安徽农业大学学报, 2017, 44(4): 558-562
Journal of Anhui Agricultural University
[DOI] 10.13610/j.cnki.1672-352x.20170811.001　网络出版时间：2017/8/11　11:53:57
[URL] http://kns.cnki.net/kcms/detail/34.1162.S.20170811.1153.002.html

四球茶转录组 SSR 位点信息分析

黎瑞源，邢　辉，申　铁
(贵州师范大学贵州省信息与计算科学重点实验室，贵阳 550001)

摘　要：对基于高通量测序拼接获得的 99 741 条四球茶嫩叶 Unigenes 进行了 SSR（simple sequence repeats）位点分析，共揭示了 23 732 个 SSR 位点，分布于 18 552 条 Unigenes 中，SSR 发生频率为 23.79%；含有 SSR 位点的 Unigenes 的总长度为 226 188 bp，SSR 位点间的平均距离为 1/2.07 kb，重复长度平均长度为 9.53 bp。在四球茶转录组 SSR 中，二核苷酸重复和三核苷酸重复是主要的重复类型，分别占总 SSRs 的 47.53%和 25.92%；在 181 种重复基元类型中，优势重复基元分别是 AG/CT、A/T、ACC/GGT，分别占总 SSRs 的 41.38%、22.10%和 6.49%，共占总 SSRs 的比例达 69.97%。

关键词：四球茶；转录组；SSR
中图分类号：S571.1　　**文献标识码：**A　　**文章编号：**1672−352X (2017)04−0558−05

Analysis of SSR information in *Camellia tetracocca* Zhang transcriptome

LI Ruiyuan, XING Hui, SHEN Tie
(Key Laboratory of Information and Computing Science Guizhou Province，Guizhou Normal University，Guiyang 550001)

Abstract: In this study, a total of 99 741 unigenes were *de novo* assembled based on the bud transcriptome sequencing of *Camellia tetracocca* Zhang. Totally, 23 732 SSRs were detected in 18 552 unigenes, accounting for 23.79% of all the unigenes. The total length of unigenes containing SSR loci was 226 158 bp. The mean distance between SSRs was 1/2.07 kb and the mean length of the repeat motif was 9.53 bp. Di-nucleotide repeats were the most abundant followed by tri-nucleotide repeats, making up about 47.53% and 25.92% of the total SSRs, respectively. Among the 181 repeat motifs, AG/CT, A/T and ACC/GGT were the most frequent motifs in di-, mono- and tri-nucleotide repeats, accounted for 41.38%, 22.10% and 6.49%, respectively, and which totally accounted for 69.97% of the total SSRs. These results will provide abundant sequences for SSR markers development on *Camellia tetracocca Chang.*

Key words: *Camellia tetracocca* Zhang；transcriptome；SSR

四球茶（*Camellia tetracocca* Zhang）是山茶属中的一个种[1]，为贵州所特有，生长在贵州黔西南州普安县境内海拔 1 700～1 950 m 的原始森林中，能够忍受－6.0 ℃以下的低温，在冰雪环境条件下都能较好的生存，具有很强的环境适应能力，对不良自然环境表现出很强的抗逆性。四球茶具有高抗寒性基因源，可作为远缘杂交、基因转移聚合等茶树育种材料，是实现突破性育种和选育优良茶树品种不可替代的基因源。同时，四球茶还具有祖先种的原始的特征，在茶树起源、演化和分类研究上具有重要的学术价值。

SSR 标记又叫做微卫星标记，是利用 SSR（简单重复序列）差异进行个体区分的一种 DNA 分子标记技术。它是一种由 1～6 个碱基组成的核心序列，其长度通常在 10～100 bp 左右。与其他标记相比，SSR 标记具有明显的优点，如多态性高、可重复性好及多等位性高等。目前，SSR 标记已经广泛应用到大麦、大豆、水稻、玉米和小麦等物种的遗传多样性分析、遗传图谱构建等研究上，在种质资源保护利用、基因定位和克隆以及优异品种开发上

收稿日期: 2016-10-22
基金项目: 国家自然科学基金地区项目（31460233）和贵州师范大学博士基金（11904/0514027）共同资助。
作者简介: 黎瑞源，博士，讲师。E-mail：1055454430@qq.com

发挥了重要作用。

目前还没有关于四球茶分子标记开发的研究报道。因此研究四球茶转录组信息，挖掘并分析转录组 SSR 位点，是大规模开发四球茶 SSR 分子标记最简便可行的的途径，对四球茶种质资源遗传多样性分析、遗传图谱构建、种质资源保护和合理利用等具有非常重要的意义。本研究基于四球茶转录组测序数据拼接得到的 Unigenes 序列进行了 SSR 位点挖掘，分析了其分布特点，旨在为四球茶 SSR 分子标记的开发提供理论依据及序列信息。

1　材料与方法

1.1　供试材料

选用采集贵州省普安县的四球茶为转录组测序的研究材料。

1.2　四球茶 Unigenes 的拼接和 SSR 位点的扫描

取新鲜的四球茶嫩叶，迅速放入液氮中，使用干冰运输到北京天一辉远生物科技有限公司进行测序，测序平台使用的是 Illumina NextSeq 500，原始数据量是 7.2 Gb，使用 Trinity 软件拼接获得总长度为 49 179 157 bp 的 99 741 条 Unigenes。使用 est timmer perl 程序对拼接的 Unigenes 进行过滤，参数设置为：-amb=2，50；-tr5=T，5，50；-tr3=A，5，50；-cut=100，700；再使用 CD_HIT 去除冗余序列，参数设置为：-c 1.00 -n 5 -M 2000。

使用软件 MISA 搜索 Unigenes 上的 SSR 位点，筛选标准为：单核苷酸重复 10 次以上，二核苷酸重复 6 次以上，三核苷酸重复 5 次以上，四核苷酸重复 5 次以上，五核苷酸重复 5 次以上。

1.3　含 SSR 位点的 Uunigenes 的抗旱性分析

对含有 SSR 位点的 Unigenes，基于 Nr 和 Nt 数据库，通过 Blast 程序（http://www.ncbi.nih.nlm.gov），在默认参数下，进行了抗耐寒性注释分析。

2　结果与分析

2.1　四球茶转录组 SSR 位点数量和重复类型分布

对拼接所得到的 99 741 条 Unigenes 进行 SSR 位点搜索，找到含有 SSR 位点的 Unigenes 共 18 552 条，占总 Unigenes 的 18.60%。其中，含有单个 SSR 位点的 Unigenes 有 14 453 条；含有多个 SSR 位点的 Unigenes 有 4 099 条。这 18 552 个 Unigene 共包括了 23 732 个的 SSR 位点，SSR 位点发生的频率为 23.79%，平均每 2.07 kb 出现一个 SSR 位点。不同的重复基元中，发生频率和平均距离差异很大（表 1）。

表 1　四球茶转录组 SSR 位点的分布

Table 1　Distribution of SSR loci in *Camellia tetracocca* Zhang transcriptome

重复类型 Repeat type	数目 Number	比例/% Proportion	频率/% Frequency	平均距/kb Average distance	总长度/bp Total length	平均长度/bp Average length
单核苷酸 Mononucleotide	5 528	23.30	5.54	8.90	73 812	13.35
二核苷酸 Dinucleotide	11 281	47.53	11.31	4.36	109 681	9.72
三核苷酸 Trinucleotide	6 153	25.92	6.17	7.99	38 313	6.23
四核苷酸 Tetranucleotide	214	0.90	0.21	229.80	1 186	5.54
五核苷酸 Pentanucleotide	158	0.67	0.16	311.26	813	5.15
六核苷酸 Hexanucleotide	398	1.68	0.40	123.57	2 383	5.99
总计 Total	23 732	100.0	23.79	2.07	226 188	9.53

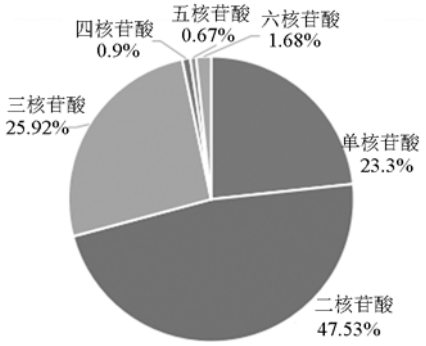

图 1　四球茶转录组各重复类型 SSR 的比例

Figure 1　Proportion of different SSR repeat type in *Camellia tetracocca* Zhang transcriptome

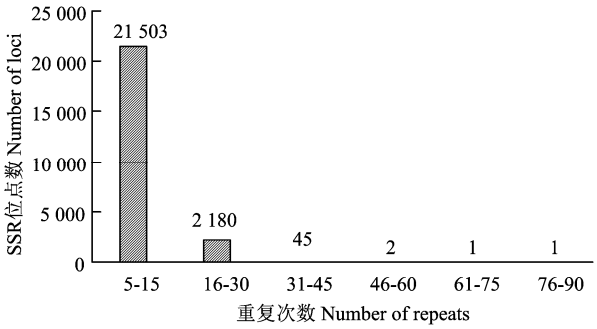

图 2　四球茶转录组 SSR 位点重复次数分布

Figure 2　Distribution of the repeat number for different motifs

四球茶转录组的 SSR 种类丰富，从单核苷酸到到六核苷酸重复都有，并且各种重复类型所占比例差异较大。其中，二核苷酸重复类型数目最多，为 11 281 个，占总 SSRs 的 47.53%；其次为三核苷酸和单核苷酸重复类型，分别为 6 153 个和 5 528 个，占总 SSR 的 25.92%和 23.30%，这 3 种重复类型，共占总 SSR 的 96.75%（图 1）。

2.2 四球茶转录组 SSR 不同基序总长度和重复次数的分析

四球茶转录组中，SSR 位点重复基序总长度是 226 188 bp，平均长度是 9.53 bp，其中二核和单核苷酸重复基序占有明显的优势，长度分别为 109 681 和 73 812 bp，其次是三核苷酸和六核苷酸重复基序，长度分别为 38 313 和 2 383 bp，最短的是四核苷酸和五核苷酸重复，长度分别为 1 186 和 813 bp。单到六核苷酸重复基序的平均长度分别为 13.35、9.72、6.23、5.54、5.15 和 5.99 bp（表 1）。

从基元重复次数来看，所检测到的 SSR 中，6 次重复最多，共有 4 255 个，其次是 5 次重复，共有 3 170 个，其余从高到低依次是 10 次、7 次和 8 次重复，分别有 2 630、2 564 和 2 000 个。重复次数在 5～15 之间的 SSR 位点，共有 21 503 个，占到总数的 90.61%，而重复次数在 31 次以上的出现频率很低，在 60 次以上，可以忽略不计（图 2）。

2.3 四球茶转录组 SSR 不同基序出现次数的比较分析

在四球茶转录组 23 732 个 SSR 位点中，共有 181 种重复基元。其中，出现次数最多的是六核苷酸重复基元，有 108 种，其次是五核苷酸重复基元，共有 33 种。四核苷酸、三核苷酸、二核苷酸以及单核苷酸重复基元分别为 22 种、10 种、4 种和 4 种（表 2）。

表 2 四球茶转录组中 SSR 重复基元的类型和数量

Table 2 Repeat type and number of SSR motifs in *Camellia tetracocca* Zhang transcriptome

重复类型 Repeat type	重复基元 Repeat motif	数量 Number	比例/%Proportion
单核苷酸	A/T	5 243	22.09
Mononucleotide	C/G	285	1.20
二核苷酸	AC/GT	900	3.79
Dinucleotide	AG/CT	9 821	41.38
	AT/AT	538	2.27
	CG/CG	22	0.09
三核苷酸	AAC/GTT	450	1.90
Trinucleotide	AAG/CTT	1 213	5.11
	AAT/ATT	165	0.70
	ACC/GGT	1 541	6.49
	ACG/CGT	188	0.79
	ACT/AGT	103	0.43
	AGC/CTG	441	1.86
	AGG/CCT	1 010	4.26
	ATC/ATG	655	2.76
	CCG/CGG	387	1.63
四核苷酸	AAAC/GTTT	38	0.16
Tetranucleotide	AAAG/CTTT	30	0.13
	AAAT/ATTT	24	0.10
	Others	122	0.51
五核苷酸	AAAAC/GTTTT、AAAAG/CTTTT	44	0.19
Pentanucleotide	Others	114	0.48
六核苷酸	ACCGCC/CGGTGG	64	0.27
Hexanucleotide	AACCCT/AGGGTT	21	0.09
	Others	313	1.32

单核苷酸重复类型中，A/T 重复基元是优势基元，占单核苷酸重复类型的 94.84%；二核苷酸重复类型中，AG/TC 重复基元所占比例最大，达 87.06%；三核苷酸重复类型中，ACC/GGT 重复基元出现次

数最多，其次是 AAG/CTT 和 AGG/CCT，所占比例分别为 21.07%、19.71%和 16.41%。四到六核苷酸重复类型中，各重复基元出现的频率都很较低（表 2）。AG/CT、A/T、ACC/GGT，AAG/CTT 和 AGG/CCT 这 5 种重复基元在所检测到的重复基元中占有明显优势，分别占总 SSRs 的 41.38%、22.10%、6.49%、5.11%和 4.26%，占总 SSRs 的 79.34%（图 3）。

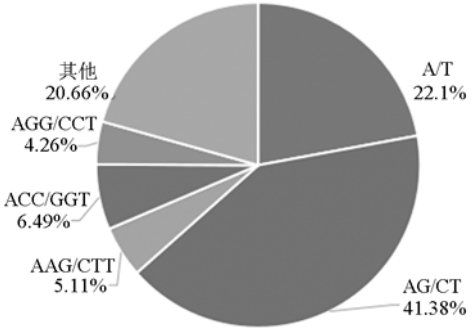

图 3　四球茶转录组 SSR 主要重复基元所占比例

Figure 3　Proportion of the major SSR motifs in *Camellia tetracocca* Zhang transcriptome

2.4　四球茶转录组 SSR 可用性的评价

SSR 分子标记使用性能的判断依据主要是多态性。对 SSR 多态性有重要影响的因素是其长度。SSR 位点的长度（*L*）与多态性之间的关系为：当 *L*≥20 bp 时，多态性高；当 20 bp＞*L*≥12 bp 时，多态性中等；当 *L*＜12 bp 时，多态性低。本研究中，四球茶转录组 SSR 长度在 10～20 bp 的有 15 847 个（占 SSR 总数的 66.78%），20～30 bp 的有 5 881 个（占总数的 24.78%），长度在 30 bp 以上的有 2 004 个（占总数的 8.44%），合计长度在 20 bp 以上的 SSR 共占有 33.22%（表 3），据此推测本研究检测到的 SSR 位点具有的较高的多态性和应用潜力。

表 3　SSR 位点长度分布

Table 3　Distribution of SSR motif length

SSR 长度/bp Length	SSR 数量/个 Number	比例/% Proportion
10～20	15 847	66.78
20～30	5 881	24.78
＞30	2 004	8.44

2.5　含 SSR 位点 Unigenes 耐寒性分析

考虑到四球茶较好的抗耐寒性，本研究基于 NR 和 NT 数据库，对包含 SSR 位点的 18 522 条 Unigenes 进行了比对注释，在默认的比对参数下，有 12 966（约占 18 522 条 Unigene 的 70%）能显著比对上 NR 或 NT 数据库，但比对结果没有明确注释为抗耐寒性相关。

3　讨论

四球茶是重要茶树种质资源，在高耐寒性等性状方面具有优异的表现。SSR 等分子标记的开发和评价在种质资源的保护、挖掘和利用方面有着重要作用。关于茶组物种转录组 SSR 标记，杨华等[2]于 2011 年已经做过相关的研究报道，但其所用材料是龙井 43 等茶种，目前尚未见针对四球茶 SSR 位点的分析报告。本研究基于转录组测序，首次对四球茶转录组 SSR 位点进行了分析，将对四球茶种质资源的保护和利用发挥重要的作用。

表 4　各茶种（品种）SSR 位点的信息比较

Table 4　Comparison of SSR loci information among tea species (cultivars)

种（品种） Species（Cultivars）	出现频率/% Frequency	平均距离/kb Mean distance	平均长度/bp Mean length	主要重复类型 Major repeat type
龙井 43 Longjin43	9.64	3.68	16	二核苷酸
舒茶早 Shuchazao	37.19	2.85	16	二核苷酸
四球茶 *Camellia tetracocca* Zhang	23.79	2.07	9.53	二核苷酸

本研究对测序拼接得到的 99 741 条四球茶转录组 Unigenes 进行了 SSR 位点搜索，有 18 552 条 Unigenes 包含 SSR 位点，占总 Unigenes 的 18.60%，这明显高于洋葱[3]（*Allium cepa*）（5.57%）、野三七[4]（*Panax vietnamensis* var.*fuscidiscus*）（16.86%）、党参[5][*Codonopsis pilosula* (Franch.) Nannf]（12.22%）、灯盏花[6][*Erigeron breviscapus*(vant.) Hand.Mazz.]（6.99%）和鱼腥草[7]（*Houttuynia cordata* Thunb.）（7.51%）等物种。本研究所获得 SSR 位点 23 732 个，平均距离为 2.07 kb，明显高于洋葱[3]（1/14.1 kb）、党参[5]（1/4.52 kb）、浮萍[8]（*Lemna minor*）（1/6.57 kb）和鱼腥草[7]（1/9.04 kb）等物种。因此，无论是发生频率还是平均距离出现率，四球茶转录组 SSR 都相对较高，这说明四球茶的 SSR

种类和数量都比较丰富。在 SSR 位点的长度分布上，四球茶嫩叶转录组 SSR 长度在 20 bp 以上的占总 SSR 的 33.22%，高于其他植物如党参[5]（11.52%）、洋葱[3]（18.11%）、野三七[4]（11.98%），也高于杨华等[2]对龙井 43 茶的分析结果（19.77%），这表明四球茶转录组 SSR 具有较高的可用性和多态性。

相对杨华等[2]对龙井 43 芽转录组 SSR 位点和陈琪等[9]对舒茶早芽叶转录组 SSR 位点的分析结果，四球茶芽转录组 SSR 在 SSR 位点的出现频率、平均距离出现率的方面表现出较大的差异（表 4），但核苷酸重复类型与其他 2 种茶树品种一样都是二核苷酸重复，这说明不同茶叶之间的位点信息不同，进化程度不同[10-12]。

本研究基于四球茶转录组的高通量测序信息拼接得到的 Unigenes，首次进行了 SSR 位点的搜索和分析，结果表明四球茶的 SSR 种类丰富、可用性高。研究结果可为四球茶功能基因 SSR 分子标记开发、种质资源的保护、遗传图谱的构建和茶树的育种等研究利用方面提供理论支撑。

参考文献：

[1] 张宏达. 茶树的系统分类[J]. 中山大学学报 (自然科学版), 1981(1): 88-99.

[2] 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, 2011, 38(6): 882-886.

[3] 李满堂, 张仕林, 邓鹏, 等. 洋葱转录组 SSR 信息分析及其多态性研究[J]. 园艺学报, 2015, 42(6): 1103-1111.

[4] 李翠婷, 张广辉, 马春花, 等. 野三七转录组中 SSR 位点信息分析及其多态性研究[J]. 中草药, 2014, 45(10): 1468-1472.

[5] 王东, 曹玲亚, 高建平. 党参转录组中 SSR 位点信息分析[J]. 中草药, 2014, 45(16): 2390-2394.

[6] 陈茵, 李翠婷, 姜倪皓, 等. 灯盏花转录组中 SSR 位点信息分析及其多态性研究[J]. 中国中药杂志, 2014, 39(7): 1220-1220.

[7] 黎晓英, 刘胜贵, 王丹, 等. 鱼腥草转录组 SSR 位点信息分析及其多态性研究[J]. 中草药, 2016, 47(10): 1762-1767.

[8] 孙蛟龙, 方扬, 靳艳玲, 等. 浮萍转录组数据 SSR 位点的生物信息学分析[J]. 应用与环境生物学报, 2015, 21(3): 401-405.

[9] 陈琪, 杨华, 韦朝领, 等. 基于茶树芽叶转录组序列的 EST-SSR 分布特征研究[J]. 安徽农业大学学报, 2016, 43(2): 170-175.

[10] TÓTH G, GÁSPÁRI Z, JURKA J. Microsatellites in different eukaryotic genomes: survey and analysis[J]. Genome Res, 2000, 10(7): 967-981.

[11] SIA E A, KOKOSKA R J, DOMINSKA M, et al. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes[J].Mol Cell Biol, 1997, 17(5): 2851-2858.

[12] HARR B, SCHLÖTTERER C. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation[J]. Genetics, 2000, 155(3): 1213-1220.