

# Structural characterization and mapping of functional EST-SSR markers in *Theobroma cacao*

Olivier Fouet · Mathilde Allegre · Xavier Argout · Mélanie Jeanneau ·  
Arnaud Lemainque · Sylvana Pavék · Anne Boland · Ange Marie Risterucci ·  
Gaston Loor · Mathias Tahí · Xavier Sabau · Brigitte Courtois · Claire Lanaud

Received: 6 August 2010 / Revised: 25 January 2011 / Accepted: 15 February 2011 / Published online: 25 March 2011  
© Springer-Verlag 2011

**Abstract** *Theobroma cacao* L. is a major cash crop for tropical countries, providing incomes for 14 million small farmers. Establishing sustainable disease resistance and maintaining cocoa qualities are among the major objectives of breeding programs. To enrich the high-density genetic map, useful for all cocoa genetic studies, with gene-based markers, a recently produced large EST resource was mined to develop expressed sequence tag-based simple sequence repeat markers (EST-SSRs) defined in genes with a putative known function. A set of 174 polymorphic EST-SSRs was identified from a selection of 314 non-redundant EST-SSRs with a putative known function. Of them, 115 loci were mapped on the cocoa reference map. This new map contains 582 codominant markers arranged in ten linkage

groups corresponding to the haploid number of chromosomes. An average interval between markers of 1.3 cM was found, with approximately one SSR every 2 cM. This new set of EST-SSRs includes 14 candidate genes for plant resistance or cocoa qualities. The percentage of polymorphic SSRs varied depending on the different gene regions from which they originated, with respectively 54%, 69%, and 82% of polymorphic EST-SSRs originating from coding sequences, and from the non-coding untranslated 5' UTR and 3'UTR regions. This new map contains a set of 384 SSR markers that are easily transferable across different mapping populations and useful for all genetic analyses in *T. cacao*. The new set of EST-SSRs will be a useful tool for studying the functional diversity of populations and for carrying out association mapping studies.

Communicated by A. Dandekar

O. Fouet (✉) · M. Allegre · X. Argout · M. Jeanneau ·  
A. M. Risterucci · X. Sabau · B. Courtois · C. Lanaud  
CIRAD-Biological Systems Department, UMR DAP,  
TA 40/03,  
34398 Montpellier, France  
e-mail: olivier.fouet@cirad.fr

A. Lemainque · S. Pavék · A. Boland  
CNG,  
2 rue Gaston Cremieux,  
CP 5721, 91057 Evry, France

G. Loor  
EET-Pichilingue, INIAP,  
Código Postal 24 Km 5 vía Quevedo,  
El Empalme, Ecuador

M. Tahí  
CNRA,  
BP 1740, Abidjan 01,  
Abidjan, Ivory Coast

**Keywords** *Theobroma cacao* · EST-SSR · Genetic map ·  
Candidate genes

## Introduction

*Theobroma cacao* L., which belongs to the Malvaceae family (Whitlock et al. 2001), is a naturally growing tree in South America. *T. cacao* is a diploid species ( $2n=2x=20$ ) with a small haploid genome size of 390 Mb (Lanaud et al. 1992; Figueira et al. 1992).

*T. cacao* germplasm was classified into three morphogenetic groups corresponding to the first cultivated cocoa varieties (Cheesman 1944), thus distinguishing between Forastero, Criollo, and Trinitario (hybrid form between the first two groups). However, the diversity of the Forastero genetic pool is very high (Lerceteau et al. 1997; Laurent et al. 1993a, b), and a new classification in ten major groups has been proposed very recently in relation to geographic

and genetic differences between the Forastero populations (Motamayor et al. 2008).

Cocoa is cultivated in humid tropical regions (Africa, Latin America, and Asia), where it is one of the most economically important perennial crops. Cocoa is the third-largest product on the world market for tropical commodities after sugar and coffee. Cocoa is mainly produced on smallholdings and, according to the World Cocoa Foundation (Cocoa market update 2009), 40 to 50 million people depend on it for their livelihood worldwide. About 3.5 million tons of cocoa are produced annually, 70% coming from Africa, and the demand for cocoa is increasing. This crop is subject to several fungal diseases, including *Moniliophthora* sp. (Aime and Phillips-Mora 2005) and *Phytophthora* sp., causing major yield losses (globally, 30%) in all regions. A major objective of breeding programs is to provide new varieties with sustainable resistance to fungal diseases, by accumulating several sources of resistance in the newly created varieties.

Several linkage maps based on codominant or dominant markers (restriction fragment length polymorphisms (RFLPs), SSRs, random amplification of polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs) and isoenzymes) have already been published for *T. cacao* (Lanaud et al. 1995; Crouzillat et al. 1996; Risterucci et al. 2000). A high-density linkage map, with codominant markers only, was developed (Pugh et al. 2004) with the addition of 201 simple sequence repeat markers (SSRs), resulting in a mean genome coverage of one SSR every 3 cM.

SSRs are present in the majority of eukaryotic genomes and consist of DNA sequences of simple motifs (mono- to hexa-nucleotides) that are repeated in tandem (Litt and Luty 1989; Beckman and Soller 1990). SSR length variation is mainly due to the slippage of DNA polymerase or unequal crossover leading to the insertion or deletion of tandem repeats (Ellegren 2004). Allelic variation in SSR loci can readily be detected by PCR using specific primers flanking the SSR motif.

Among the different classes of molecular markers, SSRs have been extensively used in many crop species for numerous applications in plant genetics and breeding because of their simplicity, reproducibility, high degree of polymorphism, codominant inheritance, relative abundance, and good genome coverage (Powell et al. 1996). Moreover, SSRs are technically easy to analyze because of the small amount of DNA required and their suitability for multiplexing on automated systems. SSR markers are widely used in cultivar fingerprinting, genetic diversity assessment, molecular mapping, QTL detection, and marker-assisted selection.

The standard method for developing SSR markers is based on the creation of a small-insert genomic library, subsequent hybridization with tandem repeated oligonucleotides (SSR nucleotides) and sequencing of candidate

clones, making the process quite time-consuming and laborious (Thiel et al. 2003). Furthermore, the SSRs acquired are limited to the probed SSR motifs (Chen et al. 2006). However, this approach has been widely used in the past for many species. With the increased public availability of genomic and EST sequences it is now possible to mine these sequence collections directly for SSR identification.

In cocoa, 273 SSR markers were developed and mapped following the construction of genomic libraries enriched with SSRs (Lanaud et al. 1999; Risterucci et al. 2000; Pugh et al. 2004). These SSRs were already useful tools for the identification of QTLs related to morphological or disease resistance traits (Crouzillat et al. 2000; Flament et al. 2001; Risterucci et al. 2003; Clement et al. 2003; Queiroz et al. 2003; Brown et al. 2007; Brown et al. 2005). The high level of polymorphism for these SSR markers also helped to create a composite map by mapping SSRs (bridge markers) common to three different maps (Brown et al. 2008), and led to the establishment of a consensus map derived from ten different progenies, which has enabled a meta-analysis of QTLs related to resistance traits in cocoa (Lanaud et al. 2009). The SSRs defined in genomic sequence often correspond to non-coding sequences and thus cannot reflect the functional diversity of cocoa populations. Alternatively, expressed sequence tag-based SSRs (EST-SSRs) can be identified in sequence databases using data mining pipelines composed primarily of SSR search and primer design programs. SSRs derived from EST libraries offer the advantage, over genomic SSRs, of being intrinsically associated with coding sequences within the genome (Eujayl et al. 2002), thus providing functional information about genome regions linked to trait variations.

To date, EST-SSRs have been identified, developed, and used in a variety of studies for several plant species. Firstly, the use of EST libraries to develop polymorphic SSR markers was investigated in some cereals such as barley (Thiel et al. 2003), durum wheat (Eujayl et al. 2002) and rye (Hackauf and Wehling 2002) and more recently in some trees such as apricot (Decroocq et al. 2003), prunus (Ogundiwin et al. 2009) and in some tropical species such as sugar cane (Cordeiro et al. 2001; Pinto et al. 2004), cotton (Han et al. 2004; Guo et al. 2007) or coffee (Poncet et al. 2006; Aggarwald et al. 2007).

Given their high level of transferability from one species to another (Gupta and Prasad 2009; Luro et al. 2008), EST-SSRs constitute a useful tool for functional diversity studies and comparative mapping between species (Varshney et al. 2005). Although genomic SSRs are more polymorphic than genic SSRs (Cho et al. 2000; Lee et al. 2004), the latter have some important advantages: (1) the estimated frequency of genic SSRs is higher than genomic SSRs (Morgante et al. 2002), (2) such genic markers are functional markers that can contribute to

“direct allele selection”, if they are shown to be completely associated with, or even responsible for, the variation of a trait of interest (Sorrells and Wilson 1997), (3) EST-SSR markers provide a set of readily shared markers that can be used to unify different genetic maps and establish consensus maps and (4) they can be more easily transferred between related species if they are derived from putative candidate genes related to traits of interest that are relatively well conserved among taxa (Gupta and Rustgi 2004).

A first set of 34 EST-SSRs was published for cocoa (Borrone et al. 2007; Riju et al. 2009). These markers were identified in the first ESTs (6581 sequences) from leaf and bean cDNA libraries (Jones et al. 2002), and from suppressive subtractive hybridization (SSH) from leaves treated with inducers of the defense response (Verica et al. 2004), and shoot tips inoculated with witches’ broom caused by *Moniliophthora perniciosa* (Leal et al. 2007). More recently, Lima et al. (2010) identified 49 EST-SSRs from full length and SSH libraries constructed from shoots and pods elicited with *M. perniciosa*.

Recently, a large collection of cocoa ESTs was generated from various tissues under various conditions (Argout et al. 2008). In total, 149,650 valid EST sequences were generated, corresponding to 48,594 unigenes, of which 29,849 unigenes shared significant homology with public sequences from other species. This constitutes an available resource to search for new SSRs in large quantities.

In order to find potentially mappable loci, the allelic diversity of some loci was tested on 8 genetically very different genotypes, including parents of offspring used in the construction of the reference map (Pugh et al. 2004). In this work, a subset of 314 EST-SSRs was analyzed for the potentiality to reveal polymorphism and 115 of them were mapped on the cocoa reference map.

## Materials and methods

### SSR detection

EST-SSR markers were detected in a *T. cacao* unigene set generated from the large EST collection produced under international collaboration and stored in the ESTtik database (Argout et al. 2008). SSRs were identified using the MicroSATellite identification tool (MISA). To consider as true SSRs, a minimum repeat unit of six was used for dinucleotide repeats and five for tri-, tetra-, penta-, and hexanucleotide repeats. The maximum interruption between two SSRs was set at 100 nucleotides. Primer3 software (Rozen SaS 2000) was then used to design three pairs of primers. The product size range was set between 100 and 280 nucleotides, the optimum, minimum and maximum

primer sizes were set at 17 and 23 nucleotides respectively, and the optimum melting temperature at 56°C (ranging from 50°C to 63°C).

The localization of SSRs in the coding sequence was estimated using the prot4EST pipeline (Wasmuth and Blaxter 2004). ESTs were first translated into high quality predicted polypeptides and SSRs were classified according to the position found in the proteins produced: in the coding sequence, in 5’UTR or in 3’UTR.

### EST functional annotation

Functional annotation of the EST-SSRs was carried out using the Blast2GO program (Conesa et al. 2005) which uses definitions of Blast assigned by gene ontology (GO) annotation (Ashburner et al. 2000). EST-SSR Blast searches were completed against the NCBI non-redundant protein database (Wasmuth and Blaxter 2004) with an *E* value set at 1e-6. Ten maximum hits were retained per result. The Gene Ontology mapping step was performed against a locally built Blast2GO database and the Annotation Cutoff was set at 75.

### Polymorphism detection

The polymorphism of SSRs was revealed by genotyping eight genotypes originating from contrasting genetic groups: LAN28 (Criollo), SCA6 (Upper-Amazon Forastero), MAT 1–6 (Lower-Amazon Forastero), SA16, SNA1001, SNA1003 (Ecuadorian Nacional varieties) and two parents of the reference map: UPA402 (Upper-Amazon Forastero) and UF676 (Trinitario).

PCR reactions were performed on an MJ Research PTC Thermal cycler (MJ research, Waltham, Mass.) in a 20- $\mu$ l volume containing 10 ng of cocoa DNA, 0.25  $\mu$ M of forward primer, 0.25  $\mu$ M of reverse primer, 2 mM MgCl<sub>2</sub>, 50 mM KCL, 10 mM Tris–HCl (pH 8.3), 0.2 mM dNTP mix, incorporation of 0.34 nMol of  $\alpha$ -(<sup>33</sup>P) CTP (2  $\mu$ Ci) and 1 U Taq polymerase (Eurobio, France). The PCR reaction began with an initial denaturation step (95°C, 5 min), followed by ten cycles of 30 s at 95°C, primer annealing by touch-down (45 s from 55°C to 50°C), then 45 s at 72°C. The program continued with 25 cycles of 30 s at 95°C, primer annealing of 45 s at 50°C and 45 s at 72°C and was completed with final extension at 72°C for 8 min.

Twenty microliters of loading buffer (98% deionized formamide, 10 mM EDTA, bromophenol blue and xylene cyanol) was added to individual reactions. Samples were denatured at 94°C, and 5  $\mu$ L of each sample was subjected to electrophoresis at 55 W on 5% denaturing polyacrylamide gels containing 75 M urea in 0.5X TBE buffer (pH 8.0). The gels were dried and exposed for 48–72 h to X-ray film (Fermentas). The potentially mapped loci were

revealed with the polymorphism given by the parents of the reference map.

### Mapping population

EST-SSRs were mapped on the same mapping population used to establish the previous cocoa reference map (Risterucci et al. 2000; Pugh et al. 2004). One hundred twenty-five progeny individuals, planted in Ivory Coast, were analyzed. This population resulted from a cross between two heterozygous cacao clones, UPA 402 (♀) an Upper-Amazon Forastero and UF676 (♂) a Trinitario. In this case, there were three possibilities of segregation, the loci that were homozygous for one parent and heterozygous for the other, segregation (1:1), and those that segregated in both parents (1:2:1 or 1:1:1:1). The latter situation made it possible to establish bridge markers needed to integrate each parental dataset and construct a consensus linkage map (Grattapaglia and Sederoff 1994).

### Genotyping of the mapping population

Several panels were designed for genotyping with the selected polymorphic EST-SSR markers. Forward or reverse primers were labeled at the 5' terminus with a fluorescent dye (Fam, Hex or Dragonfly Orange™). Simplex PCR assays with fluorescently labeled primer pairs were carried out in 10 µl reactions using one of two standard PCR protocols. The PCR reactions were performed with 3.3 pmole of each primer, 25 nmoles of MgCl<sub>2</sub>, 2.5 nmoles of each nucleotide, and 0.225 U of Taq Polymerase (BioTaq™, Bionline). The PCR products were pooled according to the defined panels.

Plates for the PCR and pooling steps were prepared with a robot pipetting station. The fluorescent PCR products were then analyzed on a MegaBACE™ 1000 Sequencer (Amersham Biosciences). The internal size standard was fluorescently labeled with a fourth dye (ET-ROX™ 400, Amersham Biosciences). The raw data were then analyzed using Genetic Profiler software (version 1.1).

The PCR reaction was initiated with a denaturation step at 95°C for 12 min, followed by ten cycles of 15 s at 94°C, 15 s at 50°C, 30 s at 72°C and 20 cycles of 15 s at 89°, 15 s at 50°C and 30 s at 72°C, then final extension at 72°C for 10 min.

### Linkage mapping and map construction

The map was constructed using JoinMap software, version 4.0 (Van Ooijen 2006), by integrating the new EST-SSR loci in addition to the genomic SSRs, defense and resistance gene analogs, RFLPs, and isoenzymes previously mapped (Pugh et al. 2004). The segregation of each SSR

marker was tested with a chi-square test for goodness-of-fit to the expected Mendelian segregation ratio. JoinMap software is able to combine data of several segregation types to construct a genetic map. Here, we used the population type “CP”, planned for a population resulting from the cross between two heterozygous diploid parents with possibly known linkage phases. A logarithm of the odds (LOD) score of five was used to identify linkage groups. The Kosambi mapping function, with a LOD of five and jump threshold of three, was used to convert recombination frequencies into map distances (Kosambi 1944).

### Marker nomenclature

Designation of the EST-SSR loci was based on the nomenclature presented on the last reference map (Pugh et al. 2004), thus these new loci were named mTcCIR like the SSR markers that were previously identified from genomic DNA.

## Results and discussion

### Polymorphism screening

In total, 149,650 valid EST sequences were generated corresponding to 48,594 unigenes. Of them, 29,849 unigenes showed significant homology with public sequences from other species. A functional characterization of ESTs was performed to assign Gene Ontology annotations. A set of 314 EST-SSRs identified in genes with a known function was selected from part of the cocoa EST collection constructed at CIRAD (Argout et al. 2008). After primer definition, the polymorphism of SSR loci was evaluated on 8 genotypes of various genetic origins (Table 1).

Amplification was successful for 251 of the 314 EST-SSRs screened. Of them, 174 revealed polymorphism between the eight genotypes tested, representing 55.4% of the total SSRs screened. Of those polymorphic SSRs, we were able to genotype 115 SSRs with the MegaBACE™ 1000 sequencer and integrate them on the genetic map. The markers mapped on the new map therefore corresponded to 36.6% (115) of the EST-SSRs screened (Table 2) and the 59 polymorphic but unmapped EST-SSRs amounted to 18.8% (Table 3).

When polymorphism was tested for eight different genotypes, 20.1% of loci were not amplified. This result can be partly explained by the presence of introns. Indeed, we defined the primers on cDNA sequences and the amplifications were tested on genomic DNA, which may have generated a larger amplified DNA fragment than expected, due to intron length.

**Table 1** Polymorphism revealed by 314 EST-SSR markers on eight genotypes

Type of SSR repeats	Number of screened SSRs	Good amplification	Monomorphic SSRs	Polymorphic SSRs	SSRs mapped	Mean number of alleles/locus for 8 genotypes
Dinucleotide	156 (49.7%)	124 (49.4%)	27	97	67	3.5
Trinucleotide	133 (42.4%)	108 (43%)	45	63	40	2.9
Composite	17 (5.4%)	14 (5.6%)	3	11	8	3.8
Tetranucleotide	5 (1.6%)	4 (1.6%)	2	2	0	2.5
Hexanucleotide	3 (0.9%)	1 (0.4%)	0	1	0	2
Total	314	251 (79.9%)	77 (24.5%)	174 (55.4%)	115 (36.6%)	

### EST-SSR distribution

The most frequent classes of SSRs presenting good amplification (251 EST-SSRs) were dinucleotide repeats (49.4%) and trinucleotide repeats (43%), compared with tetranucleotide (1.6%) and hexanucleotide repeats (0.4%). Dinucleotide SSR repeats and composite SSRs (often composed of dinucleotides) were the most polymorphic and revealed the largest number of alleles per locus out of the eight genotypes screened. For EST-SSRs with good amplification, 78.2% of loci were polymorphic when they consisted of dinucleotide repeats and 58.3% for the trinucleotide repeats (Table 4). This observation can be explained by polymerase errors that are more frequent when the repeat element is short.

The SSR loci showed strong heterogeneity in their level of polymorphism depending on the different gene regions from which they originated (5'UTR, 3'UTR, and CDS). The SSRs originating from the CDS region were the least polymorphic, with 53.5% of polymorphic loci, compared with the 5'UTR (68.9%) and 3'UTR (82.5%) regions. This result was expected and related to better conservation of the coding region, thus providing fewer polymorphic SSRs. The SSRs originating from the 5'UTR region were less polymorphic than those originating from the 3'UTR region. This observation can be explained by the presence, in the 5' UTR region, of important regions involved in the regulation of gene expression, which need to be more conserved.

The distribution of SSR repeat patterns was very different depending on the EST regions considered. In the CDS region, the main repeats were of the trinucleotide or hexanucleotide type (83.7% of SSRs). This percentage was much higher than in the 5'UTR and 3'UTR non-coding regions, with a 37.2% and 32.1% trinucleotide pattern respectively. This result seems logical because the trinucleotide and hexanucleotide repeat retains the correct reading phase for coding sequences, causing minimum disruption.

### Linkage analysis and map construction

Of the 115 EST-SSRs mapped, 74 (64.4%) were heterozygous for UF 676 only, nine (7.8%) were heterozygous for

UPA 402 only, and 32 (27.8%) were heterozygous for both parents (bridge markers) with two, three, or four alleles. These percentages were similar to those found for SSRs mapped by Pugh et al. (2004).

In order to check good co-linearity of the parents, two parental maps were constructed separately (data not shown) with JoinMap 4.0 software. The large number of bridge markers, 154 (amounting to 26.4% of total markers), led to easier construction of the consensus map between the two parents. However, the number of heterozygous loci in both parents was found to be very heterogeneous between the linkage groups, ranging from 0 for LG8 to 34 for LG9. LG8 and a large share of LG10 showed markers segregating parent UF676 only. The resulting order of the loci and map distances between markers are shown graphically in the integrated linkage map (Fig. 1).

The complete map contains 582 codominant markers (115 EST-SSRs, 269 SSRs from genomic DNA, 177 RFLPs, five isoenzyme loci, and 16 Rgenes-RFLPs) arranged in ten linkage groups for a LOD of 5.0, corresponding to the haploid chromosome number of *T. cacao*. The total length of the map was estimated at 779.2 cM, comparable to the last reference map (Pugh et al. 2004) which was 782.8 cM in length. The average distance between markers on this map was evaluated at 1.3 cM against 1.7 cM before. A large variation was found in the number of mapped loci between linkage groups, ranging from 32 in LG7 to 80 in LG9. Marker density between the different linkage groups was relatively homogeneous, from 1.1 cM for LG5 to 1.8 cM for LG8. Only LG10 showed lower marker density with an average of 2.2 cM between markers. The current level of genome SSR coverage was estimated approximately at one SSR every 2 cM. A genetic map was constructed with the 384 SSRs and EST-SSRs only (data not shown); it accounted for 93.8% (730.6 cM) of the total length of the map.

EST-SSR markers were distributed throughout all linkage groups, but their distribution was not random. The number of EST-SSRs per group ranged from two in LG7 to 21 in LG9. The average spacing between EST-SSRs was heterogeneous, ranging from 4.3 cM for LG9 to 8.3 cM for

**Table 2** Characteristics of 115 EST-SSR loci mapped on the reference map

Marker name	EST accession number	LG location	Forward primer (5'-3')	Reverse primer (5'-3')	UPA402 alleles	UF676 alleles	Allele number	SSR type	SSR Location	Putative gene function
mTeCIR 292	CU535239	2	TCCCCACGCAACTACAA	CTCTTCCACACCACCA	245/245	245/249	3	(TC)12	5'UTR	Defense-related protein
mTeCIR 293	CU481076	9	GAAAGCCATATTGATGCT	CTAATTCACACTCAAITCCA	268/270	260/272	4	(TA)8	3'UTR	Pathogenesis-related protein
mTeCIR 294	CU607912	9	GGGAGACACAGAGAGCTA	GCCACTTCTCCATCGT	116/116	112/116	4	(AG)14	5'UTR	Protein kinase
mTeCIR 297	CU533717	9	AAACGAGATTTTCGAAGCCA	TAGGGACATGACGGGAGAAAG	275/275	272/275	2	(GGT)5	3'UTR	Oleosin low molecular weight isoform
mTeCIR 298	CU598057	5	CAGCGATTATTCTAGGGA	AACACTGACAGCGAAAA	226/228	228/228	3	(ATT)6	3'UTR	Aspartyl protease family protein
mTeCIR 299	CU543634	6	CCTTTTGTTCCCTTCATTTT	CCGGCTTCTCTCTCC	141/144	141/144	2	(TGT)5	5'UTR	Transducin family protein wd-40 repeat family protein
mTeCIR 301	CU472276	6	AAAFCTGCCTCTCTTCTTTT	AGGTTGGTTTAAAGTGGGG	98/98	98/107	4	(AG)9	3'UTR	Leucine-rich receptor-like protein kinase
mTeCIR 302	CU494939	1	CTGTAACCAACCCTAAAGAAAA	AGTTGAATGTTGGCCCTT	259/279	281/301	6	(AD)8	3'UTR	Calcium-dependent protein kinase CP4
mTeCIR 303	CU584157	8	GGCTTCCAGGCTTTACC	CCACTCCGAAATCCCT	258/258	258/261	3	(TGA)7	5'UTR	Serine threonine protein
mTeCIR 304	CU597590	1	GTGCTGGTGTGAAGGA	GGCTTCGAGGAAAGGT	221/221	219/221	4	(TA)8	CDS	7S vicilin
mTeCIR 305	CU531763	9	TTCGCTTCTTCTTGCT	GGGACGACGAAACC	173/181	173/177	4	(CT)8	5'UTR	Chloroplast acyl-ACP thioesterase
mTeCIR 306	CU596449	9	AGACGACTGAGATTGCC	TAGCCAGCGATGAAGATG	272/272	268/272	2	(CTT)6	5'UTR	3-ketoacyl-ACP synthase
mTeCIR 308	CU602588	9	GAGGCATGTTACACAGCA	GACGCTTGACCTTTGGA	237/249	247/249	5	(AG)7	5'UTR	SNARE protein
mTeCIR 309	CU596405	4	CGCATGAGTTTCGGTTT	ATGATCGACTCCCTCC	185/197	185/185	3	(GAA)6	CDS	Glutelin
mTeCIR 310	CU573228	8	GGGGATACTTGACCTCTGTT	GCTGGACCTAAACGCA	174/174	166/174	4	(CT)12	5'UTR	24-sterol C-methyltransferase
mTeCIR 312	CU477743	4	GCAACGCAACATTTCTCC	ACGCAAAAACAAAACACA	147/147	145/147	2	(TC)6	5'UTR	UDP-glucose pyrophosphorylase
mTeCIR 314	CU494559	1	GGATCTTACAGTCCCAAAA	ATCCAGGTAGAGGAGAAATG	249/259	249/257	6	(AT)7N(AD)6	3'UTR	MADS-box protein
mTeCIR 315	CU482694	6	CCAAATGTTTACAAAGGG	TCATGCAGCAACAAGAGA	260/260	250/260	3	(GA)11(CA)7	3'UTR	Nucleic acid binding
mTeCIR 316	CU544991	3	AAAAGCATCTCTCCCA	AGCTGTCCATGCTCCTT	141/143	139/143	4	(CT)15	5'UTR	myb-like transcription factor 2
mTeCIR 317	CU530579	9	GGTCCGCCCTCTCTGCTA	CCAFTTCGTTTTCTCCGAA	252/252	252/256	4	(CT)10	5'UTR	Transcription factor IIa large subunit
mTeCIR 318	CU495912	2	TTCGGGGTTCACCTGG	CAGCATCGCAAGACA	195/207	191/191	4	(AG)9	5'UTR	Floral homeotic protein apetalal
mTeCIR 319	CU532243	9	TCTATTCTAFTGATGGTGTTC	CTGCTTTTGTGTTGTGG	204/207	207/207	3	(AGA)6	5'UTR	bZIP transcription factor
mTeCIR 320	CU596254	9	AGCAACACTGCCGCTAC	CCCCATGCTCATATCCA	157/169	153/157	4	(TC)8	5'UTR	AP2/EREBP transcription factor
mTeCIR 322	CU584757	5	TTACTGTTTCTCACCACTGTTA	CTGCTTCAITGCAITCT	275/289	265/295	5	(CT)12N(TC)8N(TG)9	5'UTR	Phantastica transcription factor A
mTeCIR 324	CU542017	3	CGAAACTCTCTTCTTCGCT	GGCAGTGGGTTGGTTG	251/251	249/251	2	(CT)6	5'UTR	Scarecrow-like transcription factor 8
mTeCIR 325	CU492823	6	TCATCACCAACCATGACC	TGGTAAACCCCTCCCAATC	197/206	206/212	5	(ACC)6	5'UTR	Urease accessory protein G
mTeCIR 326	CU594176	5	ACACCATAAATAAATACTCCCTC	CITGGCGTGTGTTGCT	204/204	188/198	5	(TC)9	3'UTR	Phosphate phosphoenolpyruvate translocator-like protein
mTeCIR 327	CU627045	7	TTTTGAGTGGAAACAATGGC	CCGATCCAAAAGCAAAAG	153/163	153/153	3	(TAA)8	5'UTR	Phosphoenolpyruvate carboxylase
mTeCIR 329	CU518760	8	AACCTAAGAAAAACGCAAGG	TGGTCAACAATGGGCTT	259/259	263/273	3	(AD)8	3'UTR	Lysine and histidine specific transporter protein
mTeCIR 331	CU530671	1	CCCCTCCATTGGTTTC	CATGACCTCCGCTGT	224/224	221/224	3	(CT)12	5'UTR	Homeodomain-leucine zipper protein 56
mTeCIR 333	CU571479	1	CAGCACCAACCCTGAAAA	GGTTTCGGAGAAATGGG	121/121	121/123	2	(CAC)5	5'UTR	Nucleotide sugar epimerase-like protein
mTeCIR 336	CU536100	3	AGTGGAGGAACAGTATGGC	TAAACCGTGTCCACCAACA	157/157	151/165	4	(TCC)7	CDS	Plastid-lipid associated protein
mTeCIR 337	CU541857	6	ACGAAAGCCGTAACCTTGG	TGCAGGACTCTCTGTCACT	212/212	212/215	3	(GGA)5	CDS	Fibrillin precursor-like protein

mTeCIR 339	CU494243	9	CCAAAGCTCCAAGGAAAC	GTCCCAAAGTTCTCCCAA	152/155	152/152	3	(TC)9	5'UTR	GL1 protein
mTeCIR 341	FP101929	1	CCGCCGTCTTCTTCTT	TTTTACCTGGGGTTGGG	258/264	258/258	2	(CCA)6	5'UTR	Guanine nucleotide exchange factor P532
mTeCIR 342	CU480025	1	CCGGAAAACATTGAACAC	TGGGCTCACCCCTTCT	220/220	220/236	4	(CT)9	5'UTR	3-methyl-2-oxobutanoate hydroxymethyltransferase
mTeCIR 343	CU584671	4	GC TTG C C C T T T C T C T C T	AGCACTGAACCCGAGCAA	265/267	263/267	4	(AD)10	5'UTR	Protein arginine N-methyltransferase family protein
mTeCIR 344	CU483060	4	GTGCAGAAAAGCAGGG	GGGAGATGGAAGGGAA	240/242	238/242	4	(AT)9	5'UTR	Late embryogenesis-abundant protein
mTeCIR 348	CU617697	8	TGTGGCATCTCAATCC	GGAGCGAAGAAAGAGG	262/262	256/262	3	(TCT)6	CDS	One-helix protein
mTeCIR 349	CU572974	9	CATGGGAGTTGAAAAAGAA	TTGGATTCGAGGGTGG	242/247	246/247	4	(AGA)5	3'UTR	Arabinogalactan protein 2
mTeCIR 350	CU535201	1	CGACGACGACAACAACA	AAGGCAGAGGCTGAGGT	204/210	204/210	3	(CAA)7	CDS	Unnamed protein product ( <i>G. hirsutum</i> SSR)
mTeCIR 351	CU603065	5	CCAAAAGAAACAGGAAAGCAA	GAAGTTAGGGCATGGG	269/269	263/275	4	(AC)10	3'UTR	Lipoxygenase
mTeCIR 352	FP101931	3	CGATTCTTCGGHTTCCC	CCATCTGTTGCGGTTGT	156/164	160/162	5	(GA)10	5'UTR	Unnamed protein product ( <i>G. hirsutum</i> SSR)
mTeCIR 353	CU482102	3	TGAGTTGAATCACATTTGGCT	TGTTGGACGGGTTGTTT	250/250	247/250	2	(TGT)6	3'UTR	Progesterone-binding protein homolog
mTeCIR 354	CU597304	1	AACCCCGAAAAGCACATCT	TTAAGGTCGCTCCACGGTT	274/278	274/276	4	(CT)10N(CAA)6	5'UTR	Tyrosine specific protein phosphatase family protein
mTeCIR 355	CU593782	4	TCTGGTGAAGAGCGGAAA	CAAAGAAAACGGCCAAA	190/194	194/198	5	(AGA)5	5'UTR	Glycosyl transferase family 17 protein
mTeCIR 356	CU576411	1	GGGGTTGCTTCTGCT	AAAACCTGGCGAGGAG	168/168	166/168	2	(TC)6	5'UTR	Auxin-induced protein IAA9
mTeCIR 358	CU604521	5	TTGTGCTTGTGCTATGGGA	TTTTTGTAAATTAATCCCAACTTT	228/228	231/236	3	(TA)7	CDS	Storage protein (late embryogenesis abundant)
mTeCIR 359	CU546177	4	TCGAGATACGCAACGAA	TGGACATTCGCAAAACC	188/190	187/188	3	(AG)7	3'UTR	Amia calva GARS-AIRS-GART
mTeCIR 360	CU476689	2	GGAAAAGGAAAAGGAAAC	TTGCTCAGTGGGAGAA	122/122	116/122	4	(TC)7	5'UTR	Phosphatidylserine synthase
mTeCIR 361	CU476829	2	CCCTTCTGCTCTCTGCT	GGGTGGAACCTGAAACC	175/175	175/190	4	(AGG)8	5'UTR	Nac domain protein
mTeCIR 363	CU585497	9	CCCAGCCCTCTAAACCT	GCACCCAAATCAAAGGA	162/162	162/165	5	(TC)8N(TCT)7	5'UTR	Ubiquitin-conjugating enzyme
mTeCIR 364	CU585945	5	CGCCGATCTCTCTCCT	CAGTCTCCGACGAACCA	203/203	201/203	2	(GA)9	5'UTR	Importin beta
mTeCIR 365	CU593133	3	AAGGTTGCCAGAGAGGG	GGTTTGGGAGGGATGAG	167/167	165/167	2	(TC)8	3'UTR	Cyclic nucleotide-gated calmodulin-binding ion channel
mTeCIR 366	CU593926	2	GGGAAAACAACAGCAAAA	TTTTGGAGGCTTGAGG	193/193	193/197	4	(TC)8	5'UTR	Calcium-dependent protein kinase
mTeCIR 367	CU601277	6	GGAAAACCTCAATCTCTCTTC	TGAGTTGTGGAGGTGGA	132/132	132/134	2	(TC)11	5'UTR	Gibberellin 3 beta-hydroxylase
mTeCIR 369	CU601643	3	GGGAAAGCAACTGAAAGA	TTCCCTTGCTAAACC	275/277	253/263	6	(GA)10	5'UTR	NAC2-like protein
mTeCIR 373	CU607930	8	CAGGTACACTCATGGACTTTT	CAAAATGGAAGGATGGG	262/262	250/260	3	(TC)11	3'UTR	NAC domain protein
mTeCIR 374	CU608928	6	GGGAGTCTCTTCCCTTCC	ACTCGGAACCCTTAGCAC	122/122	122/124	4	(CT)9	5'UTR	Heat shock protein binding
mTeCIR 375	CU530455	1	GGGGTCCGACAGTGAGTT	TCACATCGGCTGACTTCTTG	306/306	296/306	3	(TC)12	5'UTR	Immunophilin
mTeCIR 376	CU530464	1	TTAGAACAATTTGAGGGGG	TCTCAGGGCAGCACTTCTT	199/201	199/203	4	(GA)9	3'UTR	ABC transporter family protein
mTeCIR 378	CU531696	9	TGGCTTCTGTTCTTGG	TCTCAGCCTGCAAAATCC	114/120	108/114	4	(GA)8	3'UTR	Plastocyanin-like domain-containing protein
mTeCIR 379	CU534268	2	CACCCCGATCTCTCCTC	CCGATTTCACTTTGCTGTT	143/143	139/143	2	(AG)10	5'UTR	RING-H2 finger protein RHG1a
mTeCIR 380	CU532201	4	CCCCGGACACCAAAG	CTGAGCTAGTGGCGGTG	193/195	193/195	4	(CT)10	5'UTR	Gluthathione reductase
mTeCIR 382	CU534930	8	AAAGAGCGAAAAGCAGGG	CCACACCAAGGAGCGGT	246/246	236/248	4	(AG)16	5'UTR	Protein disulfide isomerase
mTeCIR 383	CU534255	10	GGGGATTCTCGCT	TGTAGCTTGGGATGGTTG	159/159	149/159	4	(AG)10	5'UTR	Yippee-like protein
mTeCIR 384	CU535362	5	TCCGGTGTCTTTCATCGT	GGGGAAATGGGAAAAGA	235/235	229/239	4	(CT)7N(AGC)6	5'UTR	RAN binding protein

Table 2 (continued)

Marker name	EST accession number	LG location	Forward primer (5'-3')	Reverse primer (5'-3')	UPA402 alleles	UF676 alleles	Allele number	SSR type	SSR Location	Putative gene function
mTeCIR 387	CU472686	6	CATGACCAATTGCTTTTCAACTCT	AGCTGCCCGCGGTTTT	220/220	220/224	2	(GCT)7	CDS	STI ATP binding DNA binding DNA-directed DNA polymerase
mTeCIR 388	CU472694	10	CTCCCGAAATTCCTC	TTGGGTTCCAGTCAAAGAG	271/271	257/270	5	(AT)8	5'UTR	3-phosphoinositide-dependent protein kinase-1
mTeCIR 390	CU624659	5	TTGACATCTTGTAGCGGG	GGCCCAAAAATAAGGGAG	194/194	188/194	2	(AG)8	CDS	Villin 2
mTeCIR 391	CU629602	8	GAAAGCTGGCGGTGAA	CGGCTGGCTGTCTGT	201/201	201/210	5	(CTT)11	CDS	heat shock protein
mTeCIR 392	CU619851	3	GCAGCAACGAGTGTTCGTG	GGGGCCGTTATCTTTT	191/191	185/191	4	(CT)13N(AG)6	5'UTR	NHL repeat-containing
mTeCIR 393	CU538185	9	CGTCTCTTTTCTGTTCACAIT	CGAAAAGAAAAGGCAATCCA	167/171	157/165	5	(CT)11	5'UTR	Inorganic pyrophosphatase
mTeCIR 394	CU481769	4	GGCCAACCCAATAACA	GCACCATCACCCACAA	132/134	136/144	4	(GA)14N(TGG)5	5'UTR	Plastocyanin-like domain-containing protein
mTeCIR 397	CU593725	7	CGACCATACGACGAAGA	ATTAAACCGGTTTCGATG	267/267	267/270	2	(GAA)6	5'UTR	Outer membrane OMP85 family protein
mTeCIR 398	CU593124	6	CTCTTCTGTCTCGCA	AGTTCCTTTCGCTTTT	344/344	344/356	2	(AAG)6	CDS	Monocarboxylic acid transporters
mTeCIR 400	CU576035	9	TCAAAAACGGGAAACAGA	GTGTCCCGTTGTTTGGT	279/279	277/279	4	(ATT)5	3'UTR	CALM_SPIOL calmodulin (CaM)
mTeCIR 402	CU549279	4	AACAGAGATCGTACAACAAA	GCGGATTGAAACACAAA	258/258	258/264	2	(GAA)5	3'UTR	MADS-box protein BM5A
mTeCIR 403	CU549287	8	TCACGGTCACACACCAC	TTCGAGCTTGAATCGT	279/279	279/282	2	(CCA)7	CDS	MADS-box interactor-like
mTeCIR 404	CU549399	5	TTCCTTCCAAAAGCCCTC	CGAGGCTGCGGATTT	180/180	180/183	2	(AAT)6	3'UTR	Transcription regulator
mTeCIR 405	CU549899	3	CCGACAAATTCGCTTC	TGCGGTCTGTAACCTTCT	143/143	143/146	3	(GGA)5	5'UTR	Zinc finger protein OBP4
mTeCIR 406	CU548113	1	GCTTTCAAAAATGGCATGAA	AAGCATGTAAGGGGCAATG	313/313	299/315	3	(AG)8	3'UTR	Rab GTPase activator
mTeCIR 408	CU616178	5	TCCACAAATFACATCAITAC	GTTTTGGATTTCTGGGTT	244/247	244/250	5	(CAT)8	3'UTR	Zinc finger (C2H2 type) family protein
mTeCIR 409	CU618894	9	CGGGATCGAGTGAAA	TGCTGAAAGTTGGTGGCT	160/167	160/167	2	(TCT)10	5'UTR	Nucleoid DNA-binding protein CND41
mTeCIR 410	CU619198	3	TTTTGCCTCCCTTGTCT	TGAAATCGTTGAGCGAAAG	260/260	250/258	3	(TA)10	5'UTR	N-rich protein
mTeCIR 411	CU619479	2	TGCAGTTTCCCAACC	GCACCACCCAAATGCT	301/303	289/295	4	(AG)9	3'UTR	Somatic embryogenesis receptor kinase
mTeCIR 412	CU620454	1	GATGTTGATTCCTGTGCCT	ACACCTCAGCTCGGACA	110/118	114/118	4	(GA)7	5'UTR	Sm-D1 protein
mTeCIR 413	CU583866	6	GCCCTCCGAAAATCCT	CAACCACGAAATCTCCCTT	232/232	232/234	4	(CT)10	5'UTR	Aldose 1-epimerase
mTeCIR 414	CU595542	2	TGGATCACATCCCATCAC	TGACGAAAAGTGCAAGGAG	240/240	236/240	4	(TA)11	CDS	CCCH-type zinc finger transcription factor
mTeCIR 415	CU600680	3	CGGGACTATCCAACTCT	CGACCAGAAAAACCAAG	99/107	99/109	3	(TC)13	5'UTR	Glycosyl hydrolase family 1 protein
mTeCIR 416	CU571882	1	CCTCCAAATCAAGCCC	GTTCCGGTTCTGGGGAG	233/233	233/250	4	(CAG)10	CDS	AP2 domain containing protein
mTeCIR 418	CU529232	6	ATCCAATGCCAAACCTACA	CGAAGCGATGAGCAGG	276/276	273/276	3	(AAT)7	3'UTR	Transcription factor
mTeCIR 419	CU547762	1	ATTTCCTTCTCGCTC	CAGCAGCACGCGCTAAAC	244/244	244/250	2	(CCG)8	5'UTR	Branched-chain amino acid aminotransferase
mTeCIR 420	CU547857	5	GGGCTTTCCTTCTTTGG	GATCAACGGCTGACTCTATG	274/274	274/283	3	(AAG)7	5'UTR	Nucleic acid binding
mTeCIR 421	CU549693	2	TTCAAATCTCTTCCCTCTC	TCAATCGCTTGAITGGCT	134/134	130/134	3	(TC)7	5'UTR	Mitotic checkpoint protein
mTeCIR 422	CU550190	1	ACATCTTTTCTCTGCCTTT	CCCTTCCCATCCCTCT	223/227	209/213	9	(GA)18	5'UTR	MYB transcription factor MYB92
mTeCIR 423	CU628249	2	TTTGCCAGGAATCTGTTG	CCTTGGGTGGGCTTC	286/286	274/282	6	(GA)11	3'UTR	Receptor like protein
mTeCIR 424	CU628687	9	TCTACCAGTCAAAAATGCC	CTTCCCTCATCCCAAC	227/237	237/237	4	(GA)8	3'UTR	CGS1 mRNA stability 1
mTeCIR 425	CU631787	9	TGGCCGAAGTCATAAACA	TCGAGGAGGAGGAGAGA	162/168	162/168	4	(AAC)14	5'UTR	Serrate transcription factor
mTeCIR 426	CU531204	1	CTTCCCTTCTCCGAAAC	CCCAGAAAGCTGGAA	244/244	239/244	3	(CAG)9	CDS	ATP binding/ATP-dependent helicase
mTeCIR 428	CU539536	8	TGGCTGGTCCCTACCTT	AGACTTGCCCGGTGTTT	268/268	262/268	4	(TCT)10	3'UTR	Small heat shock protein



mTeCIR 429	CU593925	9	GGGGATACATCTTCTGCTCT	TTTATTGGTTTCCCGTCC	169/175	167/179	5	(CT)12	5'UTR	Dihydroflavonol reductase
mTeCIR 430	CU592712	2	CGAAGAAAGGAAGTAGAATAAG	CGTGGCTTCGATGAGTT	218/218	206/208	5	(AG)10	3'UTR	Lipase-like protein
mTeCIR 431	CU592928	3	GCTGCCAATCGTATCCC	TTGGTGAAGCCCTTGGT	236/236	234/236	3	(GA)10	3'UTR	Protein kinase
mTeCIR 432	CU519970	8	GGGAAACTTGA AACCAAA	CGGTGGTGGCTCTTTC	220/220	220/226	4	(CT)8	5'UTR	Cytosolic factor
mTeCIR 433	CU518308	10	AGCATTTGCCTTCGCTT	TGAAACAGTCTGCGCA	356/356	365/393	5	(ATG)11	3'UTR	Pathogen-induced calmodulin-binding protein
mTeCIR 434	CU520113	2	TGCAAGCAITGTCAAGTAT	GAATAAGAGGGGGCGTT	241/247	241/249	4	(GA)12	5'UTR	Glyoxalase extradiol ring-cleavage dioxygenase
mTeCIR 436	CU586512	6	GATCCGAGGAGTTTGGG	TTTTAGGGAGTGGTGGGA	137/137	133/137	2	(CT)12	5'UTR	Translation initiation factor (eIF-4A)
mTeCIR 438	CU531891	3	TCGTAAGGGTACAAACGAA	CCGCTTGCTCTCCAC	118/118	118/122	2	(GA)7	5'UTR	Luminal binding protein
mTeCIR 439	FP101930	6	CCGATCTGGAGGACTTG	GTTGCCTCGAATTGACC	99/102	102/102	5	(CTT)7	5'UTR	WRKY family transcription factor
mTeCIR 441	CU530607	3	GGCTGTGATTTGCGTTTCTT	CTGATACCCAGTATCGTCAA	125/125	127/133	4	(GT)8	3'UTR	Nitrilase-associated protein
mTeCIR 444	CU546233	8	TGAACCGGATTGTTGGA	GGGACTTAATCTGGACATGC	121/121	113/121	3	(TC)15	5'UTR	Trypanothione-dependent peroxidase
mTeCIR 445	CU600969	9	GTCCCTTCTCCACTCT	CCAGAATCGCAATCAA	364/392	358/364	5	(AG)8	5'UTR	GTP-binding protein RAB11
mTeCIR 446	CU601147	1	TCTCCAAACCCTAGCCC	CATCGTTATCACTGCCCTT	123/123	123/125	4	(CT)8	5'UTR	Heterogeneous nuclearrribonucleoprotein A2
mTeCIR 447	CU571514	9	GGGGCCTCTCAGATCAA	CTGCCAAGTGGTTGTGG	213/213	213/216	3	(CAG)7	5'UTR	NPH4 transcription factor

LG2. LG7 and LG10 were very poor in EST-SSRs, the average spacing between two EST-SSRs in these groups being more than 24 cM. These two chromosomes were also those that incorporated fewer markers (32 and 33 markers, respectively).

In many plants, the level of polymorphism of EST-SSRs has been reported to be lower than that of genomic SSRs (Cho et al. 2000; Eujayl et al. 2001). In our study, the polymorphism revealed by the EST-SSRs (55.4% of loci) was similar to that found for genomic SSRs (52%). EST-SSR loci represented from 14% to 30% (with an average of 19.6%) of total loci on chromosomes. Only two chromosomes (LG7 and LG10) contained a very small proportion of EST-SSRs (6.3% and 9.1% of all markers respectively). One hypothesis explaining this situation would be that these chromosomes are less rich than the rest of the genome in coding regions or they have large homozygous regions that we cannot detect on the genetic map due to a lack of polymorphism.

The addition of new markers enabled us to increase the density of the previous map, but did not lead to the saturation of LG10. The non-homogeneous distribution of the polymorphism on the chromosomes can be explained by the dynamics of domestication (Pugh et al. 2004). Such selection can fix some segments of chromosome (Temnykh et al. 2000). LG8 was an exception because all the markers were mapped from UF676 only. The explanation is probably linked to the UPA402 pedigree. Indeed, UPA402 is an Upper-Amazon Forastero clone derived from a full-sib cross and probably includes several homozygous chromosome regions.

### Segregation distortion

Of the 115 EST-SSR loci mapped, 88.5% followed the Mendelian ratio expected from the genotypes of the parents. A total of 67 markers (11.5%), 19 RFLPs, two isoenzymes, one R gene, 24 SSRs, and 21 newly included EST-SSRs, significantly deviated from the expected ratio at  $P < 0.05$ .

After calculation of the chi-square test at  $P < 0.05$ , several segregation distortion regions (SDRs) appeared on the genetic map. An examination of the direction of segregation distortion showed that the loci were skewed for the male parent (UF 676) only, or for the female parent (UPA402) only.

A large proportion of the EST-SSR loci (18.3%) displayed distortion when compared with the expected Mendelian ratio. These new loci were twice as distorted as SSRs from genomic libraries (8.9%) and significantly more distorted than all other markers (11.4%). All these EST-SSR loci were clustered in SDRs in the genome. The genetic map constructed with new markers highlighted all the SDRs found on the previous map (Pugh et al. 2004). There

**Table 3** Characteristics of 59 unmapped EST-SSR loci but revealing polymorphism on eight genotypes

Marker name	EST accession number	Forward primer (5'-3')	Reverse primer (5'-3')	Expected size	Allele number	SSR type	SSR location	Putative gene function
mTeCIR 300	CU473134	CAAAGAGCTGCAATACAAAGAA	CCTGACCCCAAAATCACA	274	4	(CT)10	3'UTR	Protein kinase-like
mTeCIR 307	FP101932	GCGCTTTGCTGTAGTGG	AAACACCCGACCCAATC	254	2	(CT)7	5'UTR	Dihydrolipoamide acetyltransferase
mTeCIR 313	CU598258	GGATGGCAACAAGCAAG	AGGGGACAAAAGGCAAC	241	4	(AT)12	No protein	YABBY2-like transcription factor YAB2
mTeCIR 321	CU596170	CGCAAACACCAATAAGAGAA	GAACCGAGCGAGTCCA	136	3	(GTG)7	5'UTR	Basic helix-loop-helix family protein
mTeCIR 328	CU610642	GGCAAGTACGCTCAGTATC	CCTCAATGTGCTGCGT	122	5	(GA)11	3'UTR	Storage protein
mTeCIR 330	CU609095	CGTAAAGAAATGAGGATGAGG	AGAGAGCTTTGCGGTTG	129	4	(AT)10	3'UTR	Lysyl-tRNA synthetase
mTeCIR 334	CU481680	AATACCAGCAGGGGTTTG	TGGGTACCTGTTTGG	167	2	(GCAGGA)5	CDS	Nitrilase-associated protein
mTeCIR 335	CU495748	TTCCCTTGTGATGTTGTTGG	CAAGATTTGAGGGGAAA	185	2	(CT)9	No protein	GPI-anchored protein
mTeCIR 338	CU571572	TTGCCTGGGGTTTCTT	TGTGGGTCAAAGGGTGAA	158	4	(CT)8	5'UTR	Protease inhibitor seed storage lipid transfer protein family protein
mTeCIR 340	CU475541	GGACTTTGCCAATTTCTC	AGCGAACCGACCTTTG	114	3	(TCT)6	CDS	Mature anther-specific protein LAT61
mTeCIR 346	CU471351	AAAAGAGAAGAAAGCGAAG	ACTCACCGAAGAAATCCA	249	3	(TC)7	5'UTR	Serine threonine protein phosphatase pp2a-2 catalytic subunit
mTeCIR 347	CU591918	GCACCTTTTTCGCCTCT	CTGTGGCGGTTGTTGTT	267	3	(CAA)5	CDS	vq motif-containing protein
mTeCIR 357	CU572691	CCAAAACACCTTGCCTC	CAACATAGCAATACTAAGCAGA	270	5	(AT)8	3'UTR	Glutathione S-transferase GST 18
mTeCIR 371	CU604481	TTGCAAGGACCGGTTTTTGAT	GTTTTTCTCGGGCCAATTTT	218	3	(GA)8	5'UTR	Eukaryotic translation initiation factor 5A isoform iv
mTeCIR 372	CU607927	TGGTACACTGGGGAAGG	CAAGCAGTATATCGAGGCA	229	2	(TCC)8	3'UTR	Tubulin beta chain variant 1
mTeCIR 381	CU532791	GGGGTTCACATTAGGCA	ACGACGGAGTCAAGGG	129	3	(GA)13	5'UTR	Protein transport protein SEC61 gamma
mTeCIR 385	CU531665	ACTCGCTCTCTTGGCT	ATGGAGCCTTCTGTATAGAG	247	4	(CT)14	CDS	Nuclear transport factor 2
mTeCIR 386	CU477439	TGCGATTGATGGTATTGG	GACCAATTAGCAAACCGAA	181	3	(CTT)8	5'UTR	Cadmium-induced protein
mTeCIR 389	CU472871	CGGCGACAAAGCAAAA	CCTTCCCGATCCATCTC	194	5	(GA)9	5'UTR	Dnaa-1 heat shock n-terminal domain-containing protein
mTeCIR 395	CU540724	TCCCCTTTTCCATTTCC	GTTTTGCCAAACAGCACCC	188	3	(CTC)5	5'UTR	RAS-related small GTP-binding protein
mTeCIR 399	CU600396	GGTGCCTCATCTTTC	CCCAAGCTAATCTCCC	227	2	(AGAA)6	5'UTR	Pyrrhline-5-carboxylate synthetase
mTeCIR 435	CU536268	ACAACCAAAAGGCCAA	CTTCCAGTTTTGGTGGT	163	3	(TC)6N(AT)11	5'UTR	MADS-box protein
mTeCIR 437	CU586127	TGAGCAACACAGCCA	CCCTTATTAACCAACCCA	135	4	(AT)8	3'UTR	Basic PR-1 protein precursor
mTeCIR 448	CU541384	CCCCACCTCACCA	ATGGAATTTGTGGTTGGA	247	2	(CT)9N(TTTCT)6	CDS	COPI
mTeCIR 449	CU548134	TTGAAATCAAATGCAAGGCAG	CAGCAGGCCAATTTGTAGA	182	3	(GTAT)5	3'UTR	Glycine hydroxymethyltransferase
mTeCIR 450	CU476693	CAGCTTCCAAAAGTGTC	GGGTATAGTATCTTCCCGA	241	5	(AT)9	5'UTR	Ribulose biphosphate carboxylase oxygenase small subunit
mTeCIR 451	CU475493	CACCCGCGATTATTITG	TACAGGGGAGGGAAA	132	2	(TCT)5	5'UTR	Protein phosphatase 1
mTeCIR 452	CU476755	TAACACAAAATGCGGGAA	ACAGACTCCAGACTAGAACACA	152	2	(GT)8	3'UTR	Unknown protein
mTeCIR 453	CU533740	TGAGCAAAAATAAGAGGATGAGA	CCTCTCTCTTTCAACTTCCC	207	3	(GA)6	3'UTR	Plasma membrane intrinsic protein
mTeCIR 454	FP101941	TGGGCGTTTGTCTATTCT	TCAGAAAGGAGCAGGAACA	104	2	(AG)10	5'UTR	CA2+-dependent membrane-binding protein annexin
mTeCIR 455	CU530864	TCCGCTTACACATTTT	AACCCAGTGTGAGAGAAG	168	2	(TCT)5	5'UTR	Transaldolase-like protein
mTeCIR 456	CU584414	AAACAGAAAGAAAACAAGGG	TCCGTCGAGATTGCCT	158	2	(GAG)6	5'UTR	Unknown protein
mTeCIR 457	CU544452	ATGCCCTGACCCCTTTTC	GCAACCTTTCCCAGCTAC	151	2	(AGA)5	3'UTR	Bax inhibitor-1

mTeCIR 458	CU475752	CTCGTACAAATTTCTAGGGTTTT	TCGACTGGGGAAAGGT	117	2	(TC)6	5'UTR	Unknown protein
mTeCIR 459	CU585233	AACAGCCAAAAGGAGGG	TGTGTTCTCGCTTGCT	224	2	(AT)8	3'UTR	Type-II NADH dehydrogenase
mTeCIR 460	CU608049	CTCATCTTCAAACCCAAA	ATCGGATGCCAAGAAAAC	216	3	(TTC)8	3'UTR	Ethphon-induced protein
mTeCIR 461	CU533217	GGGGATTCATAITTTCTGTCT	GGAGACGAGCAGAGGTG	178	2	(CT)9	5'UTR	TPA: putative phytosulfokine peptide precursor
mTeCIR 462	CU534235	TTTCTGGCTCTGCTTGG	GACGAGGAGGAAGTGAA	190	2	(AGA)7	CDS	Dehydrin 4
mTeCIR 463	CU615262	CCGAGGAGTCTGGGAAA	GGGAGTGGTGGATTTG	123	2	(CT)8	5'UTR	Translation initiation factor (eIF-4A)
mTeCIR 464	CU493934	TGCAAGTTGAAAACCCCA	GGCGAAGCAACAGTCAA	244	2	(TGC)7	CDS	Ankyrin repeat family protein
mTeCIR 465	CU619827	CGTCTTCTCTCGCAATGTT	AGTGATGCAGGAGGATTTGG	250	2	(CTG)8	5'UTR	1-interacting protein 1
mTeCIR 466	CU481417	AGGGCAGAAAATTGGAG	CTGTGTGTCGCGAGGAG	238	2	(AG)8	5'UTR	CA2+-transporting ATPase-like protein
mTeCIR 467	CU574086	TCTCCCAACGAAAACAA	CCTGGCAAGAGATGTGG	170	2	(CTC)7	CDS	Omithine carbamoyltransferase
mTeCIR 468	CU574287	GGGGCAGCCTAAGAGTAA	GGCAAAGGCAAAAGGGT	259	2	(AG)7	5'UTR	Nodulin-like protein
mTeCIR 469	CU601978	TCCTCTCTTTCCCTTTCCCT	TCTCTCAAAGCTCTTCCAC	256	2	(AG)7	5'UTR	Growth-regulating factor
mTeCIR 470	CU535860	CGGGGCTCTCTGTCTGT	CCCCGTTTTCTTCATCTTC	232	5	(TC)7	No protein	Spermidine synthase
mTeCIR 471	CU536493	TGCAATGGGAGAGGAGA	TCGGAAAGTTTGGGGAA	210	2	(CT)9	3'UTR	Unknown protein
mTeCIR 472	CU537422	AAGAGGAATGGGCGGT	AGTTGGAGGGCGAAAAG	178	2	(GCA)8	5'UTR	1-interacting protein 1
mTeCIR 473	CU531228	TTTTCCGCAATATGCT	CCTTGCCATGTAATCAA	206	2	(TA)7	3'UTR	Auxin-responsive family protein
mTeCIR 474	CU518331	GGCAAAAACAAGAGCAACA	TGCTTGAATTTGGCCTC	199	4	(AAG)8	5'UTR	Heat shock protein
mTeCIR 475	CU572781	GCATGACCACCCACCATC	ACCAAAAACCCGAGAGGA	184	2	(AGA)6	5'UTR	IAA-alanine resistance protein
mTeCIR 476	CU537953	GGGGAACGCAACTAAAAC	TGCTTGAATTTGGCCTC	273	4	(AAG)5	5'UTR	Heat shock protein
mTeCIR 477	CU550551	CGGTTCTGTGTGGT	TGCAAAACCCTAACCCGAA	195	2	(CAG)5	5'UTR	Poly-binding protein
mTeCIR 478	CU547742	CGTAAACAGCCGCCAA	TGGGTTTGCATTGAGG	267	2	(AGC)5	5'UTR	Calmodulin-like protein
mTeCIR 479	CU547880	CGGGATCTCTACTGATACAA	ATCAAGCACTTCACCCA	223	2	(AAG)6	5'UTR	S-adenosyl-methyltransferase
mTeCIR 480	CU535237	TTGACCCAGCAAAACGA	TCAAAACAGCTTGGCCTC	166	2	(TTA)8	3'UTR	S-adenosyl-L-methionine: beta-alanine N-methyltransferase
mTeCIR 481	CU474951	CCTATTACTATCTGCTTTCGGTT	ATAGGCTGCCCTGCTAAA	191	2	(CT)6	No protein	GDP-mannose pyrophosphorylase
mTeCIR 482	CU576143	GAAGAACAAGACAGGGCA	AGAGGAGGTGGGATTTGG	216	2	(TC)6N(ATC)5	5'UTR	LIM domain-containing protein
mTeCIR 483	CU479288	CAAGGGTTTTTGGAGAGGA	GAGAGGCTTGGAGGGAA	240	2	(CT)11	5'UTR	Transcription factor

**Table 4** Distribution and polymorphism of different types of SSR repeats depending on their position in the gene (5'UTR, CDS, and 3'UTR)

Type of SSR repeats	EST region						Total	Total polymorphic
	5'UTR		CDS		3'UTR			
	Total	Polymorphic	Total	Polymorphic	Total	Polymorphic		
Dinucleotide	83	63 (75.9%)	6	5 (83.3%)	35	29 (82.9%)	124	97 (78.2%)
Trinucleotide	55	32 (58.2%)	35	16 (45.7%)	18	15 (83.3%)	108	63 (58.3%)
Others	13	9 (69.2%)	2	2 (100%)	4	3 (75%)	19	14 (73.7%)
Total	151	104 (68.9%)	43	23 (53.5%)	57	47 (82.5%)	251	174 (69.3%)

The different classes of SSRs and their polymorphism levels (with eight different genotypes) are compared for 251 EST-SSRs presenting good PCR amplification

were five major SDRs at  $P < 0.05$ , from UF676 in LG2, LG3, LG5, LG6, and from UPA402 in LG9. The segregation distortions from UPA402 were only found in LG9. Some markers distorted at  $P < 0.01$  were detected around regions of high distortion and at the end of chromosomes (LG3, LG7, and LG9). Two other small SDRs were observed in LG8 and LG4 around the cluster of R genes. These non-random SDRs, due to preferential recombination in some regions of the genome, seemed specific to the crosses. The map constructed by Brown et al. (2005), using an F2 population produced from the cross between Sca6 and ICS1, showed very different SDRs. Only a distortion in a region of LG3 was common to both maps between mTcCIR 254 and mTcCIR 81 (12.1 cM). This SDR has been recognized as the largest for both maps.

The rate of segregation distortion on our map (11.5% of total markers) was slightly lower than on the map of Brown et al. (2005), where they found 14.8% of skewed markers at  $P < 0.05$ . These rates were relatively low compared with those found in other species such as potato (Gebhart et al. 1989) with 25.5% of skewed markers, or maize with 19% to 36% (Lu et al. 2002). The origin of these distortions in *T. cacao* remains unknown. The hypothesis most often proposed is the direct selection of genes closely linked to lethal or sub-lethal genes, or the presence of incompatibility alleles (Pugh et al. 2004). The presence of loci in segregation distortion in one or both parents could reveal chromosomal regions containing gametophytic factors subjected to selection (Lu et al. 2002). Distortions of this type have already been observed in other tree species (Barreneche et al. 1998; Dettori et al. 2001). Segregation distortions are generally very high for interspecific crosses, (Paterson et al. 1988), but can also be high in some intraspecific crosses (Hall and Willis 2005). The extent of segregation distortion seems to be related to the type of cross and more particularly to the degree of genome divergence from the parents. In *T. cacao*, structural changes have never been reported and the genes responsible for gametic selection are not known.

#### Classification of EST-SSRs according to gene ontology

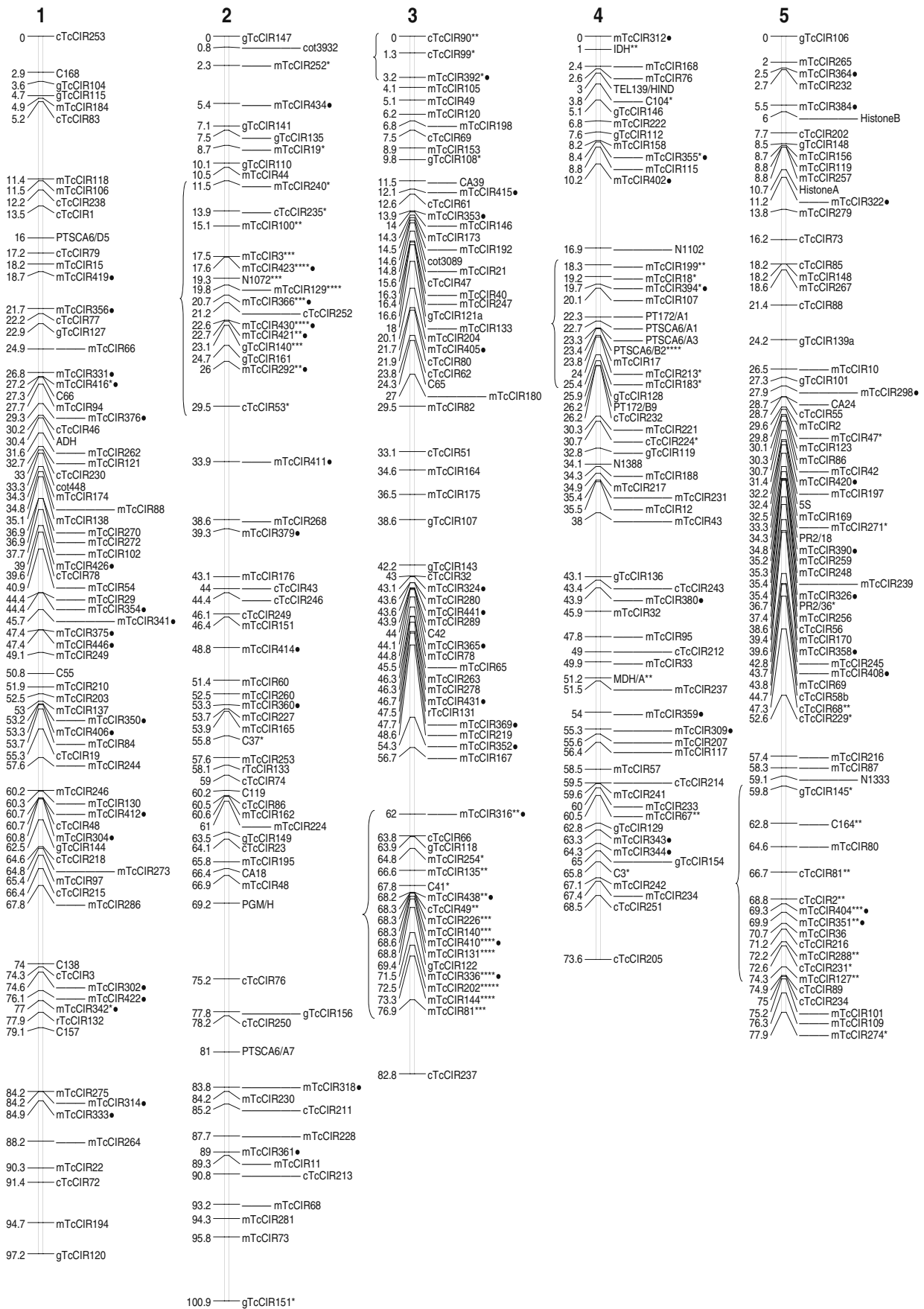
The mapped EST-SSRs were classified by comparing their similarity to *Arabidopsis* known function genes according to the gene ontology tool (Conesa et al. 2005). Genes were distributed according to three criteria: cellular component, biological process, and molecular function. For each criterion, genes were grouped into different classes (ontology) for each level. The higher the level, the more precise is the gene classification. A gene may be represented in several classes. Here, a large panel of genes was represented in the different criteria from level 2 or 3 (Fig. 2).

#### Mapping of EST-SSRs involved in cocoa quality traits or plant defense

Improving or maintaining the aromatic qualities of chocolate and providing sustainable resistance to very devastating diseases, especially those caused by *Moniliophthora* or *Phytophthora* pathogens are among the main goals of breeding programs. Of the 115 new EST-SSRs mapped, 14 constitute good candidate genes that could participate in plant defense/resistance mechanisms or cocoa quality elaboration.

Eight EST-SSRs were putatively involved in biochemical processes important for the quality of chocolate. mTcCIR 297 (oleosin), 305 (acyl-ACP thioesterase), 306

**Fig. 1** Linkage map of *Theobroma cacao* containing 582 codominant markers including 115 new EST-SSR loci. This linkage map is based on cross UPA 402 × UF676 and contains 582 codominant markers (115 EST-SSRs, 269 SSRs, 177 RFLPs, 16 Rgene-RFLPs, and five isoenzymes). EST-SSRs are designated by a black dot. Polymorphic markers for UPA 402 only, for UF 676 only and for both parents are designated on the right, left and in the middle respectively. Markers showing distorted segregation ratios are denoted with a circle for UPA402 or an asterisk for UF676; \* $P < 0.1$ ; \*\* $P < 0.05$ ; \*\*\* $P < 0.01$ ; \*\*\*\* $P < 0.005$ ; \*\*\*\*\* $P < 0.001$ . The segregation distortion regions (SDRs) are identified with braces



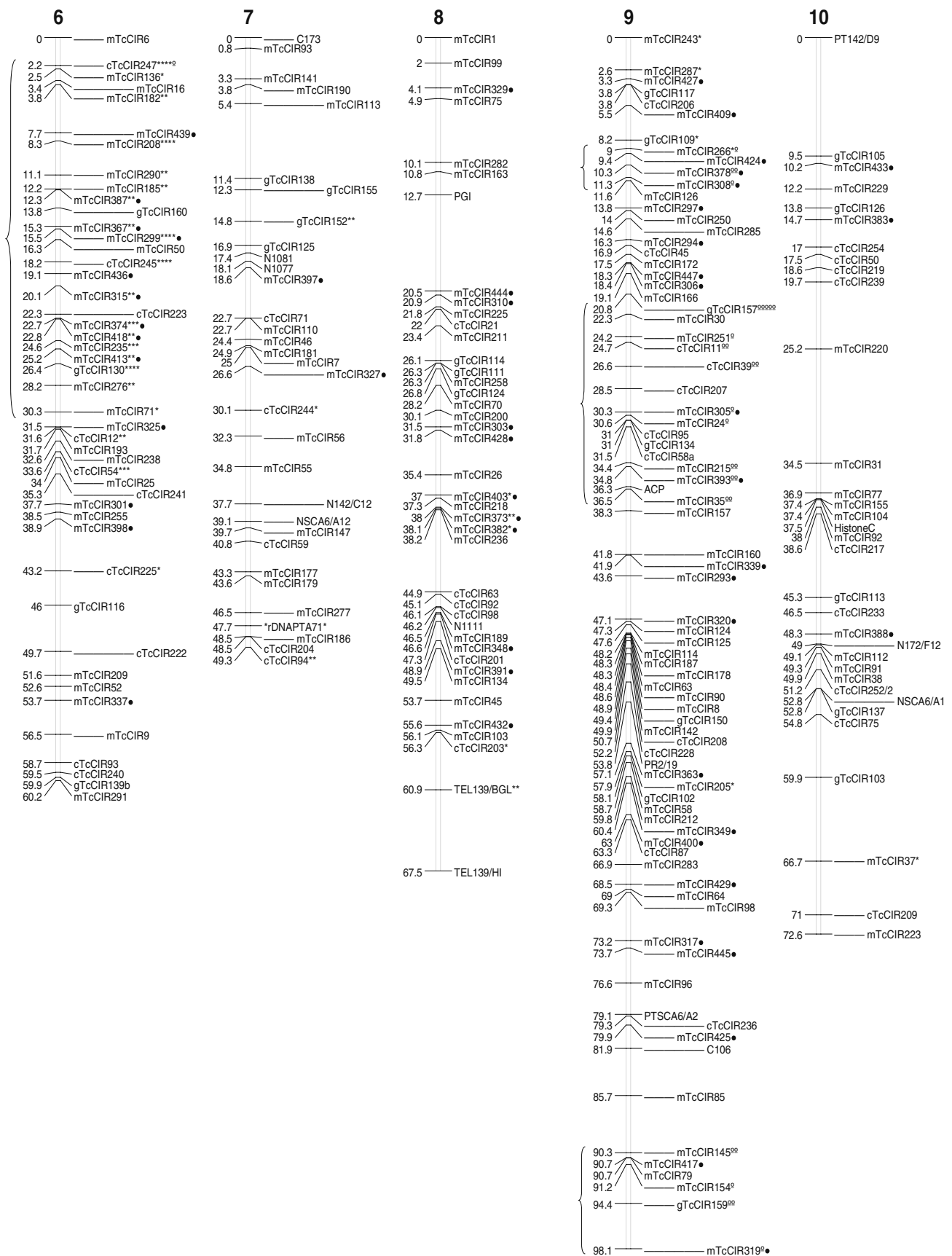
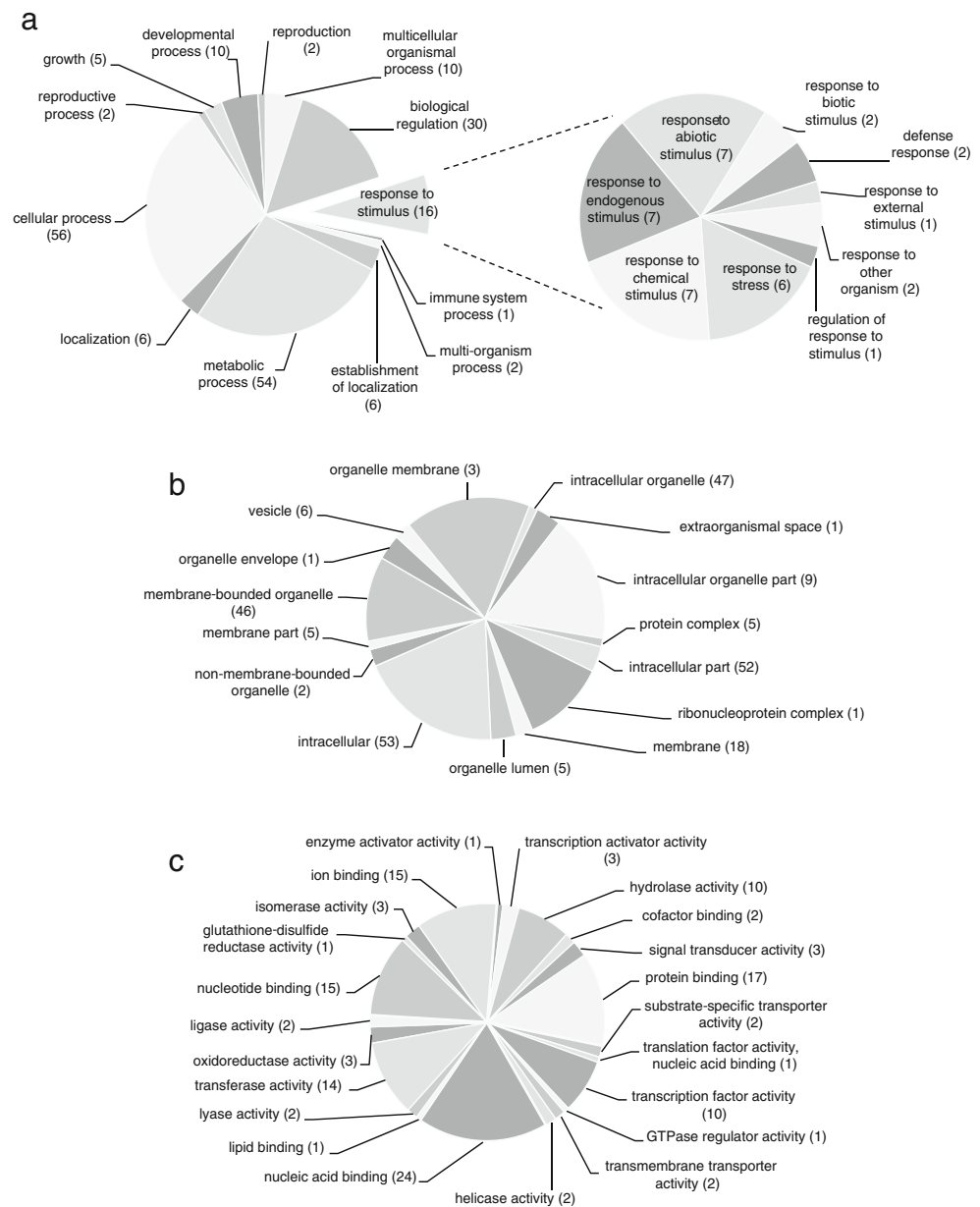


Fig. 1 (continued)

**Fig. 2** Classification of EST-SSRs in three criteria according to gene ontology. **a** Biological process (level 2+3). **b** Cellular component (level 3). **c** Molecular function (level 3)



(3-ketoacyl-ACP synthase), and 310 (2.4-sterol C-methyltransferase) were involved in lipid biosynthesis and seed storage. Cocoa seed storage lipids account for 50% of dry seed weight and participate in enhancing cocoa flavor. mTcCIR 298 and mTcCIR 304 were located in vicilin and aspartic proteinase genes, respectively. Vicilin is one of the major seed storage proteins. During fermentation, and after an increase in acidity, aspartic proteinase genes participate in the proteolytic digestion of vicilin, enabling the constitution of cocoa-specific aroma precursors (Voigt et al. 1994). Another important class of biochemical compounds involved in chocolate quality is the class of flavonoid compounds. It is a diverse group that plays an important role in plant development (Pourcel et al. 2007). In cocoa, proanthocyanidins have been shown to be

beneficial to human health by improving cardiovascular condition (Rimbach et al. 2009). The EST-SSR approach has led to the mapping of a putative dihydroflavonol reductase enzyme and a WD-repeat family transduction factor involved in phenylpropanoid biosynthesis regulation (mTcCIR 299 and 429).

Another important goal for cocoa growing and production is a reduction in yield losses due to fungal diseases. Three pathogens, *Phytophthora* sp. (black pod), *M. perniciosa* (witches' broom), and *Moniliophthora roreri* (frosty pod) cause yield losses of 20% to 80% in the different production regions worldwide.

This study led to the mapping of 7 EST-SSRs located in genes putatively involved in plant defense or resistance. Six of them were involved in conventional reactions that lead to

**Table 5** EST-SSRs from genomic sequences

Marker name	EST Accession number	Forward primer (5'-3')	Reverse primer (5'-3')	Expected size	SSR type	Linkage group	Putative gene function	Reference
mTcCIR4	Y16979	CGACTAAAACCCAAACCATCAA	AATATTAGGCAACCCGAACTT	259	(TCTCTG) <sub>2</sub> (TC) <sub>8</sub>	Unmapped	F-box family protein-like	Lanaud et al. (1999)
mTcCIR9	Y16983	ACCATGCTTCCTCCTTCA	ACATTTATACCCCAACCA	274	(CT) <sub>8</sub> N15(CT) <sub>5</sub> N9(TC) <sub>10</sub>	6	Unknown protein	Lanaud et al. (1999)
mTcCIR22	Y16995	ATTCTCGAAAAAACTTAG	GATGGAAAGGAGTGTAATAG	289	(TC) <sub>1</sub> J2N146(CT) <sub>10</sub>	Unmapped	Putative nuclear antigen homolog	Lanaud et al. (1999)
mtcCIR71	AJ566419	CGACTAAACAGCAGAAAAC	CTCCCTCCTCCTCCAT	170	(GA) <sub>10</sub>	6	Unknown protein	Pugh et al. (2004)
mtcCIR78	AJ566424	TGAAAAATACGTTCTGTCTGA	CAAAAAAGTTTTCTGAAAAGTC	159	(TC) <sub>2</sub> T (TC) <sub>9</sub>	3	Ubiquitin	Pugh et al. (2004)
mtcCIR79	AJ566425	ATTTTTCTTTAGCGCACT	TAACTACCTTCCCACTC	108	(TC) <sub>8</sub>	9	Putative Fe(II) transport protein	Pugh et al. (2004)
mtcCIR87	AJ566432	TAAGGGGCAACATAAAT	CAAAATAGCGCAGAGACAAT	145	(AG) <sub>2</sub> I	5	Unknown protein	Pugh et al. (2004)
mtcCIR113	AJ566455	GGAAAAGTTACAGCAAAGAGAGA	ACAAAGCCCGGTGAAAGG	142	(AG) <sub>9</sub>	7	Unknown protein	Pugh et al. (2004)
mtcCIR136	AJ566473	GAGGAGGTGAGAGCCA	GGTTTGTATTTTTTGATTTGAG	232	(GA) <sub>7</sub> GC (GA) <sub>7</sub>	6	Unknown protein	Pugh et al. (2004)
mtcCIR141	AJ566477	TGTTGCATAAAAACACGAGTTC	CCTAAAATCCTTCTCAACAGC	217	(TC) <sub>1</sub> J4	7	Unknown protein	Pugh et al. (2004)
mtcCIR180	AJ566508	ATGGTTTCGATTTGTCTGT	CAAAAATCTAAGCTGATFAAAAC	186	(GT) <sub>9</sub>	3	Unknown protein	Pugh et al. (2004)
mtcCIR202	AJ566527	TCTCTCATAAGCTCAAAGCA	CCTGAGTCAAAGTGCTCT	172	(TG) <sub>7</sub> (GA) <sub>9</sub>	3	Unknown protein	Pugh et al. (2004)
mtcCIR203	AJ566528	GTGGATTTGGGTGGGAT	ATTGTGTTTTGGCTAIGTTC	217	(AC) <sub>8</sub>	1	Unknown protein	Pugh et al. (2004)
mtcCIR223	AJ566544	GGTCCACACTCAAACACT	TTATTCATTTTCATTTACT	202	(TC) <sub>4</sub> GC (TC) <sub>2</sub> GC (TC) <sub>1</sub> 5 10	10	Unknown protein	Pugh et al. (2004)
mtcCIR226	AJ566547	TAACCCAAAATCAAAGTC	TTTCAACAGGCTCATCT	246	(TC) <sub>1</sub> I	3	NAD synthetase	Pugh et al. (2004)
mtcCIR234	AJ566555	TTGTGTCGGTTTGATTTC	GAAAGAGAGGGAAAAGTGA	123	(TC) <sub>9</sub>	4	Unknown protein	Pugh et al. (2004)
mtcCIR266	AJ566586	TCGTGCCCATCATAAGA	GTCGTTATTCGGAGTTCA	192	(CT) <sub>1</sub> 5	9	Unknown protein	Pugh et al. (2004)
mtcCIR285	AJ566603	TACTACCTTACCCTCTTGT	ATAAAITCCTTCCCTTCT	216	(AG) <sub>1</sub> 6	9	TINY-like protein	Pugh et al. (2004)
mtcCIR287	AJ566605	TCCTTTCTGTTTGTCTCT	TTATCCGTGTCTCCTTCT	301	(TC) <sub>9</sub>	9	Zinc finger protein-like	Pugh et al. (2004)

Already published genomic sequences containing an SSR and having complete homology with a cacao EST. These SSRs also correspond to EST-SSRs



defense activation in plants, such as glutamine synthase (mTcCIR 292), pathogenesis-related protein (mTcCIR 293) (Stintzi et al. 1993), leucine-rich repeats protein (mTcCIR 301) involved in protein/protein interactions and which could play a role in plant resistance to pathogens (Fritz-Laylin et al. 2005), bacterial induced lipoxigenase (mTcCIR 351) activated in plant-pathogen interactions (Melan et al. 1993; Véronési et al. 1996), calmodulin-binding (mTcCIR 433), and a putative WRKY transcription factor (mTcCIR 439). The other one (mTcCIR 304) was located in a gene encoding vicilin, a 7S globulin protein stored in the seed. This protein is very important for cocoa aromatic quality and is also involved in seed defense as an antifungal agent in some plants (Marcus et al. 1999; Ribeiro et al. 2007) and this role has also been described in *Malva palmiflora*, belonging to the same family as *T. cacao* (Wang et al. 2001).

Interestingly, the majority of these markers (mTcCIR 292, 293, 301, 304, and 439) were colocalized with quantitative trait loci (QTL) related to resistance identified in a meta-QTL analysis of disease resistance traits recently conducted in *T. cacao* (Lanaud et al. 2009).

These EST-SSRs could be considered as potential candidate genes for studying resistance and quality traits.

#### EST-SSRs from genomic sequences

To increase the number of EST-SSRs, we compared the SSRs (mTcCIR 1 to mTcCIR 291) previously identified in genomic sequences and already published (Lanaud et al. 1999; Pugh et al. 2004) with cocoa ESTs present in international databases. In total, 249 genomic sequences containing an SSR were analyzed with the FASTA program (Pearson and Lipman 1988). Fifty-six sequences, or 22.5% of the genomic sequences tested, revealed similarity with cocoa ESTs. The sequence homology was often partial, which may have been due to a common domain between two genes or to the presence of introns in the genomic sequence. Nineteen (7.6%) genomic sequences (genomic SSRs) were found to have complete homology with a cacao EST (Table 5).

#### Conclusions

With the development and mapping of a new set of 115 EST-SSR markers which were integrated with a set of previously mapped codominant markers (SSRs, RFLPs, R genes-RFLPs, and isoenzymes), a new linkage map was obtained with 582 codominant markers divided into ten linkage groups. This map contains 134 EST-SSRs (23% of total markers), 115 from EST sequences and 19 from genomic sequences. This map includes a set of 384 SSRs,

molecular markers easily transferable to other laboratories, especially in cocoa-producing countries. This large number of codominant markers will be useful for all genetic analyses, such as fingerprinting, mapping, QTLs, diversity analyses, and marker-assisted selection in *T. cacao*. The increased density of SSR markers on the genetic map will facilitate association mapping studies aimed at the identification of marker/trait associations from natural populations or collections for which a larger number of generations of recombination occurred between ancestors, compared with a controlled progeny. EST-SSRs also represent useful tools to study the functional diversity of populations. Indeed, they could have a role in gene expression or function, and depending on their position (5' UTR, coding region, 3'UTR), they could affect gene transcription, activate or delete genes, and be responsible for gene silencing or transcription slippage (Varshney et al. 2005). Due to the conservation between species belonging to the same family, they are also useful tools for comparative genomic and phylogenetic studies.

**Acknowledgements** We thank Mars and Agropolis foundation for their financial contribution to this project.

#### References

- Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor Appl Genet* 114(2):359–372
- Aime MC, Phillips-Mora W (2005) The causal agents of witches' broom and frosty pod rot of cacao (chocolate, *Theobroma cacao*) form a new lineage of Marasmiaceae. *Mycologia* 97(5):1012–1022
- Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci A-M, Da Silva C, Cascardo J, Allegre M, Kuhn D, Verica J, Courtois B, Llor R, Regis B, Sounigo O, Ducamp M, Gultinan M, Ruiz M, Alemanno L, Machado R, Phillips W, Schnell R, Gilmour M, Rosenquist E, Butler D, Maximova S, Lanaud C (2008) Towards the understanding of the cocoa transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* generated from various tissues and under various conditions. *BMC Genomics* 9(1):512
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 25(1):25–29
- Barreneche T, Bodenes C, Lexer C, Trontin JF, Fluch S, Streiff R, Plomion C, Roussel G, Steinkellner H, Burg K, Favre JM, Glossl J, Kremer A (1998) A genetic linkage map of *Quercus robur* L. (pedunculate oak) based on RAPD, scar, microsatellite, minisatellite, isozyme and 5S rDNA markers. *Theor Appl Genet* 97(7):1090–1103
- Beckman JS, Soller M (1990) Toward a unified approach to the genetic mapping of eukaryotes based on sequence-tagged microsatellite sites. *Biotechnology* 8:930–932

- Borrone JW, Brown S, Khun DV, Motamayor JC, Schnell RD (2007) Microsatellite markers developed from *Theobroma cacao* L. Expressed sequence tags. *Mol Ecol Notes* 7:236–239
- Brown J, Sautter R, Olano C, Borrone J, Kuhn D, Motamayor J, Schnell R (2008) A composite linkage map from three crosses between commercial clones of cacao, *Theobroma cacao* L. *Trop Plant Biol* 1(2):120–130
- Brown JS, Phillips-Mora W, Power EJ, Krol C, Cervantes-Martinez C, Motamayor JC, Schnell RJ (2007) Mapping QTLs for resistance to frosty pod and black pod diseases and horticultural traits in *Theobroma cacao* L. *Crop Sci* 47(5):1851–1858
- Brown JS, Schnell RJ, Motamayor JC, Lopes U, Kuhn DN, Borrone JW (2005) Resistance gene mapping for witches' broom disease in *Theobroma cacao* L. In an f-2 population using SSR markers and candidate genes. *J Am Soc Hortic Sci* 130(3):366–373
- Cheesman EE (1944) Notes on the nomenclature, classification and possible relationships of cocoa populations. *Trop Agric* 21:144–159
- Chen JF, Hu JG, Vick BA, Jan CC (2006) Molecular mapping of a nuclear male-sterility gene in sunflower (*Helianthus annuus* L.) using trap and SSR markers. *Theor Appl Genet* 113(1):122–127
- Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and genbank sequences in rice (*Oryza sativa* L.). *Theor Appl Genet* 100(5):713–722
- Clement D, Risterucci AM, Motamayor JC, N'Goran J, Lanaud C (2003) Mapping QTL for yield components, vigor, and resistance to *Phytophthora palmivora* in *Theobroma cacao* L. *Genome* 46(2):204–212
- Cocoa market update (2009). Available at: <http://www.Worldcocoa.foundation.Org/learn-about-cocoa/cocoa-market.Html>
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18):3674–3676
- Cordeiro G, Casu R, McIntyre C, Manners J, Henry R (2001) Microsatellite markers from sugarcane (*saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Science* 160(6):1115–1123
- Crouzillat D, Lerceteau E, Petiard V, Morera J, Rodriguez H, Walker D, Phillips W, Ronning C, Schnell R, Osei J, Fritz P (1996) *Theobroma cacao* L.: a genetic linkage map and quantitative trait loci analysis. *Theor Appl Genet* 93(1–2):205–214
- Crouzillat D, Menard B, Mora A, Phillips W, Petiard V (2000) Quantitative trait analysis in *Theobroma cacao* using molecular markers—yield QTL detection and stability over 15 years. *Euphytica* 114(1):13–23
- Decroocq V, Favé MG, Hagen L, Bordenave L, Decroocq S (2003) Development and transferability of apricot and grape EST microsatellite markers across taxa. *Theor Appl Genet* 106(5):912–922
- Dettori MT, Quarta R, Verde I (2001) A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. *Genome* 44(5):783–790
- Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5(6):435–445
- Eujayl I, Sorrells M, Baum M, Wolters P, Powell W (2001) Assessment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. *Euphytica* 119(1–2):39–43
- Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the a and b genomes of wheat. *Theor Appl Genet* 104(2–3):399–407
- Figueira A, Janick J, Goldsbrough P (1992) Genome size and DNA polymorphism in *Theobroma cacao*. *J Am Soc Hortic Sci* 117(4):673–677
- Flament MH, Kebe I, Clément D, Pieretti I, Risterucci AM, N'Goran JA, Cilas C, Despréaux D, Lanaud C (2001) Genetic mapping of resistance factors to *Phytophthora palmivora* in cocoa. *Genome* 44(1):79–85
- Fritz-Laylin LK, Krishnamurthy N, Tör M, Sjölander KV, Jones JDG (2005) Phylogenomic analysis of the receptor-like proteins of rice and arabidopsis. *Plant Physiol* 138(2):611–623
- Gebhart C, Ritter E, Debener T, Schachtschabel U, Walkemeier B, Urig H, Salamini F (1989) RFLP analysis and linkage mapping of *Solanum tuberosum*. *Theor Appl Genet* 78:65–75
- Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* 137(4):1121–1137
- Guo W, Cai C, Wang C, Han Z, Song X, Wang K, Niu X, Wang C, Lu K, Shi B, Zhang T (2007) A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in gossypium. *Genetics* 176(1):527–541
- Gupta PK, Rustgi S (2004) Molecular markers from the transcribed/expressed region of the genome in higher plants. *Funct Integr Genomics* 4(3):139–162
- Gupta S, Prasad M (2009) Development and characterization of genic SSR markers in *Medicago truncatula* and their transferability in leguminous and non-leguminous species. *Genome* 52(9):761–771
- Hackauf B, Wehling P (2002) Identification of microsatellite polymorphisms in an expressed portion of the rye genome. *Plant Breed* 121(1):17–25
- Hall MC, Willis JH (2005) Transmission ratio distortion in intraspecific hybrids of *Mimulus guttatus*: implications for genomic divergence. *Genetics* 170(1):375–386
- Han ZG, Guo WZ, Song XL, Zhang TZ (2004) Genetic mapping of EST-derived microsatellites from the diploid gossypium arboreum in allotetraploid cotton. *Mol Genet Genomics* 272(3):308–327
- Jones PG, Allaway D, Gilmour DM, Harris C, Rankin D, Retzel ER, Jones CA (2002) Gene discovery and microarray analysis of cacao (*Theobroma cacao* L.) varieties. *Planta* 216(2):255–264
- Kosambi DD (1944) The estimation of map distance from recombination values. *Ann Eugen* 12:172–175
- Lanaud C, Hamon P, Duperray C (1992) Estimation of nuclear-DNA content of *Theobroma cacao* L. by flow-cytometry. *Café Cacao* The 36(1):3–8
- Lanaud C, Risterucci AM, Ngoran AKJ, Clement D, Flament MH, Laurent V, Falque M (1995) A genetic linkage map of *Theobroma cacao* L. *Theor Appl Genet* 91(6–7):987–993
- Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJ (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. *Mol Ecol* 8(12):2141–2143
- Lanaud C, Fouet O, Clement D, Boccara M, Risterucci AM, Surujdeo-Maharaj S, Legavre T, Argout X (2009) A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L. *Mol Breed* 24(4):361–374
- Laurent V, Risterucci AM, Lanaud C (1993a) Chloroplast and mitochondrial DNA diversity in *Theobroma cacao*. *Theor Appl Genet* 87(1–2):81–88
- Laurent V, Risterucci AM, Lanaud C (1993b) Variability for nuclear ribosomal genes within *Theobroma cacao*. *Heredity* 71:96–103
- Leal GA, Albuquerque PSB, Figueira A (2007) Genes differentially expressed in *Theobroma cacao* associated with resistance to witches' broom disease caused by *Crinipellis pernicioso*. *Mol Plant Pathol* 8(3):279–292
- Lee JM, Nahm SH, Kim YM, Kim BD (2004) Characterization and molecular genetic mapping of microsatellite loci in pepper. *Theor Appl Genet* 108(4):619–627
- Lerceteau E, Robert T, Petiard V, Crouzillat D (1997) Evaluation of the extent of genetic variability among *Theobroma cacao*

- accessions using RAPD and RFLP markers. *Theor Appl Genet* 95(1–2):10–19
- Lima LS, Gramacho KP, Pires JL, Clement D, Lopes UV, Carels N, Gesteira A, Gaiotto FA, Cascardo JC, Micheli F (2010) Development, characterization, validation, and mapping of SSRs derived from *Theobroma cacao* L.–*Moniliophthora perniciosa* interaction ESTs. *Tree Genet Genomes* 4:1–14
- Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 44(3):397–401
- Lu H, Romero-Severson J, Bernardo R (2002) Chromosomal regions associated with segregation distortion in maize. *Theor Appl Genet* 105(4):622–628
- Luro FL, Costantino G, Terol J, Argout X, Allario T, Wincker P, Talon M, Ollitrault P, Morillon R (2008) Transferability of the EST-SSRs developed on nules clementine (*Citrus clementina* hort ex tan) to other citrus species and their effectiveness for genetic mapping. *BMC Genomics* 9:287
- Marcus JP, Green JL, Goulter KC, Manners JM (1999) A family of antimicrobial peptides is produced by processing of a 7 s globulin protein in *Macadamia integrifolia* kernels. *Plant J For Cell Mol Biol* 19(6):699–710
- Melan MA, Dong X, Endara ME, Davis KR, Ausubel FM, Peterman TK (1993) An *Arabidopsis thaliana* lipoxygenase gene can be induced by pathogens, abscisic acid, and methyl jasmonate. *Plant Physiol* 101(2):441–450
- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30(2):194–200
- Motamayor JC, Lachenaud P, da Silva E, Mota JW, Loor R, Kuhn DN, Brown JS, Schnell RJ (2008) Geographic and genetic population differentiation of the amazonian chocolate tree (*Theobroma cacao* L.). *PLoS ONE* 3(10):e3311
- Ogundiwin EA, Peace CP, Gradziel TM, Parfitt DE, Bliss FA, Crisosto CH (2009) A fruit quality gene map of *Prunus*. *BMC Genomics* 10:587
- Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335:721–726
- Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85(8):2444–2448
- Pinto LR, Oliveira KM, Ulian EC, Garcia AAF, de Souza AP (2004) Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome/National Research Council Canada* 47(5):795–804
- Poncet V, Rondeau M, Tranchant C, Cayrel A, Hamon S, de Kochko A, Hamon P (2006) SSR mining in coffee tree EST databases: Potential use of EST-SSRs as markers for the coffee genus. *Mol Genet Genomics* 276(5):436–449
- Pourcel L, Routaboul JM, Cheynier V, Lepiniec L, Debeaujon I (2007) Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci* 12(1):29–36
- Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci* 1(7):215–222
- Pugh T, Fouet O, Risterucci AM, Brottier P, Abouladze M, Deletrez C, Courtois B, Clement D, Larmande P, N’Goran JAK, Lanaud C (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theor Appl Genet* 108(6):1151–1161
- Queiroz VT, Guimaraes CT, Anhard T, Schuster I, Daher RT, Pereira MG, Miranda VRM, Loguercio LL, Barros EG, Moreira MA (2003) Identification of a major QTL in cocoa (*Theobroma cacao* L.) associated with resistance to witches’ broom disease. *Plant Breed* 122(3):268–272
- Ribeiro SFF, Agizzio AP, Machado OLT, Neves-Ferreira AGC, Oliveira MA, Fernandes KVS, Carvalho AO, Perales J, Gomes VM (2007) A new peptide of melon seeds which shows sequence homology with vicilin: partial characterization and antifungal activity. *Sci Hortic* 111(4):399–405
- Riju A, Rajesh MK, Sherin PTPF, Chandrasekar A, Apshara SE, Arunachalam V (2009) Mining of expressed sequence tag libraries of cacao for microsatellite markers using five computational tools. *J Genet* 88:217–225
- Rimbach G, Melchin M, Moehring J, Wagner AE (2009) Polyphenols from cocoa and vascular health—a critical review. *Int J Mol Sci* 10(10):4290–4309
- Risterucci AM, Grivet L, N’Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. *Theor Appl Genet* 101(5–6):948–955
- Risterucci AM, Paulin D, Ducamp M, N’Goran JAK, Lanaud C (2003) Identification of QTLs related to cocoa resistance to three species of *Phytophthora*. *Theor Appl Genet* 108(1):168–174
- Rozen SaS H (2000) Primer3 on the www for general users and for biologist programmers. *Meth Mol Biol* 132:365–386
- Sorrells ME, Wilson WA (1997) Direct classification and selection of superior alleles for crop improvement. *Crop Sci* 37(3):691–697
- Stintzi A, Heitz T, Prasad V, Wiedemannmerdinoglu S, Kauffmann S, Geoffroy P, Legrand M, Fritig B (1993) Plant pathogenesis-related proteins and their role in defense against pathogens. *Biochimie* 75(8):687–706
- Temnykh S, Park WD, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theor Appl Genet* 100(5):697–712
- 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(3):411–422
- Van Ooijen JW (2006) Joinmap, software for the calculation of genetic linkage maps, version 4, Wageningen, The Netherlands
- Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol* 23(1):48–55
- Verica JA, Maximova SN, Strem MD, Carlson JE, Bailey BA, Guiltinan MJ (2004) Isolation of ESTs from cacao (*Theobroma cacao* L.) leaves treated with inducers of the defense response. *Plant Cell Rep* 23(6):404–413
- Véronési C, Rickauer M, Fournier J, Pouénat ML, Esquerré-Tugayé MT (1996) Lipoxygenase gene expression in the tobacco-*Phytophthora parasitica nicotianae* interaction. *Plant Physiol* 112(3):997–1004
- Voigt J, Heinrichs H, Voigt G, Bieh B (1994) Cocoa-specific aroma precursors are generated by proteolytic digestion of the vicilin-like globulin of cocoa seeds. *Food Chem* 50(2):177–184
- Wang X, Bunkers GJ, Walters MR, Thoma RS (2001) Purification and characterization of three antifungal proteins from cheeseweed (*Malva parviflora*). *Biochem Biophys Res Commun* 282(5):1224–1228
- Wasmuth JD, Blaxter ML (2004) Prot4est: translating expressed sequence tags from neglected genomes. *BMC Bioinform* 5:187
- Whitlock BA, Bayer C, Baum DA (2001) Phylogenetic relationships and floral evolution of the Byttnerioideae (Sterculiaceae or Malvaceae s.l.) based on sequences of the chloroplast gene, ndhF. *Syst Bot* 26(2):420–437