

## **SSR mining in coffee tree EST databases: potential use of EST-SSRs as markers for the Coffea genus**

Valérie Poncet, Myriam Rondeau, Christine Tranchant, Anne Cayrel, Serge Hamon, Alexandre De Kochko, Perla Hamon

► **To cite this version:**

Valérie Poncet, Myriam Rondeau, Christine Tranchant, Anne Cayrel, Serge Hamon, et al.. SSR mining in coffee tree EST databases: potential use of EST-SSRs as markers for the Coffea genus. Molecular Genetics and Genomics, Springer Verlag, 2006, 276 (5), pp.436-449. <ird-01223821>

**HAL Id: ird-01223821**

**<http://hal.ird.fr/ird-01223821>**

Submitted on 6 Nov 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# SSR mining in coffee tree EST databases: potential use of EST–SSRs as markers for the *Coffea* genus

Valérie Poncet · Myriam Rondeau ·  
Christine Tranchant · Anne Cayrel · Serge Hamon · Alexandre K  
ochko · Perla Hamon

**Abstract** Expressed sequence tags (ESTs) from *Coffea canephora* leaves and fruits were used to search for types and frequencies of simple sequence repeats (EST–SSRs) with a motif length of 1–6 bp. From a non-redundant (NR) EST set of 5,534 potential uni-genes, 6.8% SSR-containing sequences were identified, with an average density of one SSR every 7.73 kb of EST sequences. Trinucleotide repeats were found to be the most abundant (34.34%), followed by di- (25.75%) and hexanucleotide (22.04%) motifs. The development of unique genomic SSR markers was optimized by a computational approach which allowed to eliminate redundancy in the original EST set and also to test the specificity of each pair of designed primers. Twenty-five EST–SSRs were developed and used to evaluate cross-species transferability in the *Coffea* genus. The orthology was supported by the amplicon sequences similarity and the amplification patterns. The >94% identity of flanking sequences revealed high sequence conservation across the *Coffea* genus. A high level of polymorphic loci was obtained regardless of the species considered (from 75% for *C. libericato* to 86% for *C. canephora*). Moreover, the polymorphism revealed by EST–SSR was similar to that exposed by genomic SSR. It is concluded that *Coffea* ESTs are a valuable resource for microsatellite mining. EST–SSR markers developed from *C. canephora* sequences can

be easily transferred to other *Coffea* species for which very little molecular information is available. They constitute a set of conserved orthologous markers, which would be ideal for assessing genetic diversity in coffee trees as well as for cross-referencing transcribed sequences in comparative genomics studies.

**Keywords** SSR mining · EST–SSR · *Coffea* · Transferability · Genetic variation

## Introduction

Microsatellite markers derived from anonymous genomic sequences have been extensively used over the last decade because of their highly interesting properties (Ellegren 2004). These generally co-dominant and locus-specific Wc markers have revealed high polymorphism levels. Good transferability between species and sometimes between genera has often been reported (Dirlewanger et al. 2002; Rallo et al. 2003; Gonzalez-Martinez et al. 2004; Poncet et al. 2004). These features have proved to be of great interest for genetic diversity studies, genetic and comparative mapping (Wu and Tanksley 1993; Gonzalo et al. 2005; Saha et al. 2005; Vigouroux et al. 2005). However, despite these advantages, they often correspond to non-coding sequences and thus cannot help in seeking candidate genes.

Expressed sequence tags (ESTs) are sequenced portions of complementary DNA copies of mRNA—they represent part of the transcribed portion of the genome in given conditions. As expected, they mainly correspond to relatively conserved sequences. Techniques have been developed to reveal polymorphism associated with such sequences (Cato et al. 2001), and studies

---

V. Poncet (&) · M. Rondeau · C. Tranchant · A. Cayrel · S. Hamon · A. de Kochko · P. Hamon  
UMR 1097 Diversité et Génomes des Plantes Cultivées (DGPC), IRD, Institut de Recherche pour le Développement, 911 avenue Agropolis, BP 64501, 34394 Montpellier Cedex 5, France; e-mail: poncet@mpl.ird.fr  
URL: <http://www.dgpc.org/index.html>

based on intron polymorphism (less conserved genic regions) have also been carried out (Lemand Lallemand 2003). However, EST sequence analyses revealed appro

ximately 1.5–7.5% of sequences containing microsatellite motifs in cereals (Kantety et al. 2002; Thiele et al. 2003). Among dicotyledonous species, the frequency of ESTs containing SSRs was found to range from 2.65 to 16.82% (Kumpatla and Mukhopadhyay 2005). Therefore, regardless of the plant considered, an ever-increasing number of EST sequences provides a complementary source for microsatellite marker identification. Although the conserved nature of coding sequences may limit their polymorphism, it should facilitate cross-amplification of loci among phylogenetically related species (Scott et al. 2000) and even genera. Moreover, they have a high probability of being associated with functional portions of the genome. Among their many applications, these EST-derived markers (EST-SSR) can be used to cross-reference genes between species for enhancing the resolution in comparative genomics studies and identifying conserved genomic regions among species and genera (Brown et al. 2001; Decroocq et al. 2003; Gupta et al. 2003; Saha et al. 2004; Yu et al. 2004; Park et al. 2005; Varshney et al. 2005b).

In the *Coffea* genus, SSR markers have recently become more available through the construction of enriched genomic microsatellite libraries (Rovelli et al. 2000; Dufouret al. 2001; Baruah et al. 2003). Microsatellite markers were further evaluated for amplification among *Coffea* species (Baruah et al. 2003; Moncada and McCouch 2004; Poncet et al. 2004). Good cross-species transferability and high genetic diversity were generally observed. However, only a handful of linkage maps have been constructed, including these markers (Lashermes et al. 2001; Coulibaly et al. 2003; N'Diaye et al. 2006). These maps were mainly constructed using AFLP markers but the irrelevance for coffee tree breeding programs is limited. Moreover, current *Coffea* maps derived from different *Coffea* species hardly ever share an adequate number of common (anchor) markers to be able to bridge maps.

Public accessibility to *Coffea* EST databases is limited to an SSH library of 527 non-redundant (NR) EST sequences associated with the rust fungus (Fernandez et al. 2004). These resources were very recently enhanced by the generation of 13,175 unigenes from *C. canephora* (Lin et al. 2005). Only nine EST-SSR markers have been developed to date (Bhat et al. 2005).

We recently developed two *C. canephora* EST sequences in our laboratory from cDNA isolated from leaves and fruits at different development and

maturation stages (total of 10,420 sequences, unpublished data). The first objective of the present study was to identify and characterize microsatellites present in the NR set of our ESTs to evaluate its potential as a source for marker development. The second aim was to develop a set of highly polymorphic markers that could cross-amplify and distinguish between *Coffea* species.

Here we report the development of EST-SSR markers based on *C. canephora* sequences and their cross-species transferability to six *Coffea* species. Locus orthology was monitored by analyzing amplification patterns and by sequencing some amplicons. Relevant information about the mutation model and the evolution of these loci was also noted. Polymorphisms detected within and between a set of *Coffea* species was also analyzed to assess whether these markers could be useful for diversity studies and distinguishing between *C. coffea* species. We identified a set of anchor markers, most of them with genes of functional relevance, which would be helpful for functional and comparative mapping within the *Coffea* genus.

## Materials and methods

### Plant DNA

CoVee trees were maintained in a tropical greenhouse at the IRD research center in Montpellier (France).

Total DNA from fully developed leaves was extracted according to Ky et al. (2000). Seven species (*C. canephora* Pierre, known as Robusta (CAN), *C. eugenioides* Moore (EUG), *C. heterocalyx* Stolten (HET), *C. liberica* Portères (LIB), *C. dewevrei* Portères (DEW), *C. sp.* Moloundou (MOL), and *C. pseudozanguebariae* Bridson (PSE)) representing the three African main geographical clades (Lashermes et al. 1997) were analyzed. Polymorphism was assessed in 12 CAN, 10 DEW, 10 PSE and 8 LIB, representative of the genetic diversity of the four species, plus 2 EUG, 2 MOL, and the only known HET individual.

### Data mining for SSR markers

The coffee EST databases used in this project were developed in our laboratory (de Kochko et al., unpublished data). They contained 10,420 sequences derived from fruit (5,814 sequences) and leaf (4,606) cDNA libraries (valid sequences submitted to GenBank under accession numbers E191792–EE200565). The raw chromatograms were processed using ESTdb software

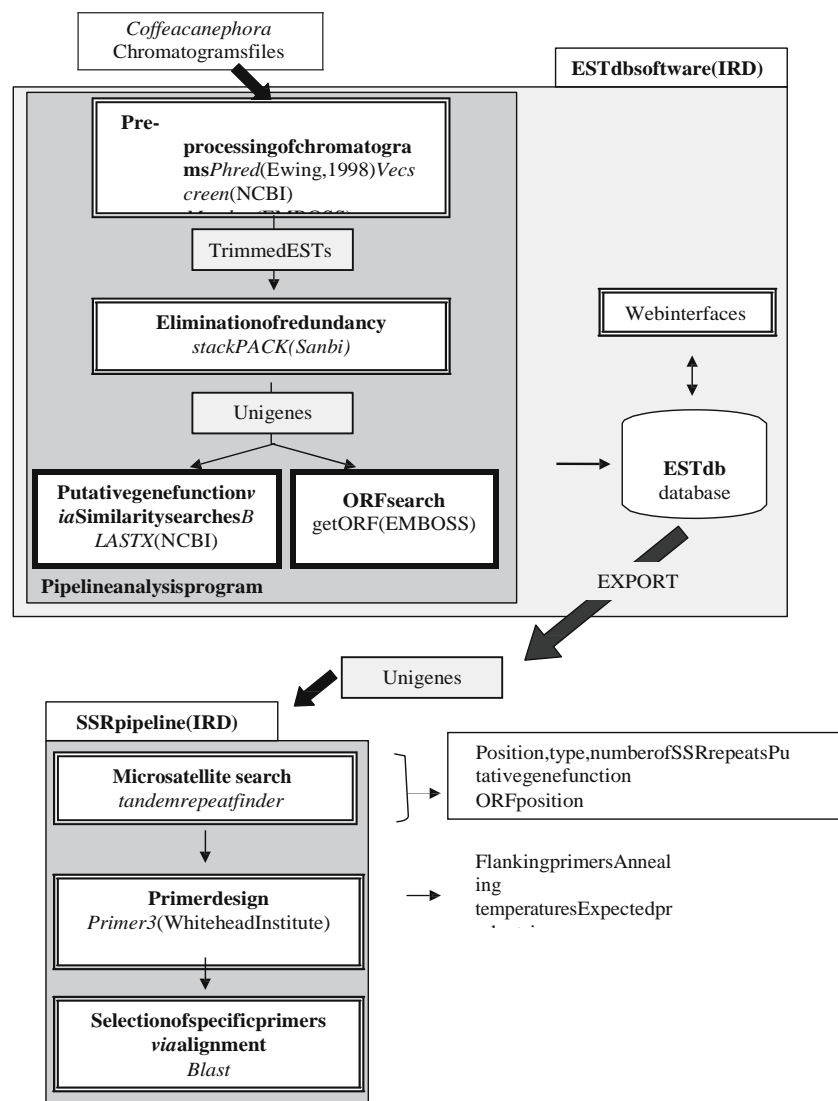
(<http://www.mpl.ird.fr/bioinfo/>). ESTdb is a set of analytical procedures that automatically verify, cleanse, store and analyze ESTs generated in our laboratory. ESTdb has three major components: (1) a pipeline analysis program based on public software integrated by an in-house developed script, (2) the results and related information restored in a relational database accessible through (3) a web interface (Fig. 1). The functions of the EST sequences were predicted through similarity searches from protein sequence GenBank databases (<http://www.ncbi.nlm.nih.gov>) using BLASTx (Altschul et al. 1997). Potential unigenes (contigs and singletons from all EST sequences processed simultaneously) were identified after clustering for a NEST set.

These unigene sequences were screened for the presence of perfect and imperfect microsatellites using SSR pipeline, a Perl script developed by Dubreuil-Tranchant (SSR pipeline is publicly available at <http://www.mpl.ird.fr/bioinfo/>). This program integrated three

public software packages: (1) a Perl program developed by Cartingour (<http://www.gramene.org>) which detects perfect microsatellites, (2) Repeat Tandem Finder, an SSR repeat finder (<http://www.tandem.bu.edu>), and (3) Primer3, a PCR primer design program (Rozen and Skaletsky 2000). Moreover, this program allowed us to check the specificity of primer pairs by blasting against the EST sequences. This tool screened each sequence for SSRs. The parameters were reset for detection of mono-, di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of 15, 9, 6, 5, 4, and 3 repeats, respectively. The major primer design parameters were reset as follows: primer length from 18 to 21 bp (optimum 20), PCR products size from 100 to 300 bp, optimum annealing temperature 60°C.

A set of 25 EST-SSR primers was further analyzed (sequences submitted to GenBank under accession numbers DQ778713–DQ778737). They were selected based on the feature of these sequences: we kept in priority

**Fig. 1** Flowchart of the clustering and annotation of ESTs via the IRD pipeline ESTdb, data mining of SSR-containing ESTs and primer design for a non-redundant SSR set



singletonESTs. Every forward primer was 5-tailed with the M13 sequence 5-CACGACGTTGTAAAACGAC-3. The primers were synthesized by MWG-Biotech AG (Ebersberg, Germany).

Polymerase chain reaction (PCR) amplification and visualisation of amplicons

Polymerase chain reaction amplifications were carried out as described in Coulibaly et al. (2003) and Poncet et al. (2004) using a touchdown PCR profile optimised for each set of primers: touchdown 60–55°C or touch-down 55–50°C. PCR products were detected on an IR<sup>2</sup> Automated DNA Sequencer (LI-COR, model 4200L-2, Lincoln, NE, USA) using an M13 primer coupled to the infrared tag IRD700 or IRD800 and after migration on 25 cm 6.5% KB plus (LI-COR, CAT#827-05607) polyacrylamide gels.

The gel images were processed by SAGAGT™ software (LICOR Biotech) to estimate the size of amplicons according to a 50–350 bp size standard (LI-COR, CAT#829-05343, 829-05344).

Microsatellite cloning and sequencing

Genomic DNAs were amplified with the appropriate forward and reverse primers but without the infrared fluorescent M13. The products were purified using EZNA Cycle-Pure (OMEGA Bio-Tek, Doraville, GA, USA) and cloned onto the pCR®4-Topo plasmid using the TOPOTA cloning kit (Invitrogen, Groningen, The Netherlands) according to the manufacturer's instructions.

Cloning efficiency was checked through PCR amplification of several colonies. The resulting products were run on polyacrylamide gels, with the initial genomic DNA amplification outcome as control. The resolution used at this step was especially important to distinguish between two alleles from heterozygous individuals.

After plasmid DNA purification (Sambrook et al. 1989), the cloned PCR fragments were sequenced by MWG-Biotech (<http://www.mwg-biotech.com/html/all/index.php>). These sequenced data have been submitted to GenBank under accession numbers DQ787368–DQ787384. Sequences were edited and analyzed using the DNASTAR software package (Lasergene, Madison, WI, USA).

Data analysis

Number of alleles, observed heterozygosity (H<sub>o</sub>), gene diversity

information content (PIC), null allele frequency (ra) and heterozygote deficiency or excess (Fis) were calculated for each locus and for the four species represented by more than five genotypes, as indicated in Poncet et al. (2004), using PowerMarker (Liu and Muse 2005). For each species, polymorphic loci at 0.05 threshold frequency and mean allele number per polymorphic locus were evaluated.

## Results

### EST unigenesets

A total of 9,820 (94%) valid ESTs were obtained from all *C. canephora* chromatogram files. The average length was 602 bp. After clustering and assembly, 5,534 unigenes were identified, including 3,747 singletons and 1,787 contigs. The BLAST results revealed that about 22.3% of these NRESTs showed no similarity to any GenBank sequences. Less than 0.03% of the ESTs was predicted to be from plastid cDNA.

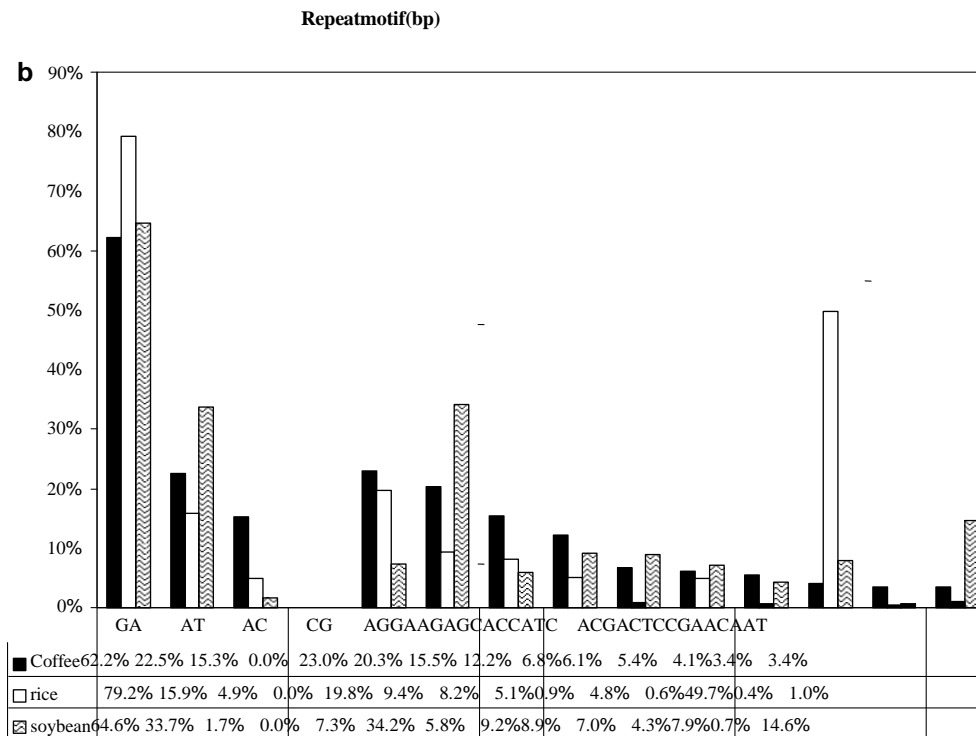
### Frequency and distribution of *C. canephora* EST-SSR

EST-SSRs were mined from a NREST set of 5,534 potential unigenes. According to the search criteria adopted, 431 unique SSRs were found within a total of 376 unigenes. The chance of finding an SSR-containing sequence in the NREST database was thus 6.8%, with an average density of one microsatellite every 7.73 kb.

Tri-nucleotide motifs were the most abundant (34.34%), followed by di- (25.75%), and hexa-nucleotide repeats (22.04%) (Fig. 2). The most abundant trinucleotide repeat motif was AGG/TCC (23%), followed by AAG/TTC (20.3%) and AAC/TTG and AAT/TTA were the least abundant motifs (3.4% each) (Fig. 2). Among the dinucleotide motifs, GA/CT was the most abundant (62.2%) and no GC/CG motif was found. For all repeat classes, the mean SSR length was 20.6 bp, but was higher for dinucleotides (24.1 bp), with a maximum of 94 bp.

Out of the 25 EST-SSRs used for designing primers, nine were found in translated regions, eight in the 5' untranslated terminal region (5' UTR) and four in the 3' UTR. The four other sequences generated no hits in the similarity search (Table 1).

### Conservation of orthologous SSR loci among *Coffea* species



**Fig.2** SSR frequency in the unigene *Coffea canephora* EST set (rice and soybean data from Gao et al. 2003) **a** according to the repeat motif length, **b** according to the motif of di- and trinucleotide repeats

cross-amplification of 25 primer pairs on seven species representing three different *Coffea* phylogenetic clades. Out of 25, 24 primer pairs amplified at least one species (only ES21 could not) (Table 2). Two microsatellites (ES3, ES15) did not amplify well, and the amplified products were substantially larger than expected. Finally, cross-species transferability within the *Coffea* genus was very high since each primer pair yielded a PCR product in an average of 6.1 species (out of seven tested) and 14 (56%) gave amplification with all tested *Coffea* species.

One or two bands per genotype were observed in all but one of the readable amplification patterns obtained, suggesting that

there was selective amplification of one locus. The ES11 primer pair amplified 3–6 products with some *C. liberica* genotypes, indicating a possible

selective amplification



triplicate sequence—  
this was considered as missing data for the subsequent analyses.

Given that primers were redesigned on the basis of *C. canephora* EST sequences, amplicons produced from *C. canephora* genomic DNA had the expected size in most cases (Table 2), consistent with locus-specific amplification. In all other *Coffea* species, the allele sizes also did not substantially differ from the expected size, which was consistent with selective orthologous amplification. The microsatellite allele size distributions, expressed in base pairs, fitted the stepwise model in most cases (Table 2)—with allelic differences, which are multiples of the repeat motif even if some intermediate allelic states were missing. Some “jumps” in the allele size distribution, associated with mixed distribution patterns, might be attributed to the additional



**Table1**

Expressed sequence tag (EST), putative gene function based on BLAST search of GenBank sequences (analysis on December 2005), SSR position in the cDNA sequence (corresponding repeated amino acid when translated)

<i>Coffea</i> EST-SSR				Homologous sequences				
Locus	Repeat motif	Expected ampliWcon size(bp)	SSR position	GenBank accession No.	Species	BLAST Evalue	Identity (%)	Putative function
ES1	(GCA) <sub>8</sub>	308	ORF(Q)	AAD02552	<i>Petunia hybrida</i>	2e106	64	PGPS(P-glycoprotein)
ES2	(A) <sub>14</sub> (CA) <sub>19</sub>	349						Nohit
ES3	(TTTCT) <sub>9</sub>	371	5 UTR	CAB75429	<i>Nicotiana plumbaginifolia</i>	1e141	90	Oligouridylate binding protein
ES4	(CT) <sub>14</sub>	337	3 UTR	CAA64545	<i>Hordeum vulgare</i>	4e121	73	Xylose isomerase
ES5	(TTC) <sub>24</sub>	309	ORF(S)	AAL65125	<i>Glycine max</i>	2e131	66	GT-2 transcription factor
ES6	(CCA) <sub>2</sub> (CCT) <sub>11</sub>	228	5 UTR	AAN71762	<i>Solanum tuberosum</i>	1e141	75	Cinnamoyl CoA reductase 2
ES7	(TCA) <sub>18</sub>	298						Nohit
ES8	(TC) <sub>26</sub>	323	5' UTR	AAB84350	<i>Arabidopsis thaliana</i>	3e131	72	Heat shock transcription factor
ES9	(TTC) <sub>10</sub>	205	5 UTR	AAD55726	<i>Vitis riparia</i>	7e174	75	Galactinol synthase
ES10	(ACC) <sub>11</sub> /(TCC) <sub>6</sub>	403	ORF(T/S)	AAP40485	<i>Arabidopsis thaliana</i>	7e-47	56	AP2 domain transcription factor
ES11	(TTC) <sub>10</sub>	240	5 UTR	AAN28269	<i>Gossypium hirsutum</i>	1e163	94	myb-Like transcription factor
ES12	(AAT) <sub>4</sub> (CAG) <sub>13</sub>	156	ORF(Q)	AAM54033	<i>Populus tremula</i>	8e169	67	PIN1-like auxin transport protein
ES13	(AG) <sub>16</sub>	235	5 UTR	AAW28999	<i>Antirrhinum majus</i>	3e144	53	1-Deoxy-D-xylulose-5-phosphate synthase
ES14	(CCA) <sub>5</sub> /(CCA) <sub>10</sub>	289	ORF(T)	AAF01534	<i>Arabidopsis thaliana</i>	1e153	56	Protein kinase
ES15	(CT) <sub>12</sub>	308	5 UTR	AAD42941	<i>Catharanthus roseus</i>	2e126	81	Ubiquitin-conjugating enzyme E2
ES16	(TA) <sub>15</sub>	186	3 UTR	AAC62396	<i>Ricinus communis</i>	2e154	72	Cysteine endopeptidase precursor
ES17	(TG) <sub>16</sub>	300	5 UTR	CAG18177	<i>Arabidopsis thaliana</i>	1e174	83	UDP-galactose transporter
ES18	(GGATCA) <sub>5</sub>	202	ORF(SG)	AAM67530	<i>Arabidopsis thaliana</i>	1e164	58	Unknown protein
ES19	(CA) <sub>13</sub> (TA) <sub>6</sub>	259	ORF(TH)	AAQ14193	<i>Solanum tuberosum</i>	2e162	67	Calcium homeostasis regulator
ES20	(AAC) <sub>9</sub>	197	ORF(N)	AAO45753	<i>Cucumis melo</i>	1e151	57	RING/C3HC4/PHD zinc finger protein
ES21	(T) <sub>28</sub>	254	3 UTR	AAO33591	<i>Arachis hypogaea</i>	7e150	65	Early light induced protein
ES22	(TTTTTTC) <sub>4</sub>	207	3 UTR	P25317	<i>Nicotiana tabacum</i>	7e150	65	glutathione S-transferase (auxin-regulated protein)
ES23	(AG) <sub>16</sub>	230						Nohit
ES24	(TCTCC) <sub>7</sub>	224						Nohit
ES25	(CCT) <sub>9</sub>	207	ORF(P)	AAG45420	<i>Chlamydomonas reinhardtii</i>	6e110	33	Vegetative cell wall protein gp1

**Table 2** Number and molecular size ranges of 25 EST-SSR tested on eight *Coffea* species: *C. canephora* CAN, *C. dewevrei* DEW, *C. pseudozanguebariae* PSE, *C. liberica* LIB, *C. heterocalyx* HET, *C. eugenioides* EUG, and *C. sp. Moloundou* MOL (sample size given in brackets)

EST-SSR	Expected size (bp)	Allele size range <sup>a</sup> (bp)/alleles no.							Total no of alleles	Type of allele size distribution
		CAN(12)	DEW(10)	PSE(10)	LIB(8)	HET(1)	EUG(2)	MOL(2)		
ES1	327	318–327 3	306–318 4	306–321 2	306–318 3	318 1	318 2	327 1	7	Stepwise
ES2	368	356–371 6	356–363 4	342–356 3	342–358 4	342 1	346 1	349 1	10	Stepwise
ES3	390	>2,000 –	>2,000 –	>2,000 –	>2,000 –	>2,000 –	>2,000 –	>2,000 –	–	–
ES4	356	344–372 8	338–396 8	338–362 5	346–372 6	350 1	362–368 3	– –	15	Mixed
ES5	328	309–315 3	291–306 5	300–309 4	297–315 4	294 1	291–294 2	300 1	9	Stepwise
ES6	247	256 1	238–268 4	244–259 3	256–298 4	238–256 2	238–256 3	262–259 2	9	Mixed
ES7	317	300–324 10	302–315 6	288–290 2	288–307 6	307 1	288–302 3	332 1	19	Nearly continuous
ES8	251	236–248 5	246–258 4	222–234 3	236–258 5	228 1	226–234 2	224 1	13	Stepwise
ES9	224	204–230 6	210–218 3	214 1	204–214 4	– –	215 1	– –	9	Nearly stepwise
ES10	422	395–415 7	392–420 5	404 1	389–404 5	– –	– –	– –	11	Nearly stepwise
ES11	259	246–266 8	240–269 6	254–260 2	240–266 9	246 1	234 2	243–263 2	14	Nearly stepwise
ES12	175	160–178 5	169–178 3	151–166 2	169–178 4	154 1	160–163 2	169 1	9	Stepwise
ES13	254	235–255 10	237–241 2	233–237 2	233–241 2	237 1	239–241 2	– –	12	Stepwise
ES14	308	297 1	291–297 3	294–297 2	291–294 2	297 1	– –	297 1	3	Stepwise
ES15	327				Nonscorable					–
ES16	205	180–222 11	206–212 4	195–212 2	207–236 8	– –	196–198 2	196 1	17	Mixed
ES17	319	»640	»640	»640	»640	»640	»640	»640		
ES18	221	218–222 2	212 1	206–218 2	212–218 2	206 1	206 1	206–218 1	4	Stepwise
ES19	278	256–284 9	250 1	250–252 2	244 1	252 1	252 1	242 1	13	Stepwise
ES20	216	204–219 4	212 1	215–229 7	209–212 2	212 1	212–219 2	212 1	11	Mixed
ES21	273	–	–	–	–	–	–	–	0	–
ES22	226	213–232 6	– –	– –	213–220 2	– –	230 1	– –	6	Mixed
ES23	249	236–262 8	238–240 2	240–270 12	236–260 5	246 1	236 1	244–246 2	16	Stepwise

**Table 2** continued

EST-SSR	Expected size(bp)	Allele size range <sup>a</sup> (bp)/alleleno.												Total no of alleles	Type of allele size distribution	
		CAN(12)	DEW(10)	PSE(10)	LIB(8)	HET(1)	EUG(2)	MOL(2)								
ES24	243	228-247	225-230	230-240	225-235	242	235	230								Mixed
ES25	226	224-227	218-221	230-239	218-221	230	239	-								Stepwise
No. markers		21	20	20	20 <sup>c</sup>	17	19	15								21
Total alleleno.		119	70	62	73 <sup>c</sup>	18	33	18								220
Average alleleno./polymorphic loci		5.7	3.5	3.1	3.7 <sup>c</sup>	1.1	1.7	1.2								10.5
		18	16	16	15 <sup>c</sup>											
		6.4	4.1	3.5	4.1 <sup>c</sup>											

<sup>a</sup>M13-tatime detected in the size

<sup>b</sup>Polymorphic loci at the 5% level for sample size equal or over 10 in individuals

<sup>c</sup>ES11 locus not included (multilocus marker in LIB)

contribution of insertions/deletions (indels) in the amplified region. This pattern was observed between species at the ES16 locus. However, some substantial allele size differences were observed with three primer pairs (ES3, ES15 and ES17), with amplicons much larger than expected, irrespective of the species. For instance, the observed allele size for ES17 was about 640 bp for all species versus 319 bp expected.

Products obtained by amplification with ES2, ES4, ES6, ES16, ES17, ES19, and ES25 primer pairs were cloned and sequenced to confirm the orthology. They were also analyzed to identify the origin of within- and between-species polymorphisms as well as discrepancies between the observed and expected allele sizes (Table 3). The genotypes chosen for cloning presented allele sizes at the extremes of the distribution of a given marker. The relative sequence sizes were in agreement with the sizes evaluated after migration, with a mean SAGA software underestimation of

6.72 bp. The genomic DNA sequences clarified three main points:

(1) there was strong sequence homology all along the regions flanking the repeat motif between amplicons derived from different species (mean 96.9% identity), in agreement with the amplification of orthologous loci; (2) the observed polymorphism mainly resulted from variations in repeat number, although some more complex mutation patterns (ES6, ES16) were also detected, involving additional variation of indels in the flanking sequences; (3) the large size of ES17 genomic fragments was due to amplification of a 426 bp intron in all species, although the polymorphism still resulted from a different number of repeats

#### Within- and between-species polymorphism

The number and molecular size ranges of alleles obtained with each EST-

SSR are shown in Table 2 together with the number of polymorphic loci. One locus appeared to be monomorphic for *C. liberica* (ES19), two for *C. canephora* (ES6, ES14) and *C. pseudozanguebariae* (ES9, ES10), and three for *C. dewevrei* (ES18, ES19, ES20). For the remaining loci, up to 12 alleles per locus were recorded (ES23 amplification in *C. pseudozanguebariae*). For the three species represented by one (*C. heterocalyx*), two (*C. sp. Moloundou*) or three (*C. eugenioides*) genotypes, 18 alleles at 17 and 15 loci were detected for *C. heterocalyx* and *C. sp. Moloundou*, respectively, while 33 alleles at 19 loci were observed for *C. eugenioides*.

When considering markers with di-, tri-, and more microsatellite repeated motifs, the average number of alleles detected per locus was 14.3 (86 alleles, 6 loci), 9.7 (107 alleles, 11 loci) and 5.7 (17 alleles, 3 loci),

**Table 3** Allele sequence analysis at seven microsatellite loci for *Coffea* species genotypes (homozygotes or heterozygotes) related to the *C. canephora* EST sequence (*ESTdb*): total length,

microsatellite repeat stretches, introns, indels (insertions/deletions), and % similarity of the flanking sequences

Locus	<i>Coffea</i> species	Individual	Cloneno.	Size(bp)	Polymorphism	Percentage similarity of flanking sequences
ES2	CAN	<i>ESTdb</i>		349	(A) <sub>14</sub> (CA) <sub>19</sub>	100
	PSE	P-8056	ES2-11	341	(A) <sub>14</sub> (CA) <sub>15</sub>	97.3
	CAN	C-IF182	ES2-21	357	(A) <sub>18</sub> (CA) <sub>21</sub>	98.0
ES2-22			351	(A) <sub>18</sub> (CA) <sub>18</sub>	98.3	
ES4	CAN	<i>ESTdb</i>		337	(CT) <sub>14</sub>	100
	DEW	D5765	ES4-15	349	(CT) <sub>20</sub>	97.4
ES6	CAN	<i>ESTdb</i>		228	(CCA) <sub>2</sub> (CCT) <sub>11</sub>	100
	PSE	P-55	ES6-12	227	(CCA) <sub>3</sub> (CCT) <sub>8</sub>	95.5
			ES6-11	220	(CCA) <sub>3</sub> (CCT) <sub>6</sub>	95.5
	LIB	L-A12	ES6-25	255	(CCA) <sub>4</sub> (CCT) <sub>10</sub> +20bp	93.9
ES6-26			247	(CCA) <sub>2</sub> (CCT) <sub>10</sub>	95.5	
ES16	CAN	<i>ESTdb</i>		186	(TA) <sub>15</sub>	100
	CAN	C-IF72	ES16-12	199	(TA) <sub>21</sub>	96.1
			ES16-25	165	(TA) <sub>4</sub>	95.9
	LIB	L-A22	ES16-15	206	(TA) <sub>17</sub> +17bp	95.5
ES17	CAN	<i>ESTdb</i>		300	(TG) <sub>16</sub>	100
	PSE	P60	ES17-10	718	(TG) <sub>10</sub> +intron426pb	97.9
ES19	CAN	<i>ESTdb</i>		260	(CA) <sub>13</sub> (TA) <sub>6</sub> (T) <sub>2</sub>	100
	CAN	C-IF461	ES19-15	266	(CA) <sub>10</sub> (TA) <sub>11</sub> (T) <sub>4</sub>	98.6
			ES19-16	246	(CA) <sub>2</sub> (TA) <sub>7</sub> (T) <sub>9</sub>	98.2
	LIB	L-A11	ES19-26	233	(CA) <sub>1</sub> (TA) <sub>0</sub> (T) <sub>9</sub>	99.1
ES25	CAN	<i>ESTdb</i>		207	(CCT) <sub>9</sub>	100
	DEW	D-5462	ES25-11	200	(CCT) <sub>6</sub>	98.3
	PSE	P-1	ES25-24	219	(CCT) <sub>13</sub>	96.0

CAN: *C. canephora*, DEW: *C. dewevrei*, PSE: *C. pseudozanguebariae*, LIB: *C. liberica*

respectively. When considering the location of the microsatellite motif in relation with the gene annotation, an average of 8.1 (73/9), 12.7 (11/3), and 11.4 (57/5) alleles per locus were obtained for ORF, 3'UTR and 5'UTR locations, respectively.

Clearly readable loci were analyzed to evaluate genetic diversity parameters, while taking into account only simple locus markers which gave an amplification with more than five genotypes per species, between 16 (*C. liberica*) to 21 (*C. canephora*) (Table 4). A high level of polymorphic loci was observed whatever the species considered (from 75% (*C. liberica*) to 86% (*C. canephora*)). The mean allele number per polymorphic locus was highest for *C. canephora* (6.4) while lower but similar values were obtained for *C. dewevrei*, *C. liberica*, and *C. pseudozanguebariae* (4.1, 4.1, and 3.5, respectively). The observed heterozygosity was lowest for *C. pseudozanguebariae* (0.28) and highest for *C. liberica* and *C. canephora* (0.52 and 0.51, respectively). The PIC value ranged from 0 (monomorphic locus) to 0.91 (ES23 for *C. pseudozanguebariae*, highly polymorphic locus) across the four species, with the lowest average value obtained for *C. pseudozanguebariae* (0.40) and similar average values for *C. canephora* and *C. liberica*

(0.59 and 0.57). There was no significant correlation between interspecific PIC values except for the *C. liberica*/*C. dewevrei* combination (0.78;  $P=0.003$ ).

Although variations between loci were observed, a global heterozygote deficit ( $F_{is}=0.10-0.26$ ) and the presence of null alleles ( $r_b=0.03-0.09$ ) could be estimated for all species. The lowest values were obtained for *C. canephora* and *C. liberica*, while the highest were obtained for *C. dewevrei* and *C. pseudozanguebariae*.

#### Diagnostic alleles and locus identification

Allele sizes identified for the cultivated species *C. canephora* (CAN) were compared to the others. DEW/PSE species comparison was also considered since this couple was involved in the first interspecific map (Ky et al. 2000). From 4 to 16 loci could be identified as diagnostic markers, i.e. markers for which there were no shared alleles between species (Table 4). The smallest number of such loci was obtained for the CAN/LIB comparison and the greatest for CAN/MOL. Moreover, within these markers, some were charact

er-ized by a non-  
overlapping allele size distribution for both species: four for  
CAN/PSE, seven for CAN/HET,

**Table 4** Diversity statistics for 21 EST-SSR in *Coffea*

species as described by the expected ( $H_e$ ) and observed ( $H_o$ ) heterozygosities, the polymorphism information content (PIC), null allele frequency (rb), and heterozygote deficiency or excess (Fis)

EST-SSR	Heobs				PIC				Null allele frequency (rb)				Fis				Diagnostic loci species combination					
	CAN	DEW	PSE	LIBHET	EUG	MOL	CAN	DEW	PSE	LIB	CAN	DEW	PSE	LIB	CP	CH	CL	CE	CM	PD		
ES1	0.17	0.75	NA	0.43	0.1/2	0	0.41	0.73	NA	0.52	0.17	0.01	NA	0.06	0.59	0.02	NA	0.18				
ES2	0.92	0.6	0.3	0.83	0.00	0.78	0.47	0.47	0.65	0.08	0.09	0.11	0.11	0.19	0.29	0.35	0.28		++	++		
ES4	0.77	0.78	0.5	1	0	2/3	-	0.74	0.84	0.75	0.81	0.02	0.07	0.14	0.11	0.04	0.16	0.33	0.24	*		
ES5	0.17	0.33	0.43	0.83	0.00	0.64	0.58	0.58	0.64	0.29	0.16	0.1	0.12	0.74	0.43	0.26	0.3		++	++		
ES6	0	0.2	0.1	0.33	1/1	1/2	2/2	0	0.51	0.19	0.69	0		0.21	0.07	0.21	NA*	0.61	0.46	0.52		
ES7	0.92	0.5	0.3	0.5	0	1/3	0	0.86	0.55	0.5	0.78	0.04	0.03	0.13	0.16	0.08	0.09	0.39	0.36			
ES8	0.77	0.4	0.22	0.57	0.1/2	0	0.74	0.73	0.2	0.71	0.02	0.19	0.01	0.08	0.04	0.45	0.08	0.2		++		
ES9	0.77	0.25	NA	0.33	-	0/1	-	0.75	0.59	NA	0.65	0.01	0.22	NA	0.19	0.03	0.58	NA	0.49			
ES10	0.25	0.4	NA	0.33	-	-	-	0.78	0.72	NA	0.78	0.3	0.18	NA	0.25	0.68	0.44	NA	0.57			
ES11	0.85	0.6	0.3	trip	0	1/2	1/2	0.7	0.73	0.46	trip	0.09	0.08	0.11	trip	0.21	0.18	0.34	trip			
ES12	0.83	0.22	0.1	0.83	0.1/2	0	0.68	0.63	0.1	0.65	0.09	0.25	0		0.11	0.23	0.65	0.05	0.28			
ES13	0.77	0.2	0.5	0.17	0.3/3	-	0.75	0.48	0.46	0.38	0.01	0.19	0.03	0.15	0.03	0.58	0.1	0.56		+		
ES14	0	0	0.11	0.5	0	-	0.00	0.46	0.11	0.38	0			0.32	NA	0.09	NA*	1	NA*	0.33		
ES16	0.92	0.75	0	0.83	-	1/2	0/1	0.88	0.62	0.28	0.85	0.03	0.08	0.22	0.01	0.05	0.22	1	0.02			
ES18	0.1	0	0.37	0.5	0.2	0	2/3	0	0.1	0	0.43	0.18	0	0.04	0.02	0.05	NA*	0.12	0.11			
ES19	0.62	NA	0	0.01	1/2	0	0.83	NA	0.41	0.01	NA	0.29	0		0.25	NA	1	NA*				
ES20	0.1	0	0.89	0.33	0	0	0/1	0.51	0	0.76	0.44	0.07	0		0.07	0.08	0.2	NA*	0.17	0.25		
++ES22			0.77	-	-	-	NA	-	0	-	0.72	-	-	NA	0.03	-	-	NA				
*		*																				
ES23	0.46	0.5	0.78	NA	0	0.1/2	0.84	0.38	0.91	NA	0.2	0.09	0.07	NA	0.45	0.33	0.14	NA				
ES24	0.38	0.13	0.33	NA	0.0/1	0/10	0.38	0.12	0.29	NA	0			0.01	0.03	NA	0.01	0.07	0.15	NA		
ES25	0.2	0.38	0.33	NA	0.0/1	-	0.32	0.43	0.28	NA	0.09	0.04	0.04	NA	0.38	0.13	0.2	NA				
Average	0.51	0.37	0.28	0.52			0.59	0.50	0.40	0.57	0.03	0.09	0.06	0.04	0.10	0.26	0.23	0.11	Total	9+ 1*8+ 4*4+ 8+ 2*10+ 6*7+		

Diagnostic loci: with no overlapping allelic range between species of each couple

NA not applicable (less than  $W$  in individuals analysed or division by zero), - indicates no amplification, trip denotes triplicated locus, ++ indicates no overlapping alleles sized distributions, + indicates no common alleles between the two species, \* indicates no amplification with the non-

CAN species, C: *C. canephora*, P: *C. pseudozanguebariae*, H: *C. heterocalyx*, E: *C. eugenioides*, M: *C. sp.* Moloundou, D: *C. dewevrei*, L: *C. liberica*

three for CAN/LIB, six for CAN/EUG, seven for CAN/MOL, and Wve for PSE/DEW comparisons.

## Discussion

ESTs are a rich source of SSRs in *Coffea*

A total of 431 unique SSRs were identified from 5,534 potential unigenes. These results clearly demonstrated that *Coffea* ESTs are a valuable resource for mining SSR markers.

Random sequencing with cDNA libraries leads to a high proportion of redundant ESTs. In our study, redundancy was eliminated prior to analysis in order to reduce the data set size. The advantage of using NREST is to avoid overestimation of the EST–

SSR frequency. For example, after redundancy elimination, Kumpatla and Mukhopadhyay (2005) observed 37.3% loss in the number of SSR–EST for *Arabidopsis thaliana*.

For coffee trees, we found that 6.8% potential unigenes contained microsatellite motifs, with an average of one microsatellite every 7.73 kb of EST sequence. This SSR–EST frequency was in the 2.65–

10.62% range obtained by Kumpatla and Mukhopadhyay (2005) for dicot species. It was slightly higher than the 1.5–4.7% range reported by Kantety et al. (2002) for monocots. The overall frequency and the frequency of the di- and tri-nucleotide repeat motifs are known to be dependent on the presence

or not of redundancy, but also related to the criteria used to identify SSR in the database mining. In general, when the minimum repeat length is 20 bp, microsatellites of various plant species are present in about 5% of ESTs (Varshney et al. 2005a).

Our results were somewhat in agreement with this estimation although our overall SSR–EST frequency might have been inflated by the detection of hexanucleotide repeat motifs in the analysis, a factor that is seldom considered in other studies (see below).

In *C. canephora* ESTs, trinucleotide repeats were found to be the most abundant (34.34%), followed by di- (25.75%) and hexanucleotide motifs (22.04%).

Trinucleotide repeats are generally the most common motif found in both monocots (54–78% among cereals, Varshney et al. 2005a) and dicots (for example 51.5% in *Medicago truncatula*, Eujaylet al. 2004). On Lyon report (Kumpatla and Mukhopadhyay 2005) has described the prevalence of dinucleotide repeats in most of the dicots investigated. However, they suggested that their results might have been due to the over-representation of untranslated regions (UTRs) compared with open reading frames (ORFs). Indeed, because of the absence of frame shift mutations when there are length

variations in tri- and hexanucleotide repeats, these motifs are found in excess in both coding and noncoding sequences, but other repeat types are much less frequent in coding regions than in UTRs (Metzgar et al. 2000). Our results confirmed this distribution (Table 1): 7/9 microsatellites found in coding sequences were trinucleotide repeats and one was a hexanucleotide repeat, whereas microsatellites in UTR were mainly dinucleotide repeats (6/12) and trinucleotide repeats (3/12).

As most microsatellite libraries used for marker development are generally enriched in di-, tri- and tetra-nucleotide repeats, computational mining of EST database mainly involves these types of motifs (Kantety et al. 2002; Morgante et al. 2002; Eujaylet al. 2004; Pinto et al. 2004; LaRota et al. 2005). However, a non-negligible abundance of mono- and hexanucleotide SSRs was observed in our study, i.e. 12.30 and 22.04%, respectively (Fig. 2). These data were in close agreement with those of Gao (2003), and supported the results obtained by Kumpatla and Mukhopadhyay (2005) and Morgante et al. (2002) on the abundance of mononucleotide SSRs. Moreover, when considering *Coffea* microsatellites with mono- to hexanucleotide repeats, their relative distribution matched that noted in soybean (Gao et al. 2003; Fig. 2).

The GA/CT motif was the most abundant dinucleotide motif (62.2%) in our *Coffea* ESTs. These motifs were also the most frequently observed SSRs in plants (Scott et al. 2000; Gao et al. 2003; Thiele et al. 2003; Sah et al. 2004). The most abundant trinucleotide repeat motif detected in the present study was AAG/TTC (23%), closely followed by AAG/TTC (20.3%). These results are in agreement with other reports on dicot species (Scott et al. 2000; Eujaylet al. 2004; Kumpatla and Mukhopadhyay 2005). The rare CCG/GGC frequency compared with rice (Fig. 2) confirmed the general trend noted in monocots, i.e. they have more CG-rich trinucleotide repeats than dicots (Morgante et al. 2002).

Ortholog amplification and cross-species transferability

Our computational strategy to detect NRSSRs and develop unique EST–SSR markers appeared efficient since nearly all the markers appeared to be single locus specific.

The use of a set of NR sequences was important for the development of unique genic SSR markers. The specificity of the designed primer pairs was then checked by blast against the EST sequences. However, the presence of paralogs (sequences derived from duplica-

tion events) together with orthologs (sequences derived from

the same common ancestor), might still be possible and it was analyzed on four *Coffea* species (2 to 9 alleles). A comparison of genetic parameters estimated from this previous study (Poncet et al. 2004) and the

the source of difficulties in interpreting between-species comparisons. The >94% identity of amplicon sequences in SSR flanking regions and the maintenance of repeat motifs confirmed the cross-species sequence conservation and the primer specificity. Finally, the orthology was further confirmed by an analysis of amplification patterns (size and number of amplification products).

Cross-species transferability within the *Coffea* genus was very high since each primer pair yielded a PCR product in an average of 6.1 species (out of seven tested) and 14 (56%) gave amplification with all the tested *Coffea* species, regardless of the phylogenetic relationships. This is an important feature of genic SSR markers, which are transferable among distantly related species or even genera (Decroocq et al. 2003; Liewlaksaneeyanawin et al. 2004; Varshney et al. 2005b; Sethy et al. 2006). Compared to previous results based on genomic microsatellites (Poncet et al. 2004), EST-SSR appeared to be more transportable markers (62.5–92 vs. 61.7%). Similarly, the transferability of EST-SSR markers from *Pinus taeda* (loblolly pine) to *P. contorta* sp. *latifolia* was total, while it was only less than a third for non-EST derived microsatellite markers (Liewlaksaneeyanawin et al. 2004).

### EST-SSR polymorphism

A high level of polymorphic loci was noted whatever the species considered. Although the conserved nature of EST-SSR promoted transferability, it could also have limited polymorphism. This has been suggested in several reports where the level of EST-SSR polymorphism was lower than that with SSR derived from genomic libraries (Choet al. 2000; Eujaylet al. 2001; Gupta et al. 2003). However, some recent studies reported high levels of polymorphism with EST-SSR markers (Eujaylet al. 2004; Fraser et al. 2004; Saha et al. 2004), with cases where EST-SSR markers were associated with the equivalent or even higher level of polymorphism than genomic SSR (e.g. Liewlaksaneeyanawin et al. 2004; Varshney et al. 2005a). Our results on seven *Coffea* species also supported these observations. Three to 19 alleles per polymorphic locus were recorded among the set of species tested. This range is larger than that reported by Bhat et al. (2005)—7–13 alleles detected using nine EST-SSR—although a wider spectrum of species was tested (14 *Coffea* and four *Psilanthus*). Interestingly, it is also larger than that noted by Poncet et al. (2004), with around 60 genomic microsatellite



present results could be illustrated with two examples. First, the cultivated species *C. canephora* displayed high values with EST-SSR than with genomic micro-satellites, i.e. mean allele number per locus (5.7 vs. 3.6), observed heterozygosity (0.51 vs. 0.38) and PIC (0.59 vs. 0.48). Secondly, comparable values for these three parameters and for the mean Fis value were obtained for *C. pseudozanguebariae* independently of the type of SSR marker.

A comparison of genetic parameters obtained for each species illustrated that EST-SSR markers were ideal for assessing genetic diversity in coffee trees. For example, *C. pseudozanguebariae* appeared to be less polymorphic than *C. canephora* when both genomic microsatellite and EST-SSR markers were used and less polymorphic than *C. dewevrei* and *C. liberica* when using EST-SSR markers. This observation is in agreement with the more restricted geographical distribution of *C. pseudozanguebariae*. When considering the two related species *C. liberica* and *C. dewevrei*, they were differentiated earlier when using morphological traits, molecular markers (AFLP) and male fertility of  $F_1$  hybrids (N'Diaye et al. 2005). The *C. dewevrei* individuals analyzed in the present study were formerly cultivated genotypes collected in the Central African Republic. The *C. liberica* sample corresponded to a mix of geographic origins comprising cultivated and wild forms. Under these conditions, it was not surprising to observe higher levels of PIC and heterozygosity associated with lower Fis values, and a null allele frequency (ra).

Another important feature of our EST-SSRs was their efficiency in distinguishing individuals from pairs of related species. 40 to 52% diagnostic loci were identified when the 16 interspecific combinations (CAN/EUG, CAN/HET, CAN/LIB, CAN/PSE and PSE/DEW) were considered.

Finally, our results demonstrated that: (1) *C. canephora* EST-SSR markers can be easily transferred to wild *Coffea* species for which no information is available on their DNA sequences; (2) they are good candidates for the development of conserved orthologous markers for genetic analysis across *Coffea* species. This high degree of transferability between species will facilitate comparative mapping and homologous gene cloning.

**Acknowledgments** This work was partly supported by EU grant No. ICA4-CT-2001-10068. The authors wish to thank I. Mougenot, C. Fizames, B. Piegou, A. Wissocq, F. Lechauve, F. Moreews, X. Argout, F. Chevalier, and many Genetop researchers for their involvement in the development of EST db, and M. Lorieux for his help in developing the SSR script (<http://www.mpl.ird.fr/bio-info/>). Thank to Dr. Santiago C. González-Martínez for his valuable comments on the manuscript.

## References

- AltschulSF, Madden TL, SchaVerAA, ZhangJ, ZhangZ, MillerW, LipmanDJ (1997) GappedBLASTand dPSI-BLAST: a new generation of protein database search programs. *NucleicAcidsRes* 25:3389–3402
- BaruahA, Naik P, Hendre S, Rajkumar R, RajendrakumarP, AggarwalRK (2003) Isolation and characterization of nine microsatellite markers from *Coffea arabica* L., showing wide cross-species amplifications. *MolEcolNotes* 3:647–650
- BhatPR, KrishnakumarV, HendrePS, RajendrakumarP, VarshneyRK, AggarwalRK (2005) Identification and characterization of expressed sequence tags-derived simple sequence repeats, markers from robusta coffee variety 'CXR' (an inter-specific hybrid of *Coffea canephora* & *Coffea congensis*). *MolEcolNotes* 5:80–83
- BrownGR, KadelIEIII, BassoniDL, KiehneKL, TemesgenB, van BuijtenenJP, SewellMM, MarshallKA, NealeDB (2001) Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrative pine genomics. *Genetics* 159:799–809
- CatoSA, GardnerRC, KentJ, RichardsonTE (2001) A rapid PCR-based method for genetically mapping ESTs. *TheorApplGenet* 102:296–306
- ChoYG, IshiiT, TemnykhS, ChenX, LipovichL, McCouchSR, ParkWD, AyresN, CartinhourS (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *TheorApplGenet* 100:713–722
- CoulbalyI, RevolB, NoirotM, PoncetV, LorieuxM, Carasco-Lacombe C, Minier J, DufourM, HamonP (2003) AFLP and SSR polymorphisms in *Coffea* interspecific backcross progeny [(*C. heterocalyx* & *C. canephora*) & *C. canephora*]. *TheorApplGenet* 107:1148–1155
- DecroocqV, FaveMG, HagenL, BordenaveL, DecroocqS (2003) Development and transferability of apricot and grape EST microsatellite markers across taxa. *TheorApplGenet* 106:912–922
- DirlewangerE, CossonP, TavaudM, AranzanaJ, PoizatC, ZannettoA, ArusP, LaigretF (2002) Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.). *TheorApplGenet* 105:127–138
- DufourM, HamonP, NoirotM, RistrerucciAM, BrottierP, VicoV, LeroyT (2001) Potential use of SSR markers for *Coffea* spp. genetic mapping. In: ASIC (ed) 19th international science colloquium on coffee, Trieste, Italy
- EllegrenH (2004) Microsatellites: simple sequences with complex evolution. *NatRevGenet* 5:435–445
- EujayI, SledgeMK, WangL, MayGD, ChekhovskiyK, ZwonitzerJC, MianMA (2004) Medicago truncatula EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *TheorApplGenet* 108:414–422
- EujayI, SorrellsM, BaumM, WoltersP, PowellW (2001) Assessment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. *Euphytica* 119:39–43
- FernandezD, SantosP, AgostiniC, BonMC, PetitotAS, SilvaMC, GuerraGuimaraesL, RibeiroA, ArgoutX, NicoleM (2004) Coffee (*Coffea arabica* L.) genes early expressed during infection by the rust fungus (*Hemileia vastatrix*). *MolPlantPathol* 5:527–536
- FraserLG, HarveyCF, CrowhurstRN, DeSilvaHN (2004) EST-derived microsatellites from Actinidia species and their potential for mapping. *TheorApplGenet* 108:1010–1016
- GaoLF, TangJF, LiHW, JiaJZ (2003) Analysis of microsatellite in major crops assessed by computational and experimental approaches. *MolBreed* 12:245–261

- Gonzalez-Martinez SC, Robledo-Arnuncio JJ, Collada C, Diaz A, Williams CG, Alia R, Cervera MT (2004) Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. *Theor Appl Genet* 109: 103–111
- Gonzalo MJ, Oliver M, Garcia Mas J, Monfort A, Dolcet Sanjuan R, Katzir N, Arus P, Monforte A (2005) Simple-sequence repeat markers used in merging linkage maps of melon (*Cucumis melo* L.). *Theor Appl Genet* 110: 802–811
- Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol Genet Genomics* 270: 315–323
- Kantety RV, LaRota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol* 48: 501–510
- Kumapatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48: 985–998
- Ky CL, Barre P, Lorieux M, Trouslot P, Aka Vou S, Louarn J, Charrier A, Hamon S, Noirot M (2000) Interspecific genetic linkage map, segregation distortion and genetic conversion in coffee (*Coffea* sp.). *Theor Appl Genet* 101: 669–676
- LaRota M, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMCG e-nomics* 6: 23
- Lashermes P, Combes MC, Prakash NS, Trouslot P, Lorieux M, Charrier A (2001) Genetic linkage map of *Coffea canephora*: effects of segregation distortion and analysis of recombination rate in male and female meioses. *Genome* 44: 589–596
- Lashermes P, Combes MC, Trouslot P, Charrier A (1997) Phylogenetic relationship of coffee trees species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. *Theor Appl Genet* 94: 947–955
- Lemp P, Lallemand J (2003) Grass consensus STS markers: a new approach for detecting polymorphism in *Lolium*. *Theor Appl Genet* 107: 1113–1122
- Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. *Theor Appl Genet* 109: 361–369
- Lin C, Mueller LA, Carthy JM, Crouzillat D, Petiard V, Tanksley SD (2005) CoVe and tomato share common genes: repeats as revealed by deep sequencing of seed and cherry transcripts. *Theor Appl Genet* 112: 114–130
- Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21: 2128–2129
- Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res* 10: 72–80
- Moncada P, McCouch S (2004) Simple sequence repeat diversity in diploid and tetraploid *Coffea* species. *Genome* 47: 501–509
- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30: 194–200
- N'Diaye A, Noirot M, Hamon S, Poncet V (2006) Genetic basis of species differentiation between *Coffea liberica* Hiern and *C. canephora* Pierre: analysis of an interspecific cross. *Genet Resour Crop Evol* (in press)
- N'Diaye A, Poncet V, Louarn J, Hamon S, Noirot M (2005) Genetic differentiation between *Coffea liberica* var. *liberica* and *C. liberica* var. *Dewevrei* and comparison with *C. canephora*. *Plant Syst Evol* 253: 95–104
- 5) An SSR- and AFLP-based genetic linkage map of tall fescue (*Festuca arundinacea* Schreb.). *Theor Appl Genet* 110: 323–336
- Park YH, Alabady MS, Ulloa M, Sickler B, Wilkins TA, Yu J, Stelly DM, Kohel RJ, el-Shihy OM, Cantrell RG (2005) Genetic mapping of new cotton Wber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. *Mol Genet Genomics* 274: 428–441
- Pinto LR, Oliveira KM, Ulian EC, Garcia AA, de Souza AP (2004) Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome* 47: 795–804
- Poncet V, Hamon P, Minier J, Carasco-Lacombe C, Hamon S, Noirot M (2004) SSR cross-amplification and variation with incoVe trees (*Coffea* spp.). *Genome* 47: 1071–1081
- Rallo P, Tenzer I, Gessler C, Baldoni L, Dorado G, Martin A (2003) Transferability of foliv microsatellite loci across the genus *Olea*. *Theor Appl Genet* 107: 940–946
- Rovelli P, Mettullo R, Anthony F, Anzueto F, Lashermes P, Graziosi G (2000) Microsatellites in *Coffea arabica* L. In: Sera T, Socol CR, Pandey A, Roussos S (eds) *CoVe bio-technology and quality*. Kluwer, Netherlands, pp 123–133
- Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) *Bioinformatics methods and protocols: methods in molecular biology*. Humana Press, Totowa, pp 365–386
- Saha MC, Mian MA, Eujayl I, Zwonitzer JC, Wang L, May GD (2004) Tall fescue EST-SSR markers with transferability across several grass species. *Theor Appl Genet* 109: 783–791
- Saha MC, Mian R, Zwonitzer JC, Chekhovskiy K, Hopkins AA (200

- Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning—laboratory manual. Cold Spring Harbor Laboratory edn. Cold Spring Harbor
- Scott KD, Egger P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. *Theor Appl Genet* 100:723–726
- Sethy NK, Choudhary S, Shokeen B, Bhatia S (2006) Identification of microsatellite markers from *Cicer reticulatum*: molecular variation and phylogenetic analysis. *Theor Appl Genet* 112:347–357
- Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 106:411–422
- Varshney RK, Graner A, Sorrells ME (2005a) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 23:48–55
- Varshney RK, Sigmund R, Borner A, Korzun V, Stein N, Sorrells ME, Langridge P, Graner A (2005b) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci* 168:195–202
- Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, Smith JS, Jaqueth J, Smith OS, Doebley J (2005) An analysis of genetic diversity across the maize genome using microsatellites. *Genetics* 169:1617–1630
- Wu KS, Tanksley SD (1993) Abundance, polymorphism and genetic mapping of microsatellites in rice. *Mol Gen Genet* 241:225–235
- Yu JK, La Rota M, Kantety RV, Sorrells ME (2004) EST derived SSR markers for comparative mapping in wheat and rice. *Mol Genet Genomics* 271:742–751