 doi: 10.1111/j.1471-8286.2004.00839.x

PRIMER NOTE

Identification and characterization of expressed sequence tags-derived simple sequence repeats markers from robusta coffee variety 'CxR' (an interspecific hybrid of *Coffea canephora* × *Coffea congensis*)

PRASANNA R. BHAT,* V. KRISHNAKUMAR,* PRASAD S. HENDRE,* P. RAJENDRAKUMAR,* RAJEEV K. VARSHNEY† and RAMESH K. AGGARWAL*

**Centre for Cellular and Molecular Biology, Uppal Road, Habsiguda, Hyderabad-500007, India, †Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, D 06466, Gatersleben, Germany*

Abstract

SSR (simple sequence repeats) markers derived from ESTs (expressed sequence tags), commonly called EST-SSRs or genic SSRs provide useful genetic markers for crop improvement. These are easy and economical to develop as by-products of large-scale EST resources that have become available as part of the functional genomic studies in many plant species. Here, we describe for the first time, nine genic-SSRs of coffee that are developed from the microsatellite containing ESTs from a cDNA library of moisture-stressed leaves of coffee variety, 'CxR' (a commercial interspecific hybrid between *Coffea congensis* and *Coffea canephora*). The markers show considerable allelic diversity with PIC values up to 0.70 and 0.75 for *Coffea arabica* and *Coffea canephora*, respectively, and robust cross-species amplification in 16 other related taxa of coffee. The validation studies thus demonstrate the potential utility of the EST-SSRs for genetic analysis of coffee germplasm.

Keywords: Coffea, expressed sequence tags, genetic diversity, microsatellites, simple sequence repeats

Received 12 August 2004; revision accepted 06 October 2004

Coffee, an important beverage crop belongs to the genus *Coffea* in the Rubiaceae family. Among *ca*. 100 known species, the commercial cultivation of coffee relies mainly on two species, the tetraploid *Coffea arabica* L. ($2n = 44$) and the diploid *Coffea canephora* Pierre ($2n = 22$). Genetic improvement of coffee suffers from a number of constraints, which can be surmounted by deploying DNA-based markers like microsatellites that provide high genetic resolution (Baruah *et al*. 2003). Despite the obvious advantages of SSR (simple sequence repeats) markers in germplasm characterization and crop improvement, only *ca*. 150 microsatellite markers have been reported to date for coffee (Combes *et al*. 2000; Rovelli *et al*. 2000; Baruah *et al*. 2003; Moncada & McCouch 2004), signifying the need to expand the repertoire of these highly informative markers.

The development of SSR markers following conventional approach of genomic library construction, is time consuming and resource intensive. In contrast, harvesting of SSRs making use of publicly available DNA sequence databases provides an attractive, time and cost-effective alternative. This approach has gained particular importance in developing a new class of SSR markers i.e. EST-SSRs, which provide additional advantage over genomic-SSRs in being 'functional markers', using the EST (expressed sequence tags) databases being developed as part of the functional genomics efforts in a number of plant species. In recent years, abundance of SSRs in EST sequences and development of EST-SSR markers for various applications have been reported in many plant species (reviewed by Varshney *et al*. 2005). However, in case of coffee, only few EST sequences are available in the public domain and no EST-SSR has been developed to the best of our knowledge. Here, we describe for the first time, nine new informative EST-SSRs using in-house generated EST database, useful for genetic studies on coffee.

In an ongoing coffee genomics program in our laboratory, we are generating ESTs from leaves of moisture-stressed plants of robusta coffee variety 'CxR' (a commercial interspecific hybrid of *Coffea congensis* and *C. canephora*). In a pilot

Correspondence: Ramesh K. Aggarwal, Tel.: 00-91-40-27192643; Fax: 00-91-40-27160591/27160311; E-mail: rameshka@ccmb.res.in

© 2004 Blackwell Publishing Ltd

Table 1 Details of the EST-derived microsatellite markers developed in the study

Locus	Repeat motif	Primer sequence (5′–3′)	*C. arabica* (n = 15)					*C. canephora* (n = 8)						
			Size range (bp)	N_A	H_O	H_E	PIC	Size range (bp)	N_A	H_O	H_E	PIC	GenBank accession	Putative function
CofEST-SSR01*	$(TC)_{13}$	F: TCGCGGTTTTGTTGCTTCCCAG R: GCAGCATGGCAGAAAAACCTCAACTT	150–154	3	0.27	0.35	0.31	146–154	5	0.14	0.76	0.66	AY705497	Sulfite reductase
CofEST-SSR02*	$(AAG)_4$ + $(GAA)_6$	F: CTGCGAGGAGGAGTTAAAGATACCAC R: GCCGGGAGTCTAGGGTTCTGTG	138–150	2	1.00	0.52	0.38	141–156	3	0.25	0.24	0.22	AY705498	Unknown protein
CofEST-SSR04*	$(AC)_{10}$	F: GGTCGCTCACTCATATCTTCCAG R: GCCTGGAAAGCAAACGTCTCA	137–151	8	1.00	0.76	0.70	133–165	6	0.67	0.85	0.75	AY705500	Heat shock protein 83
CofEST-SSR05	$(GA)_4CA$ $(GA)_4$	F: AAGGGGCATTGAACAGTTTTGAC R: TTTCCGATTCTGCCTTGCTTCTT	104	1	0.00	0.00	0.00	104	1	0.00	0.00	0.00	AY705501	Nuclear transport
CofEST-SSR06	$(AAAGG)_4$	F: CAGGCACAGAAGGAATGAAGAGC R: TGGTGGTATGGAAAACAGGAAGG	129–133	2	0.20	0.19	0.16	130–141	5	0.75	0.70	0.60	AY705502	No significant similarity
CofEST-SSR07*	$(GGT)_6$	F: GGCCGCCTCGTATCTGCTATGG R: GGTAACCAGAGCCCCGTGAACT	193–199	3	0.93	0.59	0.49	193	1	0.00	0.00	0.00	AY705503	Unknown protein
CofEST-SSR08	$(AAG)_5$	F: TCCATAACACAGAAATTTCCAGAGAGA R:ACCGTAACCTCCGTCTTCGAACTG	142–145	2	0.00	0.13	0.12	145–163	2	0.13	0.13	0.11	AY705504	Hypothetical protein
CofEST-SSR11*	$(CA)_5(TA)_4$	F:GGCCGAGGGGAAAAAGAAGC R:GGAAACCTCACGAGAAGATTACACAA	96–100	3	0.07	0.4	0.35	98–100	2	0.25	0.50	0.36	AY705507	Hypothetical protein
CofEST-SSR12*	$(TTC)_{12}$	F:CGCTGCTACTCCCTCTCCTCTCACT R:GAGGCTGAGGCTTGAAGGAAATTAAT	114–120	2	0.93	0.52	0.37	117–129	2	0.25	0.23	0.20	AY705508	No significant similarity found

N_A: Total number of alleles; H_O: Observed heterozygosity; H_E: Expected heterozygosity; PIC: Polymorphic information content; *Locus showing significant departure from the Hardy–Weinberg equilibrium (see text); PCR conditions for all the loci: 95 °C for 10 min, followed by 35 three-step cycles of 94 °C/1 min, 57 °C/1 min, 72 °C/1 min, and final extension of 72 °C for 5 min; Composition of PCR reaction: 10 ng of genomic DNA, 1 pmole of each primer, 1.5 mM $MgCl_2$, 100 μM dNTPs, 1 × PCR buffer II (500 mM KCl, 100 mM Tris-HCl pH 8.3) and 0.5 U *AmpliTaq*Gold DNA polymerase (Applied Biosystems). All the PCR reactions were performed on GeneAmp 2700 thermocycler (Applied Biosystems).

© 2004 Blackwell Publishing Ltd, *Molecular Ecology Notes*, 10.1111/j.1471-8286.2004.00839.x

study, around 500 ESTs generated in-house were surveyed manually for the presence of SSR motifs (minimum repeat length of 15 nucleotides) for developing EST-SSR markers. The primer pairs were designed for 14 such identified ESTs using GENETOOL version 2.0 (http://www.biotools.com/products/genetool.html), and were commercially synthesized from Bioserve Biotechnologies (http://www.bioserve.com) with forward primer of each primer pair having 6-FAM fluorescence-tag at its 5′-end. The primer pairs were tested for their utility as potential genetic markers on 23 elite genotypes of *C. arabica*, and *C. canephora* (common name robusta coffee) and also on 16 representative DNA samples belonging to 12 other species of *Coffea* and four species of *Psilanthus*. PCR amplifications and GENESCAN analyses for all the potential EST-SSRs were done as described by Baruah *et al*. (2003). The genetic parameters were calculated using CERVUS version 2.0 (Marshall *et al*. 1998).

Out of the 14 primer pairs tested, nine could be validated as useful microsatellite markers. PCR conditions and other attributes of the new EST-SSR markers are summarized in Table 1, whereas their cross-species status is shown in Table 2. The BLASTX analysis (Altschul *et al*. 1997) of the EST sequences from which SSR markers were developed, suggests a putative function for at least 30% of them (Table 1), unlike the genomic-SSRs with no putative function reported by us earlier (Baruah *et al*. 2003). Of the nine markers described here, one marker, CofEST-SSR05 was found monomorphic across the tested material (Tables 1 and 2). BLASTX search revealed it to be part of an important gene NTF-2 (nuclear transport factor-2), indicating its possible utility in functional studies on expression profiling/isolation of the gene. For the remaining eight EST-SSRs, the number of alleles varied from seven to 13 in the total tested coffee germplasm (Tables 1 and 2), which were comparable to that seen with the genomic-SSRs (Baruah *et al*. 2003). Compared to the other related taxa, the allelic variability was moderate across the eight loci for the 15 arabica (two to eight alleles; mean PIC 0.320) and eight robusta coffee genotypes (two to six alleles; mean PIC 0.322). This relatively low level of polymorphism in the arabica and canephora genotypes was as expected; autogamous *C. arabica* genepool is now widely demonstrated to have a narrow genetic base, whereas, the robusta genotypes tested here had overlapping pedigrees.

Allelic diversity across the eight EST-SSRs was tested for Hardy–Weinberg equilibrium (HWE) and Linkage Disequilibrium (LD) for the 15 arabica and eight robusta genotypes separately. The HWE and LD tests were done using Markov chain algorithm and Fisher's exact test using ARLEQUIN

Table 2 Cross-species amplification status of the CofEST-SSR markers developed from robusta coffee variety 'CxR'

			Microsatellite alleles (in bp) observed across the nine CofEST-SSRs								
Coffee species/ Related taxa*	Coffee type‡	Geographical distribution	CofEST-SSR01	CofEST-SSR02	CofEST-SSR04	CofEST-SSR05	CofEST-SSR06	CofEST-SSR07	CofEST-SSR08	CofEST-SSR11	CofEST-SSR12
C. canephora†	E	Ethiopia	146–154	141–156	133–165	104	130–141	193	145–163	98–100	117–129
C. arabica†	E	Ethiopia	150–154	138–150	137–151	104	129–133	193–199	142–145	96–100	114–120
C. congensis	E	WCA	152	150,156	167	104	133, 141	193	145, 163	100	111, 117
C. eugenioides	*M*	*CA*	140	138	143, 149	104	133	196, 199	145	114	111
C. kapakata	M	CA	140	138, 147	139, 155	104	137	196	145	100	111
C. racemosa	M	EA	140	129, 135	139, 147	104	131	178	145, 155	112, 114	111
C. salvatrix	M	EA	170, 190	120, 129	137	104	NA	178	145, 160	108, 110	108
C. stenophylla	Me	WA	156	138, 147	141	104	133	187	148	100	114
C. excelsa	P	WA	144, 156	150, 159	137	104	131	193	145	98	114
C. liberica	P	WCA	154	NA	145	104	135	199	145	96	120
C. abeokutae	P	WCA	148, 156	150, 159	137	104	131, 133	193	145, 160	98	114
C. dewevrii	P	CA	154	150, 156	135, 143	104	135, 141	193	145, 163	98	111, 117
C. arnoldiana	P	CA	146, 156	159	137	104	133, 139	187, 193	145, 160	96	114, 120
C. aruweniensis	P	CA	146, 156	150, 159	137	104	133	193, 199	145, 148, 160	98	114, 117
P. bengalensis	Pa	India	156	138, 147	139, 155	104	129, 133	169	175, 179	94	105
P. wightiana	Pa	India	NA	138, 150	NA	104	133	175, 178	166, 169, 178	114	117
P. khasiana	Pa	India	146, 150	138, 150	NA	104	135	169	175, 181	92	108, 117
P. travencorensis	Pa	India	150	129, 150	143, 149	104	129, 133	184, 187	166, 178	104, 122	108, 117
Range of allele size (in bp) over species			140–190	120–159	133–167	104	129–141	169–199	142–181	92–122	105–129
Total alleles scored over all the species			10	9	13	1	9	8	12	11	7

*One representative sample was used for each species except for *C. arabica* and *C. canephora*; †The data for *C. canephora* and *C. arabica* are based on eight and 15 samples, respectively (see Table 1); N_A: No amplification detected; ‡E: Erythrocoffea; M: Mozambicoffea; P: Pachycoffea; Me: Melanocoffea; Pa: Paracoffea; WCA: West and Central Africa; CA: Central Africa; EA: East Africa; WA: West Africa.

© 2004 Blackwell Publishing Ltd,

version 2.0 (Schneider *et al*. 2000) and their significance was tested after applying Bonferroni correction at 0.05% level. Analysis revealed significant ($P < 0.05$) deviations from HWE at five microsatellite loci for arabicas and at one locus for robustas (Table 1). Similarly, significant LD was seen for seven pairs of markers (Cof EST-SSR02 with CofEST-SSR04, CofEST-SSR07, Cof EST-SSR12; CofEST-SSR04 with Cof EST-SSR07, Cof EST-SSR11, Cof EST-SSR12, and Cof EST-SSR07 with Cof EST-SSR12) for arabica and one pair (Cof EST-SSR07 with Cof EST-SSR08) in the case of robusta genotypes. These results were rather expected, as the analysed samples did not represent an interbreeding population but comprised only diverse genotypes (elite, mostly unrelated varieties), which are more akin to a structured population. Further, the later was less apparent in the case of robusta genotypes, probably because many of them shared their pedigree (although not a part of interbreeding population).

Studies to ascertain the cross-species transferability of EST-SSRs described here, across 18 species belonging to *Coffea* and *Psilanthus* revealed robust amplifications in most of the species (Table 2) with high levels of polymorphism. The data thus suggest wider potential of the new markers in genetic analysis of coffee germplasm and comparative genomic studies.

The present study is an initiative in the direction of development of functional coffee-specific microsatellite markers that can be used for genetic improvement of coffee utilizing both primary and secondary genepool. It is hoped that with increasing emphasis on functional genomics, large EST resources will become available for coffee in near future that can be gainfully utilized for developing many more such markers.

Acknowledgements

The authors thank the Department of Biotechnology, New Delhi, India for the financial support, Dr Lalji Singh, Director, CCMB, Hyderabad for the facilities, Dr R Naidu, Director Research, Central Coffee Research Institute, Chikamagalur for the coffee germplasm and Dr M. Udayakumar, University of Agricultural Sciences, Bangalore for leaf materials from the moisture stressed coffee plants.

References

Altschul SF, Madden TL, Schaffer AA, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Baruah A, Naik V, Hendre PS, Rajkumar R, Rajendrakumar P, Aggarwal RK (2003) Isolation and characterization of nine microsatellite markers from *Coffea arabica* (L.) showing wide cross-species amplifications. *Molecular Ecology Notes*, **3**, 647–650.

Combes MC, Andrzejewski S, Anthony F (2000) Characterization of microsatellite loci in *Coffea arabica and* related coffee. *Molecular Ecology*, **9**, 1178–1180.

Marshall TC, Slate J, Kruuk L, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655. [http://helios.bto.ed.ac.uk/evolgen/cervus/cervus.html]

Moncada P, McCouch S (2004) Simple sequence repeat diversity in diploid and tetraploid *Coffea* species. *Genome*, **47**, 501–509.

Rovelli P, Mettulio R, Anthony F (2000) Microsatellites in *Coffea arabica* L. In: *Coffee Biotechnology and Quality* (eds Sera T, Soccol CR, Pandey A, Roussos S), pp. 123–133. Kluwer Academic Publishers.

Schneider S, Roessli D, Excoffier L (2000) *ARLEQUIN ver 2000: a Software for Population Genetics Data Analysis*. Genetics and Biometry Laboratory. University of Geneva, Switzerland. [http://lgb.unige.ch/ar/equin/software/]

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers: their characteristics, development and application to plant breeding and genetics. *Trends Biotech*, in press.