小于 150 bp 的 EST-SSR 设计引物，每条序列产生 5 对引物，共有 1 484 条党参 EST-SSR 序列获得 7 420 对引物。通过去除不符合条件的引物，并将产物序列利用 SSRFinder 校验，检验结果是否与 MISA 结果一致，最终筛选出引物。结果表明，共有 848 条党参 EST-SSR 序列成功设计出 4 329 对引物（表 3），占到党参 EST-SSR 总数的 8.6%。其中 20 bp 以上 EST-SSR 序列且包含低级基元一、二、三核苷酸共设计 226 对引物。

表 3 党参 EST-SSR 引物序列

Table 3 EST-SSR primer sequences in *C. pilosula*

重复基元及类型	上游引物 (5’-3’)	下游引物 (5’-3’)	产物长度 / bp
TCA (3*5)	TAAAGAATCACTTGGGGAGTCTG	ATGGACAGAACATTGCAAAACTT	127
AG (2*9)	GAGAATTATGACCTTGAGAAGCG	GATTCTGCGCTACAATCAAAATC	128
TGA (3*6)	ATAAGGACAATACATCCACAGGG	CTTCAAGTTCATTCAACTCGTCC	129
AACGAA (6*4)	AGGACCCGTTCTTTGATGATATT	CAATAGCAATGGCTGACTTATCC	154
GAA (3*5)	GCACTTTATGGATATGATGCTCC	GTGCTTCTGCTTCATCTCTGTTT	153
AG (2*9)	GCGACATGAACTTGAAGAACTTT	TGATCTAGTCATTCATGCTCTTCC	135
GAAGGA (6*4)	GGTAGAAAAGACCTCCACCCTAA	ACCCTTTTCTCTTTCCTTCTCCT	147
TCC (3*5)	GGAATTCATCCTTACCATCAACA	AAGAAGATCGTCCTAACCTCGTC	149
TCA (3*6)	ACTTAGACTGCTCCACTTCATGC	GGGTCCTAGAACCAACTTGAGAG	133
GA (2*9)	AGTAAGTGGGATGATGAAAAGCA	CGATTTCCTCCATTTCCTTCTAT	116
GAG (3*6)	TTTTCTAAGGACAGTTGTGGTGG	CTTATCATCAGTGCTGACCTGTG	132
TGG (3*5)	GGAAGTAGACCAAGAGTGGGAGT	TTGATTCTCAAACAAGTGTCACG	132
CAC (3*6)	TCATCCATTGCAACCTAATCAGT	TTAGAACTAGGAGCTGCACCATT	138
GA (2*10)	TCTTGCTTCTCAAAGATACGACC	CTAGCAGGTGAAAGCAAACACTA	134

3 讨论

党参转录组中共搜索到 7 327 个 SSRs，平均出现频率为 1/4 520。与其他药用植物比较，党参 SSR 出现的频率低于丹参（1/2 100）[11]，高于人参[10]（1/5 800）、金银花[13]（1/7 490）、西洋参[12]（1/7 500）和杜仲[18]（1/11 610）；同时与模式植物比较，也高于拟南芥（1/13 830）、大豆（1/7 400）、杨树（1/14 000）和棉花（1/20 000）[19]等植物，这表明党参转录组中 SSR 数量很丰富。

根据已有报道，大多数植物的 EST-SSRs 以三核苷酸和二核苷酸重复类型为主，但主要的重复基元类型有所差异[15]。研究发现党参 EST-SSR 以二核苷酸为最多的重复基元，其次是三核苷酸重复，这与人参、丹参以及茶树 EST-SSR 中的优势重复类型是一致的。从出现的频率来看，各种不同的重复基元出现最多的是 AG/CT，其次是 AC/GT 和 A/T。GC 重复基元在多数植物中很难见到，但出现频率都极低[15]。在党参转录组中仅找到 5 个 GC 重复，出现频率也极低。此外，党参转录组中长度在 20 bp 以上的 SSR 达到 844 条，占 SSR 总数的 11.52%，并且其中有 66.82%是多态性潜能高的低级基元一、二、三核苷酸 SSR，这部分 SSR 可能具有较高的利用价值。

对遗传学背景研究取决于合适的遗传分析工具

的选择。而分子标记技术中，RAPD 技术重复性差、不稳定；AFLP 虽然稳定可靠，但成本高、对 DNA 质量要求高，实验程序复杂、工作量大；ITS 条形码仅可用于药用植物及其易混伪品的鉴定。相比之下 SSR 标记具有共显性、高度可重复、多态性丰富、对 DNA 质量要求低、可通过 PCR 快速检测，是遗传学背景研究非常有效的工具。缺点是 SSR 标记具有特异性，必须进行 PCR 检测，存在引物开发问题[20]。本实验对所获得的党参 EST-SSR 利用软件设计引物，方法快速、通量大、成本低，并通过筛选共获得 4 329 对 SSR 引物，为党参的 SSR 分析奠定数据基础。

本研究结果表明党参转录组数据是 SSR 重要来源之一，与基因组 SSR 相比，EST-SSR 反映了基因组的编码区域，可直接获得基因表达的信息，省去了 SSR 引物开发过程中的克隆和测序步骤，充分利用了现有测序数据，降低了开发成本。同时也明确了党参 EST-SSR 的基本特征，为进一步开发新的党参功能基因 SSR 标记奠定了基础。党参 EST-SSR 对于党参功能基因资源的开发利用、遗传资源评价、丰富其分子标记、种质资源改良和比较基因组学研究都具有重要的价值。

参考文献

[1] 中国药典 [S]. 一部. 2010.

[2] 张向东, 高建平, 曹铃亚, 等. 中药党参资源及生产现状 [J]. 中华中医药学刊, 2013, 31(3): 496-498.

[3] 张建清, 苏 雪, 吴 琼, 等. 药用植物党参的 RAPD 分析 [J]. 中药材, 2006, 29(5): 417-420.

[4] 李忠虎, 刘晓东, 王小琦, 等. 党参微卫星引物筛选及群体遗传多样性研究 [J]. 中草药, 2013, 44(2): 210-214.

[5] 赵 莎, 辛天怡, 侯典云, 等. 党参药材及其混伪品的 ITS/ITS2 条形码鉴定研究 [J]. 世界科学技术—中医药现代化, 2013, 15(3): 421-428.

[6] He J Y, Zhu S, Komatsu K, *et al*. Genetic polymorphism of medicinally-used *Codonopsis* species in an internal transcribed spacer sequence of nuclear ribosomal DNA and its application to authenticate *Codonopsis Radix* [J]. *J Nat Med*, 2014, 68(1): 112-124.

[7] 苏 强, 杨 静, 王 哲, 等. 党参 DNA-AFLP 分析体系建立及条件优化 [J]. 药物生物技术, 2010, 17(4): 326-330.

[8] Liu T, Zhu S, Fu L, *et al*. Development and characterization of 1 827 expressed Sequence tag-derived simple sequence repeat markers for ramie (*Boehmeria nivea* L. Gaud) [J]. *PLoS One*, 2013, 8(4): e60346.

[9] 张利达, 唐克轩. 植物 EST-SSR 标记开发及其应用 [J]. 基因组学与应用生物学, 2010, 29(3): 534-541.

[10] Li C, Zhu Y, Guo X, *et al*. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* CA Meyer [J]. *BMC Genomics*, 2013, 14(1): 245.

[11] 邓科君, 张 勇, 熊丙全, 等. 药用植物丹参 EST-SSR 标记的鉴定 [J]. 药学学报, 2009, 44(10): 1165-1172.

[12] 杨维泽, 金 航, 赵振玲, 等. 西洋参 EST 资源的 SSR 信息分析 [J]. 西南农业学报, 2011, 24(1): 275-278.

[13] 蒋 超, 袁 媛, 刘贵明, 等. 基于 EST-SSR 的金银花分子鉴别方法研究 [J]. 药学学报, 2012, 47(6): 803-810.

[14] Lu X, Wang H, Liu B, *et al*. Three EST-SSR markers associated with QTL for the growth of the clam Meretrix meretrix revealed by selective genotyping [J]. *Mar Biotechnol*, 2013, 15(1): 16-25.

[15] 李 珊, 周天华, 赵桂仿, 等. 马蹄香表达序列标签资源的 SSR 信息分析 [J]. 中草药, 2010, 41(3): 464-468.

[16] 杨 华, 陈 琪, 韦朝领, 等. 茶树转录组中 SSR 位点的信息分析 [J]. 安徽农业大学学报, 2012, 38(6): 882-886.

[17] Dreisigacker S, Zhang P, Warburton M L, *et al*. SSR and pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different megaenvironments [J]. *Crop Sci*, 2004, 44(2): 381-388.

[18] 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, 49(5): 176-181.

[19] Cardle L, Ramsay L, Milbourne D, *et al*. Computational and experimental characterization of physically clustered simple sequence repeats in plants [J]. *Genetics*, 2000, 156(2): 847-854.

[20] 李明芳, 郑学勤. 开发 SSR 引物方法之研究动态 [J]. 遗传, 2004, 26(5): 769-776.