RESEARCH ARTICLE

Mining of expressed sequence tag libraries of cacao for microsatellite markers using five computational tools

AIKKAL RIJU[1], M. K. RAJESH[1], P. T. P. FASILA SHERIN[1], A. CHANDRASEKAR[1], S. ELAIN APSHARA[2] and VADIVEL ARUNACHALAM[1]*

[1]*Central Plantation Crops Research Institute, Kasaragod 671 124, India*
[2]*Central Plantation Crops Research Institute, Regional Station, Vittal 574 243, India*

Abstract

Expressed sequence tags (ESTs) provide researchers with a quick and inexpensive route for discovering new genes, data on gene expression and regulation, and also provide genic markers that help in constructing genome maps. Cacao is an important perennial crop of humid tropics. Cacao EST sequences, as available in the public domain, were downloaded and made into contigs. Microsatellites were located in these ESTs and contigs using five softwares (MISA, TRA, TROLL, SSRIT and SSR primer). MISA gave maximum coverage of SSRs in cacao ESTs and contigs, although TRA was able to detect higher order (>5-mer) repeats. The frequency of SSRs was one per 26.9 kb in the known set of ESTs. One-third of the repeats in EST-contigs were found to be trimeric. A few rare repeats like 21-mer repeat were also located. A/T repeats were most abundant among the mononucleotide repeats and the AG/GA/TC/CT type was the most frequent among dimerics. Flanking primers were designed using Primer3 program and verified experimentally for PCR amplification. The results of the study are made available freely online database (http://riju.byethost31.com/cocoa/). Seven primer pairs amplified genomic DNA isolated from leaves were used to screen a representative set of 12 accessions of cacao.

[Riju A., Rajesh M. K., Sherin P. T. P. F., Chandrasekar A., Apshara S. E. and Arunachalam V. 2009 Mining of expressed sequence tag libraries of cacao for microsatellite markers using five computational tools. *J. Genet.* **88**, 217–225]

Introduction

Expressed sequence tags or ESTs provide researchers with a quick and inexpensive way of discovering new genes, analyse gene expression and regulation, and to construct genome maps. Microsatellites or simple sequence repeats (SSRs) are stretches of DNA containing tandem repeats of mono, di, tri, tetra and above nucleotide units ubiquitously distributed throughout the genome. SSRs originate from unequal crossing over or replication errors resulting in the formation of unusual secondary structures such as hairpins or slipped strands (Pearson *et al*. 1998). Molecular markers developed from ESTs are known as genic markers, are highly useful in developing linkage maps and in marker assisted breeding programmes (Varshney *et al*. 2005; Yasodha *et al*. 2008).

Cacao (*Theobroma cacao* L.) is an important perennial crop of the tropics and source of chocolates and confectionery dishes. It is a shade loving companion crop in orchards, coconut and arecanut gardens. The cacao tree is a diploid plant with a genome of 390 Mb (Couch *et al*. 1993) in 10 pairs of chromosomes. EST-derived-SSR markers are easily developed using computational methods and have applications in molecular genetics of plants for detecting polymorphism and /or linkage mapping (Tang *et al*. 2006). Linkage maps of cacao (Pugh *et al*. 2004) were saturated with many microsatellite markers that are obtained by enriching genomic libraries for dimeric or trimeric repeats. Vast EST resources from cacao plant (TIGR gene index; Quackenbush *et al*. 2001; Verica *et al*. 2004) have been made available in recent years. These could be harnessed to obtain a large number of new microsatellite markers. So far ESTs of cacao cover only a fraction of the cacao genome (0.0067%). In this study, we have used five SSR mining tools (MISA, TRA, TROLL, SSRIT and SSR primer) to maximize the chances of locating microsatellites. SSR motifs are grouped into unique classes based on the property of DNA base complementarity. EST-SSR markers have shown potential in linkage map-

*For correspondence. E-mail: vadivelarunachalam@yahoo.com.

Keywords. genome; *in silico*; simple sequence repeats; Malvales; *Theobroma cacao* L.

ping (Borrone *et al*. 2007) and resistance breeding (Lima *et al*. 2008) of cacao. EST-derived-SSR markers usually cross amplify related taxa (Tang *et al*. 2006) and hence can be useful in comparative genomics. Cacao belongs to Malvales, so they could also be useful in related crops like cotton. The main objective of the present work is to locate the SSRs in 6581 EST sequences of cacao by using five softwares, and to design primers using the flanking sequences and verify them by PCR amplification.

Materials and methods

ESTs of cacao were retrieved from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) of NCBI (GenBank accession numbers AM 117760–AM 117768, DN 237949–DN 237957, CK 144293–CK 144298, CF 972636–CF 974749 and CA 794213–CA 798660). These 6581 (dbEST release 012006) tissue-specific cacao EST sequences were grouped into seven tissues. Most of the cacao EST sequences were from *Forastero cacao* cv. Comum (Verica *et al*. 2004). These EST sequences were then processed to minimize sequencing errors and avoid redundant sequences, and were grouped into contigs. Contigs or assembled sequences are longer and potentially contain more interpretable coding sequences than their individual component ESTs. These virtual cDNA sequences may extend to the 5′ end of the mRNA, greatly facilitating cloning of genes in the laboratory. We used Phrap (Ewing and Green 1998) to assemble the EST sequences into contigs. Individual ESTs contain part of a few noncoding regions including long stretches of poly A tails especially in 3′UTR (untranslated regions). Phrap constructs the contig sequence as a mosaic of the highest quality read segments and makes extensive assembly information to assist in trouble-shooting assembly problems. We have manually sorted the SSRs likely to correspond to poly A/T tails from ESTs.

We identified three types of SSRs: (i) perfect SSRs, with an exact repeat of any of the motif, e.g., $(AT)_{15}$, (ii) imperfect repeats, those sequences having at least two or more exact simple repeats separated by non-repeated nucleotides varying in size, shown by $(CGTAT)_{10}$-GATATA-$(AGAAG)_{15}$ where GATATA nucleotides are notrepeated and (iii) compound repeats, combinations of two or more repeated motifs with length 20 or more, e.g., $(CA)_{16}$ $(TC)_{10}$. We located SSRs from both individual ESTs as well as contigs. For mononucleotides, although A, T, C and G are possible, A and T are grouped into a single category, since an A repeat on a strand is the same as a T repeat on the opposite strand, and a C on a strand is the same as a G on the opposite strand, resulting in two unique classes of mononucleotides, A/T and C/G (Katti *et al*. 2001). Similarly, all dinucleotide motifs were grouped into the following four unique classes: (i) AT/TA, (ii) AG/GA/CT/TC, (iii) AC/CA/TG/GT, and (iv) GC/CG. The trinucleotide repeats are grouped into 10 unique classes as per the SSR classification, for example the repeat group AAG/TTC contains AAG/AGA/GAA/CTT/TTC/TCT SSRs (Jurka and Pethiygoda 1995; Katti *et al*. 2001).

Putative orthologs and functions of the contigs were identified using NCBI - blastx / tblastn (www.ncbi.nlm.nih.gov) and tblastn from ExPASy (http:/expasy.org/tools/blast). Microsatellites located in this study were compared with the existing cacao microsatellites (Pugh *et al*. 2004; Araujo *et al*. 2007; Lima *et al*. 2008) using BLAST to check their uniqueness.

We tried to assess the comparative efficiency of five software tools viz. MISA (Perl script, Thiel *et al*. 2003), TRA (visual C++ program, Bilgen *et al*. 2004), TROLL (Tcp/Tk script, Martins *et al*. 2006), SSRIT (Perl script, Temnykh *et al*. 2001) and SSR primer (Perl script, Robinson *et al*. 2004). We identified class 1, or hypervariable markers, consisting of SSRs $\geq$ 20 bp. Primer3 (Whitehead Institute, Cambridge, USA) was used to design flanking primers for the detected microsatellites.

Eleven SSRs (nine perfect and two imperfect) found in contigs were verified by wet lab experiments (table 1). Details of the cocoa accessions used in the present study are given in table 2. These included seven cacao collections, one improved cacao clone and four-biclonal hybrids of the cacao gene bank at the Regional Station, Central Plantation Crops Research Institute, Kasaragod, Vittal, South Kannara, Karnataka, India. A duplicate set of these cacao collections, clone and hybrids were evaluated at the experimental farm at CPCRI Kasaragod, Kerala, India. DNA was extracted from young leaves of 12 cacao accessions from the Kasaragod farm using the DNeasy mini kit (Qiagen, Duesseldorf, Germany). The annealing temperature for each primer pair was determined first using gradient PCR. Once optimized, PCR reactions were conducted in volumes of 20 μL containing 35-μg genomic DNA, 0.2 μM each of forward and reverse primers, 50 μM of each dNTPs (Bangalore Genei, Bangalore, India), 1× buffer (10 mM tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$) and 0.3 unit of *Taq* DNA polymerase. PCR amplifications were performed on an Eppendorf gradient thermal cycler (Eppendrof, Hamburg, Germany) with a PCR profile of 94°C for 5 min followed by 30 cycles of 1 min at 94°C, 2 min at the different annealing temperatures standardized for the individual SSR locus, and 2 min at 72°C, with a final extension for 5 min at 72°C. After amplification, a volume of 3 μL of loading buffer (98% formamide, 10 mM EDTA, 0.005% each of xylene cyanol and bromophenol blue as tracking dyes) was added to each amplified product and the amplicons were separated on a 3% agarose gel and visualized after staining with ethidium bromide. Each band generated by SSR primers was considered as an independent allele. Clearly resolved, unambiguous bands were scored visually for their presence or absence with each primer. The scores were obtained in the form of a matrix with '1' and '0', which indicate the presence and absence of bands, respectively in each plant. Based on the number of polymorphic bands, percentage polymorphism was calculated for each primer.

Table 1. List of EST-SSR primers used in the present study, the repeat motif and expected amplicon size.

Primer set	Sequence	Amplified (yes/no)	Repeat motif	Expected amplicon size
TH1	F: CCAACTTAATATCTCCGCCT R: GGCTGGTTTCTTATTCACAG	Yes	$(CT)_{11}$	217
TH2	F: TCTTCGTAACTATCGGAAACA R: CTAGACGTGGAACTTTGAGG	Yes	$(AT)_{11}$	391
TH3*	F: TTTAAACAATCAGTTTAACCC R: ATTGTTGGTCGGATTACAAG	No	$(TC)_{11}$	254
TH4	F: AACTGGAGGTCAGAGTCAAA R: CTTGTCGTCCTTCTTGTTTC	Yes	$(GAA)_8$	362
TH5	F: TCTCTTCTTCTTGATTGATCG R: GTCGTAGTCGGAGTCAAGTC	No	$(AAG)_{10}$	163
TH6	F: ACAATCGCAGTCGTAATCTC R: CTCTCCAGTGAAATTGCTTC	No	$(AAC)_7$	372
TH7*	F: AATTGGCACGAGTATCTTTG R: GTCACTTGAACTCCGATCAT	No	$(CTG)_7$	247
TH8	F: TCGACAACAACATTCTTTGA R: TATTGCTGCTTCCTTCTTGT	Yes	$(CTG)_6CT$	148
TH9	F: GGATTTCTACTGCAAGCAAC R: AACAAGATATCGAGGCAGAA	Yes	$(ATAA)_5$	192
TH10	F: TTTGATCTCAAAGCAGGTTTA R: GGTATGCATGTGCTAGTGTG	Yes	$(GTAT)_5$	208
TH11	F: CGAGGCTTAACTGAAACAGA R: GCCGTTATTAGTGACCAAAG	Yes	$(ACCAA)_4$	109

*TH3 and TH7 were imperfect repeats.

Table 2. List of cacao accessions used in the present study.

Sl. no.	Name of the accession	Institute no.
1	1-14	VTLC-1
2	II-67 × NC-42/94	VTLCH-4
3	II-67 × NC-29/66	VTLCH-3
4	I-56	VTLC-11
5	NC-42/94	VTLC-56
6	NC-45/53	VTLCC-1
7	NC-42/94 × III-35	VTLCP-22
8	Jerangau Red Axil	VTLC-1A
9	II-67	VTLC-5
10	ICS-6	VTLC-17
11	SCA-6	VTLC-22
12	NC-23/43 × III-35	VTLCP-23

VTLC, CPCRI RS vittal cacao collection; VTLCC, vittal cacao clone; VTLCH, vittal cacao hybrid; VTCP, vittal cacao progeny.

The average polymorphism information content (PIC) was calculated by applying the formula given by Powell *et al*. (1996) and Smith *et al*. (1997).

$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} f_i^2,$$

where f_i is the frequency of the ith allele, n represents the number of alleles. The number of alleles refers to the number of scored bands. The frequency of an allele was obtained by dividing the number of accessions where it was found by the total number of accessions. We constructed the dendrogram using the binary data scored. The genetic associations between varieties were evaluated by calculating the Dice (1945) similarity coefficient for pair-wise comparisons based on the proportions of shared bands produced by the primers. A similarity matrix was generated using the NTSYS-PC software, version 2.0 (Rohlf 1998). The similarity coefficients were used for cluster analysis and a dendrogram was constructed by the unweighted pair-group method (UPGMA) (Sneath and Sokal 1973).

Results and discussion

Contigs from cacao ESTs

About 6581 EST sequences of cacao were retrieved from dbEST, which are derived from different libraries and tis-

sue and/or conditions such as bean and leaf, defense related (leaves), differential display, immature zygotic embryo, mature zygotic embryo, young red leaves and somatic embryo. The majority of ESTs belonged to bean and leaves. Phrap was used to assemble ESTs to contigs and also to remove redundant sequences. A total of 769 assembled sequences (contigs) and 2015 singletons were generated after using Phrap. The putative function of assembled contigs was interpreted using the blastx and tblastn tools of NCBI and ExPASy (table 3).

Frequency of SSRs in cacao ESTs and contigs

A total of 87 perfect, 10 imperfect and 37 compound repeats were found in individual ESTs of cacao using all the five tools (table 4). The estimated frequency of SSRs in cacao ESTs was one per 26.9 kb, based on the total size of the examined sequences (2616452 bp). The frequency of SSRs in ESTs of other crops reported earlier were one per 11.8 kb in rice; one per 23.8 kb in soybean, one per 17.24 kb in wheat and one per 28.32 kb in maize (Gao *et al*. 2003) and one per 2.5 kb in eucalyptus (Rabello *et al*. 2005). But the SSR frequency can follow a unique pattern in a plant species where seven-fold variation was seen in 24 plants of widely differing taxa with a mean of 114.7 SSRs per Mb (von Stackelberg *et al*. 2006) with a standard deviation of +/− one per 60.1 Mb. They also noticed a very low frequency of microsatellites in onion and *Pinus* that have only 38.9 and 41.5 SSRs per Mb, respectively. The SSR frequency reported in

Table 3. Blast results of contigs containing validated SSRs.

Primer	Contig names	Total	Base-pair putative gene information	*E* value
TH1	Beanandleaf.fasta.Contig320	$(CT)_{11}$	HMGB2 (HIGH MOBILITY GROUP B 2) *Arabidopsis thaliana*	6.00E-44
TH2	Beanandleaf.fasta.Contig284	$(AT)_{11}$	LON peptidase N-terminal domain and ring finger 2, isoform CRA_b *Homo sapiens*	4.3
TH3	Defencerelated.fasta.Contig57	$(TG)_{10}$*	No similar sequences found (in ncbi and expasy)	
TH4	Beanandleaf.fasta.Contig316	$(GAA)_8$	Os10g0400200 *Oryza sativa* (japonica cultivar-group)	4.00E-10
TH5	Beanandleaf.fasta.Contig314	$(AAG)_{10}$	Nucleic acid binding *Arabidopsis thaliana*	2.00E-69
TH6	Beanandleaf.fasta.Contig72	$(AAC)_7$	Os06g0306300 *Oryza sativa* (japonica cultivar-group)	4.00E-14
TH7	Beanandleaf.fasta.Contig280	$(CTG)_7$*	Hypothetical protein Lxx13950 (Leifsonia xyli subsp. xyli str. CTCB07)	4.3
TH8	Beanandleaf.fasta.Contig376	$(CTG)_6$	Os06g0574500 *Oryza sativa* (japonica cultivar-group)	1.00E-43
TH9	Defenceralated.fasta.Contig262	$(ATAA)_5$	Nitrate/nitrite transporter-like protein *Salinispora tropica* CNB-440	0.22
TH10	Defencerelated.fasta.Contig52	$(GTAT)_5$	Os07g0179300 *Oryza sativa* (japonica cultivar-group)	3.00E-62
TH11	Beanandleaf.fasta.Contig205	$(CCAAA)_4$*	Os06g0247500 *Oryza sativa* (japonica cultivar-group)	1.00E-41

Table 4. Type of simple sequence repeats located using five tools.

Tool	Type	Mono	Di	Tri	Tetra	Penta	Hexa	Above deca	Compound
MISA	ESTs	21	17	19	4	0	1	0	37
	Contigs	2	3	3	2	0	0	0	6
TRA	ESTs	21	19	22	0	1	2	1	–
	Contigs	0	3	4	0	0	0	0	–
TROLL	ESTs	20	17	15	0	1	0	0	–
	Contigs	1	3	3	2	0	0	0	–
SSRIT	ESTs	–	10	16	2	0	4	0	–
	Contigs	–	3	3	0	0	0	0	–
SSR primer	ESTs	–	10	20	4	2	0	0	–
	Contigs	–	3	4	2	1	0	0	–

SSRs located by different software are listed. SSRIT and SSR primer have no provision to detect monomer repeats. TRA have detected maximum SSRs from ESTs and contigs except monomers. TRA also detected one above decamer repeats. SSRIT detected higher number of hexanucleotide repeats. Overall tri-35 nos (36%) mono-25 nos (25.7%) and di-21 nos (21.6%) reported in ESTs.

the present study would change for cacao when more EST sequences are made available. At present, it represents only a fraction of the cacao genome.

Type of SSRs and motif groups in cacao ESTs and contigs

Results of the SSRs mined from EST libraries of cacao using MISA, TRA and TROLL showed mononucleotide repeats as the most abundant SSR type. Among mononucleotides, the repeat A/T occurred 21 times (95% of the mononucleotide) as identified by MISA as well as TRA. The program TROLL detected 19 A/T repeats among the 20 monomeric repeats. Only one C/G mononucleotide was seen. SSRIT, which is designed to detect microsatellite dimers to decamers, showed trinucleotide repeats as the most abundant repeats in EST (50%), followed by dinucleotide (31.25%). SSR primer also showed similar trends of abundance of trinucleotide repeats (58.7%) followed by dinucleotide (21.74%, including imperfect repeats). SSR primer would also find perfect and imperfect repeats; we observed 10 imperfect repeats with length of 20 bp or more. We identified 134 SSRs (87 perfect, 10 imperfect and 37 compound repeats) by pooling the results of all the five tools.

Imperfect repeats included one dinucleotide, seven trinucleotide, one tetranucleotide and one pentanucleotide repeats. Pentameric and above repeats were detected only by SSR primer. Among the 37 compound repeats, 24 compounds were formed with the combination of mononucleotides and eight of them with trimeric repeats. Apart from these 134 SSRs, 11 monomers were reported at either the start or the end positions of a sequence. These were considered to correspond to poly A/T tails and hence were not used for further studies like designing of primers. Trinucleotide (35 nos) repeats (36%) were mainly seen followed by 25 mononucleotide (25.7%). Twenty-one dimer, six tetramer, three pentmer, six hexamer and one above decamer (21 mer) repeats were observed from this study. The mononucleotide repeat A/T occurred 24 times and among the trinucleotide repeats, the AAG/AGA/GAA/CTT/TTC/TCT group was more abundant than other classes (58.6%). Among dinucleotides, the AG/GA/TC/CT group was the most prevalent (65%). The frequency of cacao EST-SSRs classified by repeat unit size is given in table 4, and the distribution of motifs in EST-SSRs is given in table 5.

Five different tools detected a total of 12 unique perfect SSRs, and two imperfect sites and six compound repeats in 769 contigs. The result of detected SSRs from cacao contigs by MISA program showed the dinucleotide and trinucleotide repeats as more abundant SSRs. Trinucleotide (57.14%) and dinucleotide (42.86%) repeats were found as most frequent when TRA was used. The most abundant motif was AGG/GAG/GGA/TCC/CTC/CCT (42.86%). The program TROLL could detect a total of nine SSRs in the contigs, dinucleotide and trinucleotide repeats were the most abundant

Table 5. Type of motifs of SSRs in ESTs of cacao.

Motif type	No.
A/T	24
G/C	1
AG/GA/TC/CT	13
AT/TA	7
CA/AC/GT/TG	1
AAC/ACA/CAA/TTG/TGT/GTT	3
AAG/AGA/GAA/TTC/TCT/CTT	18
AAT/ATA/TAA/ATT/TTA/TAT	1
ACC/CCA/CAC/TGG/GGT/GTG	1
CAT/ATC/TCA/GTC/TCG/CGT	3
CTG/TGC/GCT/GAC/ACG/CGA	7
GCA/CAG/AGC/CGT/GTC/TCG	2
ATAA	3
GTAT	2
TTCT	1
(ACCAA)4	2
(ATAAC)4	1
(CATCAC)4	3
(CTCCCT)6	1
(GCACAC)4	1
(GTGATG)4	1
21(CAAATCTGGTGCTAATGCCTC)6	1
Total	97

types (33.33%). Among the trimerics, the most abundant repeat were of group AGG/GAG/GGA/TCC/CTC/CCT, which were repeated twice (22.22%). The software tool SSRIT also showed that dimeric and trimeric repeats were as equally distributed (50%) in cacao contigs with the most abundant motif being AGG/GAG/GGA/TCC/CTC/CCT (33.33%). One pentameric repeat $(ACCAA)_4$ was identified by using the tool SSR primer and two tetranucleotide repeats ($ATAA_5$ and $GTAT_5$) were located by using TROLL and SSR primer in contigs.

The result of cocoa contigs by different software showed that mono, di, tri, tetra and penta nucleotide constituted 16.66%, 25%, 33.33%, 16.66% and 8.3%, respectively. This clearly shows that trinuleotide repeats are the most abundant repeats in cocoa ESTs. Trinucleotide repeats were found to be the most abundant SSR class recovered from plant ESTs (Cardle *et al*. 2000; Temnykh *et al*. 2001; Kantety *et al*. 2002; Varshney *et al*. 2002; Gao *et al*. 2003; Kumpatla and Mukhopadhyay 2005). Among the trinucleotide repeats, AGG/GAG/GGA/TCC/CTC/CCT class were dominant than other trinucleotide classes in cacao ESTs. The former class was found to be the most frequent in dicot plants (Kumpatla and Mukhopadhyay 2005). Trimerics of type AGG/GAG/GGA/TCC/CTC/CCT were found to be the most frequent SSR in rice, barley, soybean (Gao *et al*. 2003) and in *Arabidposis* ESTs (Cardle *et al*. 2000). The CCG/CGG type of trimerics was the most abundant in cereals (Varshney *et al*. 2002). Most EST sequences consist of exonic regions which are under strong selection against frame-shift mutations as

they will be translated into proteins. As codons are functional units of three nucleotides, indel mutations causing a shift in three nucleotides will not perturb the current reading frame of a given gene (Metzgar *et al*. 2000). For this reason, trinucleotide repeats are expected to be the most abundant SSR class found in ESTs. The distribution and frequency of SSRs is a result of many factors such as mutation, selection and DNA repair mechanisms (Sreenu *et al*. 2007). The computational ability of different tools used to find cacao contigs and individual EST sequences is listed in table 4.

In the present study, we found trinucleotide repeats as the most frequent repeat class in cacao ESTs (32.4%) as well as contigs (33.33%). The most abundant repeat was the A/T type and G/C repeats were rare. This is an agreement with genomic DNA analysis of *Arabidopsis* (Lawson and Zhang 2006), where A/T motifs were abundant among the monomer repeats in EST libraries. Similar results were also seen in eucalyptus (Rabello *et al*. 2005) and in many other crop species (Gao *et al*. 2003). In mycobacterial genomes stretches of long monomer repeats were noticed as they act as contingency loci (Sreenu *et al*. 2007). Among the dimerics, the AG/TC type was more common than other groups in cacao ESTs. AG/TC was the predominant motif among the dinucleotide repeats in other dicot plants such as apple (Newcomb *et al*. 2006), coffee (Aggarwal *et al*. 2007) and eucalyptus (Rabello *et al*. 2005). But AC/GT was found abundant in primitive plant forms such as mosses, algae and conifers (von Stackelberg *et al*. 2006). Genomes of dicot plants such as cacao are rich in A+T content, hence the microsatellite motifs identified in these species are A+T rich (Kumpatla and Mukhopadhyay 2005). In contrast the monocot plant genomes are rich in G+C content hence the SSR motifs in these plants are rich in G/C (Varshney *et al*. 2002). We have also found three pentanucleotide repeats (one imperfect repeat included) by using TRA and SSR primer, six hexanucleotide repeats (by using MISA, TRA, TROLL and SSRIT) and one 21-mer repeat $(CAAATCTGGTGCTAATGCCTC)_6$ using program TRA in cacao EST libraries. These rare repeats can be potential markers in detecting polymorphism and mapping studies.

So far, SSR markers with motifs of compound repeats, di, tri, tetra and one penta repeats have been found in cacao either by genome enrichment (Araujo *et al*. 2007) or electronically sorted from EST libraries (Borrone *et al*. 2007; Lima *et al*. 2008). Among the experimentally tested microsatellite motifs, most of those with (AAG)n show potential in linkage mapping in cacao (Borrone *et al*. 2007) and (AG)n repeats and (CAA)n repeats show polymorphism (Araujo *et al*. 2007). The definition of SSR varies by size and type of repeat and some authors do not consider monomer repeats as SSRs. Many authors target class II SSRs also with size of 10 bp and above. Few authors consider only SSR if the repeat motif larger than 20 bp (Varshney *et al*. 2002). Hence comparisons of SSR size and type of repeat are difficult to discuss. Kumpatla and Mukhopadhyay (2005) targeted four classes of repeats (viz. mono, di, tri and tetra) with a default settings for repeat as 15 for mono nucleotides and five for di-, tri-, or tetra-nucleotides. Our study found trimeric repeats to be predominant in individual ESTs followed by monomeric repeats, because we looked mainly for SSR that are larger than 20 bp.

Among the five tools used, MISA program gave the maximum coverage of SSRs in both ESTs and contigs. It is powerful to detect mono to hexa repeats and compound repeats, but it is limited by its inability to detect above hexamer repeats. SSR primer shows both perfect and imperfect repeats. It is useful for finding SNPs in microsatellite and automated primer designing. We found TRA highly useful to molecular biologists interested in locating hexamer and above sized repeats, even minisatellites. However, those interested in locating maximum repeats but less than hexamer size should use the MISA Perl script.

We identified putative functions of the ESTs using blastx, tblastn tools of NCBI and ExPASy servers. The tblastn was used in the case where blastx analysis gave no similarity. Six frames were created with NCBI ORF finder http://www.ncbi.nlm.nih.gov/gorf/gorf.html and lengthy sequence was chosen to perform tblastn. The tool has blast options like blastp and tblastn. The database selected was non-redundant sequence database of Viridiplantae (taxid: 33090). Putative functions such as PGK (phosphoglycerate kinase), gibberellin 20-oxidase 1, flavanone 3-hydroxylase, DNA-directed RNA polymerase II subunit RPB2 (RNA polymerase II subunit B2, RNA polymerase II subunit 2, DNA-directed RNA polymerase II 135-kDa polypeptide) etc. were identified with blast *E* value exactly 0 (table 3).

The resulting microsatellites were compared with existing cocoa microsatellites (Pugh *et al*. 2004; Araujo *et al*. 2007; Lima *et al*. 2008) but no similarity could be found. Hence, the present study gives additional set of microsatellites from ESTs for use in cacao genome analysis. We intended to test all the designed primers from cocoa contigs because we could design only 11 primer pairs from the 20 SSRs reported. The remaining nine SSRs were either at the beginning or the end of the contig sequences and, consequently, we did not further design primers for them. Eleven primers were tested for amplification of cacao genomic DNA using PCR. Nine of these primers (viz. TH1, TH2, TH4, TH6, TH7, TH8, TH9, TH10 and TH11) produced amplification (figure 1). Out of nine EST-SSR markers, only seven markers (viz. TH1, TH2, TH4, TH8, TH9, TH10 and TH11) produced amplicons of the expected size. Two of the primers viz. TH6 and TH7 produced amplicons, that were larger than the expected size. The outcome that some of the obtained amplicons were larger than expected size could be explained by indels (insertion/deletions; Yu *et al*. (2004)) or the incorporation of small introns (Thiel *et al*. 2003). EST sequences do not include the intron(s) and we have used the genomic DNA for validation of primers. When the sequences contain very large introns, the expected amplicon size

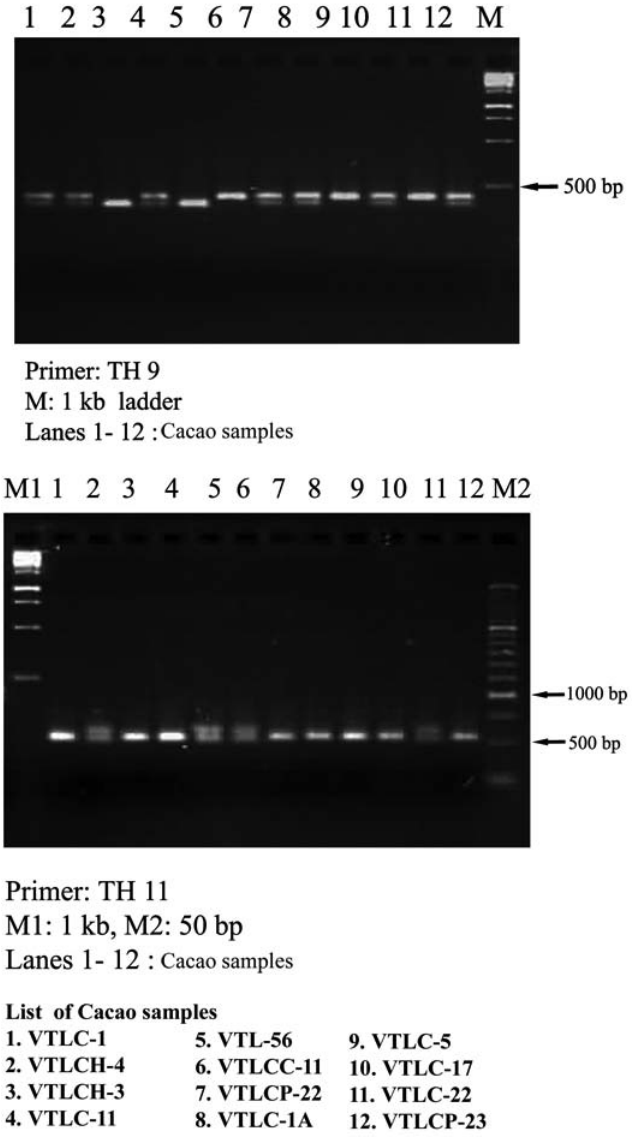

Figure 1. EST-SSR marker profile of cacao samples.

including introns also might get larger than that calculated based on EST sequence. Sometimes such large amplicons, if exceeding 2-kb size, are not obtained in normal PCR reactions. Nonamplification for the two primers (TH3 and TH5) could be the result of presence of large introns. Another reason for nonamplification problems could be the result of incorrect sequence assembly especially in contig 57 (table 3). The primer TH3 designed from this contig did not give any amplification. This contig has no match with any known gene (table 3).

The seven markers selected were used to screen a representative set of 12 cocoa accessions to test their level of polymorphism detection. The seven EST-SSR markers produced 27 polymorphic alleles in the 12 cocoa accessions ranging from two to six alleles per locus with an average of 3.85 alleles per locus (table 6). The TH2 primer pair was observed as the most polymorphic, with six alleles per locus, whereas the least polymorphic EST-SSR marker was TH1, with two alleles per locus. The average PIC value, calculated from the average frequency of polymorphic bands across all genotype was 0.57. The primer TH8 generated the greatest PIC value of 0.802 and the lowest was observed in TH4 with PIC value 0.156 (table 6).

A similarity index, based on the Dice (1945) coefficient, was obtained after pair-wise comparison among 12 cocoa accessions. The highest similarity index of 0.80 was observed between accessions I-14 and VTLC-1, and lowest similarity index (0.111) was observed between NC42/94X III 35 and VTLCC-1. The dendrogram (figure 2) generated separated the 12 cocoa accessions into two major clusters at 35% similarity level. The first major cluster had five sub-clusters and included nine accessions. Accessions I-14 and VTLC-1 exhibited 80% similarity. The second major cluster had two sub-clusters. VTLCC-1 formed a distinct accession.

In the present study, five computational tools were compared for their efficiency in locating microsatellite repeats in ESTs and contigs of cacao. MISA software tool gives the maximum coverage of SSRs detected. It is a widely used tool by many researchers for the purpose. SSR primer gives both perfect and imperfect repeats. It is useful for finding SNPs in microsatellite and automated primer designing. We found TRA highly useful for molecular biologists

Table 6. Name of the primer with polymorphic bands, per cent polymorphism and polymorphism information content.

Primer	Total number of bands produced	Polymorphic bands	Per cent polymorphism	PIC value
TH1	3	3	100	0.569
TH2	6	6	100	0.760
TH4	3	3	100	0.156
TH8	5	5	100	0.802
TH9	3	3	100	0.642
TH10	5	5	100	0.753
TH11	2	2	100	0.277

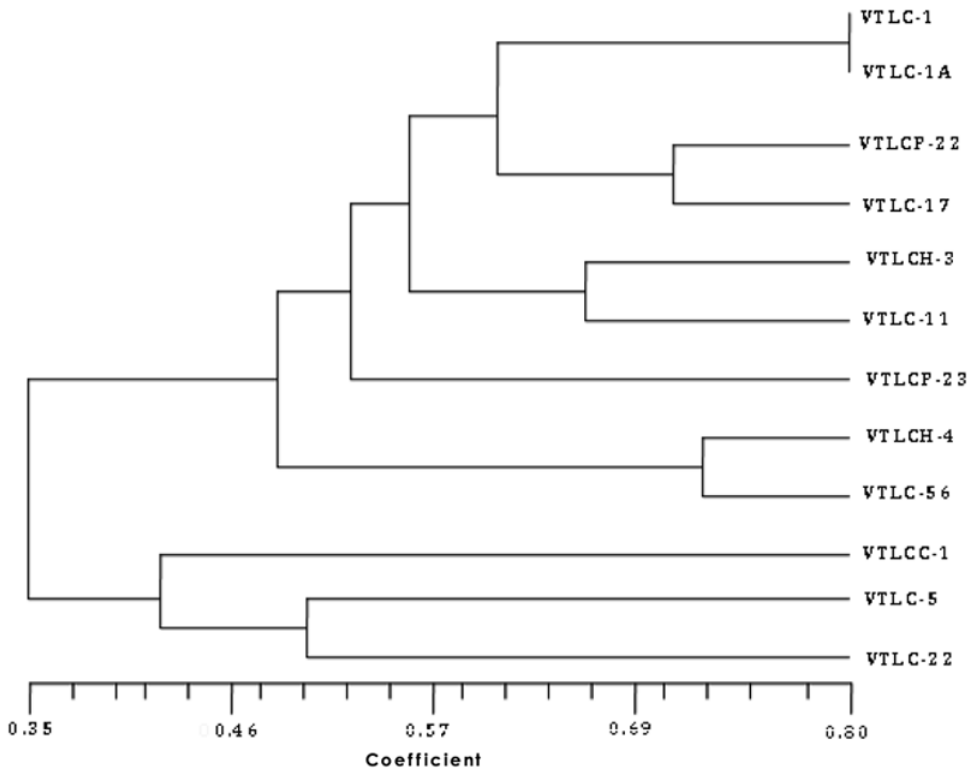

Figure 2. UPGMA dendrogram of the 12 cacao accessions based on Dice co-efficient.

interested in locating specifically large size (hexamer and above) repeats. SSRs are found at a frequency of one per 26.9 kb in expressed sequences of cacao plant. One-third of repeats are trimers protecting the frame shift mutations in coding regions. Our study has practical implications in using appropriate tools for analysing EST resources for marker development. Seven primer pairs designed from contigs of EST-SSR amplified a representative set of 12 cacao accessions. These seven unique polymorphic SSR loci identified in the study are useful in analysing the cacao genome. We provide the results of the study as a public domain database.

Database availability

The result of the study is compiled as a public domain database (http://riju.byethost31.com/cocoa/). User can use query for different SSR finding tools to access the SSRs in cacao ESTs. The database also gives all the experimentally validated primers, including those from this study, and other supplementary information.

Acknowledgements

This work was supported by a grant from Department of Biotechnology (BTISnet), Government of India. We are grateful to Dr V. Rajagopal, former Director, and Dr George V. Thomas, Director, CPCRI, Kasaragod for the encouragement, guidance and facilities.

References

Aggarwal R. K., Prasad S. H., Varshney R. K., Prasanna R. B., Krishnakumar V., Lalji S. *et al*. 2007 Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor. Appl. Genet.* **114**, 359–372.

Araujo J. S., Intorne A. C., Pereire M. G., Lopes U. V. and deSouza Filko G. A. 2007 Development and characterization of novel tetra, tri and dinucletotide repeat microsatellite markers in cacao (*Theobrama cacao* L.) *Mol. Breed.* **20**, 73–81.

Bilgen M., Karaca M., Onus A. and Ince A. 2004 A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. *Bioinformatics* **20**, 3379–3386.

Borrone J. W., Brown J. S., Kuhn D. N., Motamayor J. C. and Schnell R. J. 2007 Microsatellite markers developed from *Theobroma cacao* L. expressed sequence tags. *Mol. Ecol. Notes* **7**, 236–239.

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

Couch J. A., Zintel H. A. and Fritz P. J. 1993 The genome of the tropical tree *Theobroma cacao* L. *Mol. Gen. Genet.* **237**, 123–128.

Dice L. R. 1945 Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302.

Ewing B. and Green P. 1998 Basecalling of automated sequencer traces using phred II. error probabilities. *Genome. Res.* **8**, 186–194.

Gao L. F., Tang J. F., Li H. W. and Jia J. Z. 2003 Analysis of mi-

crosatellites in major crops assessed by computational and experimental approaches. *Mol. Breed.* **12**, 245–261.

Jurka J. and Pethiygoda C. 1995 Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* **40**, 120–126.

Kantety R. V., La Rota M., Matthews D. E. and Sorrells M. E. 2002 Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum, and wheat. *Plant Mol. Biol.* **48**, 501–510.

Katti M. V., Ranjekar P. K. and Gupta V. S. 2001 Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18**, 1161–1167.

Kumpatla S. P. and Mukhopadhyay S. 2005 Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* **48**, 985–998.

Lawson M. J. and Zhang L. 2006 Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol.* **7**, R14.

Lima L. S., Gramacho K. P., Gesteira A. S., Lopes U. V., Gaiotto F. A., Zaidan H. A. *et al*. 2008 Characterization of microsatellites from cacao-Moniliophthora perniciosa interaction expressed sequence tags. *Mol. Breed.* **22**, 315–318.

Martins W., De Sousa D., Proite K., Guimaraes P., Moretzsohn M. and Bertioli D. 2006 New softwares for automated microsatellite marker development. *Nucleic Acids Res.* **34**, e31.

Metzgar D., Bytof J. and Wills C. 2000 Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* **10**, 72–80.

Newcomb R. D., Crowhurst R. N., Gleave A. P., Rikkerink E. H. A., Allan A. C., Beuning L. L. *et al*. 2006 Analysis of expressed sequence tags from apple. *Plant Physiol.* **141**, 147–166.

Pearson C., Sinden E. and Richard R. 1998 Trinucleotide repeat DNA structure: Dynamic mutations from dynamic DNA. *Curr. Opin. Structl. Biol.* **36**, 884–889.

Powell W. W., Machery G. C. and Provan J. 1996 Polymorphism revealed by simple sequence repeats. *Trends Genet.* **1**, 76–83.

Pugh T., Fouet O., Tisterucci A. M., Brottier P., Abouladze M., Deletrez C. *et al*. 2004 A new cacao linakage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theor. Appl. Genet.* **108**, 1151–1161.

Quackenbush J., Cho D., Lee F. L., Hott I., Karamychera S. and Parizi B. 2001 The *TIGR* gene indices: analysis of gene transient sequences in highly sample eukaryotic species. *Nucleic Acids Res.* **29**, 159–164.

Rabello E., Nunes de Souza A., Saito D. and Tsai S. M. 2005 *In silico* characterization of microsatellites in *Eucalyptus* spp.: abundance, length variation and transposon associations. *Genet. Mol. Biol.* **28**, 582–588.

Robinson A. J., Love C. G., Batley J., Barker G. and Edwards D. 2004 Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics* **20**, 1475–1476.

Rohlf F. J. 1998 On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Sys. Biol.* **47**, 147–158.

Smith J. S. C., Chin E. C. L., Shu H., Smith O. S., Wall S. J., Senior M. L. *et al*. 1997 An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPs and pedigree. *Theor. Appl. Genet.* **95**, 163–173.

Sneath P. H. A. and Sokal R. R. 1973 *Numerical taxonomy, the principles and practice of numerical classification.* W. H. Freeman, San Francisco.

Sreenu V. B., Kumar P., Nagaraju J. and Nagarajaram H. A. 2007 Simple sequence repeats in mycobacterial genomes. *J. Biosci.* **32**, 3–15.

Tang J. F., Gao L., Cao Y. and Jia J. 2006 Homologous analysis of EST-SSRs and transferability of wheat SSR-EST markers across barley, rice and maize. *Euphytica* **151**, 87–93.

Temnykh S., DeClerck G., Lukashova A., Lipovich L., Cartinhour S. and McCouch S. 2001 Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* **11**, 1441–1452.

Thiel T., Michalek V. and 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 R. K., Graner A. and Sorrells M. E. 2005 Genic microsatellite markers in plants: features and applications. *Trends Biotech.* **23**, 48–55.

Varshney R. K., Thiel T., Stein N., Langridge P. and 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.

Verica J., Maximova S., Strem M., Carlson J., Bailey B. and Guiltinan M. 2004 Isolation of ESTs from cacao (*Theobroma cacao* L.) leaves treated with inducers of the defense response *Plant Cell Rep.* **23**, 404–413.

von Stackelberg M., Rensin S. A. and Reskhi R. 2006 Identification of genic moss SSR markers and a comparative analysis of 24 algal and plant gene indices reveal specific rather group specific characteristics of microsatellites. *BMC Plant Biol.* **6**, 9.

Yasodha R., Sumathi R., Chezhian P., Kavitha S. and Ghosh M. 2008 Eucalyptus microsatellites mined *in silico*: survey and evaluation. *J. Genet.* **87**, 21–25.

Yu J. K., La Rota M., Kantety R. V. and Sorrells M. E. 2004 EST derived SSR markers for comparative mapping in wheat and rice. *Mol. Gen. Genomics* **271**, 742–751.

Received 4 August 2008, in final revised form 24 March 2009; accepted 25 March 2009

Published on the Web: 30 July 2009