Mol Breeding (2015) 35:31
DOI 10.1007/s11032-015-0247-z

SHORT COMMUNICATION

New EST–SSR markers of *Coffea arabica*: transferability and application to studies of molecular characterization and genetic mapping

Luís Felipe V. Ferrão · Eveline T. Caixeta · Guilherme Pena · Eunize M. Zambolim · Comes D. Cruz · Laércio Zambolim · Maria Amélia G. Ferrão · Ney S. Sakiyama

Received: 1 March 2014 / Accepted: 29 September 2014

Abstract Microsatellite markers (SSR) have broad utility in genetic studies due to a high rate of polymorphisms, a codominant nature and multiallelism. EST–SSRs are markers derived from the expressed sequences of a genome and represent transcribed genes. Despite the importance of the genus *Coffea*, only a small number of EST–SSR markers are currently available. Thus, this study was designed to mine and develop a new set of EST–SSR markers from the Brazilian Coffee Genome Project. We investigated 130,792 expressed sequence tags (ESTs), from which 24,031 DNA sequences with microsatellites were identified. After stability and amplification testing, 101 new EST–SSR markers were developed and analyzed in different coffee species. The average rate of transferability was 88 %, showing that these markers are useful in genetic studies across the genus *Coffea*. Polymorphism levels and the degree of diversity were consistent with the evolutionary history of the species. All coffee genotypes were discriminated, even the *C. arabica* genotypes that have known narrow genetic basis. It was also possible to locate 14 EST–SSRs into different linkage groups of the *C. arabica* genetic map, which demonstrate that these markers can be useful in QTL mapping studies and in molecular-assisted selection.

Keywords *Coffea* sp. · Coffee · Microsatellites · Brazilian Coffee Genome Project · Molecular markers

The coffee tree is a member of the *Rubiaceae* family and belongs to the genus *Coffea*, of which 103 species have been currently described. Among them, two are commercially grown: *Coffea arabica* ($2n = 4x = 44$), grown in the highlands and in mild climates, and *Coffea*

Electronic supplementary material The online version of this article (doi:10.1007/s11032-015-0247-z) contains supplementary material, which is available to authorized users.

L. F. V. Ferrão · G. Pena · E. M. Zambolim
BIOAGRO, BIOCAFÉ, Universidade Federal de Viçosa, Viçosa, MG 3650–000, Brazil

E. T. Caixeta (✉)
Embrapa Café, BIOAGRO, BIOCAFÉ, Universidade Federal de Viçosa, Viçosa, MG 3650–000, Brazil
e-mail: eveline.caixeta@embrapa.br

C. D. Cruz
Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG 3650–000, Brazil

L. Zambolim
Departamento de Fitopatologia, Universidade Federal de Viçosa, Viçosa, MG 3650–000, Brazil

M. A. G. Ferrão
Embrapa Café/Incaper, Rua Afonso Sarlo, 160 Bento Ferreira, Viçosa, MG 3650–000, Brazil

N. S. Sakiyama
Departamento de Fitotecnia, Universidade Federal de Viçosa, Viçosa, MG 3650–000, Brazil

Published online: 21 January 2015

canephora ($2n = 2x = 22$), grown at low altitudes and in warmer weather. Global efforts have been made to improve the production and quality of coffee. In this sense, breeding programs stand out mainly for improving agronomic aspects involved in the production and marketing of coffee beans.

In this context, the use of molecular markers becomes imperative because it allows access to reliable and clear information at the DNA level. Among the different types of markers, microsatellites or simple sequence repeats (SSRs) stand out for important features, e.g., high rates of polymorphism, a codominant nature, multiallelism, locus specificity, reproducibility, ease of automation and relative low cost in comparison with SNP. These characteristics make this marker a favorite for genetic mapping studies, marker-assisted selection (MAS), genetic diversity surveys, QTL analysis and germplasm maintenance. Despite these advantages, the number of SSR markers for the genus *Coffea* is limited mainly by the difficulties in developing specific primers.

Advances in sequencing techniques over the past few decades have led to the sequencing of thousands of DNA expressed sequences (ESTs) that are then deposited into public and private DNA databases. ESTs represent regions of the genome transcribed under certain physiological conditions. This scenario has led to the development of SSR markers with greater ease and lower costs. Thus, EST–SSR markers are developed with the advantage of being related to functional portions of the genome and providing a high degree of transferability between related species (Kalia et al. 2010; Guichoux et al. 2011).

A database with over 200,000 ESTs for the genus *Coffea* was made available from the Brazilian Coffee Genome Project (Vieira et al. 2006). Thousands of clones randomly selected from cDNA libraries of *C. arabica*, *C. canephora* and *C. racemosa* were sequenced, representing specific stages of development of cells and tissues.

In this study, only *C. arabica* sequences (130,792 ESTs) were investigated. The development of new EST–SSRs for this specie was performed in three steps: (1) data mining in the Brazilian Coffee Genome database and primer design; (2) analysis of polymorphisms and transferability; and (3) validation in genetic studies.

The data mining was conducted through the bioinformatics platform of the Brazilian Coffee Genome Project (http://lge.ibi.unicamp.br/cafe). The identification and position of microsatellites were determined by the Gramene software (http://gramene.org/db/searches/ssrtool). EST–SSR mining was performed considering perfect SSRs with a minimum size of 12 bp. The Primer3 (Rozen and Skaletsky 2000) and Primer Select (Plasterer 1997) software were used for primer design and stability evaluation, respectively. From this analysis, 24,031 microsatellite sequences were identified, in which 4,380 (18.23 %) are DNRs (dinucleotides), 8,811 (36.67 %) are TNRs (trinucleotides) and 10,840 (45.11 %) are TtNRs (tetranucleotides). The $(AGGG)_n$ and $(AGG)_n$ stand out as the most frequent repetitive elements (Supplementary Table S1).

From the EST–SSR mined, we designed a set of 146 pairs of specific primers. Across the primer pairs tested, 101 (36.6 % DNRs; TNRs 37.6 %, 25.8 % TtNRs) showed clear amplification in *C. arabica* species. The sequence of these EST–SSR primers is described in Supplementary Table S2. To validation and analysis of the polymorphism levels, these markers were tested in 12 samples of *C. arabica* species, five *C. canephora* species and three Híbrido de Timor (*C. arabica* × *C. canephora*). The TETRASAT (Markwith and Scanlon 2007) and the PowerMarker v3.25 (Liu and Muse 2005) software were used for polyploidy (*C. arabica* and Híbrido de Timor) and diploid (*C. canephora*) analyses, respectively. The following parameters were considered in these analysis of polymorphic levels: (1) number of alleles accessed by each marker (Na); (2) heterozygosity (He); (3) Shannon–Weiner diversity index (H′); (4) average polymorphic information content (PIC); and (5) discriminating power (DP).

Null values of He, H′ and PIC were attributed to the monomorphic markers, i.e., those that showed the same fragments size for all genotypes in the gel. The ability of EST–SSRs to discriminate genotypes was measured according to the parameter DP. The frequency of markers that did not work (NW) and duplicate loci (DL) were also measured between the coffee species. A summary of the comparative analysis of polymorphism levels is shown in Fig. 1. The parameters He, H′, PIC and DP are influenced by the ploidy because the number and frequency of alleles are taken into account during the calculation of these parameters. For comparison, in Fig. 1, the mean values of *He, *H′, * PIC and *DP are shown with a correction for the ploidy of each organism. Details about the polymorphic levels of each EST–SSR marker are presented in Supplementary Table S3.

The *C. arabica* and Híbrido de Timor genotypes had the same number of duplicate loci (4) and contained 31

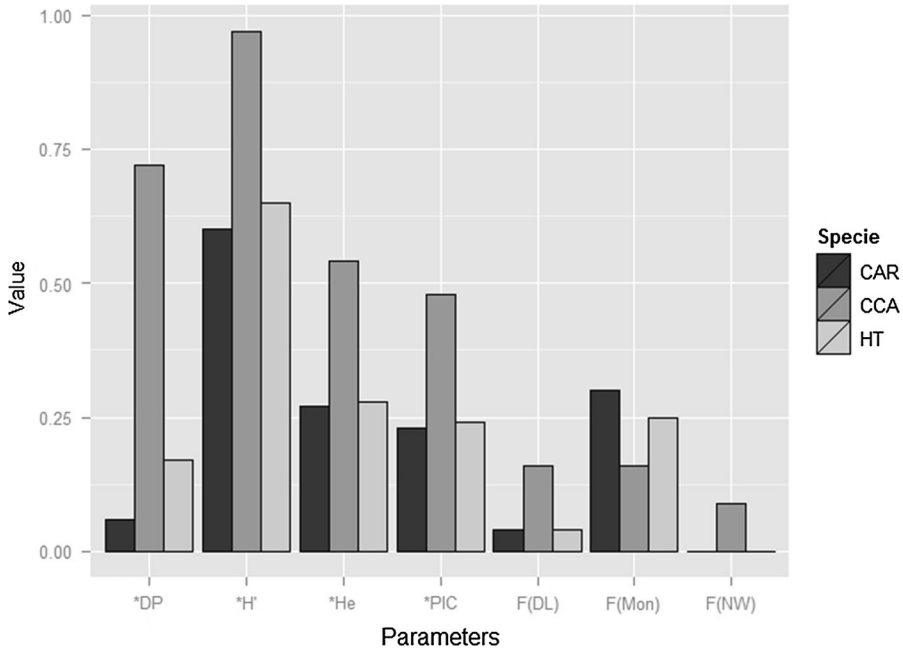

Fig. 1 Comparative analysis of the levels of polymorphism considering the frequency of duplicate loci (DL), monomorphic loci (Mon) and markers that did not work (NW), and the corrected mean values for ploidy of heterozygosity (*He), Shannon–Weiner diversity index (H'), polymorphic information average content (PIC*) and discriminant power (*DP) in the genotypes of *Coffea arabica* (CAR), Híbrido de Timor (HT) and *Coffea canephora* (CCA)

and 25 monomorphic loci, respectively. All of the markers worked for interspecific hybrids and showed good amplification patterns with clear bands. The *C. canephora* specie showed the lowest number of monomorphic loci (16); however, it showed a higher frequency of duplicate loci and unamplified markers (15 and 9, respectively). These results for *C. canephora* were expected because the EST–SSRs were developed from *C. arabica*.

The Na values of *C. arabica* and *C. canephora* ranged from 1 to 6, with an average of 2.11 and 2.78 alleles per locus, respectively. The moderate levels of PIC, H, He, and DP observed for *C. arabica* and Híbrido de Timor revealed lower genetic and allelic diversity, whereas the *C. canephora* species presented a high degree of diversity. Similar results were described by Lashermes et al. (1999), Poncet et al. (2006) and Hendre et al. (2008) and are consistent with the mode of reproduction, the genome evolution and the process of domestication for coffee.

Besides the analysis of level of polymorphism, the rate of cross-taxa transferability (T_{MAR}) of the new EST–SSR markers was measured. The markers were analyzed in four other species of the genus *Coffea* (*C. canephora*, *C. eugenioides*, *C. congensis* and *C. racemosa*) and two natural interspecific hybrids: triploid originated by natural crossed among *C. arabica* × *C. racemosa* and Híbrido de Timor originated by natural crossed among *C. arabica* × *C. canephora*. The transferability was calculated as a proportion of primers showing successful amplification for each species. The following formula was used: $T_{MAR} = {}^{SA_{esp}}/_{NM_{esp}}$, where SA_{esp} is the successful amplification for each species and NM_{esp} is the total number of primers that were used in the analyses of each species.

The results of transferability rate showed an average T_{MAR} of 88 %. Emphasis is given to the natural hybrids (Híbrido de Timor and Triploid), which show high values of T_{MAR} (1.0 and 0.97, respectively). The species *C. racemosa* and *C. congensis* showed the lowest values (0.70 and 0.71, respectively), while *C. canephora* (0.89) and *C. eugenioides* (0.87) had intermediate values (Supplementary Table S4).

The transferability analysis indicates that EST–SSR markers have potential for use in genetic studies for other species of the genus *Coffea*. Overall, the EST–SSRs have a higher transfer rate between related species when compared with traditional genomic SSRs (Ellis and Burke 2007). The explanation for this is that EST primers flank the coding regions of the genome that are more likely to be conserved across species. The new EST–SSR markers were also validated for molecular characterization and genetic mapping. The effectiveness of new EST–SSRs in studies on diversity and taxonomy was measured in genotypes of *C. canephora*, *C. eugenioides*, *C. congensis, C. racemosa* and natural

hybrids (Híbrido de Timor and Triploid) using two different approaches. The first, implemented in the STRUCTURE software (Pritchard et al. 2000), was the Bayesian approach. In these analyses, K values ranged from 1 to 10 with genetic mixture model and 15 repetitions were considered. Each run was implemented with a period of 50,000 burn-in and 100,000 MCMC interactions. The true number of genetic groups was estimated using the StructureHarvester program (Earl and vonHoldt, 2012), by calculating ΔK (Evanno et al. 2005). In this methodology, the genotypes were classified into two subsets ($k = 2$) according to the value of ΔK, indicating that the higher level of the hierarchy is associated with the genotype ploidy. Thus, the tetraploids (Híbrido de Timor and *C. arabica*) and the triploid were placed into one group, while the diploids (*C. canephora*, *C. eugenioides*, *C. congensis* and *C. racemosa*) formed another distinct group (Fig. 2a). Similar results were found by López-Gartner et al. (2009), in which *C. arabica* and the Híbrido de Timor showed differential clustering of the diploid species.

In a second way, the genotypes were evaluated considering an integrated analysis of genetic similarity, based on Jaccard similarity index and the Bayesian clustering. Accordingly, we identified the formation of seven subsets ($k = 7$), four of which correspond to the diploid species (*C. canephora*, *C. eugenioides*, *C. congensis* and *C. racemosa*) and three formed by polyploidy genotype (*C. arabica* and hybrids) (Fig. 2b). The results of the molecular characterization using the new markers were efficient in discriminate coffee species, showing a potential application for taxonomic analysis. Another promising result observed was the possibility to use the EST–SSRs in genotypic differentiation of the *C. arabica* genotypes despite the narrow genetic basis. All the coffee accession were distinguishable, except CAR.2 and CAR.3. For the *C. canephora* genotypes, they were able to distinguish the CCAN3 that belongs to the group of Conilon and the others that belong to the Robusta group. This reinforces the importance of these markers, especially for parental selection in breeding

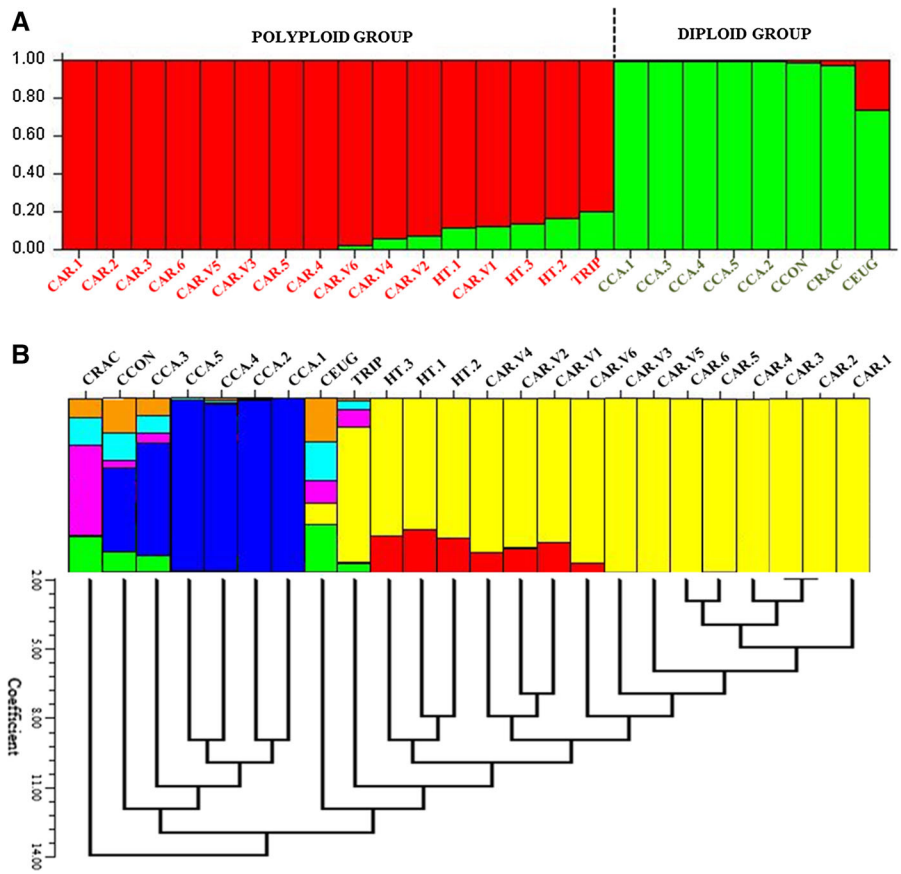

Fig. 2 **a** *Bar plot* of the STRUCTURE software used to study the diversity in genotypes of the *Coffea arabica* (CAR), Hybrid of Timor (HT), *Coffea canephora* (CCA), natural triploid (TRIP), *Coffea racemosa* (CRAC), *Coffea eugenioides* (CEUG) and *Coffea congensis* (CCON). These 24 genotypes were divided into two groups ($k = 2$) according to ploidy level and were grouped according to the value of Q (coefficient of similarity). **b** Analysis of genetic diversity of the same 24 genotypes, considering an integrated analysis of the STRUCTURE software ($k = 7$) and the NJ dendrogram based on the Jaccard similarity index

programs and the management of genetic resources in germplasm banks.

In genetic mapping studies, EST–SSRs were positioned on *C. arabica* linkage map. The genetic map for *C. arabica* was previously described by Pestana (2010) and has 94 markers (AFLP, RAPD and SSR) mapped in 11 linkage groups. Of this total, 14 correspond to the new EST–SSRs and were mapped in seven linkage groups. The small number of EST–SSRs located on the map is due to the low level of polymorphism between the mapping population parents. As previously discussed, this is a result of the narrow genetic base derived from the domestication process of the species. In the context of plant breeding, these results are important, especially for the identification and quantification of genomic regions that control traits of interest.

This study provides the scientific community with a set of 101 new EST–SSRs mined from the Brazilian Coffee Genome Project for *C. arabica* species. The characterization and validation of these markers demonstrate their usefulness in studies of molecular characterization, genetic diversity, taxonomic classification and genetic mapping. The transferability rate was measured and showed that these studies can be extended to other species within the genus *Coffea*, including for evolutionary and molecular phylogeny studies. In breeding programs, the EST–SSRs emerge as an important tool for parent selection, management of genetic resources, assisted selection and elucidation of complex traits (QTLs). Thus, this study aims to increase the number of molecular tools available to the genus *Coffea*, in order to enhance research for coffee farming. Finally, we emphasize that a database with 24,000 sequences containing microsatellites is available for the development and validation of new EST–SSR markers.

Acknowledgments The authors thank Dr. Antônio A. Pereira and Antonio Carlos B. de Oliveira for providing samples of *C. arabica*. This work was financially supported by Brazilian Coffee Research and Development Consortium (Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café), Minas Gerais State Foundation for Research Aid (FAPEMIG) and National Council of Scientific and Technological Development (CNPq).

References

Earl D, vonHoldt B (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

Ellis JR, Burke JM (2007) EST–SSR as a resource for population genetic analyses. Heredity (Edinb) 99:125–132

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11:591–611

Hendre PS, Phanindranath R, Annapurna V, Lalremruata A, Aggarwal RK (2008) Development of new genomic microsatellite markers from robusta coffee (*Coffea canephora* Pierre ex A. Froehner) showing broad cross-species transferability and utility in genetic studies. BMC Plant Biol 8:51

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2010) Microsatellite markers: an overview of the recent progress in plants. Euphytica 177:309–334

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A (1999) Molecular characterisation and origin of the *Coffea arabica* L. genome. Mol Gen Genet 261:259–266

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

López-Gartner G, Cortina H, McCouch SR, Moncada MDP (2009) Analysis of genetic structure in a sample of coffee (*Coffea arabica* L.) using fluorescent SSR markers. Tree Genet Genomes 5:435–446

Markwith SH, Scanlon MJ (2007) Multiscale analysis of Hymenocallis coronaria (Amaryllidaceae) genetic diversity, genetic structure, and gene movement under the influence of unidirectional stream flow. Am J Bot 94:151–160

Pestana KN (2010) Caracterização fenotípica e molecular da resistência do cafeeiro Hibrio de Timor a Hemileia vastatrix. Universidade Federal de Viçosa, Viçosa

Plasterer TN (1997) PRIMERSELECT. Primer and probe design. Methods Mol Biol 70:291–302

Poncet V, Rondeau M, Tranchant C, Cayrel A, Hamon S, de Kochko A, Hamon P (2006) SSR mining in coffee tree EST databases: potential use of EST–SSR as markers for the *Coffea* genus. Mol Genet Genomics 276:436–449

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945–959

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Vieira LGE, Andrade AC, Colombo CA, Moraes AHDA, de Oliveira AC, Labate CA, Marino CL, Vitorello CBM, Monte DC, Giglioti E, Kimura ET, Romano E, Kuramae EE, Lemos EGM, Almeida ERP, Jorge EC, Albuquerque EV, da Silva FR, Vinecky F, Sawazaki HE, Dorry HFA, Carrer H, Abreu IN, Batista JAN, Teixeira JB, Kitajima JP, Xavier KG, de Lima LM, de Camargo LEA, Pereira LF, Coutinho LL, Lemos MVF, Romano MR, Machado MA, Costa MMC, de Sá MFG, Goldman MHS, Ferro MIT, Tinoco MLP, Oliveira MC, Filho MAG, Shimizu MM, Maluf MP, Eira MTS, de Oliveira OLBC, Harakava R, Balbao SF, Tsai SM, Formighieri SMZ, Carazzolle MF, Pereira GAG (2006) Brazilian coffee genome project: an EST-based genomic resource. Braz J Plant Physiol 18:95–108