安徽农业大学学报, 2011, 38(6): 882-886
Journal of Anhui Agricultural University
[DOI]CNKI：34-1162/S.20111031.1102.001　网络出版时间：2011-10-31　15:39:06
[URL]http://www.cnki.net/kcms/detail/34.1162.S.20111031.1102.001.html

茶树转录组中 SSR 位点的信息分析

杨　华[1,2]，陈　琪[1]，韦朝领[1]，史成颖[1]，方从兵[1,3]，宛晓春[1*]
(1. 安徽农业大学农业部茶及药用植物安全生产重点开放实验室，合肥 230036；
2. 安徽农业大学理学院，合肥 230036；3. 安徽农业大学园艺学院，合肥 230036)

摘　要：利用茶树全转录组的高通量测序获得的 127 094 条 Unigenes 来发掘茶树转录组 SSR 功能性标记。在这些序列中共搜索出 12 242 个 SSRs，分布于 10 325 条 Unigenes 中，出现频率为 9.63%。茶树转录组 SSRs 的平均长度为 16 bp，平均分布频率是 1/3.68 kb。在茶树转录组的 SSRs 中，二核苷酸重复是主要的类型，占总 SSRs 的 63.78%。茶树转录组 SSRs 共包含 181 种重复基元，二核苷酸重复基元 CT/AG 和 TC/GA 是优势重复基元类型，分别占总 SSRs 的 23.84%和 23.58%。同时对这些 SSR 的可用性进行了评价。

关键词：茶树；转录组；SSR 信息
中图分类号：S571.1　　**文献标识码**：A　　**文章编号**：1672−352X (2011)06−0882−05

Analysis on SSR information in *Camellia sinensis* transcriptome

YANG Hua[1, 2], CHEN Qi[1], WEI Chao-ling[1], SHI Cheng-ying[1], FANG Cong-bing[1, 3], WAN Xiao-chun[1]
(1. Key Laboratory of Tea & Medicinal Plants and Product Safety, Ministry of Agriculture, Anhui Agricultural University, Hefei 230036;
2. School of Science, Anhui Agricultural University, Hefei 230036;
3. School of Horticulture, Anhui Agricultural University, Hefei 230036)

Abstract: A total of 127 094 unigenes derived from deep sequencing of *camellia sinensis* transcriptome were used for the development of functional SSR molecular markers. Overall, 12 242 SSRs distributed in 10 325 unigenes were detected, accounting for 9.63% of all the unigenes. The average length and distribution distance of the transcriptomic SSRs are about 16 bp and 3.68 kb, respectively. There are 181 kinds of repeat motifs existing in the tea plant transcriptome, and the dinucleotide repeats are the main types, accounting for 63.78% of all the SSRs. CT/AG and TC/GA are the most frequent motifs, accounting for 23.84% and 23.58% in all the SSR repeat motifs, respectively. The potential of the transcriptomic SSRs for further usage and research was assessed.

Key words: tea plant; transcriptome; SSR information

简单重复序列(Simple sequence repeats，SSR)又称微卫星 DNA、短串联重复序列，一般为以 1～6 个碱基为核心序列。SSR 遗传位点广泛分布于所有原核生物和真核生基因组中，具有分布丰富性、遗传共显性和技术简单性等特点，而且多态性十分丰富，已广泛用于遗传和物理图谱的构建、品种鉴定、基因定位、遗传多样性、植物分类和进化及比较基因组等方面的研究[1-2]。根据序列性质不同．SSR 标记主要分为基因组 SSR(Genomic SSR，gSSR)和表达序列标签 SSR(Expressed sequence tag SSR，EST-SSR)两种。与 gSSR 标记相比，EST-SSR 标记源于基因的转录区，其多态性可能与基因功能直接相关，因此，比 gSSR 标记具有更高通用性[3]。目前，已有利用 NCBI 数据库中登陆的茶树 EST 数据进行 EST-SSR 标记开发和应用的相关报道[4-6]，但是相对于拟南芥、水稻、小麦、大麦等模式植物、作物来说，茶树功能基因 SSR 的标记开发还非常有限。随着新一代高通量测序技术的成熟以及测序成

收稿日期：2011-09-20
基金项目：“十二五”国家科技支撑计划（2011BAD01B01）和教育部、农业部重点实验室开放基金项目（ltbb201102042）共同资助。
作者简介：杨　华，女，博士研究生，讲师。
* **通讯作者**：宛晓春，男，博士，教授，博士生导师。E-mail：xcwan@ahau.edu.cn

本的急剧下降，带来了各种组学研究的兴起。例如，利用新一代测序技术可以对全基因组范围内的转录本进行大规模的高通量测序，并能产生较之 EST 测序更为海量的转录组数据[7]，这为功能基因组 SSR 标记的开发提供了更丰富和极有价值的可利用资源[8]。为此，作者基于本课题组在国际上首次利用新一代的高通量 Illumina 测序技术获得的茶树全转录组数据[9]，利用计算机的辅助进行大规模转录组 SSR 标记的发掘，同时对其组成、分布及特征进行了分析，以期为开发新的茶树功能基因组 SSR 标记提供理论依据。

1　材料与方法

1.1　茶树转录组数据来源

茶树转录组的数据来自于本课题组前期对茶树品种龙井 43 [*Camellia sinensis* (L.) O. Kuntze cv. *Longjing* 43]进行的全转录组的 Illumina 高通量深度测序[9] (http://www.biomedcentral.com/1471-2164/12/131)，共有 45.07 Mbp 的数据量，含有 127 094 条 Unigenes。

1.2　茶树转录组 SSR 的筛选

利用软件 SSRFINDER（http://www.maizemap.org/bioinformatics/SSRFINDER/）对茶树转录组中 Unigene 的 cDNA 序列数据进行 SSR 搜索，筛选的标准为：重复单元长度 2～6 bp，单核苷酸重复的次数在 16 次或 16 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。

2　结果与分析

2.1　茶树转录组中 SSR 位点的数量与分布

利用软件 SSRFINDER 对茶树转录组中 127 094 条 Unigenes 的 cDNA 序列数据进行搜索，共在 10 325 条 Unigenes 中找到符合条件的 SSR，发生频率(含有 SSR 的 Unigene 数目与总 Unigene 数目之比值)为 8.12%。其中，8 786 条 Unigene 含单个 SSR 位点，另外的 1 539 条 Unigenes 含 2~7 个 SSR 位点。共检出 12 242 个 SSR 位点，在整个茶树转录组中的出现频率(检出的 SSR 个数与总 Unigene 数目之比值)为 9.63%。从分布情况看，茶树转录组中平均每 3.68 kb 就出现 1 个 SSR，即平均距离(茶树转录组 Unigene 长度与 SSR 数目之比值)，但不同重复种类间大相径庭(表 1) 。

表 1　SSR 在茶树转录组中的出现频率

Table 1　Occurrence of SSRs in *Camellia sinensis* transcriptome

重复类型 Repeat type	数目 Number	各类型比例/% Proportion	频率/% Frequency	平均距离/kb Average distance	总长度/bp Total length	平均长度/bp Average length
单核苷酸 Mononucleotide	1 424	11.63	1.12	31.65	30 002	21
二核苷酸 Dinucleotide	7 808	63.78	6.52	5.77	115 824	15
三核苷酸 Trinucleotide	2 771	22.64	2.18	16.26	46 467	17
四单核苷酸 Tetranucleotide	140	1.14	0.11	321.93	2 916	21
五核苷酸 Pentanucleotide	45	0.37	0.04	1 001.56	1 165	26
六核苷酸 Hexanucleotide	54	0.44	0.04	834.63	1 620	30
总计 Total	12 242	100.00	9.64	3.68	197 994	16

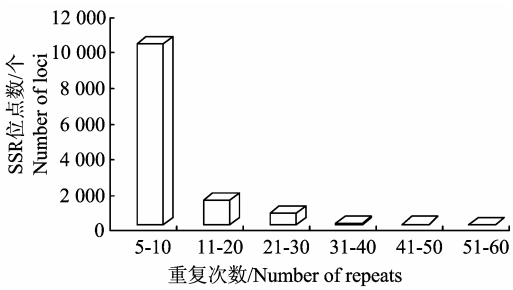

图 1　重复次数分布

Figure 1　Distribution of the number of repeats

茶树转录组 SSR 种类较为丰富，一至六核苷酸重复类型都能看到，但各种类型的出现频率差异较大（表 1）。主要集中在一至三核苷酸重复上，占总 SSR 的 98.05%。其中，二核苷酸重复类型的数量最多，占 63.78%，其次是三核苷酸和单核苷酸重复类型，分别占 22.64%和 11.63%，四、五、六核苷酸重复类型的数量很少，总计不足 2%（表 1）。

茶树转录组 SSR 位点的序列总长达到 197 994

bp，其中单核苷酸重复基元、二核苷酸重复基元、三核苷酸重复基元、四核苷酸重复基元、五碱基重复基元和六核苷酸重复基元的 SSR 位点的碱基总长分别是 30 002、115 824、46 467、2 916、1 165 和 1 620 bp；SSR 位点的平均长度是 16 bp，各类型 SSR 位点的平均长度分别是 21、15、17、21、26 和 30 bp（表 1）。

表 2 重复单元的类型及数量

Table 2 Type and number of repeat motifs

重复类型 Repeat type	重复基元 Repeat motif	数量/个 Number	发生频率/% Frequency	比例/% Proportion
单核苷酸 Mononucleotide	A/T	1 299	1.02	10.61
	C/G	125	0.10	1.02
二核苷酸 Dinucleotide	CT/AG	2 918	2.30	23.84
	TC/GA	2 887	2.27	23.58
	AT/TA	1 340	1.05	10.95
	TG/CA	354	0.28	2.89
	AC/GT	295	0.23	2.41
	GC/GC	14	0.01	0.11
三核苷酸 Trinucleotide	GAA/TTC	302	0.24	2.47
	TCT/AGA	258	0.20	2.11
	AAG/CTT	216	0.17	1.76
	GAT/ATC	179	0.14	1.46
	TGA/TCA	163	0.13	1.33
	CCA/TGG	143	0.11	1.17
	CAC/GTG	133	0.10	1.09
	GGT/ACC	122	0.10	1.00
	CAA/TTG	112	0.09	0.91
	CTC/GAG	110	0.09	0.90
	AAT/ATT	100	0.08	0.82
	GGA/TCC, ATG/CAT, CCT/ACG, TTA/TAA, AAC/GTT, ATA/TAT, CAG/CTG, TGT/ACA, GCT/AGC, TGC/GCA, CGC/GCG, CGG/CCG, AGT/ACT, GAC/GTC, TCG/CGA, CTA/TAG, GGC/GCC, TAC/GTA, CGT/ACG	963	0.76	7.87
四单核苷酸 Tetranucleotide	AAAT/ATTT, GAAA/TTTC, AAAG/CTTT, TTTA, AGAA/TTCT, TTTA/TAAA, AACA/TGTT, ATAT, ATAC/GTAT, GTGA/TCAC, TATC, AACC/GGTT, AAGA/TCTT, AATA/TATT, ACAA, ACCA/TGGT, ACGA, AGTG/CACT, ATAA/TTAT, ATAG/CTAT, ATCC, ATGG/CCAT, ATGT/ACAT, GGGA/TCCC, CCCT, GTTT/AAAC, TTTG/CAAA, ACTC/GAGT, AGGC/GCCT, CTTC/GAAG, GCTA/TAGC, AATC, AATG, AGAC, AGAT/ATCT, AGGA, ATCA, ATCG, ATGA, ATTC, ATTG, CCTT, GGAA/TTCC, CTCA, CTCC,GATC, GATG, TACC,TAGA, TAGG, TATG, TCTG, TGGA, TGTA, TGTG, TTAA	140	0.11	1.41
五核苷酸 Pentanucleotide	AAAAG/ CTTTT, TTTTC, AAAGA, AAAAC, AGAAG/CTTCT, AAACA, ACACC, AGAAA, AGAGA, ATCCC, ATGAG, ATTTT, CAAAC, CAACC, CAATA, CAATC, CACAA, CCAAA, GATGG, GCTCG, GGAGA, GGATT, GTGCT, TGTTG, TTATT, TTCTC, TTCTT, TTGGA, TTGGT, TTGTG, TTTTA	45	0.04	0.37
六核苷酸 Hexanucleotide	AAAAAG, AAAATA, AAACCA, AAGCAA, AATACT, ACAGGG, AGGCAA, AGGGAG, AGGGTT, ATGAAG, ATTCCT, ATTGTT, ATTTCG, CAAACC, CAACAT, CAACCA, CAACCT, CACCAT, CAGACT, CAGCAT, CCACCG, CCCTAA, CTCAGG, CTCCTG, GATTTT, GCTCCT, GCTGTG, GGCGGA, GGCGGT, GGCTTT, GGTGGC, GGTTAG, GTTGAT, GTTGGA, GTTTGG, TAGGGT, TCAAAA, TCACAC, TCATCT, TCCCAA, TCGTCC, TCTCCT, TCTCGT, TCTGGT, TCTTCC, TGAGGT, TGCAGA, TGGCTG, TGGTGT, TGGTTG, TTAGGG, TTCTGA, TTTATT, TTTTCT	54	0.04	0.44

从重复次数看，发现重复基元以重复 6 次出现的频率最高，有 3 356 个，占总 SSRs 的 27%，其次为 7 次、5 次和 8 次重复，出现频率均在 1 292~1 990 个之间。统计 5~10 次重复的 SSR 位点有 10 152 个，占 82.93%；11~20 次重复的位点有 1 364 个，占 11.14%；21~30 次重复的位点有 632 个，占 5.16%；31~40 次重复的位点有 68 个，占 0.56%；41~50 重复的位点有 20 个，占 0.16%；51~60 次重复的位点有 6 个，占 0.05%（图 1）。

2.2　茶树转录组 SSR 的特性

在 12 242 个茶树转录组 SSR 位点中，共观察到 181 种重复基元，其中一、二、三、四、五及六核苷酸重复基元分别有 4、6、30、56、31 和 54 种（表 2）。

从出现的频率来看，占优势的前 3 种重复基元类型是二核苷酸重复基元 CT/AG、TC/GA/和 AT/TA，占总 SSRs 的 23.84%、23.58%和 10.95%，三者总计占总 SSRs 的 58.37%；其次是单核苷酸重复基元 A/T，占总 SSRs 的 10.61%。二核苷酸重复基元中，CT/AG、TC/GA/和 AT/TA 出现的数量最多，三者共占二核苷酸 SSRs 的 91.51%(图 2)。在三核苷酸重复基元中，以 GAA/TTC、TCT/AGA 和 AAG/CTT 为主，但出现的数量差别不大，最高为 GAA/TTC，有 302 个（表 2）。其他四核苷酸、五核苷酸和六核苷酸重复基元类型分布相对分散，出现频率均较低（表 2）。

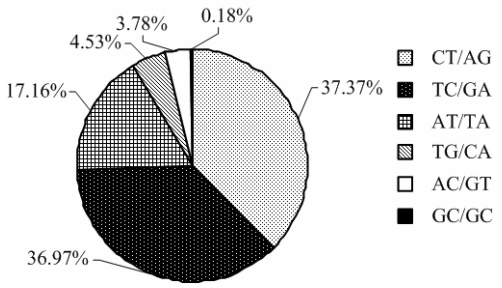

图 2　二核苷酸 SSRs 不同重复基元比例

Figure 2　Percentage of different motifs in dinucleotide SSRs

2.3　茶树转录组 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。根据 Temnykh 等[10]的研究，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 之间的 SSR 多态性中等，而长度在 12 bp 以下时多态性极低。因此，本研究在 SSR 筛选过程中就已经将多态潜能很低的 l2 bp 以下的 SSR 过滤去除。茶树转录组 SSRs 的长度主要集中在 12～54 bp 范围，其中长度在 12～20 bp 的 SSR 有 9 822 条，占 SSR 总数的 80.23%，这类 SSR 具有中等多态性；而长度在 20 bp 以上的 SSR 达到 2 420 条，占 SSR 总数的 19.77%，此类 SSR 具有较高多态性。此外，根据 Dreisigacker 等[11]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 以上的茶树转录组 SSR 中，90.12%都是多态性潜能高的低级基元 SSR，包含低级基元一、二、三核苷酸 SSR 共 2 181 条。可以预计这部分多态性潜能高的 SSR 在茶树上应具有较高的利用价值。

3　讨论

从茶树转录组 127 094 条 Unigenes 中搜索出 12 242 个 SSRs，分布于 10 325 条 Unigenes 中，平均出现频率是 1/3.68 kb，略低于金基强[4]报道的茶树 EST-SSR 的分布频率(1/2.6 kb)，可能是与本文搜索 SSR 所用的茶树转录组序列的数量远远大于茶树 EST-SSR 开发所用的 ESTs 的数量有关，或者与搜索 SSR 的算法(如参数设定)等多种因素有关。例如，本研究在搜索五、六核苷酸重复时限定的重复次数标准是 5 次，高于金基强设定的 4 次，因此所获得的五、六核苷酸重复类型占所有 SSRs 的比例很小。但是和其他物种相比，茶树转录组 SSRs 出现频率高于小麦[12] (1/15.6 kb)、大麦[13] (1/6.3 kb)、拟南芥(1/13.83 kb)、番茄(1/11.1 kb)、棉花(1/20.0 kb)、大豆(1/7.4 kb)、杨树(1/14.0 kb) [14]和柑橘[15](1/ 5.7 kb)等植物，这表明茶树转录组中 SSR 数量很丰富。

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。研究发现茶树转录组 SSR 重复基元以二核苷酸为最多，占所有 SSRs 的 63.78%，其次是三核苷酸重复，占所有 SSRs 的 22.64%，这与金基强报道的茶树 EST-SSR 中的优势重复类型是一致的[4]，并且和木本植物猕猴桃[16]也是相似的。但与 Cardle 等[14]在大豆、番茄、棉花、杨树和拟南芥以及 Varshney 等[17]在大麦、小麦、玉米、燕麦、黑麦、高粱和水稻等主要的禾谷类作物上的研究结果却不同，这些植物物种 SSR 的主要类型是三核苷酸重复。

研究发现，茶树转录组 SSR 重复基元以二核苷酸为最多。从出现的频率来看，各种不同的重复基元出现最多的是 CT/AG，其次分别是 TC/GA/和 AT/TA，3 种重复基元占总 SSRs 的 58.37%；其次是单核苷酸重复基元 A/T，占所有 SSRs 的 10.61%。这与金基强报道的茶树 EST-SSR 中的优势重复基

元类型是 CT/AG 也是一致的[4]。GC 重复基元在多数植物中很难见到，如在拟南芥、杏树、桃树[18]、水稻、玉米、大豆[19]等植物中都未发现，而在小麦、油菜和猕猴桃中虽有发现，但出现频率都极低。如 Gao 等[19]报道小麦 EST-SSR 中 GC 重复只以很低的频率(0.09/100 kb)出现，在油菜[20]的研究中发现 GC 出现的频率也很低（0.016/100 kb)。试验在茶树转录组 SSR 中也筛查到了 14 个 GC 重复，出现频率也极低（0.031/100 kb)，这是在以前报道的茶树 SSR 中未发现的，可能是与本研究利用的是转录组数据库有着很重要的关系，转录组数据库较之以前的 EST 数据涵盖了整个基因组的转录本，包含的功能基因更多更全面。从上述分析，可以推测植物对 GC/GC 重复有明显的偏倚性，但也不缺乏这种重复序列。

从搜索到的茶树转录组 SSR 中，发现重复位点的长度在 20 bp 以上的 SSR 达到 2 420 条，占 SSR 总数的 19.77%，并且这其中的 90.12％都是多态性潜能高的低级基元一、二、三核苷酸 SSR。可以预计这部分多态性潜能高的 SSR 在茶树上应具有较高的利用价值。

综上所述，茶树转录组 SSR 不但出现频率高，而且类型丰富；从多态性潜能的角度考虑，搜索到的这些 SSR 也具有较高的可用性。因此，本研究的结果为进一步开发新的茶树功能基因 SSR 标记奠定了基础，这种标记的建立对于加速茶树功能基因资源的开发利用、丰富其分子标记类型、遗传资源评价、绘制遗传图谱、实现特定性状的辅助选择和进行比较基因组学研究都具有重要的意义。

参考文献：

[1] Powell W, Machray G C, Provan J. Polymorphism revealed by simple sequence repeats [J]. Trends Plant Science, 1996, 1: 215-222.

[2] Morgante M, Olivieri A M. PCR-amplified microsatellites as markers in plant genetics [J]. The Plant Journal, 1993, 3: 175-182.

[3] Eujayl I, Sorrells M, Banm M, et a1.Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat [J]. Theor Appl Genet, 2002, 104(2): 399-407.

[4] Jin J Q, Cui H R, Chen W Y, et a1. Data mining for SSRs in ESTs and development of EST-SSR marker in tea plant (*Camellia sinensis*) [J]. J Tea Sci, 2006, 26: 17-23.

[5] Sharma R K, Bhardwaj P, Negi R, et a1. Identification, characterization and utilization of unigene derived microsatellite markers in tea (*Camellia sinensis* L.) [J]. BMC Plant Biology, 2009, 9: 53.

[6] Yao M Z, Ma C L, Qiao T T, et al. Diversity distribution and population structure of tea germplasms in China revealed by EST-SSR markers [J/OL]. Tree Genetics & Genomes, 2011 (DOI: 10.1007/s11295-011-0433-z).

[7] Simon S A, Zhai J, Nandety R S, et al. Short-read sequencing technologies for transcriptional analyses [J]. Annu Rev Plant Bio, 2009, 60: 305-333.

[8] Graham I A, Besser K, Blumer S. The genetic map of *Artemisia annua* L. identifies loci affecting yield of the antimalarial drug artemisinin [J]. Science, 2010, 327(5963): 328-331.

[9] Shi C Y, Yang H, Wei C L, et a1. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds [J]. BMC Genomics, 2011, 12:131.

[10] Temnykh S, DeClerck G, Lukashova A, et al. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.) [J]. Genome Res, 2001, 11: 1441-l452.

[11] 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 Science, 2004, 44(2): 381-388.

[12] Kantety R V, Rota M L, Matthews D E, et al. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat [J]. Plant Mol Biol, 2002, 48: 501-510.

[13] Thiel T, Michalek W, Varshney R K, et al. Exploiting EST database for the development and characterization of gene-derived SSR markers in barley (*Hordeum vulgare* L.) [J]. Theor Appl Genet, 2003, 106: 411-422.

[14] 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.

[15] 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.

[16] 姜春芽, 徐小彪, 廖娇. 猕猴桃 EST 序列的 SSR 信息分析[J]. 中国农学通报, 2009, 25(13): 37-39.

[17] Varshney R K, Graner A, Sorrells M. Genic microsatellite markers in plants: features and applications [J]. Trends in Biotechnology, 2005, 23 (1): 48-55.

[18] 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.

[19] 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: 245-261.

[20] 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.