Theor Appl Genet (2007) 114:359–372
DOI 10.1007/s00122-006-0440-x

ORIGINAL PAPER

Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species

Ramesh K. Aggarwal · Prasad S. Hendre · Rajeev K. Varshney · Prasanna R. Bhat · V. Krishnakumar · Lalji Singh

Received: 8 August 2006 / Accepted: 19 October 2006 / Published online: 18 November 2006

Abstract Genic microsatellites or EST–SSRs derived from expressed sequence tags (ESTs) are desired because these are inexpensive to develop, represent transcribed genes, and often a putative function can be assigned to them. In this study we investigated 2,553 coffee ESTs (461 from the public domain and 2,092 in-house generated ESTs) for identification and development of genic microsatellite markers. Of these, 2,458 ESTs (all >100 bp in size) were searched for SSRs using *MISA*—search module followed by stackPACK clustering that revealed a total of 425 microsatellites in 331 (13.5%) non-redundant ESTs/consensus sequences suggesting an approximate frequency of 1 SSR/2.16 kb of the analysed coffee transcriptome. Identified microsatellites mainly comprised of di-/tri-nucleotide repeats, of which repeat motifs AG and AAG were the most abundant. A total of 224 primer pairs could be designed from the non-redundant SSR-positive ESTs (excluding those with only mononucleotide repeats) for possible use as potential genic markers. Of this set, a total of 24 (10%) primer pairs were tested and 18 could be validated as usable markers. Sixteen of these markers revealed moderate to high polymorphism information content (PIC) across 23 genotypes of *C. arabica* and *C. canephora*, while 2 markers were found to be monomorphic. All the markers also showed robust cross-species amplifications across 14 *Coffea* and 4 *Psilanthus* species. The apparent broad cross-species/genera transferability was further confirmed by cloning and sequencing of the amplified alleles. Thus, the study provides an insight about the frequency and distribution of SSRs in coffee transcriptome, and also demonstrates the successful development of genic-SSRs. It is expected that the potential markers described here would add to the repertoire of DNA markers needed for genetic studies in cultivated coffee and also related taxa that constitute the important secondary genepool for coffee improvement.

Communicated by H. Nybom.

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-006-0440-3 and is accessible for authorized users.

R. K. Aggarwal · P. S. Hendre · R. K. Varshney · P. R. Bhat · V. Krishnakumar · L. Singh
Centre for Cellular and Molecular Biology (CCMB), Uppal Road, Tarnaka, Hyderabad 500007, India

R. K. Varshney
International Crops Research, Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324, India

R. K. Aggarwal (✉)
Centre for Cellular and Molecular Biology, Uppal Road, Tarnaka, Hyderabad 500007, India
e-mail: rameshka@ccmb.res.in

Introduction

Analysis of variation at DNA level is the key for modern genetics studies, which encompasses newer tools and methods like microsatellite analysis, single nucleotide polymorphism (SNP) studies and other DNA marker systems based on gross and specific DNA sequence variations. Due to their ability to reveal the unexplored enormous genetic variation in the genome, such DNA markers have become extremely important for the genetic analysis of crop plants. Among different classes of molecular markers, microsatellite or simple

sequence repeat (SSR) markers are the most favoured for a variety of applications in plant genetics and breeding because of their multi-allelic nature, reproducibility, codominant inheritance, high abundance and extensive genome coverage (Gupta and Varshney 2000).

Coffee is an important beverage and plantation crop belonging to genus *Coffea* (family Rubiaceae). Although, more than 100 species of coffee are known, its commercial cultivation relies only on two species, amphidiploid *Coffea arabica* L. ($2n = 4x = 44$) and diploid *C. canephora* Pierre ex A. Froehner ($2n = 22$). Despite the apparent advantages, development of SSR markers in this important plantation crop has been slow, as only about? 150 microsatellite markers have been reported to date (Combes et al. 2000; Rovelli et al. 2000; Baruah et al. 2003; Moncada and McCouch 2004; Bhat et al. 2005).

Microsatellites developed from ESTs, popularly known as EST–SSRs or genic SSRs, represent functional molecular markers as a 'putative function' for a majority of such markers can be deduced by database searches and other in silico approaches. Furthermore, EST–SSR markers are expected to possess high interspecific transferability as they belong to relatively conserved genic regions of the genome. With recent increasing emphasis on functional genomics, large datasets of ESTs are being developed, and with evolving bioinformatic tools it is now possible to identify and develop EST–SSR markers at a large scale in a time and cost-effective manner (Scott et al. 2000; Kantety et al. 2002; Varshney et al. 2002). Because of the above advantages of genic SSR markers, and relatively easy accessibility of large EST resources, increasing numbers of genic SSR markers are now being identified and used for a variety of applications in a number of plant species like, grapes (Scott et al. 2000), sugarcane (Cordeiro et al. 2001), and cereals such as wheat, barley, rye, rice (see Varshney et al. 2005).

For development of genic SSR markers for coffee, 461 ESTs available in public domain (as per dbEST release 073004, http://www.ncbi.nlm.nih.gov) were pooled with an interim set of 2,092 ESTs of coffee generated in-house at CCMB, Hyderabad, India and analysed with the following objectives: (1) analysis of the frequency and distribution of SSRs in the expressed portion of the coffee genome, (2) development of novel EST–SSR markers for coffee, (3) validation of developed EST–SSR markers for detection of polymorphisms in cultivated coffee germplasm, as well as their interspecific or intergeneric transferability.

Materials and methods

Plant material

For the present study, several genotypes belonging to *C. arabica* and *C. canephora* along with other related species mentioned in Table 1 were used. The leaf samples from the genotypes were collected from the coffee germplasm bank maintained at Central Coffee Research Institute, Balehonnur, Chikamagalur, India and genomic DNA was isolated as described by Aggarwal et al. (2002).

In silico analyses

Sequence data sources

The EST sequences for coffee available in the public domain were acquired through a Sequence Retrieval System (SRS version 7.1.1 release 79). In addition, we used an interim set of 2,092 coffee ESTs generated at CCMB, Hyderabad, India.

Searching the microsatellites

The identification and localization of microsatellites in ESTs was accomplished by a microsatellite search module named *MISA* (*MI*croSAtellite, http://www.pgrc.ipk-gatersleben.de/misa; Fig. 1). In the preparatory step, the raw EST sequences were processed by removing the poly-A and poly-T stretches until no stretch of $(T)_5$ or $(A)_5$ was present in a window of 50 bp on the 5′- or 3′-end, respectively. Similarly, sequences larger than 700 bp were clipped at their 3′ side to preclude the inclusion of low quality sequences. In addition, ESTs of <100 bp length were excluded. Criteria for SSR search by the *MISA* were repeat stretches having a minimum of: 10 repeat units for mononucleotide SSRs, and 4 repeat units in case of di-, tri-, tetra-, penta- and hexa-nucleotide SSRs. The microsatellites were classified considering the complementarities of the repeat motifs, e.g., AG, GA, TC and CT were considered as a single category. Finally, in order to minimize redundancy, a cluster analysis was performed on SSR containing ESTs (SSR–ESTs) using stackPACK v 2.2 program (Miller et al. 1999).

Marker development

Primer pairs for non-redundant SSR–ESTs were designed as described earlier by Varshney et al. (2002) using PRIMER3 (http://www.fokker.wi.mit.edu/primer3/),

Table 1 Plant materials used for marker validation and cross-species transferability

Name of genotype	Pedigree/source
A. *Coffea arabica* genotypes	
S288	Pureline from S 26 (*C. arabica* × *C. liberica*)
S795	S 288 × Kent
Tafarikela	Pureline from Ethiopian collections
S5A	Double cross hybrid; Devamachy (*C. canephora* × *C. arabica*); in common with S 881, S-333 *arabica* s
S7.3	Multi-step cross of San Ramon Hybrid with S795, Agaro followed by HdeT
S8	Pure line from spontaneous R × A hybrid; Introduction from Timor Island (HdeT)
S9	HdeT × Tafarikela
S10	Double Cross Hybrid; Caturra with Cioccie and S.795 (both *arabicas*)
S11	Amphidiploid, *C. liberica* × *C. eugenioides*
S12	Caturra × HdeT
S2790	HdeT × Tafarikela
S2792	Tafarikela × HdeT
BM	Blue Mountain Pure line
Kent	Pure line
Agaro-Sln4	Pure line from Ethiopian collections
B. *Coffea canephora* (robusta) genotypes	
Kaganalla	Selection
BR9	Selection
BR12	Selection
C × R	Hybrid of *C. congensis* × *C. canephora*
L1 Valley	Selection
S3329	Selection
S3334	Selection
Sln27	Pure line
C. Other *Coffea* sp., related *Psilanthus* taxa used for cross species transferability	
1. *C. congensis*	Erythrocoffea (West & Central Africa)
2. *C. excelsa*	Pachycoffea (Cylon)
3. *C. liberica*	Pachycoffea (West & Central Africa)
4. *C. abeokutae*	Pachycoffea (Ceylon)
5. *C. dewevrei*	Pachycoffea (USDA)
6. *C. arnoldiana*	Pachycoffea (SanMarino)
7. *C. aruwemiensis*	Pachycoffea (SanMarino)
8. *C. eugenioides*	Mozambicoffea (Central Africa)
9. *C. racemosa*	Mozambicoffea (East Africa)
10. *C. salvatrix*	Mozambicoffea (East Africa)
11. *C. kapakata*	Mozambicoffea (Central Africa)
12. *C. stenophylla*	Melanocoffea (West Africa)
13. *P. wightiana*	Paracoffea (India)
14. *P. khasiana*	Paracoffea (India)
15. *P. bengalensis*	Paracoffea (India)
16. *P. travancorensis*	Paracoffea (India)

and Bhat et al. (2005) using GENETOOL version 1.0 (http://www.biotools.com/products/genetool.html).

In order to identify the putative function(s) of EST–SSR markers, the corresponding SSR–ESTs were compared to the NR-PEP (non-redundant peptide) database at the DKFZ Heidelberg, Germany (see http://www.genome.dkfz-heidelberg.de/) using the BLASTX2 program (Altschul et al. 1997).

About 10% of the total designed primer pairs were used for validation studies (allelic diversity and cross-species transferability) using a panel of 15 *C. arabica* and 8 *C. canephora* genotypes, as well as 16 other related taxa of coffee (Table 1).

PCR conditions and allele sizing of microsatellites

PCR amplifications and microsatellite analysis were performed as described by Bhat et al. (2005). In brief, the EST–SSR markers were amplified on a PTC-200 Thermal-Cycler (MJ Research), and amplified alleles were resolved through GeneScan analysis on ABI-377 DNA sequencer and sized using the software Geno-

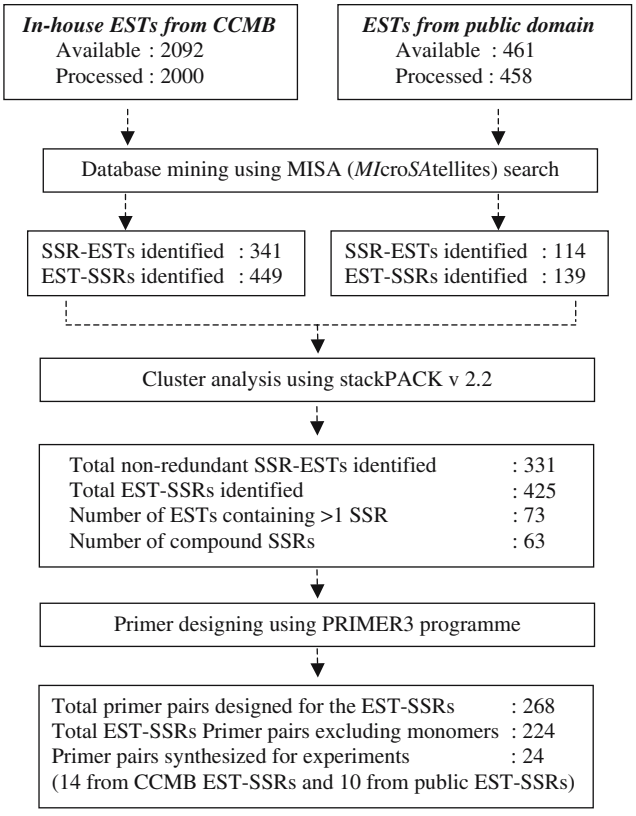

Fig. 1 Scheme used for database mining and development of genic SSR markers from coffee ESTs

typer ver. 2.5 (Applied Biosystems). The allelic data were used to calculate the number, range, and distribution of amplified alleles.

Statistical analysis

The allelic data were used to calculate PIC values as follows (Anderson et al. 1993):

$$\mathrm{PIC} = 1 - \sum_{i=1}^{k} P_i^2$$

where, k is the total number of alleles detected for a microsatellite marker and P_i the frequency of the ith allele in the set of analysed genotypes. In a few cases, where more than two alleles were observed in a given genotype, these parameters were calculated manually. The biallelic polymorphic data were also tested for Hardy–Weinberg equilibrium (HW) and linkage disequilibrium (LD) as described by Bhat et al. (2005).

Genetic diversity analysis

EST–SSR allelic data were used to ascertain the generic relationships among the tested genotypes by cluster analysis. The data were transformed to binary mode using scores 1/0 for presence/absence of allele, respectively, as was done earlier by Moncada and McCouch (2004) for SSR based clustering in coffee. The binary data were used to derive Dice coefficients (as indicator of genetic similarity) followed by phenetic clustering using the Neighbour-Joining (NJ) method. The analytical routines were carried out using NTSYSpc ver 2.02 (http://www.ExeterSoftware.com). The clustering was also tested by bootstrap analysis using Winboot program (Yap and Nelson 1996) with 1,000 iterations.

Confirmation of cross-species SSR transferability

In order to confirm the EST–SSR transferability, the microsatellite alleles amplified in related species (average for 13 species) for four of the randomly selected EST–SSR markers (CofEST–SSR01, CofEST–SSR05, CofEST–SSR06 and CofEST–SSR12; Table ESM1) were individually cloned, sequenced and examined for the conservation of the amplified targets by sequence comparison. Amplified PCR products were cloned into pMOS (Amersham) or TA (Invitrogen) plasmid vector and transformed in *E. coli* DH5α competent cells. Multiple individual colonies (average 12 clones per cloning event) were used for plasmid preparation, amplification of cloned amplicons using standard methods, followed by sequencing for both strands using M13 universal primers and BigBye terminator cycle sequencing chemistry on a 3730 Automated DNA Analyzer (Applied Biosystems). The final edited sequences belonging to each locus were compared with the original SSR–EST sequence using CLUSTAL-X (http://www.ftp-igbmc.u-strasbg.fr/pub/ClustalX/) for ascertaining the target domain/SSR conservation.

Results

Frequency and distribution of SSRs in the coffee transcriptome

A total of 2,553 coffee ESTs were used for the study (Fig. 1), of which 2458 ESTs (458 from the public domain and 2,000 developed in-house at CCMB) were selected after the initial processing involving clipping of poly-A/poly-T tails, 3′ ends and excluding 90 sequences smaller than 100 bp size for SSR search. These represented approximately 919.1 kb of putative functional coffee transcriptome, *MISA* based microsatellite search of these ESTs detected a total of 588 SSRs in 455 (18.5%) ESTs (SSR–ESTs), suggesting

an average frequency of SSR as ~1/1.56 kb and/or 1/5.4 ESTs in the coffee transcriptome analysed.

However, it may be noted that the above SSR estimates are based on a redundant EST dataset. Accordingly, to reduce overestimation, a redundancy analysis was performed on the detected SSR–ESTs using stackPACK v 2.2. The cluster analysis, thus performed revealed a total of 267 SSR–ESTs as singletons and 188 SSR–ESTs into 64 clusters. As a result, 331 non-redundant ESTs and/or consensus sequences were identified that contained a total of 425 SSRs (Table 2). Moreover, considering the redundancy correction, the average frequency of non-redundant EST–SSRs is expected to be ~1/2.16 kb of the coffee transcriptome.

Analysis of SSR motifs in the non-redundant SSR–ESTs (Fig. 1) revealed 73 (22.1%) ESTs that contained more than one SSR. Of the total 425 SSRs seen in these ESTs, 362 (85.2%) contained simple repeat motifs while 63 (14.8%) were of compound type. Moreover, most of these represented smaller repeat-unit size SSRs (Table 2): 105 (24.7%) mononucleotide repeats (MNRs), 197 (46.3%) dinucleotide repeats (DNRs), 111 (26.1%) trinucleotide repeats (TNRs), 5 (1.2%) tetranucleotide repeats (TTNRs), 2 (0.5%) pentanucleotide repeats (PNRs) and 5 (1.2%) hexanucleotide repeats (HNRs). Among the DNRs, AG motif was the most common (52.8%) followed by AT (24.8%) and AC (21.3%) motifs, whereas CG motif was the least common (1.1%) (Table 2). Similarly, among the TNRs, the motif AAG was the most common (28.8%) followed by the motifs ACT (12.6%), ACC (11.7%) and AAT (10.8%) whereas the motif CCG was the least common (2.7%). However, the TTNRs, PNRs or HNRs were found in insignificant numbers (<2%).

Development of potentially functional EST–SSR markers

The 331 non-redundant SSR–ESTs (comprising consensus sequences for 64 clusters and 267 singleton

Table 2 Frequency and distribution of different types of SSRs identified in the analysed 2,458 coffee ESTs (after considering sequence complementarities of the repeat motifs)

Repeat motif	Number of repeat units													Total repeats
	4	5	6	7	8	9	10	11	12	13	14	15	>15	
A/T	–	–	–	–	–	–	29	15	11	9	5	4	20	93
C/G	–	–	–	–	–	–	4	4	1	–	–	–	3	12
AC/GT	30	4	2	2	–	–	3	1	–	–	–	–	–	42
AG/CT	75	10	2	4	1	2	1	1	1	1	2	3	1	104
AT/AT	34	9	2	1	2	–	1	–	–	–	–	–	–	49
CG/CG	2	–	–	–	–	–	–	–	–	–	–	–	–	2
AAC/GTT	4	2	–	–	–	–	–	–	–	–	–	–	–	6
AAG/CTT	18	9	3	1	–	–	–	–	1	–	–	–	–	32
AAT/ATT	9	2	1	–	–	–	–	–	–	–	–	–	–	12
ACC/GGT	11	1	1	–	–	–	–	–	–	–	–	–	–	13
ACG/CTG	8	–	1	–	–	–	–	–	–	–	–	–	–	9
ACT/ATG	8	5	1	–	–	–	–	–	–	–	–	–	–	14
AGC/CGT	2	3	1	–	–	–	–	–	–	–	–	–	–	6
AGG/CCT	8	–	1	1	–	–	–	–	–	–	–	–	–	10
AGT/ATC	3	1	2	–	–	–	–	–	–	–	–	–	–	6
CCG/CGG	2	–	1	–	–	–	–	–	–	–	–	–	–	3
AAAG/CTTT	–	1	–	–	–	–	–	–	–	–	–	–	–	1
AAAT/ATTT	–	1	–	–	–	–	–	–	–	–	–	–	–	1
AAGT/ATTC	–	1	–	–	–	–	–	–	–	–	–	–	–	1
ACAT/ATGT	–	–	1	–	–	–	–	–	–	–	–	–	–	1
ACCT/ATGG	1	–	–	–	–	–	–	–	–	–	–	–	–	1
AAAGG/CCTTT	1	–	–	–	–	–	–	–	–	–	–	–	–	1
AACTC/AGTTG	1	–	–	–	–	–	–	–	–	–	–	–	–	1
AACGGT/ATTGCC	1	–	–	–	–	–	–	–	–	–	–	–	–	1
ACCGCT/ATGGCG	1	–	–	–	–	–	–	–	–	–	–	–	–	1
ACGCGG/CCTGCG	1	–	–	–	–	–	–	–	–	–	–	–	–	1
AGAGGG/CCCTCT	1	–	–	–	–	–	–	–	–	–	–	–	–	1
AGTATC/AGTCAT	1	–	–	–	–	–	–	–	–	–	–	–	–	1
N (MNR)	–	–	–	–	–	–	33	19	12	9	5	4	23	105
NN (DNR)	141	23	6	7	3	2	5	2	1	1	2	3	1	197
NNN (TNR)	73	23	12	2	–	–	–	–	1	–	–	–	–	111
NNNN (TTNR)	1	3	1	–	–	–	–	–	–	–	–	–	–	5
NNNNN (PNR)	2	–	–	–	–	–	–	–	–	–	–	–	–	2
NNNNNN (HNR)	5	–	–	–	–	–	–	–	–	–	–	–	–	5

SSR–ESTs) were used for primer designing. Of these, primers could be successfully designed only for 268 (80.9%) ESTs (62 clusters and 206 singletons). The remaining ESTs were inappropriate for primer modelling mainly due to short unique domains flanking the microsatellite core.

Of the 268 potential EST–SSRs, 44 contained MNRs as the SSR core, and these were excluded from the final list since practical problems related to allele sizing were expected. The details of primer sequences and expected product size with SSR motifs for the remaining 224 potential markers are described in Table ESM 1. Further, it is evident from the details in Table ESM1 that these markers are based on a total of 213 unique ESTs. This in turn would suggest that only ~8.7% of the total ESTs investigated in the study represent potential candidates for SSR-marker development.

Moreover, based on BLASTX analysis, a putative function could be assigned to 118 (52.7%) potential markers assuming a threshold of <1.00E-05 and to only 82 (36.7%) markers using a more stringent threshold of <1.00E-20 (Table ESM 1). Also, a majority of the coffee SSR–ESTs (80%) showed significant homology to the annotated proteins of dicotyledonous species (*Arabidopsis*) rather than to those of monocotyledonous species like rice, wheat, barley and maize.

Marker validation and detection of polymorphism

A total of 24 designed primer pairs (Table ESM 1) comprising 10 pairs based on public domain ESTs and 14 based on ESTs developed at CCMB were used for validation of the genic SSR markers. Of these, 18 (75%) primer pairs amplified the expected size of amplicons with considerable polymorphism (Table 3), while the remaining six tested primers pairs (CofEST–SSR10, CofEST–SSR14, DCM02, DCM03, DCM09 and DCM10; Table ESM 1) did not yield any scorable amplicon. Some of the data (mainly pertaining to the amplification conditions and PIC values) for 9 of the working EST–SSRs were presented earlier (Bhat et al. 2005), which were used in this study for ascertaining their potential in genetic diversity analysis of coffee germplasm, as well as, validation of cross-species transferability by sequencing of the cross-species alleles.

The 18 amplifiable markers revealed low to medium allelic diversity with PIC values ranging from 0–0.77 (mean 0.42 ± 0.116) to 0–0.82 (mean 0.42 ± 0.125), and expected heterozygosity (*He*) from 0–0.78 (mean 0.49 ± 0.131) to 0–0.85 (mean 0.43 ± 0.128) for *arabica* and *robusta* genotypes, respectively. Overall, a maximum of 8 alleles with an average of 3.4–3.5 alleles/marker were obtained for the tested genotypes of *C. arabica* and *C. canephora* (*robusta*) genotypes. Two markers (CofEST–SSR05, DCM08) were monomorphic in both *C. arabica* and *C. canephora*, while another marker (CofEST–SSR07) was monomorphic only for *canephora* genotypes.

For *arabica* genotypes five out of 12 polymorphic loci viz. CofEST–SSR01, CofEST–SSR06, CofEST–SSR08, DCM05 and DCM06, and in *robusta*s nine out of 11 polymorphic loci *viz*. CofEST–SSR02, CofEST–SSR04, CofEST–SSR06, CofEST–SSR08, CofEST–SSR11, CofEST–SSR12, DCM01, DCM05 and DCM07 were found to be in HW equilibrium. Linkage disequilibrium (LD) test performed for the loci in HW equilibrium revealed only one pair (DCM05 and DCM07) for *arabica* and three pairs of loci (CofEST–SSR04 and CofESR–SSR06; CofEST–SSR08 and CofEST–SSR11; CofEST–SSR08 and DCM07) for *robusta* genotypes with significant LD (at 5% level after applying Bonferroni correction).

Diversity analysis and genetic relationship

Allelic data from the working EST–SSRs were used to test their potential in genetic studies by ascertaining the genetic diversity/interrelationships in the cultivated genotypes, as well as the related taxa of coffee. The phenetic clustering based on genotypic data from 16 polymorphic markers for 15 *arabica* and 8 *robusta* genotypes resulted in an NJ tree which clearly resolved the tested germplam in two distinct clusters (as expected of their origin and genetic make up), one representing all the tetraploid *arabica*s while the other comprised all the diploid *robusta* genotypes (Fig. 2a). Similarly, a clustering analysis of the EST–SSR allelic data of 16 related species (12 *Coffea* and 4 *Psilanthus* spp.) along with 2 genotypes each of *C. arabica* and *C. canephora*, largely resolved their generic affinities (Fig. 2b) as expected based on conventional as well as earlier molecular studies. In general, the clusters appeared to support the expected origin, geographical distribution and botanical classification (Chevalier 1947) of coffee. The Erythrocoffea species *C. canephora* (represented by CxR and Kagnalla) and *C. congensis* were nearest to *C. arabica* (Tafarikela and Blue Mountain). Four of the Pachycoffea species appeared as a coherent cluster within which a strong geographical correspondence was evident. The results further validate the placement of the four related Paracoffea as the most distant to *arabica*s and *robusta*s.

Table 3 Marker validation and inter-specific/generic transferability of the 18 working genic SSR markers

Marker examined	Tm (°C)	Expected product size (bp)	Marker validation/Cross species transferability											
			Coffea Arabica (n = 15)				*Coffea canephora* (n = 8)				*Coffea* sp. (n = 12)		*Psilanthus* sp. (n = 4)	
			Allele no./ Size range (bp)	H_o	H_e	PIC	Allele no./ Size range (bp)	H_o	H_e	PIC	Amplification[a]/allele size range in bp			
CofEST–SSR01	57	150	3 150–154	0.27	0.41	0.31	5 146–154	0.14	0.86**	0.66	12	140–170	3_{12}	146–156
CofEST–SSR02	57	151	2 138–150	1	0.52**	0.38	3 141–156	0.25	0.24	0.24	114	120–159	4	129–150
CofEST–SSR03	57	153	6 138–157	0.93	0.66**	0.63	5 138–155	0.63	0.73	0.75	$9_{4,5,8}$	134–160	4	142–158
CofEST–SSR04	57	147	8 137–151	1	0.76**	0.7	6 133–165	0.67	0.89	0.75	12	135–167	$2_{12,13}$	139–155
CofEST–SSR05	57	104	1 104	0	0	0	1 104	0	0	0	12	104	4	104
CofEST–SSR06	57	143	2 129–133	0.2	0.24	0.16	5 130–141	0.75	0.7	0.6	118	131–141	4	129–135
CofEST–SSR07	57	199	3 193–199	0.93	0.59**	0.49	1 193	0	0	0	12	178–199	4	169–187
CofEST–SSR08	57	145	2 142–145	0	0.19*	0.12	2 145–163	0.13	0.24	0.11	12	145–163	4	166–178
CofEST–SSR09	57	119	5 102–128	1	0.74	0.72	8 104–130	1	0.74	0.72	12	100–122	4	142–158
CofEST–SSR11	57	100	3 96–100	0.07	0.40**	0.35	2 98–100	0.25	0.23	0.2	12	96–114	4	92–122
CofEST–SSR12	57	131	2 114–120	0.93	0.52**	0.37	2 117–129	0.25	0.34	0.2	12	108–120	4	105–117
CofEST–SSR13	57	117	6 115–139	0.87	0.78	0.77	5 104–116	0.38	0.71	0.69	12	110–138	4	108–120
DCM01	58	262	3 260–266	0.93	0.55**	0.42	6 262–282	0.63	0.69	0.61	$10_{2,9}$	246–268	3_{12}	244
DCM04	58	131	5 117–129	1	0.64	0.62	7 102–129	1	0.38	0.82	12	114–132	4	114–126
DCM05	60	163	3 161–170	0.87	0.55*	0.42	2 164–170	0.13	0.24	0.11	1_{18}	158–170	4	158–176
DCM06	58	210	4 210–214	0.6	0.73**	0.65	5 209–213	0.13	0.87**	0.69	12	209–214	4	211–215
DCM07	58	149	3 147–151	0.6	0.47	0.39	3 147–151	0.57	0.47	0.39	12	143–157	$2_{13,16}$	159–161
DCM08	58	181	1 181	0	0	0	1 181	0	0	0	12	181	4	181
Mean ± SE				0.62 ± 0.179	0.49 ± 0.131	0.42 ± 0.116		0.39 ± 0.122	0.43 ± 0.128	0.42 ± 0.125				

For primer information (and NCBI accession numbers of ESTs) for the above EST–SSR markers, see Table ESM 1

[a] Number of species showing amplification; subscripts (as per the serial number in Table 1, section C) indicate species that failed to amplify

*/**Markers (loci) that showed significant departure from HW equilibrium in the analysed samples

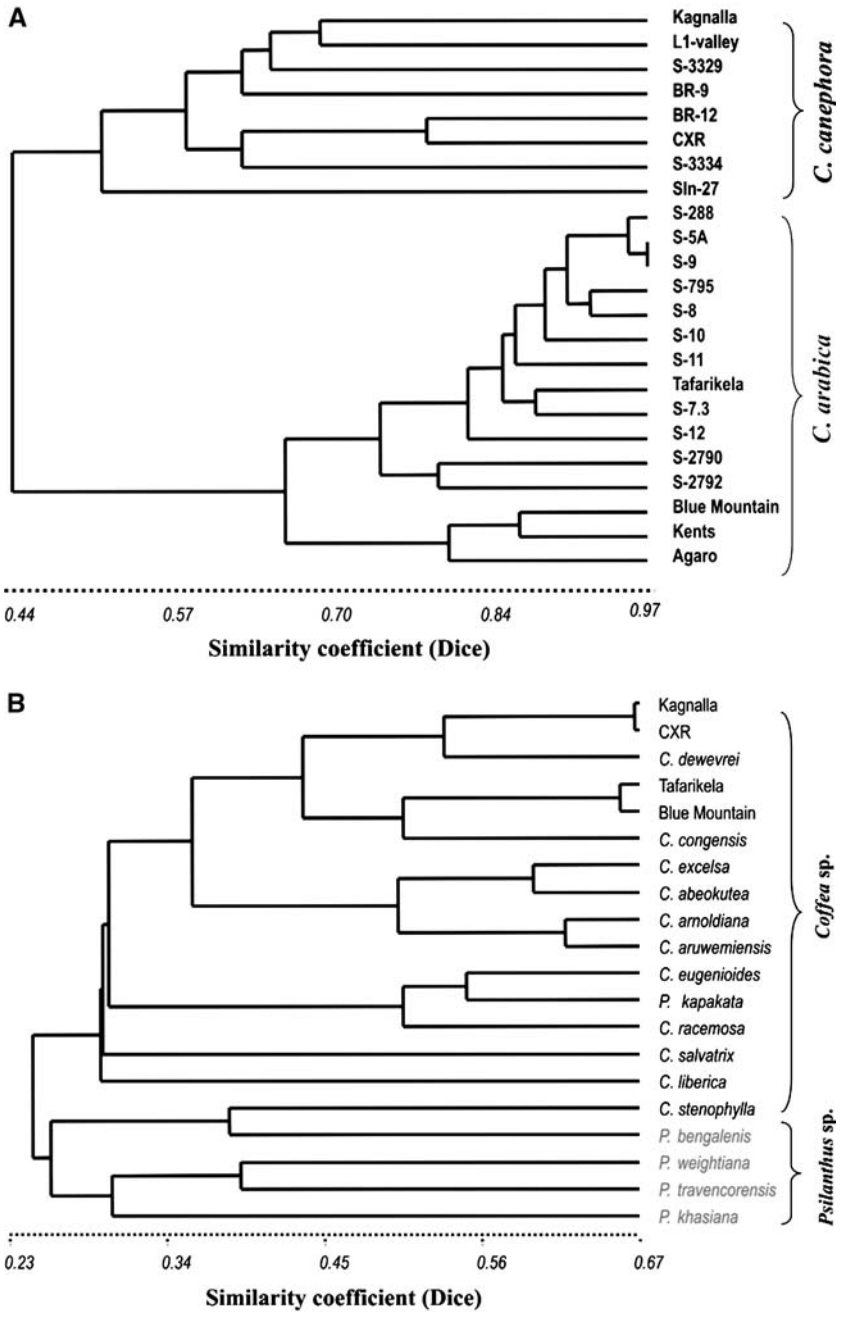

Fig. 2 Phenetic trees based on the allelic diversity (revealed by the new genic SSR markers developed in the study), showing generic relationships between: **a** genotypes of *C. arabica* and *C. canephora*, and **b** species of *Coffea* and *Psilanthus*

Cross-species/genera transferability and validation by sequencing of cross-species alleles

All the 18 working primer pairs revealed robust cross-species amplifications with alleles of comparable sizes when tested on 12 related *Coffea* species (apart from *C. arabica* and *C. canephora*) and 4 *Psilanthus* taxa (Table 3). As expected, the average transferability of the validated markers was relatively higher for *Coffea* species (96.3%) than for the species of *Psilanthus* (91.4%). Interestingly, the two markers (CofEST–SSR05, DCM08) that were monomorphic in *C. arabica* and *C. canephora*, were also monomorphic for all other *Coffea* and *Psilanthus* species.

Moreover, cloning of products from 51 PCR reactions (representing amplified alleles for 12–14 related species for four of the developed EST–SSRs), and sequencing >600 clones (12 clones per ligation event) unequivocally confirmed the cross-species conservation and transferability of the developed EST–SSR loci (Fig. 3). In general, in all cases the sequenced alleles (NCBI accession numbers DQ655733 to DQ655790) from different species were homologous to the original locus (EST sequence) from which the marker was

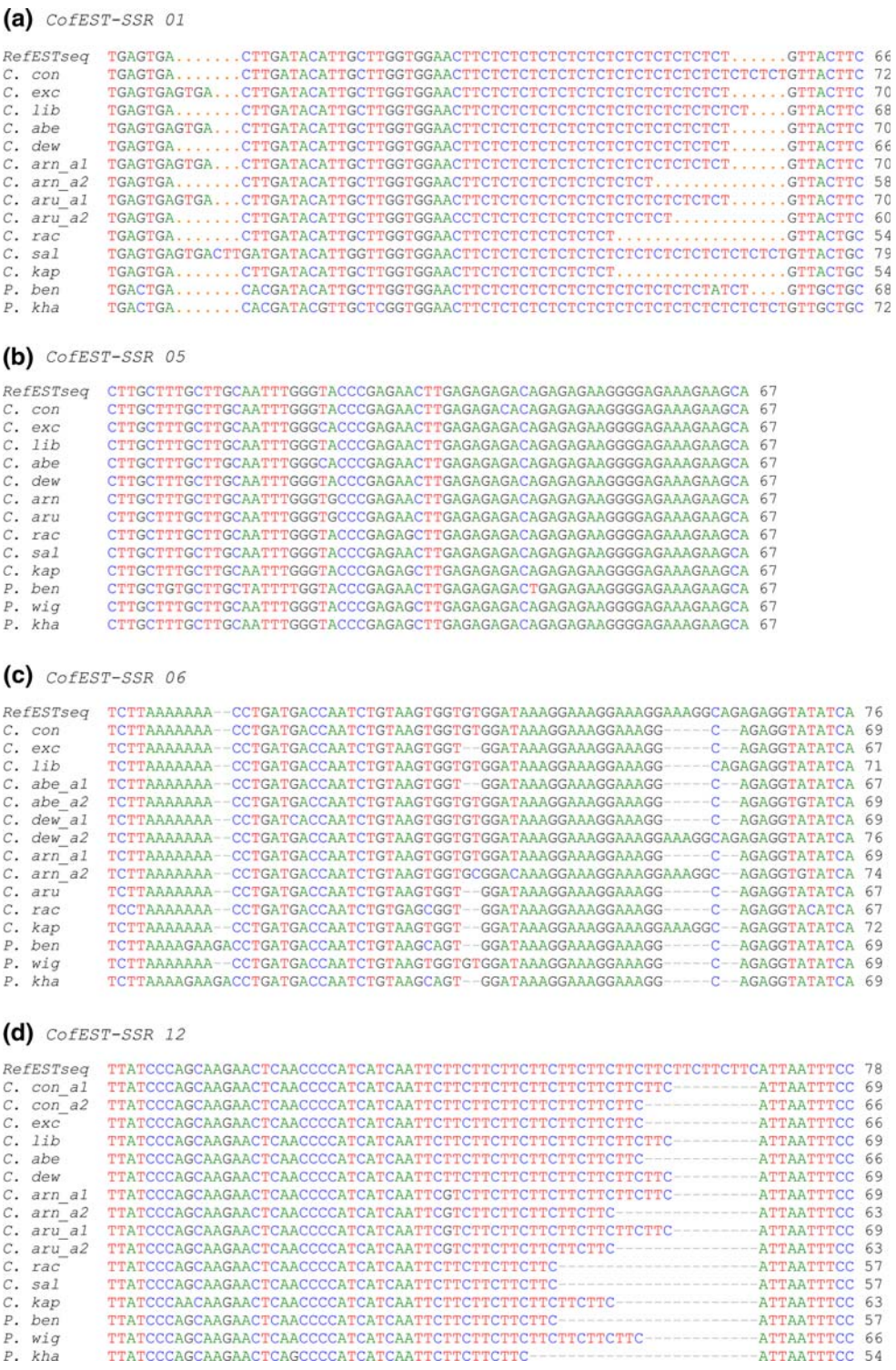

Fig. 3 Partial aligned sequence of alleles obtained in various related taxa of coffee using four of the new EST–SSRs described in the study, showing/establishing cross-species/genera conservation and transferability. The four EST–SSR markers/their reference EST sequences are: **a** CofEST–SSR 01/AY705497; **b** CofEST–SSR 05/AY705500; **c** CofEST–SSR 06/AY705501; **d** CofEST–SSR 12/AY7055505. The abbreviations: *C. con, C. exc, C. lib, C. dew, C. rac, C. abe, C. arn, C. aru, C. kap, C. sal, P. ben, P. wig*, and *P. kha*, represent the related coffee taxa: *Coffea congensis, C. excelsa, C. liberica, C. dewevrei, C. racemosa, C. abeokutae, C. arnoldiana, C. aruwemiensis, C. kapakata, C. salvatrix, Psilanthus bengalensis, P. wightiana* and *P. khasiana*, respectively. The suffix 'a1' and 'a2' in the taxa names stands for allele numbers. The 58 sequences corresponding to different cross species/genera allele used above are deposited in NCBI database under accession numbers: DQ655733 to DQ655790

developed. All the alleles that were originally seen in the GeneScan analysis of the four tested species-marker combinations were precisely correlated with SSR motif repeat length variation in the sequenced alleles. In addition, few additional point mutations/substitutions (mainly transitions, ranging from 1.4 to 1.86% of the sequenced bases) were seen in the regions flanking the SSR motifs in some of the alleles in related species amplified using three markers (CofEST–SSR 01, CofEST–SSR 05, CofEST–SSR 06), while none was seen in case of CofEST–SSR 12. Similarly, five small *indels* were also revealed in a few alleles amplified

using two markers (CofEST–SSR 01, CofEST–SSR 06). This additional variation was, in general, higher for alleles belonging to the *Psilanthus* taxa.

Discussion

SSR frequency and distribution

The frequency, distribution and abundance of SSRs can be highly variable depending on the SSR search criteria, the size of the dataset, and the database-mining tools (Varshney et al. 2005). Accordingly, the reports on SSR abundance across different EST resources for plants and animals differ significantly in their absolute values. Compared to the earlier reports for grapes (Scott et al. 2000), sugarcane (Cordeiro et al. 2001), cereals (Varshney et al. 2002; Kantety et al. 2002; Thiel et al. 2003), a relatively higher abundance of SSRs (redundant SSRs in 18.5% ESTs) was observed in the present study for coffee ESTs. This difference can be attributed to the SSR search criterion that in this study was defined as four repeat units for all types of SSR motifs except for MNRs for which the threshold was kept as ten.

Similarly, among various coffee EST–SSRs identified in this study (Table 2) the highest proportion comprised of DNRs followed by the TNRs. This is in contrast to a majority of the earlier studies which invariably report TNRs as the most abundant class of SSRs in ESTs (Scott et al. 2000; Cordeiro et al 2001; Varshney et al. 2002; Kantety et al. 2002; Thiel et al. 2003; Nicot et al. 2004), but in agreement with recent studies in *Actinidia* (Fraser et al. 2004) and *Picea* species (Rungis et al. 2004) wherein the DNRs were found to be the most abundant class of EST–SSRs. In fact, DNRs have been reported to be most abundant SSRs in the ESTs of many animal species such as, medaka, *Fundulus*, zebrafish, and *Xiphophorus* (Ju et al. 2005). These apparent differences in the relative abundance of the DNRs and TNRs can again be attributed to the differences in SSR search criteria used for EST database mining in different studies. It was noteworthy that in most of the earlier studies which showed abundance of TNRs, invariably the minimum number of repeat units for SSR identification was considered higher for DNRs (6–10 repeats) than TNRs (5–6 repeats). However in this study, same number of minimum repeat units (4) was considered for all types of SSRs (DNRs, TNRs, TTNRs, PNRs and HNRs) except MNRs. Interestingly, when this criterion was changed to 6 repeat units for DNRs and five repeat units for TNRs, TTNRs, PNRs, HNRs, we obtained a higher abundance of TNRs (22.2%) in comparison to DNRs (16.6%) (data not shown), as reported in many earlier studies. Thus our results demonstrate that the SSR search criteria used for EST database mining can significantly alter the relative estimates of frequency/distribution of EST–SSRs, supporting the opinion of Varshney et al. (2005). In turn, these data suggest the need for formulating a universally acceptable definition of SSR to obtain more meaningful estimates and avoid discrepancies in the absolute values in future comparative studies.

Furthermore, it was significant to note that in general, the GC-rich SSR motifs were less frequent in coffee ESTs (Table 2). This was most evident in the relative abundance of AG/AAG and deficiency of CG/CCG repeats motifs among the DNRs/TNRs, respectively identified in this study. Interestingly, similar differences in SSR motif in ESTs have been reported earlier, and seems to be a common feature of the dicot species (Cardle et al. 2000; Gao et al. 2003).

Novel genic microsatellite markers

In coffee, to the best of our knowledge, to date only ~150 SSR markers have been described in the literature (Combes et al. 2000; Rovelli et al. 2000; Baruah et al. 2003; Moncada and McCouch 2004; Bhat et al. 2005), warranting continuous efforts to develop additional new efficient genetic markers for desired integration and utility of DNA marker technology/tools in genetics/breeding efforts on this otherwise difficult plantation crop species. In this context, the set of 224 EST–SSR markers (Table ESM 1) identified in this study is expected to be a significant addition to the presently available relatively small repertoire of microsatellite markers. Moreover, most of the SSR markers described earlier for coffee are genomic (non-genic SSRs), which further increases the importance of the markers described in the present study. The EST–SSR markers, in addition to the merits of the conventional (genomic) SSR markers, are also expected to improve detection of marker-trait associations since they are part of the transcribed domain(s) of the genome. In fact in recent years emphasis is slowly shifting towards development of functional molecular markers instead of anonymous markers (Anderson and Luebberstedt 2003) as they have the potential for assaying the functional diversity in germplasm collection or natural population and may prove more useful for marker-assisted selection if found to be associated with a gene/QTL of interest. Other practical advantages of EST–SSR markers (expected owing to their higher sequence conservation) are the probability of fewer

null alleles and high cross species transferability (Varshney et al. 2005).

In the coffee SSR–EST dataset reported here a putative function was deduced for 36% of the markers and it is expected that this number will increase in the future as the protein databases (SWISPROT or NR-PEP) are continuously growing. The remaining SSR–ESTs when searched for a putative function resulted in ''no hit'' (18%), ''no significant homology'' (29%) or ''hypothetical protein'' (17%), and these may in fact represent the specific transcriptome of coffee, which is yet to be characterized for its putative functions. Here, it may be important to mention that a majority of the coffee SSR–ESTs matched with the known proteins of dicotyledonous species (*Arabidopsis, Solanum, Nicotiana*, etc.) and only about 10% of the candidate SSR–ESTs matched with the proteins of monocots (*Oryza, Zea*, etc.). This observation seems to be reflective of the functional diversification among dicots and monocots, and thus expected of coffee, which is a dicot species. These data thus qualify the markers identified here as potential novel functional EST–SSR markers for coffee. Furthermore, considering that ~75% of the tested EST–SSR primer pairs (Table 3) could be successfully validated (see below), it is expected that 224 EST based primer pairs designed in the study (Table ESM 1), may potentially provide about 175 novel working microsatellite markers, which can be used for detection of polymorphisms, diversity and other genetic studies.

Level of polymorphism and cross-species/genera transferability

The validated genic SSR markers displayed a low level of polymorphism in *arabica* and *robusta* genotypes. This is expected as these SSRs are located in highly conserved portions of the genome and therefore display a lower level of polymorphism (see Varshney et al. 2005). However, no major difference was observed in terms of allele numbers and PIC values for the markers between *arabica* and *robusta* genotypes. Overall the variation was lower, especially for robustas than our earlier observations using non-genic genomic SSRs (Baruah et al. 2003), suggesting that EST–SSRs being relatively conserved functional domains of the genome may be less efficient compared to genomic SSRs in detecting the intraspecific variation. Furthermore, monomorphic behaviour of two of the tested primer pairs (CofEST–SSR05 and DCM08) across all the *Coffea* and *Psilanthus* species suggest that these represent highly conserved genes with some important cellular function(s), which indeed becomes evident from their BLASTX based results, which show CofEST–SSR05 and DCM08 to be parts of ''Nuclear transport factor 2'' and ''protein phosphatase'' genes, respectively (Table ESM 1). Analysis of the CofEST–SSR05 and DCM08 sequences revealed their SSR domain (comprising of GA/CT repeats) in the immediate (within 25–35 bp) upstream and downstream untranslated regions (UTRs), respectively. It is plausible that any change in repeat length of the SSR domain in the exon and/or the UTRs (regulatory regions that are increasingly being documented to be important in gene regulation/function) of important housekeeping genes may affect the protein structure or expression adversely (Kashi and Soller 1999; Sangwan and O'Brian 2002; Pauli et al. 2004; Li et al. 2004), and thus can be under strong selection pressure making them resistant to change. The latter may be of relatively less magnitude in case of other genes allowing them to tolerate SSR variation despite the same being part of their exons (as seen in CofEST–SSR06 and CofEST–SSR07).

The low level of polymorphism detected by genic SSRs may be compensated by their higher potential for cross-species transferability as shown in the present study. Although cross-species amplification was observed with genomic SSRs as well (Baruah et al. 2003), comparatively higher rate of transferability has been observed, especially across related genera. A total of 77% of the genic SSRs investigated in the present study yielded an amplicon in four *Psilanthus* species as compared to only 37.5% of the genomic SSRs (Baruah et al. 2003). This observation is noteworthy, as successful cross-species amplification of SSRs is generally restricted to related species within the genus. Peakall et al. (1998) observed that while cross-species transferability of soybean SSRs was 65% within its own genus *Glycine*, it reduced drastically to 3–13% for other species of related genera. On the other hand, recently Wang et al. (2005) also have reported that the polymorphism level detected by EST–SSRs is almost comparable at cross-species and cross-genus level (similar to above observation in this study), again highlighting the fact that genic SSR markers have higher transferability and thus better applicability than genomic SSR markers.

Validation of cross-species amplicons/alleles

Sequences of cross-species amplicons generated by four of the randomly chosen EST–SSRs for 12–13 related taxa, unambiguously demonstrated the conservation and transferability of the developed EST–SSR loci. In general, the amplified regions were found

to be homologous to the original coffee EST sequences (from which the SSRs were developed) and their comparisons across species (Fig. 3) correlated the observed 'cross-species alleles' precisely with the expected SSR repeat length variations, which are necessary attributes for the cross-species applicability of developed markers.

Moreover, the cross-species allele sequences also revealed a few additional point mutations/substitutions (mainly transitions) in the regions flanking the SSR motifs in some of the alleles. In general, these mutations/substitutions were more common for alleles belonging to the *Psilanthus* taxa. Similar additional variation in the cross-species SSR alleles (comprising point mutations, *indels* in flanking regions, expansion of SSR motif and repeat conversion) has been reported earlier in some other studies of similar type (Peakall et al. 1998; Shepherd and Lambert 2005; Sethy et al. 2006). Such variation is expected to be due to the innate evolving nature of the genome, and thus can be indicative of the evolutionary relationships of the tested taxa. Accordingly, a closer analysis of the point mutations in cross-species alleles revealed an apparent transitional bias for *Coffea* species (closer taxa to the source coffee species from which the marker was developed), but relatively more transversions in *Psilanthus* taxa (which represented evolutionarily more distant species belonging to another genus). Shepherd and Lambert (2005) observed a similar transversional bias in the flanking regions of SSR loci across genera of penguins.

Diversity analysis and genetic relationships within/between *Coffea* and *Psilanthus* species

The EST–SSRs described here, despite revealing a relatively low level of polymorphism were able to individualize all the 23 genotypes of the two cultivated coffee species. The phenetic tree based on the allelic diversity clustered the tested genotypes as per their species status (Fig. 2a) and broadly conforming to their known pedigree. The exposed genetic diversity was higher within the 8 *robusta*s in comparison to the 15 *arabica* genotypes. Also, more loci (9 out of 11) were in HW equilibrium ($P > 0.01$) in *robusta*s than *arabica*s, and of these only a few were in LD ($P > 0.05$, after applying Bonferroni correction). These results are indeed reflective of the genetic composition and mating behaviour of the tested materials; the tested *robusta*s comprised allogamous, relatively unrelated genotypes (selections, pure lines and only one hybrid), whereas *arabica*s comprised mostly hybrid varieties/selections with overlapping/shared pedigrees and represented mainly autogamous forms. These findings are in general agreement to those obtained using genomic SSRs (Baruah et al. 2003) and various other types of nuclear markers (our unpublished data) and as reported earlier by others (Orozco-Castillo et al. 1996; Lashermes et al. 2000), thus suggesting the utility/suitability of the genic SSR markers for genetic diversity studies on coffee genepool.

Similarly phenetic analysis of 22 representative samples belonging to 16 *Coffea* and 4 *Psilanthus* species, revealed generic affinities (Fig. 2b), which were broadly in agreement with their known taxonomic relationships in terms of geographical distribution, and also botanical classification as described by Chevalier (1947). Overall, 14 of the analysed taxa were well resolved and grouped in their respective 4 distinct clusters representing: Erythrocoffea (*C. arabica, C. canephora, C. congensis*), Pachycoffea (*C. abeokutae, C. excelsa*, *C. arnoldiana, C. aruwemiensis*), Mozambicoffea (*C. racemosa*, *C. eugenioides*, *C. kapakata*), and Paracoffea (*P. bengalensis*, *P. wightiana*, *P. tranvencorensis*, *P. khasiana*). The taxonomic placement of two species (*C. salvatrix* and *C. liberica*) remained unresolved, and the status of two other species (*C. stenophylla* and *C. dewevrei*) was rather unexpected. These results are in general agrrement with the only two earlier published studies wherein coffee species relationships have been ascertained using SSR markers (Moncada and McCouch 2004; Poncet et al. 2004), and our own work using genomic SSRs (unpublished data). A close affinity between *C. kapakata* and *C. eugenioides* as seen here, was also revealed in ISSR marker-based clustering (Ruas et al. 2003). On the other hand, the exact generic affinity of *C. stenophylla* (a Melanocoffea taxon) has remained a debated issue, as it was indicated to be closer to Mozambicoffea taxon *C. eugenioides* based on RAPD analysis (Orozco-Castillo et al. 1996) but to Erythrocoffea group based on ITS2 sequence polymorphism (Lashermes et al. 1997). Similarly, the placement of *C. dewevrei* (a Pachycoffea species as per Chevalier's taxonomy) along with the Erythrocoffea group (Fig. 2b), suggest the need for further detailed studies to ascertain the exact generic affiliations between members of Paracoffea and Erythrocoffea; an enigma that has also been observed in earlier DNA polymorphism studies on coffee species relationships (Lashermes et al. 1996; Orozco-Castillo et al. 1996). Nevertheless, the above demonstrate that the EST–SSR markers are as informative as any other non-genic DNA marker approaches in exploring the taxonomic relationships of coffee species complex.

In summary, the present study describes the first effort to ascertain the frequency and distribution of

SSRs in the coffee transcriptome, and also attempts development of genic-SSRs for use in genetic studies. A set of 224 primer pairs has been developed from 213 unique SSR–ESTs and/or contings of which ~10% primer pairs were also tested for their potential use as genic-SSR markers. Overall, 75% of the tested primers pairs were successfully validated. Considering a similar success rate it is expected that the primer pairs designed in the study can potentially provide about 175 new functional microsatellite markers. Our results also demonstrate that the designed EST–SSRs show broad cross-species transferability. Thus the study provides genic-SSR markers not only for cultivated coffee species but also for genetic studies involving related species that constitute the important secondary genepool for improvement of coffee.

Acknowledgments The authors thank the Department of Biotechnology, Government of India, New Delhi, India for the financial support to RKA, Director, CCMB, Hyderabad for the facilities to undertake the study, Dr R Naidu, Director Research, Coffee Board, Bangalore and Dr M. Udayakumar of University of Agricultural Sciences, Bangalore for the drought-stressed coffee leaf materials. PSH was supported by Senior Research Fellowship of Council of Scientific and Industrial Research, New Delhi.

References

Aggarwal RK, Shenoy VV, Ramadevi J, Rajkumar R, Singh L (2002) Molecular characterization of some Indian Basmati and other elite rice genotypes using fluorescence-AFLP. Theor Appl Genet 105:680–690

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

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186

Anderson JR, Luebberstedt T (2003) Functional markers in plants. Trends Plant Sci 8:554–560

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. Mol Ecol Notes 3:647–650

Bhat PR, Krishnakumar V, Hendre PS, Rajendrakumar P, Varshney RK, Aggarwal RK (2005) Identification and characterization of gene-derived EST–SSR markers from robusta coffee variety 'CxR' (an interspecific hybrid of *Coffea canephora* × *Coffea congensis*). Mol Ecol Notes 5:80–83

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156:847–854

Chevalier A (1947) Les Cafeiers du Globe. Paul Lechevalier, Paris, p 356

Combes MC, Andrzejewski S, Anthony F, Bertrand B, Rovelli P, Graziosi G, Lashermes P (2000) Characterization of microsatellite loci in *Coffea arabica* and related coffee species. Mol Ecol 9:1171–1193

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 160:1115–1123

Fraser LG Harvey CF, Crowhurst RN, De Silva HN (2004) EST-derived microsatellites from *Actinidia* species and their potential for mapping. Theor Appl Genet 108:1010–1016

Gao LF, Tang J, Li H, Jia J (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol Breed 12:245–261

Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113:163–185

Ju Z, Wells MC, Martinez A, Hazlewood L, Walter RB (2005) An in silico mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, *Fundulus*, and *Xiphophorus*. *In Silico* Biol 5:439–463

Kantety RV, Rota ML, 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

Kashi Y, Soller M (1999) Functional roles of microsatellites and minisatellites. In: Goldstein DB, Schlotterer C (eds) Microsatellites: evolution and applications. Oxford University Press, Oxford, pp 10–23

Lashermes P, Combes MC, Trouslot P, Anthony F, Charrier A (1996) Molecular analysis of the origin and genetic diversity of *Coffea arabica* L.: implications for coffee improvement. In: Proceedings of EUCARPIA meeting on tropical plants, Montpellier, pp 23–29

Lashermes P, Combes MC, Trouslot P, Charrier A (1997) Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. Theor Appl Genet 94:947–955

Lashermes P, Andrzejewski S, Bertrand B, Combes MC, Dusseri S, Graziosi G, Trouslot P, Anthony F (2000) Molecular analysis of introgression breeding in coffee (*Coffea arabica* L.). Theor Appl Genet 100:139–146

Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: Structure, function, and evolution. Mol Biol Evol 21:991–1007

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA (1999) A comprehenssive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res 9:1143–1155

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

Nicot N, Chiquet V, Gandon B, Amilhat L, Legeai F, Leroy P, Bernard M, Sourdille P (2004) Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). Theor Appl Genet 109:800–805

Orozco-Castillo C, Chalmers KJ, Powell W, Waugh R (1996) RAPD and organellar specific PCR re-affirms taxonomic relationship within the genus *Coffea*. Plant Cell Rep 15:337–341

Pauli S, Rothnie H M, Chen G, He X, Hohn T (2004) The cauliflower mosaic virus 35 S promoter extends into the transcribed region. J Virol 78:12120–12128

Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A (1998) Cross species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol Biol Evol 15:1275–1287

Poncet V, Hamon P, Minier J, Carasco C, Hamon S, Noirot M (2004) SSR cross-amplification and variation within coffee trees (*Coffea* spp.). Genome 47:1071–1081

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

Ruas PM, Ruas CF, Rampim L, Carvaljo VP, Ruas EA, Sera T (2003) Genetic relationship in *Coffea* species and parentage determination of interspecific hybrids using ISSR (inter-simple sequence repeat) markers. Genet Mol Biol 26:319–327

Rungis D, Bérubé Y, Zhang J, Ralph S, Ritland CE, Ellis BE, Douglas C, Bohlmann J, Ritland K (2004) Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. Theor Appl Genet 109:1283–1294

Sangwan I, O'Brian MR (2002) Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA. Plant Physiol 129:1788–1794

Scott KD, Eggler 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

Shepherd LD, Lambert DM (2005) Mutational bias in penguin microsatellite DNA. J Hered 96:566–571

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 (2005) Genic microsatellite markers in plants: features and applications. Trends Biotech 23:48–55

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett 7:537–546

Wang ML, Barkley NA, Yu JK, Dean RE, Newman ML, Sorrels ME, Pederson GA (2005) Transfer of simple sequence repeat (SSR) markers from major cereal crops to minor grass species for germplasm characterization and evaluation. Plant Genet Res 3:45–57

Yap IV, Nelson RJ (1996) WinBoot: a program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms. IRRI Discussion Paper Series 14, International Rice Research Institute, Manila, Philippines