Conservation Genetics (2006) 7:13–24
DOI 10.1007/s10592-005-7568-0

Genetic diversity and natural population structure of cacao (*Theobroma cacao* L.) from the Brazilian Amazon evaluated by microsatellite markers

Maria L. Sereno[1], Paulo S.B. Albuquerque[2], Roland Vencovsky[1] & Antonio Figueira[3,*]
*[1]Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Av. Pádua Dias, 11, CP 83, Piracicaba, SP 13400-970, Brazil; [2]ERJOH, Comissão Executiva do Plano da Lavoura Cacaueira, BR 316 km 17, CP 46, Marituba, 67105-970, PA Brazil; [3]Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Av. Centenário, 303, CP 96, Piracicaba, SP 13400-970, Brazil (*Corresponding Author: Phone: +55-19-34294814; Fax: +55-19-34294610; E-mail: figueira@cena.usp.br)*

Received 18 January 2005; accepted 18 May 2005

Key words: center of diversity, cocoa, germplasm, simple sequence repeats

Abstract

A sample of 94 accessions of *Theobroma cacao* L. (cacao), representing four populations from the Brazilian Amazon (Acre, Rondônia, lower Amazon and upper Amazon) were analyzed using microsatellite markers to assess the genetic diversity and the natural population structure. From the 19 microsatellite loci tested, 11 amplified scorable products, revealing a total of 49 alleles, including two monomorphic loci. The Brazilian upper Amazon population contained the largest genetic diversity, with the most polymorphic loci, the highest observed heterozygosity; and the majority of rare alleles, thereby this region might be considered part of the center of diversity of the species. The observed heterozygosity for all the Brazilian populations ($H_o = 0.347$) was comparable with values reported for other similar upper Amazon *Forastero* cacao populations, with the Acre and Rondônia displaying the lowest values. The lower Amazon population, traditionally defined as highly homozygous, presented an unexpectedly high observed heterozygosity ($H_o = 0.372$), disclosing rare and distinct alleles, with large identity with the upper Amazon population. It was hypothesized that part of the lower Amazon population might derive from successive natural or intentional introduction of planting material from other provenances, mainly upper Amazon. Most of the loci exhibited a lower observed heterozygosity than expected, suggesting that self-pollination might be more common than usually assumed in cacao, but excess of homozygotes might also derive from sub-grouping (Wahlund effect) or from sampling related individuals. Most of the gene diversity was found to occur within groups, with small differentiation between the four Brazilian Amazon populations, typical of species with high gene flow.

Abbreviations: AFLP – amplified fragment length polymorphism; bp – base pairs; CAB – Cacao Amazon Brazil; cDNA – complementary DNA; CIRAD – "Centre de Cooperation Internationale en Recherche Agronomique pour le Developpement"; CTAB – hexadecytrimethyl ammonium bromide; dNTPs – deoxyribonucleotides; EDTA – ethylenediaminetetraacetic acid; ERJOH – "Estação de Recursos Genéticos do Cacau José Haroldo"; RAPD – random amplified polymorphic DNA; rDNA – ribosomal DNA; RFLP – restriction fragment length polymorphism

Introduction

Theobroma cacao L. (cacao), a member of the recently expanded family Malvaceae *sensu lato* (Alverson et al. 1999), is a crop with major economic importance, since cacao is grown by more than 2 million growers in more than 50 countries. Cacao fat-rich seeds are the unique source of cocoa solids and cocoa butter, fundamental raw materials for the chocolate and cosmetic industries (Pires et al. 1998). Establishing the genetic diversity and structure of cacao natural populations is critical to define strategies for long-term conservation of genetic resources of this Neotropical tree species, and to maintain the industry sustainability.

The putative center of diversity of *T. cacao* was hypothesized to be in the region located between Ecuador, Colombia and Peru (Cheesman 1944) based on reports about the morphological diversity of wild plants occurring in the Amazon basin (Pound 1943). The large genetic diversity of cacao from the upper Amazonian region has been confirmed by analyses using isozymes (Lanaud 1987; Ronning and Schnell 1994; Warren 1994); RFLPs of a rDNA gene (Laurent et al. 1993; Figueira et al. 1994) and cDNA probes (Laurent et al. 1994; N'Goran et al. 1994, 2000; Lerceteau et al. 1997); RAPD (Figueira et al. 1994; N'Goran et al. 1994; Lerceteau et al. 1997; Whitkus et al. 1998); and microsatellites (Lanaud et al. 1999; Motamayor et al. 2002). However, legal restriction to collect or acquire cacao germplasm in Brazil has limited the comprehensive analysis of the natural genetic diversity occurring in the Brazilian Amazon.

A systematic collection to obtain representation of the genetic diversity of wild and semi-wild cacao from the whole Brazilian Amazon region was conducted by the Brazilian government from 1976 to 1991 (Almeida et al. 1987, 1995). A similar approach had been previously adopted in Ecuador (Allen 1987), in contrast to preceding initiatives, which were restricted to search and collect cacao exhibiting resistance to witches' broom (Pound 1943; Baker et al. 1953; Soria 1970), a severe disease of cacao caused by the basidiomycete *Crinipellis perniciosa* (Stahel) Singer (Purdy and Smith 1996). The objectives of the Brazilian initiative were to preserve and characterize the diversity of cacao in *ex-situ* germplasm repositories in the region, in face of the increasing risk of genetic erosion, caused by colonization and deforestation of the Amazon rainforest. The collection established at the "Estação de Recursos Genéticos do Cacau José Haroldo" (ERJOH), located in Marituba (1°12′ S; 49°13′ W), Pará state, contains more than 1800 accessions, denominated as the Cacao Amazon Brazil (CAB) series, of which 940 derived from clonal propagation, while 877 derived from open-pollinated seedlings, representing 36 river basins of the 186 Brazilian Amazon basins (Almeida et al. 1995).

The conventional classification of cacao assumes three major morphogeographic groups: *Forastero*, *Criollo* and *Trinitario* (Cheesman 1944; Figure 1). Cacao populations from the Amazon basin are considered to be members of the *Forastero* group, which can be further subdivided into upper Amazonian (wild or semi-wild cacaos) and lower Amazonian *Forastero*, assumed to be a homogeneous widely cultivated cacao type (Cheesman 1944). The *Criollo* group contains populations from Central America and northern Colombia and Venezuela, while *Trinitarios* are described as hybrids between *Forastero* and *Criollo* (Motamayor et al. 2002; Figure 1). There is limited information about the genetic structure of natural *T. cacao* populations, with the pre-existing research based on non-natural populations, which were represented either by the morphogeographic groups or by country of origin (Lanaud 1987; Ronning and Schnell 1994; Lerceteau et al. 1997; N'Goran et al. 2000; Motamayor et al. 2002).

Using microsatellites, this study reports the first effort to analyze a sample of the genetic diversity and the genetic structure of CAB accessions, considered to be from the *Forastero* morphogeographic group, originally collected from 19 Amazon River basins.

Material and methods

Plant material

Ninety-four clonal CAB accessions, all considered from the *Forastero* group (Cheesman 1944), representing 19 river basins from various geographic origins of the Brazilian Amazon were analyzed (Figure 1). The accessions were grouped according to the original region of collection (Acre, Rondônia, lower Amazon and upper Amazon), considered here as populations. Accessions with sequential numbering were collected at nearby location.

DNA extraction

Leaves were collected from field-grown plants at ERJOH, Marituba, Pará, Brazil. Leaves were washed, blotted dry, and leaf disks were sampled and frozen in liquid Nitrogen (N). Leaf disks

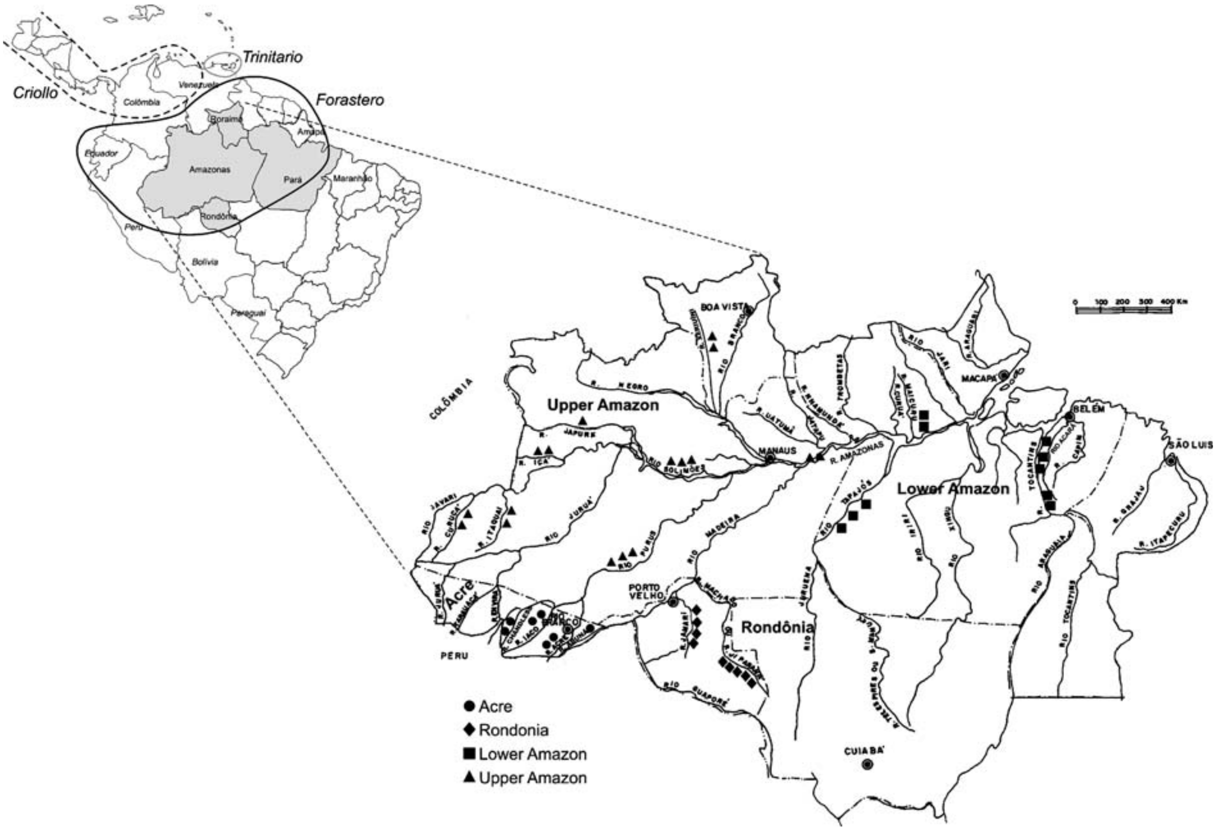

Figure 1. Brazilian Amazonian map indicating the approximate location of sample collections from the four populations or regions.

were kept at −80 °C until DNA extraction. The frozen disks were ground in liquid N, and DNA was extracted using a protocol adapted from Doyle and Doyle (1990). Briefly, ground leaf tissues were extracted with buffer (2% CTAB; 1.4 M NaCl; 20 mM EDTA pH 8.0; 1% polyvinylpyrrolidone MW 10,000; 100 mM Tris–HCl, pH 8.0; 0.2% β-mercaptoethanol; and 0.1 mg ml^{-1} of proteinase K), mixed well and incubated at 55 °C for 60 min. The solution was then extracted three times with chloroform:isoamyl-alcohol (24:1 v/v). DNA was precipitated from the aqueous phase adding cold isopropanol, centrifuged, washed with 70% ethanol and allowed to air dry. The DNA pellet was resuspended in 50 μl of TE buffer (10 mM Tris–HCl, pH 8.0; 0.1 mM EDTA) containing ribonuclease A (10 μg ml^{-1}) and incubated at 37 °C for 30 min. DNA was quantified by fluorimetry in a DyNA Quant 2000 fluorometer (Amersham Biosciences, Buckinghamshire, UK).

PCR conditions

The total reaction volume was 13 μl, containing 13 ng of genomic DNA; 50 mM KCl; 10 mM Tris–HCl (pH 8.8); 0.1% Triton X-100; 1.5 mM $MgCl_2$; 100 μM of each dNTPs; 0.2 μM of each primer and 1.2 U of *Taq* polymerase (Invitrogen do Brasil, São Paulo, Brazil). A total of 19 primer-pairs were tested, but only 11 were used (Table 1). Primers were developed using the software Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) based on sequences deposited by Dr. Claire Lanaud (CIRAD, Montpellier, France) before publication (Lanaud et al. 1999). Primers were synthesized by Invitrogen (Brazil) or by University of British Columbia (Vancouver, Canada). Amplifications were conducted on a GeneAmp 9600 or 9700 thermocycler (Applied Biosystems, Foster City, CA, USA), programmed initially with a denaturing step at 94 °C for 4 min, followed by 30 cycles of 40 s at 94 °C; 40 s at

Table 1. Description of the microsatellite loci named according to Lanaud et al. (1999), with primer-pair sequences used

Locus	Primers sequences	A	Alleles (bp)	H_o	H_e	Linkage group
mTcCIR1	5′ AGGCTCAGTGAAGCAAAGGA 3′ 5′ TGGGAGGGACAATAAGTTGG 3′	6	243, 235, 227, 203, 196, 186	0.310	0.761	8
mTcCIR2	5′ CCAGGGAGCTGTGTTATTGG 3′ 5′ CTCCCTCTGTCCCTCTCTCC 3′	1	220	0.000	0.000	5
mTcCIR6	5′ AATTCCTGTCCTTTTCCCTCT 3′ 5′ TCCCAATCAGCAATTCTAGG 3′	7	250, 240, 222, 216,208, 202, 189	0.293	0.717	6
mTcCIR7	5′ ACATGCGAATGACAACTGGT 3′ 5′ CGAATGAGGTCAGGGCTTAG 3′	3	195, 190, 184	0.058	0.266	7
mTcCIR10	5′ CCGAATTGACAGATGGCCTA 3′ 5′ CCCAAGCAAGCCTCATACTC 3′	4	235, 220, 215, 211	0.811	0.705	5
mTcCIR11	5′ CATTGCGGATTACGGTTTTT 3′ 5′ TGATTAAGCACACGAGCACTG 3′	7	**243**, **230**, **222**, 217, 210, 199, 194	0.384	0.758	2
mTcCIR12	5′ TTTCTGACCCCAAACCTGTAA 3′ 5′ TTCCAGTTAAAGCACATGAGGA 3′	8	**250**, 243, 235, 225, 218, 206, 198, 188	0.880	0.835	4
mTcCIR17	5′ CAGACAATTTGGATGCAACG 3′ 5′ GAAGCTTGGACCCATACGAG 3′	5	**235**, **220**, 208, 204, 199	0.142	0.434	4
mTcCIR18	5′ GCTAAGGGGATTGAGGAAGC 3′ 5′ TGGGTTGCAGTCAATGTCTC 3′	3	**223**, 208, 197	0.272	0.373	4
mTcCIR22	5′ CCCCAAAAATGGAAACGTAA 3′ 5′ CCTAGCCGCAAAGACAAGAG 3′	4	265, 260, 245, 240	0.517	0.628	1
mTcCIR28	5′ GTAAGCTTCGTCCCAGATGC 3′ 5′ CAGCACACCGAAGCTGAATA 3′	1	222	0.000	0.000	6
Total		49				
Average		4.45		0.334	0.497	

Observed total number of alleles per locus (A); size of each allele in base pairs (with rare alleles in bold face); observed heterozygosity (H_o); expected heterozygosity (H_e); and linkage group localization according to Risterucci et al. (2000).

55 °C; and 1 min at 72 °C, ending with a 7-min extension. Later, a touchdown PCR program was used, with identical denaturing and extension conditions, but the annealing temperatures decreased from 65 to 55 °C by 1 °C every cycle for 10 cycles, followed by 20 cycles at 55 °C for 40 s, and without the final extension period of 7 min.

Electrophoresis and polymorphism detection

Amplified products were separated either on 3% agarose gel containing 50% Metaphor (FMC Bioproducts, Rockland, ME, USA) running in TBE buffer (45 mM Tris–Borate; 1 mM EDTA, pH 8.3) or on 20-cm non-denaturing 6% polyacrylamide gel ran at 6 V cm^{-1} for 4.5 h in TBE buffer, stained with ethidium bromide (6 μg ml^{-1}).

Data analysis

Microsatellites alleles were scored in all 94 accessions using a molecular weight standard (123 bp ladder; Invitrogen, Carlsbad, CA, USA), and compared to the expected size of the microsatellite sequence. Genetix 4.02 (Belkhir 2001) was used to estimate: gene diversity statistics of Nei (1973), the average number of alleles per locus, the percentage of polymorphic loci at 95% and 99% levels of significance, the mean heterozygosity per locus, and the *F*-statistics of Wright. The estimated total gene diversity (or total heterozygosity) H_T of the whole sample was subdivided into the gene diversity within (H_S) and between (D_{ST}) sub-samples, where $H_T = H_S + D_{ST}$. The coefficient of gene differentiation G_{ST} is measured by D_{ST}/H_T, and is an estimate of the fraction of the total diversity derived from genetic differences between sub-samples. If there are only two alleles at a locus, G_{ST} becomes identical to F_{ST}, whereas in the case of multiples alleles, G_{ST} is equal to the weighed average of F_{ST} for all alleles (Nei 1973). The absolute magnitude of the genetic divergence among sub-populations can be better estimated using the parameter D_m. This parameter (D_m) is independent of the genetic diversity within sub-populations and is defined by $D_m = sD_{ST}/(s-1)$, where s = number of sub-populations or samples being compared (Nei 1987). G_{ST}, H_T and H_S were calculated using the non-biased estimates according to Nei and Chesser (1983), considering sample size, because N values are approximately proportional to number of individuals in nature. The *F*-statistics of Wright were calculated according to Weir and Cockerham (1984). The total deficit of heterozygotes (F_{IT}) was partitioned into F_{IS}, which measures the deficit of heterozygotes within a population, and F_{ST}, which measures the differentiation between populations.

Results

From the 19 microsatellite loci tested, 11 produced consistent and scorable alleles. The amplification of DNA from 94 Brazilian Amazon accessions using 11 microsatellite loci revealed a total of 49 alleles, with an average of 4.45 alleles per locus (Table 1). The number of alleles per locus ranged from 1 to 8, with an average of 81.8% of polymorphic loci (Table 1). The most polymorphic locus was *mTcCIR12* with 8 alleles, while loci *mTcCIR2* and *mTcCIR28* were monomorphic (Table 1). The polymorphic loci with the least number of alternative alleles were *mTcCIR7* and *mTcCIR18*, with 3 alleles each. According to the cacao consensus genetic map (Risterucci et al. 2000), some of the loci analyzed are linked, particularly *mTcCIR12*, *mTcCIR17* and *mTcCIR18*, all at linkage group 4, while loci *mTcCIR10* and *mTcCIR6* are linked to the monomorphic loci *mTcCIR2* and *mTcCIR28*, respectively (Table 1).

The observed heterozygosity per locus ranged from 0.00 to 0.88, with an overall average of 0.33 (Table 1). Most of the loci showed a smaller observed heterozygosity than expected, except for *mTcCIR10* (4 alleles) and *mTcCIR12* (8 alleles) (Table 1).

The loci *mTcCIR11*, *mTcCIR12*, *mTcCIR17* and *mTcCIR18* presented rare alleles, defined as alleles with a frequency of less than 0.05 in polymorphic loci (Table 1). The accessions containing the least frequent alleles and their geographic origin are presented in Table 2. The 250-bp allele of locus *mTcCIR12* was the most rare with a frequency of 0.011, appearing in a heterozygous state in accessions CAB0268 and CAB0015, originally collected in the river Solimões-low Japurá and Xeriuini, respectively, both in the Brazilian upper Amazon region. The second least frequent was the 230-bp allele of *mTcCIR11*, which was present twice in 86 cases (0.012), in accession CAB0036 and CAB0354 originally collected in the rivers Acará (lower Amazon) and Itaquai (upper Amazon), respectively. All the allelic forms, except for the 243-bp allele from locus *mTcCIR11* (Table 2) were present in the Brazilian upper Amazon population. Some alleles were only found in the upper Amazon population, such as the 250-bp allele of locus *mTcCIR12* and the 223-bp allele of *mTcCIR18*. The rare alleles present at loci *mTcCIR17* and *mTcCIR11* had a more even distribution over the four regions (Table 2).

Organization of the genetic diversity within Brazilian Amazon populations

Gene diversity statistics were calculated based on the 9 microsatellite polymorphic loci for 94 accessions, grouped into four regions or populations (Acre, Rondônia, lower Amazon and upper Amazon), according to the original location of collection. The average number of accessions considered per population ranged from 9.1 for Rondônia to 55.4 for the upper Amazon population (Table 3).

The Acre population, with an average of 12.9 accessions, presented the lowest observed heterozygosity ($H_o = 0.285$) among all populations (Table 3), for an expected heterozygosity of 0.524. The percentage of polymorphic loci of the Acre population was identical to the one estimated for the Rondônia population (Table 3). The Rondônia population exhibited a similar observed heterozygosity (0.288), from an expected heterozygosity of 0.555. The lower Amazon population had an average of 9.2 individuals, with an identical percentage of polymorphic loci (100%) as compared with the estimated for the upper Amazon region (Table 3). The lower Amazon region had an observed heterozygosity of 0.372, while the expected heterozygosity was 0.553. The upper Amazon region, with the largest number of individuals (55.4), had the highest observed and expected heterozygosities (0.445 and 0.630, respectively). Our sample reflects the fact that

Table 2. Rare alleles from various loci, with identification of size in base pairs; frequencies according the number of individual (from 86 to 92 accessions); accession; river and region of origin

Locus	Rare alleles (bp)	Overall Frequency	Accessions	River	Region (population)
mTcCIR11	243	0.023	CAB0010	Jiparaná	Rondônia
			CAB0017, CAB0019, CAB0024	Maicuru	Lower Amazon
	230	0.012	CAB0036	Acará	Lower Amazon
			CAB0354	Itaquai	Upper Amazon
	222	0.046	CAB0352	Curuçá	Upper Amazon
			CAB0354	Itaquai	Upper Amazon
			CAB0028, CAB0045, CAB0056	Solimões/Amazonas	Upper Amazon
			CAB0177	Jiparaná	Rondônia
			CAB0071	Iaco	Acre
mTcCIR12	250	0.011	CAB0268	Solimões/Low Japura	Upper Amazon
			CAB0015	Xeriuini	Upper Amazon
mTcCIR17	235	0.046	CAB0169	Acre	Acre
			CAB0068	Iaco	Acre
			CAB0376, CAB0244	Jamari	Rondônia
			CAB0196, CAB0197	Purus	Upper Amazon
			CAB0324	Solimões/Amazonas	Upper Amazon
			CAB0092	Tocantins	Lower Amazon
	220	0.028	CAB0211, CAB0212, CAB0214, CAB0130	Purus	Upper Amazon
			CAB0355	Itaquai	Upper Amazon
mTcCIR18	223	0.022	CAB0350	Curuçá	Upper Amazon
			CAB0365	Içá	Upper Amazon
			CAB0319	Solimões/Low Japura	Upper Amazon

All accessions were heterozygous, except for CAB0045, which was homozygous for the 222-bp allele, and CAB0350, homozygous for the 223-bp allele.

more accessions from the upper Amazon are preserved at the germplasm collection from ERJOH (Almeida et al. 1995).

For Nei's statistics, the within group diversity estimates (H_S) were comparable in the Acre (H_S = 0.524), Rondônia ($H_S = 0.555$) and lower Amazon populations ($H_S = 0.553$) (Table 3), whereas the upper Amazon population presented the largest value (0.630). The estimated absolute gene differentiation ($D_m = 0.031$) and the relative gene differentiation among populations ($G_{ST} = 0.040$) were low, reflecting small genetic differences between groups (Table 3). There was more diversity within each population (H_S), with an average value of

Table 3. Gene diversity statistics estimated based on the analyses of 9 polymorphic microsatellite loci

Population	N	A	P	H_o	H_S	H_T	F_{IS}	D_m (D_{ST}) G_{ST}
Acre	12.9	3.6	88.9	0.285	0.524		0.456	
Rondônia	9.1	3.0	88.9	0.288	0.555		0.316	
Lower Amazon	9.2	3.8	100.0	0.372	0.553		0.331	
Upper Amazon	55.4	5.1	100.0	0.445	0.630		0.294	
Average				0.347	0.566	0.589		0.031 (0.024) 0.040

Average number of genotypes analyzed per region (N); average number of alleles per locus (A); percent of polymorphic loci (95%) (P); observed heterozygosity (H_o); gene diversity within population (H_S); total gene diversity (H_T); inbreeding coefficient (F_{IS}); absolute gene differentiation (D_m); gene diversity between populations (D_{ST}); and relative magnitude of gene differentiation among subgroups (G_{ST}). The values correspond to average over loci. Estimates of H_S, H_T and G_{ST} are corrected for finite sample size, according to Nei and Cheseer (1983).

0.566, than between the populations ($D_{ST} = 0.024$; Table 3).

Similar results were obtained for the genetic differentiation among the four populations using the *F*-statistics of Wright. The estimated F_{IS} values for the Acre population was 0.456 (Table 3), ranging from −0.083 to 0.773 per each locus (data not shown), while for the Rondônia population, F_{IS} was 0.316 (Table 3), ranging from −0.286 to 1.000 per locus (data not shown). The F_{IS} was estimated to be 0.331 for the lower Amazon population (Table 3), ranging from −0.116 to 1.000 per locus (data not shown), while it was 0.294 for the upper Amazon (Table 3), ranging from −0.294 to 0.779 (data not shown). The high average F_{IS} levels indicated an important within population structure for all populations. The upper Amazon population tended to display less homozygosity for most of the loci (Table 3). Deviation from Hardy–Weinberg equilibrium by the excess of homozygotes was mainly detected for *mTcCIR1*, *mTcCIR6*, *mTcCIR7*, and *mTcCIR17* loci (Table 4). The locus *mTcCIR10* tended to be more heterozygous.

Based on the average values for all loci, the values for F_{IS} and F_{IT} were 0.3184 and 0.3474, respectively (Table 4), indicating a deficit of heterozygotes. The coefficient of differentiation (F_{ST}) among the four populations was estimated to be 0.0425 ± 0.0137 (95% confidence interval based on 1000 bootstrap replications) (Table 4).

The coefficient of differentiation (F_{ST}) was estimated for all pairs of populations (Table 5). The Acre and Rondônia displayed the highest F_{ST} values, as the most distinct populations analyzed.

Table 4. *F*-statistics of Wright estimated for the 9 polymorphic loci

Locus	F_{IT}	F_{ST}	F_{IS}
mTcCIR1	0.6103	0.0712	0.5805
mTcCIR6	0.5953	0.0005	0.5951
mTcCIR7	0.7833	−0.0070	0.7848
mTcCIR10	−0.1391	0.0098	−0.1504
mTcCIR11	0.5191	0.0953	0.4685
mTcCIR12	−0.0367	0.0264	−0.0649
mTcCIR17	0.6756	0.0841	0.6458
mTcCIR18	0.3040	0.0847	0.2397
mTcCIR22	0.1860	0.0119	0.1762
Overall	0.3474	0.0425	0.3184

Table 5. Coefficient of differentiation F_{ST} between all pairs of populations, averaged over all loci

Population	Acre	Rondônia	Lower Amazon	Upper Amazon
Acre	0.000	–	–	–
Rondônia	0.102	0.000	–	–
Lower Amazon	0.058	0.036	0.000	–
Upper Amazon	0.049	0.046	0.018	0.000

The lower and upper Amazon populations presented the lowest F_{ST} indicating the limited differentiation (Table 5).

Discussion

It is generally accepted that South America is the origin of *Theobroma cacao*, which is now supported by molecular evidences (Motamayor et al. 2002). The evaluation of the genetic diversity of wild cacao plants based on morphology, agronomic characters and molecular markers supports the upper Amazon region as the putative center of diversity of cacao. However, the exact location and extension of this center of diversity, as originally proposed to be located between the Caquetá, Napo and Putumayo rivers (Cheesman 1944), have been argued (e.g. Warren 1994). Genotypes from the Brazilian Amazon have been rarely evaluated for genetic diversity (Almeida and Almeida 1987; Bartley et al. 1987; Marita et al. 2001; N'Goran et al. 1994, 2000), restricting possible conclusions about the inclusion of any part of this region in the center of maximum diversity of *T. cacao*. The present analysis of the genetic diversity of cacao populations from the Brazilian Amazon based on microsatellites revealed high levels of polymorphic loci and detected rare alleles in all populations sampled (Tables 2 and 3). Among the Brazilian populations studied, the upper Amazon revealed the largest number of alleles per locus and the most polymorphic loci, in addition to the highest levels of observed heterozygosity. Most of the accessions with the least frequent alleles were originally collected in the Brazilian upper Amazon region, and presented the rare allele in a heterozygous state (Table 2). We present support to the hypothesis that the upper Amazon region contains the largest genetic diversity of *T. cacao* in the

Table 6. Observed heterozygosity, average number of alleles per locus and number of accessions analyzed reported in the literature for cacao populations, estimated by various types of molecular markers

Marker	Observed heterozygosity (H_o) (Average number alleles per locus; number of accessions)					
	Morphogeographic groups					
	Forastero			*Criollo*		*Trinitario*
		Upper Amazonian	Lower Amazonian		Ancient *Criollo*	
Isozyme (Lanaud 1987)	0.207 (2.2; $n=41$)	—	—	0.291 (1.7; $n=13$)	—	0.362 (1.6; $n=19$)
Isozyme (Ronning and Schnell 1994)	0.173 (2.5; $n=45$)	—	—	0.268 (1.9; $n=7$)	—	0.289 (2.1; $n=21$)
RFLP (Lerceteau et al. 1997)	0.188 (2.3; $n=29$)	—	—	0.301 (1.8; $n=9$)	—	0.383 (2.0; $n=29$)
RFLP (N'Goran et al. 2000)	0.180 (2.3; $n=71$)	0.230 (2.3; $n=29$)	0.126 (2.2; $n=30$)	0.250 (2.2; $n=35$)	—	0.300 (2.3; $n=45$)
RFLP (Motamayor et al. 2002)	0.180 (2.4; $n=37$)	—	—	0.470 (2.0; $n=68$)	0.002 (1.1; $n=92$)	0.430 (2.1; $n=67$)
Microsatellite (Motamayor et al. 2002)	0.340 (8.7; $n=28$)	0.480 (5.5; $n=13$)	—	0.59 (2.4; $n=13$)	0.000 (1.2; $n=41$)	0.680 (2.2; $n=14$)
Microsatellite (this work)	0.347 (4.4; $n=94$)	0.445 (5.1; $n=59$)	0.372 (3.8; $n=11$)	—	—	—

Brazilian Amazon, and is part of the center of diversity of the species.

Observed heterozygosities estimated from different types of molecular markers are not directly comparable since the number alleles per locus may vary widely among markers. In the case of cacao, there have been a few reports on estimation of observed heterozygosity employing different markers, and the average number of alleles observed using isozymes and RFLPs were similar, whereas a larger number of alleles per locus for some populations, were revealed on the application of microsatellite markers (Table 6). The observed heterozygosity here described for all 94 accessions ($H_o = 0.334$) was, in general, comparable with values reported for other *Forastero* populations, estimated by isozymes, RFLP and microsatellite analysis (Table 6). Based on isozyme analysis, the observed heterozygosity for 45 *Forastero* genotypes was estimated to be 0.173 (Ronning and Schnell 1994), while for various specific upper Amazonian *Forastero* populations, H_o values ranged from around 0.100 for populations LCT-EEN (Ecuador) and SPEC (Colombia) to 0.347 (EQX from Ecuador), while a Peruvian population had $H_o = 0.207$ (Lanaud 1987; Table 6). Similar values of H_o for upper Amazon *Forasteros* were estimated based on RFLP markers, ranging from 0.18 ($n = 37$) (Motamayor et al. 2002) to 0.23 ($n = 29$) (N'Goran et al. 2000) (Table 6). Based on microsatellite analysis (Motamayor et al. 2002), the overall observed heterozygosity for a *Forastero* population ($n = 28$) was comparable to values described here (Table 6), and similar values to the Brazilian upper Amazon population were estimated for specific populations from Peru ($H_o = 0.48$; $n = 13$) and from Colombia-Ecuador ($H_o = 0.39$; $n = 5$). On the other hand, *Trinitarios* and modern *Criollos* (as defined by Motamayor et al. 2002), have consistently presented the largest observed heterozygosity among the morphogeographic groups of cacao using various molecular markers, including isozymes (Lanaud 1987; Ronning and Schnell 1994), RFLP (Lerceteau et al. 1997; N'Goran et al. 2000; Motamayor et al. 2002) and microsatellites (Motamayor et al. 2002), probably due to their recent hybrid origin (*Criollo* × *Forastero*), including recent introgression of genes from *Forastero* (Motamayor et al. 2002).

Among the Brazilian populations, the Acre and Rondônia displayed the lowest average observed heterozygosity (Table 3), while the population from the lower Amazon presented higher values (Table 3). The low observed heterozygosity for the Acre population was not anticipated since a large phenotypic diversity for various morphological, molecular and agronomic characters have been reported for genotypes collected in this region (Pires et al. 1994, 1999). Conversely, the population from Rondônia, described by Almeida and Almeida (1987) as morphologically homogeneous with little phenotypic variability, presented similar heterozygosity as the Acre population.

The lower Amazon *Forasteros* have been typically defined as an homogeneous population, growing wild in the Guyanas and in the eastern region of the Amazon, characterized by presenting a rather uniform pod type called *Amelonado* (Cheesman 1944), which is, currently, the most prevalent cultivated cacao type world-wide. Populations from the lower Amazon might derive from a proto-domestication conducted by pre-Colombian Amazonian peoples for the aromatic pulp (Barrau 1979). Molecular analyses of cultivated genotypes classified as lower Amazon *Forastero* have consistently displayed low numbers of alleles per locus; small percentage of polymorphic loci and of observed heterozygosity (Lanaud 1987; N'Goran et al. 2000).

Thus, the observed heterozygosity here estimated for the lower Amazon population was unexpectedly high, since genotypes originated in this region have been generally considered to be highly homozygous. The Brazilian upper and lower Amazon populations appeared more similar to each other, demonstrated by the lowest coefficient of differentiation F_{ST} (Table 5), with similar levels of genetic diversity, despite the difference in the number of accessions analyzed from each population (Table 3). Accessions CAB0017, CAB0019, CAB0024, CAB0036 and CAB0092 were originally collected in various locations in the lower Amazon region (Figure 1), some in areas of traditional cacao cultivation (over 300 years), such as Alenquer, Pará state (Bartley et al. 1987). These accessions disclosed rare and distinct alleles (Table 2), suggesting they might result from various introductions over time. Because of the long history of cacao cultivation in the lower Amazon valley, dating back to the 17th century (von Martius 1930; Bartley et al. 1987); it is difficult to distinguish wild from remnant cultivated forms, and to presume the origin of these plants. Thus, it is not clear if the high levels of diversity here detected for the lower Amazon population is wild, or derived from a rather recent (<300 years) intentional introduction of planting material from other provenance, mainly the upper Amazon, or from other natural seed dispersal agents, such as migrating native peoples. The lower Amazon sample analyzed represented a distinct population from the one consistently associated with the *Amelonado* type population, occurring naturally from Guyana and Surinam to the lower reaches of the Amazon (Toxopeus 1987; Cheesman 1944). Therefore, it can be hypothesized that part of the population at the lower Amazon region may be derived from natural stands located up river, from the upper Amazon region, including Acre.

Most of the loci showed a smaller observed heterozygosity than expected, except for *mTcCIR10* (4 alleles) and *mTcCIR12* (8 alleles) (Table 1). The same trend was described for most of the microsatellite loci, including the same *mTcCIR10* and *mTcCIR12* loci, analyzed with a different set of genotypes (Lanaud et al. 1999). The deficit in heterozygotes has been a current observation in cacao populations, estimated by isozymes (Ronning and Schnell 1994) and RFLP (N'Goran et al. 2000). N'Goran et al. (2000) observed a deficit of heterozygotes in the upper and lower Amazon populations, detected by the lower observed heterozygosity and confirmed by high values for F_{IS}, except for *Trinitario* ($F_{\mathrm{IS}} = -0.01$).

Cacao appears to be a typical outbreeding species, because of its floral morphology, adapted to pollination by small midges, and the occurrence of a unique gameto-sporophytic self-incompatibility system (Cope 1976). Incompatible mating is characterized by the failure of gametic nuclei fusion at the embryo sac, resulting in flower abscission (Knight and Rogers 1955; Cope 1962). However, the incompatibility system of cacao is not absolute, but quantitative, depending on the ratio of fused to non-fused ovules, and can be overcome by employing a mixture of compatible and incompatible pollen with successful self-fertilization, or even naturally at a very low rate (Glendinning 1960; Lanaud 1987). Studies have been conducted in clonal gardens for hybrid seed production, detecting up to 96% of self-fertilization of incompatible materials, affected by environment and maternal genotype (Lanaud et al. 1987). Thus, self-pollination in cacao might be more common than assumed, but to our knowledge, mating systems have not been evaluated under natural conditions. Cope (1976) proposed that genotypes occurring near the putative center of diversity would be uniformly self-incompatible, while self-compatible trees would predominate away from this area. The occasional success in self-fertilization of incompatible trees would allow the maintenance of isolated trees under natural conditions (Cope 1976). Genotypes from the lower Amazon of the *Amelonado* type are known to be self-compatible, and inbreeding could have occurred in some cases, while for some of the other populations, the reduction of heterozygotes might derive from the occurrence of sub-grouping (Wahlund effect) or from sampling-related individuals. Some of the accessions analyzed have a sequential numbering, such as CAB00142 and CAB00143; CAB00177 and CAB00178, which could indicate that sampling could have occurred within a nearby area, collecting direct relatives.

Most of the gene diversity was found to occur within groups, rather than between the four Brazilian Amazon populations or regions here considered (Acre, lower Amazon, Rondônia, and upper Amazon) (Table 3). The allocation of most of the genetic diversity within groups rather than between groups has already been described in cacao (Ronning and Schnell 1994; Lerceteau et al. 1997; N'Goran et al. 2000), being similar to other outcrossing woody perennial species (Ronning and Schnell 1994; Hamrick and Godt 1990). It is suggested from our data that *T. cacao* has a strong intra-population structure with small differentiation between populations, typical of species with high gene flow (Hamrick et al. 1992).

Some river basins (sub-populations) had more individuals in our sampling, which could have favored the chance of identification of rare alleles. In contrast, accession CAB0015, one of the only two accessions originally collected at river Xeriuini (state of Roraima), maintained at the ERJOH germplasm collection, disclosed rare alleles at locus *mTcCIR12* (Table 2) and *mTcCIR1* (not shown). The occurrence of two rare alleles in one of two individuals from this river basin underlines the importance of analyzing the genetic diversity of germplasm collections and may assist to indicate potential areas for further collections.

The level of polymorphism detected by microsatellite markers is much larger in comparison to other methods, because of the high mutation rate, generating a large number of alleles. The present analysis of 94 accessions for 11 microsatellite loci revealed a total of 49 alleles, with an average of 4.45 alleles per locus (Table 1). For the same 11 microsatellite loci, Lanaud et al. (1999) described a total of 60 alleles, with an average of 5.4 alleles per locus, based on a different set of genotypes, and employing radiolabeled amplification products separated in denaturing polyacrylamide gels. The radioactive detection system had probably allowed an improved resolution of fragments and identification of more alleles, undistinguishable under the conditions used here. But for some cases, such as loci *mTcCIR1* and *mTcCIR17*, more alleles were detected in this study (6 and 5 alleles, respectively) against the 3 alleles reported for both cases by Lanaud et al. (1999), probably because of sample size and the geographic origin of the genotypes. The present work analyzed 94 accessions, some collected near the putative center of origin, while Lanaud et al. (1999) tested 18 or 14 genotypes for loci *mTcCIR1* and *mTcCIR17*, respectively, from various origins. Motamayor et al. (2002) identified 150 alleles for 16 microsatellite loci (average of 9.4 alleles per locus) studying 102 individuals representing the whole diversity of *T. cacao*. However, only 5.5 alleles per locus were identified for the sample of genotypes from Peru ($n = 13$), and 3.94 for 5 genotypes from Colombia–Ecuador (Motamayor et al. 2002), values comparable to the number of alleles detected in the present study. In general, more precise determination of allele sizes can be obtained using either fluorescence or radioactive labeling of amplification products, however in developing countries, the use of ethidium bromide or silver nitrate staining can be effective for detecting microsatellites at lower cost in genetic diversity studies.

In conclusion, the use of microsatellite markers was highly informative for cacao populations, presenting a high average number of alleles per locus. This work detected important levels of diversity at the Brazilian Amazon region, with the disclosure of rare alleles and elevated observed heterozygosity, especially for the Brazilian upper Amazon population, but with low genetic differentiation among populations. We have provided support to the hypothesis that the upper Brazilian

Amazon region might be part of the center of diversity of the species.

Acknowledgements

This work was financially supported by IFS (D 2814-2F); CNPq (479573/2001-8) and CEPLAC, Brazilian Ministry of Agriculture. The authors are deeply grateful to the Sereno's family for their commitment and incentive. The assistance for DNA extraction by Cristiane Cristina Borges was highly appreciated. We thank the useful comments from both reviewers.

References

Allen JB (1987) London Cocoa Trade Amazon Project. Final Report Phase 2. Cocoa Growers' Bul. 39, 94 pp. Cadbury Limited, Birmingham, UK.

Almeida CMVC, Almeida CFG (1987) Coleta de cacau silvestre no estado de Rondônia. *Rev. Theobroma*, **17**(2), 65–92.

Almeida CMVC, Barriga JP, Machado PFR, Bartley BGD (1987) Evolução do programa de conservação dos recursos genéticos de cacau na Amazônia Brasileira. Boletim Técnico no. 5. Ministério da Agricultura. Comissão Executiva do Plano da Lavoura Cacaueira, Belém, Pará.

Almeida CMVC, Machado PFR, Barriga JP, Silva FCO (1995) Coleta de cacau (*Theobroma cacao* L.) da Amazônia brasileira: uma abordagem histórica e analítica. Ministério de Agricultura e Reforma Agrária. Comissão Executiva do Plano da Lavoura Cacaueira, 92 pp. Belém, Para, Brasil.

Alverson WS, Whitlock BA, Nyffeler R, Bayer C, Baum DA (1999) Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *Am J Bot.*, **86,** 1474–1486.

Baker RED, Cope FW, Holliday PC, Bartley BGD, Taylor J (1953) The Anglo-Colombian cacao collecting expedition. Report Cacao Research. St. Augustine, Trinidad. In: *Reprint Archives of Cocoa Research* (Ed. Toxopeus H), vol. 1, pp. 127–154. American Cocoa Research Institute, USA.

Barrau J (1979) Sur l'origine du cacaoyer, *Theobroma cacao* Linne, Sterculiacées. *J. d'Agr. Trad. Bot. Appl.*, **26**(3–4), 171–180.

Bartley BGD, Machado PFR, Ahnert D, Barriga JP, Almeida CMVC (1987). Descrição de populações de cacau da Amazônia brasileira. I- Observações preliminares sobre populações de Alenquer, Pará. In: 10th International Cocoa Research Conference, Santo Domingo, Dominican Republic, pp. 665–672. Cocoa Producers' Alliance, Lagos, Nigeria.

Belkhir K (2001) GENETIX, logiciel sous WindowsTM pour la génétique des populations. Laboratoire Génome et Populations, CNRS UPR 9060, Université de Montpellier II, Montpellier (France).

Cheesman EE (1944) Notes on the nomenclature, classification and possible relationships of cocoa populations. *Trop. Agric.*, **21**(8), 144–159.

Cope FW (1962) The mechanism of pollen incompatibility in *Theobroma cacao, Heredity*, **17,** 157–182.

Cope FW (1976) Cacao, *Theobroma cacao* In: Evolution of Crop Plants (Ed., Simmonds NW), pp. 285–289. Longman, London.

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. *Focus*, **12**(1), 13–15.

Figueira A, Janick J, Levi M, Goldsbrough P (1994) Reexamining the classification of *Theobroma cacao* L. using molecular markers. *J. Amer. Soc. Hort. Sci.*, **119,** 1073–1082.

Glendinning DR (1960) Selfing of self-incompatible cocoa *Nature*, **187**(4732), 170.

Hamrick JL, Godt M (1990) Allozyme diversity in plant species. In: Plant Population, Genetics, Breeding and Genetic Resources (Ed., AHD Brown, MT Clegg, AL Kahler, BS Weir), pp. 43–63. Sinauer Associates, Sunderland.

Hamrick JL, Godt MJ, Sherman-Broyles SL (1992) Factors influencing levels of genetic diversity in woody plant species. *New Forest*, **6,** 95–124.

Knight R, Rogers H (1955) Incompatibility in *Theobroma cacao. Heredity*, **9,** 67–69.

Lanaud C (1987) Nouvelles donnes sur la biologie du cacaoyer (*Theobroma cacao* L): diversité des populations, systéme d'incompatibilité, haploides spontanées. Leurs consequence pour l'amelioration genetique de cette espece. D. Sc. Thesis, Université de Paris Sud, Centre d'Orsay, France.

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. *Mol. Ecol.*, **8,** 2141–2143.

Lanaud C, Sounigo O, Amefia YK, Paulin D, Lachenaud P, Clement D (1987) Nouvelles données sur la fonctionement du systéme d'incompatibilité du cacaoyer et ses consequences pour la selection. *Café Cacao Thé*, **31**(4), 267–277.

Laurent V, Risterucci AM, Lanaud C (1993) Variability for nuclear ribosomal genes within *Theobroma cacao. Heredity*, **71,** 96–103.

Laurent V, Risterucci AM, Lanaud C (1994) Genetic diversity in cocoa revealed by cDNA probes. *Theor. Appl. Genet.*, **88,** 193–198.

Lerceteau E, Robert T, Pétiard V, Crouzillat D (1997) Evaluation of the extent of genetic variability among *Theobroma cacao* accessions using RAPD and RFLP markers. *Theor. Appl. Genet.*, **95,** 10–19.

Marita JM, Nienhuis J, Pires JL, Aitken WM (2001) Analysis of genetic diversity in *Theobroma cacao* with emphasis on witches' broom disease resistance. *Crop Sci.*, **41,** 1305–1316.

von Martius KF (1930) Sur le cacao et les espéces du genre Theobroma *Rev. Bot. Appl. D'agric trop.*, **110,** 184–192.

Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity*, **89,** 308–386.

Nei M (1972) Genetic distance between populations *Am. Nat.*, **106,** 283–292.

Nei M (1973) Analysis of gene diversity in subdivided populations *Proc. Natl. Acad. Sci. USA*, **70,** 3321–3323.

Nei M (1987) *Molecular Evolutionary Genetics*, Columbia University Press, New York, 190 p.

Nei M, Chesser RK (1983) Estimation of fixation indexes and gene diversities. *Ann. Hum. Genet.*, **47,** 253–259.

N'Goran JAK, Laurent V, Risterucci AM, Lanaud C (1994) Comparative genetic diversity studies of *Theobroma cacao* L. using RFLP and RAPD markers. *Heredity*, **73,** 589–597.

N'Goran JAK, Laurent V, Risterucci AM, Lanaud C (2000) The genetic structure of cocoa populations (*Theobroma cacao* L.) revealed by RFLP analysis. *Euphytica*, **115,** 83–90.

Pires JL, Luz EDMN, Lopes UV, Bartley BGD (1994) Incidência natural de podridão parda em clones de cacaueiro na Bahia, Brasil. In: Proc. of the 11th International Cocoa Research Conference, Yamoussoukro, Cote d'Ivoire, pp. 871–877. Cocoa Producers' Alliance, Lagos, Nigeria.

Pires JL, Cascardo JCM, Lambert SV, Figueira A (1998) Increasing cocoa butter yield through genetic improvement of *Theobroma cacao* L: seed fat content variability, inheritance, and association with seed yield. *Euphytica*, **103,** 115–121.

Pires JL, Monteiro WR, Luz EDMN, Silva SDVN, Pinto LRM, Figueira A (1999) Cocoa breeding for witches' broom resistance at CEPEC, Bahia, Brazil. In: Proc. 2nd INGENIC International Workshop on Cacao Genetics, pp. 91–101. Salvador, Bahia, Brazil, November 24th–26th 1996.

Pound FJ (1943) Cacao and witches' broom disease (*Marasmius pernicious*) In: Reprint Archives of Cocoa Research. (Ed., H Toxopeus), pp. 73–92. American Cocoa Research Institute, USA.

Purdy LH, Smith RA (1996) Status of cacao witches' broom: biology, epidemiology, and management. *Annu. Rev. Phytopathol.*, **34,** 573–594.

Soria J (1970) The latest cocoa expeditions to the Amazon basin *Cacao*, **15,** 5–15.

Risterucci AM, Grivet L, N'Goran JAK, Pieretti L, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. *Theor. Appl. Genet.*, **101,** 948–955.

Ronning CM, Schnell RJ (1994) Allozyme diversity in a germplasm collection of *Theobroma cacao* L. *J. Hered.*, **85,** 291–295.

Toxopeus H (1987) Botany, types and populations In: Cocoa (Ed., Wood GAR, Lass RA), pp. 11–37. Longman, London.

Warren JM (1994) Isozyme variation in a number of populations of *Theobroma cacao* L obtained through various sampling regimes. *Euphytica*, **72,** 121–126.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38,** 1358–1370.

Whitkus R, de la Cruz M, Mota-Bravo L, Gómez-Pompa A (1998) Genetic diversity and relationships of cacao (*Theobroma cacao* L.) in southern Mexico. *Theor. Appl. Genet.*, **96,** 621–627.