Theor Appl Genet (1997) 95:10–19

E. Lerceteau · T. Robert · V. Pétiard · D. Crouzillat

Evaluation of the extent of genetic variability among *Theobroma cacao* accessions using RAPD and RFLP markers

Received: 10 June 1996 / Accepted: 11 October 1996

Abstract Random amplified polymorphic DNA (RAPD) and restriction fragment length polymorphism (RFLP) markers were used to evaluate genetic relationships within the *Theobroma cacao* species and to assess the organization of its genetic diversity. Genetic variability was estimated with 18 primers and 43 RFLP probes on 155 cocoa trees belonging to different morphological groups and coming from various geographic origins. The majority of the RFLP probes issued from low-copy DNA sequences. On the basis of on the genetic distance matrices, the two molecular methods gave related estimates of the genetic relationship between genotypes. Although an influence of cocoa morphological groups and geographical origins of trees was observed, a lack of gene differentiation characterized the *T. cacao* accessions studied. The continuous RFLP variability observed within the species may reflect the hybridization and introgressions between trees of different origins. Nevertheless, the Nacional type was detected to be genetically specific and different from well-known types such as Forastero, Criollo and Trinitario. Some of those genotypes were characterized by a low heterozygosity rate and may constitute the original Nacional pool. These results also provide information for the constitution of a cocoa tree core collection.

Key words *Theobroma cacao* · RFLP · RAPD · Genetic diversity

Communicated by P. M. A. Tigerstedt

E. Lerceteau · V. Pétiard · D. Crouzillat (✉)
Centre R&D Nestlé, 101 avenue Gustave Eiffel, Notre Dame d'Oé, BP9716, 37097 Tours Cedex 2, France

T. Robert
Université Paris Sud XI – CNRS (URA 1492), Laboratoire d'Evolution et Systématique des Végétaux, Bat: 362, 91405 Orsay Cedex, France

Introduction

Throughout this century a large number of samples from landraces of cultivated plants and from related wild species have been collected, and these are now available to breeders aiming at enlarging the genetic basis of this plant material. However, emphasis has been put on the general underuse of this potentially useful diversity in breeding programs (Marshall 1989). This situation is partly due to the lack of a systematic genetic evaluation of plant collections. Information on germplasm diversity and genetic relationships among plant materials is therefore essential for breeders.

Cocoa tree is not an exception to this general statement. Genetic diversity exists among the different wild *Theobroma* species, but the current *T. cacao* cultivated clones represent only a small part of this genetic pool (Cope 1976). Up to the beginning of this century the genetic basis was too narrow to develop efficient breeding programs. One of the first attempts to increase the variability of this woody and outcrossing crop was motivated by the necessity to look for resistance capacity to witches' broom disease (Lockwood 1985). Since 1938, many botanical prospectings have been carried out. The main way for increasing the value of these collections has been through the production of hybrid clones.

On the basis of the morphological traits of its pods and beans, *T. cacao* species can be divided into three morphological groups, Criollo, Forastero and Trinitario, the latter considered to be hybrid form between the first two groups. While Criollo trees usually have good quality beans, most of the time they are poor-yielding and highly susceptible plants, whereas Forastero trees, which give 80% of the world's production of cacao, are vigorous and more resistant to diseases. This morphological group is divided into Lower Amazon and Upper Amazon Forastero according to their geographical locations. However, the classification of some cocoa clones remains doubtful as, for example, the cocoa trees known as Nacional from

Ecuador [classified as Forastero by Cheesman (1944) and Soria (1970) and recently placed among Criollo (Enriquez 1992)].

Nowadays much effort is being expended to reduce the number of botanical descriptors in order to facilitate genotype characterization. From a list of 65 standardized descriptors (IBPGR 1981), Engels (1992) selected 10 characters on the basis of their agronomic and taxonomic importance. Bekele and Bekele (1996) evaluated 100 accessions using 28 quantitative and 26 qualitative morphological descriptors. However, these classifications could be biased by environmental influences.

The efficiency of molecular markers in assessing the organization of genetic variability and phylogenetic relationships in plant complex-species has already been demonstrated. On the basis of isozyme data, Upper Amazonia is considered to be the primary center of diversity (Lanaud 1987; Warren 1994). Recent studies tried to estimate the different components of isozymes gene diversity in cocoa (Ronning and Schnell 1994). Laurent et al. (1993a, b, 1994) observed mitochondrial, chloroplastic, ribosomal and cDNA polymorphisms and were able to provide some insights into the evolution of cocoa, underlining the complementary aspect of all these markers. The random amplified polymorphic DNA (RAPD) technology (Williams et al. 1990) has also been successfully used to infer genetic relationships within many species, including *Theobroma cacao* (Figueira et al. 1994; N'goran et al. 1994). However the comparison between geographical and morphological classifications has never been investigated using restriction fragment length polymorphisms (RFLPs) and RAPDs. Moreover, Nacional genotypes, including fine cacao, have been scarcely studied at the molecular level.

In the investigation reported in this paper we analyzed 155 cocoa genotypes with the RAPD and RFLP techniques. The objectives of the study were (1) to compare and evaluate the efficiency of RAPD and RFLP markers in determining relationships in *T. cacao* species, (2) to assess the organization of genetic diversity in this species and clarify evolutionary relationships between morphological groups.

Materials and methods

Plant material

The 155 *Theobroma cacao* trees used in this study are listed in Table 1. The morphological group and the geographical origin of each genotype are also indicated, when known. This information was provided by INIFAP (Instituto Nacional de Investigaciones Forestales y Agropecuarias), Mexico, CATIE (Centro Agronomico Tropical de Investigacion y Enseñanza), Costa Rica, Nestlé R&D center S.A. Quito Ecuador, INIAP (Instituto Nacional Investigaciones Agropecuarias), Ecuador, and CIRAD (Centre de Coopération Internationale en Recherches Agronomiques pour le Développement), France.

The three morphological groups, Criollo, Forastero, and Trinitario, were represented in this study. Nacional genotypes were also considered. The Ecuadorian Nacional accessions came in part from the EETP (Estacíon Experimental Tropical Pichlingue of Ecuador) germplasm collection that dates from about 1940 located in the region of Loma-long, whereas the Sebastiàn Arteaga (SA) and the Balao Chico (BCH) Nacional genotypes were collected in two 80- to 100-year-old plantations, located in the north and south of Ecuador, respectively, about 450 km from each other. These two plantations were established from a limited number of pods.

While the genotypes Venezolano amarillo and Venezolano morado are usually considered as belonging to the Trinitario group, the exact origin of these trees remains doubtful (Pound 1938). Consequently they were not included in the Trinitario group. Genotypes of different geographical origins were used, but most of them came from Central or South America, where cocoa trees are widely scattered. Genotypes with the same names but a different number came from different institutes.

RFLP analysis

Total genomic DNA was extracted from green fresh mature cocoa leaves following the method described by Crouzillat et al. (1996).

A total of 290 probes were screened on 9 cocoa genotypes representing different geographical and genetic origins, and 43 of these showing polymorphism were selected for their distinct hybridization pattern. These probes included 37 *Pst*I genomic library clones, 1 cDNA probe, and 5 probes coming from the amplification of cocoa DNA using either random, consensus or specific primers. Consensus primer sequences were obtained by GCG (Genetic Computer Group, Madison, USA) database comparisons and enabled the amplification of the Rubisco gene [TGCATGiAGGTGTGGCC and TTGTCGAAiCCGATGGA (i = inositol)] and Chalcon synthase gene (CATGATGTACCAiCAiGGiTGCTT and CTiGACATGTTiCCATACTC). Specific primers amplified the cocoa seed storage protein gene [ATGGTGATCAGTAAGTCTCCTTTC and ATAAGCGGAGGCTTTTACAGTG (Spencer and Hodge 1992)] and the cocoa chitinase gene [GCTGAGCAGTGTGGACGGC and GACCACATTCAAGGCCGCC (R. Furtek, personal communication)]. Polymerase chain reaction (PCR) amplifications were performed as described for the RAPD analysis in Crouzillat et al. (1996) except for the concentration of the primers (0.25 μ*M* each) and the magnesium chloride (2 m*M*). The annealing temperature was 45°C. PCRs were loaded on a 1% low-melting agarose gel, and the amplification products were extracted from the gel to be used as a RFLP probe. The restriction of 5 μg of each genomic DNA with *Hin*dIII (10 U/μg) and Southern blot hybridization were performed according to the supplier's recommendations (Appligène).

RAPD analysis

One hundred and forty decamer oligonucleotides purchased from Operon Technologies were screened on 5 cocoa genotypes. A set of 18 decamer oligonucleotides was selected on the basis of the high reproducibility of the patterns and the signal intensity.

Statistical analysis

For each cocoa genotype, RAPD and RFLP profiles were scored by identifying each band. Data were binary coded: 1 for the presence of a band or 0 for its absence. For RFLP and RAPD data sets and a given number of polymorphic bands, 200 genetic distance (GD) matrices were randomly constructed by bootstrap analysis (Efron and Tibshirani 1986). The genetic distances for each pairwise combination in each band subsample was estimated using the complement

Table 1 Cocoa genotypes used in the study

No.[a]	Name	O[b]	Type[c]	H[d]
a	BCH1	E	N	3
b	BCH2	E	N	6
c	BCH3	E	N	0
d	BCH4	E	N	3
e	BCH5	E	N	3
f	BCH6	E	N	0
g	BCH7	E	N	10
h	BCH9	E	N	6
i	BCH10	E	N	6
j	BCH11	E	N	6
k	BCH12	E	N	0
l	BCH13	E	N	6
m	BCH14	E	N	6
n	Catongo	B	LF	0
o	CatongoPS	B	LF	6
p	CC212	CR	F	29
q	CC222	CR	H	45
r	CC231	CR	H	45
s	CCN51	E	H	29
t	CHOCO	E	N	29
u	Chone01-2	E	N	10
v	Criollo46	Nc	C	13
w	Ebc10S401	Col	UF	6
x	EET19	E	N X VA	45
y	EET20	E	N	48
z	EET21	E	N	19
aa	EET40	E	N X VA	39
ab	EET42	E	VA	32
ac	EET43	E	VA	39
ad	EET46	E	N	55
ae	EET48	E	(VA) NXVA	48
af	EET48(2)	E	(VA) NXVA	45
ag	EET48(3)	E	(VA) NXVA	48
ah	EET53	E	N	48
ai	EET58	E	N	55
aj	EET59	E	N	45
ak	EET60	E	N	32
al	EET62	E	(VA) NXVA	39
am	EET62(2)	E	(VA) NXVA	55
an	EET63	E	VA	55
ao	EET66	E	N X VA	55
ap	EET73	E	N	35
aq	EET75	E	VA X VM	48
ar	EET76	E	N	13
as	EET90	E	N	48
at	EET95	E	(VA) NX?	52
au	EET95(2)	E	(VA) NX?	52
av	EET96	E	(VA) NX?	42
aw	EET96(2)	E	(VA) NX?	52
ax	EET103	E	(VA) NX?	52
ay	EET103(2)	E	(VA) NX?	52
az	EET105	E	N	19
ba	EET109	P	UF	23
bb	EET111	Td	T	48
bc	EET116	P	UF	26
bd	EET117	E	VA	13
be	EET141	E	N	45
bf	EET145	E	N	0
bg	EET147	E	N	45
bh	EET153	E	N	26
bi	EET155	E	N	3
bj	EET161	E	N X VA	48
bk	EET162	E	N X VA	48
bl	EET164	E	N X VA	52
bm	EET167	E	N	45

Table 1 Continued

No.[a]	Name	O[b]	Type[c]	H[d]
bn	EET173	E	N	35
bo	EET178	E	N	6
bp	EET187	E	N	3
bq	EET194	E	N	45
br	EET221	E	N	35
bs	EET233	E	VA	23
bt	EET235	E	N X VA	35
bu	EET332	E	UF	13
bv	EET333	E	U	13
bw	EET400	E	U	45
bx	F1(oxdb)		H	32
by	G8	I	C	29
bz	GU275	FG	LF	3
ca	GU293	FG	LF	10
cb	GU302	FG	LF	10
cc	GU305	FG	LF	3
cd	ICS1	Td	T	26
ce	ICS6	Td	T	26
cf	ICS8	Td	T	45
cg	ICS16	Td	T	23
ch	ICS95	Td	T	26
ci	IMC14	P	UF	23
cj	IMC23	P	UF	26
ck	IMC53	P	UF	13
cl	IMC67	P	UF	29
cm	IMC67(2)	P	UF	23
cn	IMC70	P	UF	39
co	LAFI	S	T	26
cp	Morado	Td	T	19
cq	N38	Ng	T	16
cr	NA34	P	UF	23
cs	OC61	V	T	10
ct	OC77	V	T	55
cu	Ostuacan	M	C	23
cv	PA13	P	UF	13
cw	PA35	P	T	19
cx	Pichucalco	M	C	35
cy	POR	V	C	42
cz	Porcelana	V	C	39
da	Pound7	P	UF	16
db	Pound12	P	UF	42
dc	R2	M	T	55
dd	R43	M	T	55
de	R106	M	T	55
df	RB41	B	LF	6
dg	RIM2	M	T	55
dh	RIM23	M	T	55
di	RIM24	M	T	55
dj	RIM44	M	T	55
dk	RIM68	M	C	39
dl	RIM75	M	T	55
dm	RIM76A	M	C	16
dn	RIM88	M	T	55
do	RIM105	M	T	55
dp	RIM117	M	T	55
dq	SA1	E	N	0
dr	SA2	E	N	6
ds	SA3	E	N	6
dt	SA4	E	N	0
du	SA5	E	N	6
dv	SA6	E	N	3
dw	SA7	E	N	0
dx	SA8	E	N	0
dy	SA9	E	N	6
dz	SA10	E	N	6

Table 1 Continued

No.[a]	Name	O[b]	Type[c]	H[d]
ea	SA11	E	N	3
eb	SA12	E	N	3
ec	SA13	E	N	3
ed	SA14	E	N	3
ee	SA15	E	N	3
ef	SA16	E	N	3
eg	SCA6	P	UF	13
eh	SCA6(2)	P	UF	13
ei	SCA12	P	UF	52
ej	SCA12(2)	P	UF	16
ek	SPA9	Col	T	23
el	SNK12	Ca	T	16
em	TAP1	P	UF	13
en	Tenguel 15	E	(VA) NX?	45
eo	UF29	CR	N	29
ep	UF221	CR	T	3
eq	UF296	CR	T	35
er	UF613	CR	T	35
es	UF168	Pa	T	55
et	ZEA218	V	C	35
eu	Chone 01-1	E	N	19
ev	Chone 01-2 (2)	E	N	10
ew	Chone 01-6	E	N	3
ex	Chone 02-6	E	N	16
ey	Chone 02-7	E	N	6

[a] No = Genotype code. Genotypes with the same name but a different number come from different institutes.
[b] O = Country of origin, B, Brazil; Ca, Cameroon; Col, Colombia; CR, Costa Rica; E, Ecuador; FG, French Guyana; I, Indonesia; M, Mexico; Nc, Nicaragua; P, Peru; Pa, Panama; S,; Samoa; Td, Trinidad; V, Venezuela.
[c] Type = Morphological type; C, Criollo; F, Forastero; H, hybrid; LF, low Amazonian Forastero; N, Nacional; T, Trinitario; UF, upper Amazonian Forastero; VA, Venezolano amarillo; VM, Venezolano morado
[d] H = Type of heterozygosity obtained from 31 RFLP probes

to the simple matching coefficient (Gower 1985) based on the following equation: $GD(i,j) = \sum^{N}(i \neq j)/[\sum^{N}(i \neq j) + \sum^{N}(i = j)]$, where GD is the measure of the genetic distance between trees i and j, while $\sum^{N}(i \neq j)$ and $\sum^{N}(i = j)$ are the total number of scores discordant and concordant between trees i and j, respectively, over all N bands considered. For each pair of trees, the variability among the 200 bootstrap experiments was estimated by the coefficient of variation. The mean coefficient of variation of genetic distance of all the pairwise combinations was plotted against the sample size to determine the number of bands required to obtain consistent estimates of genetic relationship between genotypes (Tivang 1992).

For 31 RFLP probes of the 43 studied, the pattern obtained enabled allele frequencies to be determined. Modified Rogers distances (Wright 1978) were then computed and dendrograms were established following the unweighted pair group procedure with arithmetic mean (UPGMA) (Sneath and Sokal 1973).

The gene diversity statistics (Nei 1973), the mean allele number per locus, the percentage of polymorphic loci at 95% and 99% levels of significance, and the observed heterozygosity were calculated using the computer program BIOSYS-1 (Swofford and Selander 1981). The total gene diversity (or total heterozygosity, H_T) in a whole sample can be divided into the gene diversities within (H_S) and between (D_{ST}) subsamples. $G_{ST} = D_{ST}/H_T$ gives the percentage of the total diversity that is due to genetic differences between subsamples. H_T and H_S were obtained using the unbiased estimates of Nei and Chesser (1983). Absolute gene differentiation (D_m) is a standardized measure of gene diversity among subsamples: $D_m = n\, D_{ST}/(n-1)$ where n is the number of subsamples (Nei 1987). For RAPD data, the 1/0 matrix was treated through a Principal Component Analysis [software Statgraphics and Unistat (Uniware)].

Results

Comparison between RFLP and RAPD analyses

The 43 selected probes showing polymorphism revealed 122 fragments, namely 2.8 per probe. Among these, 94 polymorphic bands (77%) were analyzed. The majority of the polymorphic probes (63%) detected two allelic forms. The 18 primers selected amplified 67 bands, giving an average of 3.7 DNA fragments per primer; 40 of these (60%) were polymorphic and thus included in the analysis.

Figure 1 gives the evolution of the mean coefficient of variation of the genetic distance matrices between cocoa genotypes based on the number of RFLP and RAPD bands. The mean coefficient of variation of genetic distance decreased as more bands were added to the data set, and a linear relationship was observed after natural-log transformation of the scales (data not shown). The number of bands required to reach a mean coefficient of variation of 10% was 421 for RFLP and 506 for RAPD, which corresponds, respectively, to 192 RFLP probes and 228 primers. For the RFLP and RAPD fragments studied, the mean coefficient of variation of genetic distance was 21% and 38%, respectively. Despite this variation, the correlation coefficient (R) between RFLP and RAPD genetic distance matrices established for 40 loci (maximum number of loci common to the two molecular studies) was 0.77. Thus, the two molecular methods gave related estimates of the genetic relationship among the genotypes tested.

Of the cocoa genotypes 21.5% and 32.9% remained non-differentiated on the basis of their RFLP and RAPD patterns, respectively. Forty percent of the non-differentiated trees were observed with the two types of molecular markers. Some cocoa accessions supposed to be genetically identical but coming from several

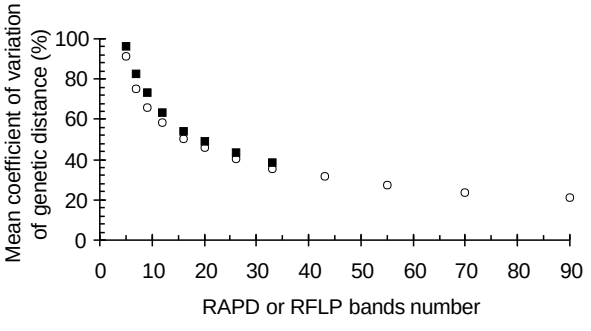

Fig. 1 Plot of the percentage of mean coefficient of variation of genetic distance among the 155 studied cocoa clones versus number of RFLP (*white circle*) and RAPD (*black square*) bands

Table 2 Gene diversity statistics estimated from RFLP data

	Mean allele number per locus	Percentage of polymorphic loci		H_O	H_S	H_T	$D_m(D_{ST})$	G_{ST}
		0.95[a]	0.99[b]					
Morphological groups						0.341	0.080 (0.060)	0.161
Nacional (59)[c]	1.9	64.5	77.4	0.161	0.190			
Forastero (29)	2.3	90.3	100.0	0.188	0.329			
Trinitario (29)	2.0	80.7	93.6	0.383	0.306			
Criollo (9)	1.8	80.7	80.7	0.301	0.313			
Mean				0.258	0.281			
Geographic origins						0.340	0.100 (0.086)	0.228
Ecuador (87)	2.2	67.7	96.8	0.244	0.246			
Mexico (15)	1.6	54.8	54.8	0.477	0.275			
Peru (18)	2.2	80.7	100	0.233	0.319			
Costa Rica (7)	1.8	83.9	83.9	0.318	0.292			
Venezuela (5)	1.8	80.7	80.7	0.361	0.352			
Trinidad (7)	1.8	83.9	83.9	0.304	0.285			
French Guyana (4)	1.1	12.9	12.9	0.065	0.065			
Mean				0.286	0.254			

H_O, Observed heterozygosity; H_T, total diversity; H_S, intragroup diversity; D_{ST}, intergroup diversity; D_m, absolute gene differentiation; G_{ST}, differentiation between groups; H_T, H_S, D_m, D_{ST}, G_{ST} corresponded to the mean values of the totality of the loci

[a, b] The locus is considered as polymorphic if the most common allele has a frequency inferior to 0.95 and 0.99, respectively

[c] Number of genotypes per group are indicated in brackets

institutes could be distinguished. This may be explained by the method of propagation, either vegetative (cuttings) or sexual (seeds).

Organization of genetic diversity within cocoa species

The gene diversity statistics (Nei 1973) evaluated from RFLP data are given in Table 2. Genotypes were classified according to their morphological groups, including Nacional, on the one hand, and their geographic origins on the other. The hybrid trees and their geographic origins, being poorly sampled in this study, were not taken into account.

The average number of RFLP alleles per loci was the highest for Forastero, and most of the loci studied were polymorphic (Table 2). All of the allelic forms detected were only found in the Forastero group, sometimes at a low frequency (<0.05). Two morphes could be considered as being specific to Forastero (allelic frequency higher than 0.05). The intragroup diversities (H_S) of Forastero, Criollo, and Trinitario were similar respectively equal to 0.329, 0.313, and 0.306, whereas the gene diversity estimate of the Nacional group was lower ($H_S = 0.190$). The level of observed heterozygosity (H_O) was higher for Trinitario (0.383) and Criollo ($H_O = 0.301$) than for Forastero (0.188) or Nacional (0.161). The number of RFLP alleles per locus was low for French Guyanese clones (1.1), moderate for Mexican ones (1.6), and higher for Peruvian and Ecuadorian trees (2.2) (Table 2). Rare alleles could be detected essentially in the Peruvian and Ecuadorian genotypes. For the trees of these two countries, all the allelic forms detected were present, except one for Ecuador and two for Peru. The geographic origin of cocoa genotypes also revealed a high heterogeneity of the diversity values. A drastically reduced diversity was observed within the French Guyanese accessions ($H_S = 0.065$), while the highest one was within the Venezuelan genotypes for which the total diversity was distributed mainly within the country ($H_S = 0.352$). The higher diversity value observed in Ecuador compared to Nacional revealed the heterogeneity of the Ecuadorian cocoas, including Nacional, Venezolano amarillo and Silecia (EET332, EET333, EET400) trees.

Discrimination between groups was better when the genotypes were classified according to their geographical origins ($G_{ST} = 0.228$) instead of morphological groups ($G_{ST} = 0.161$). Nevertheless, the intergroup diversity values were low ($D_m = 0.080$ and $D_m = 0.100$ for the morphological and geographical analyses, respectively. This genetic similarity between the three morphological groups was also observed on the dendrogram constructed from the matrix of the modified Rogers distance (Wright 1978) within the morphological groups (Fig. 2a). The Trinitario and Criollo genotypes are indeed tightly close. The Forastero type appears to be less distant from Trinitario ($D = 0.197$)

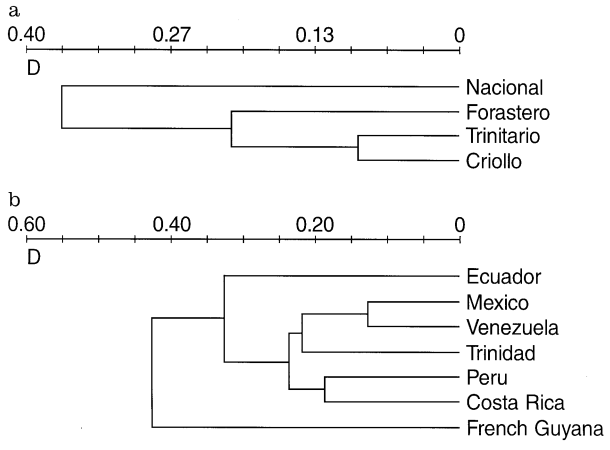

Fig. 2a, b UPGMA dendrogram constructed using the modified Rogers distance (Wright 1978) within morphological groups (**a**) and the geographical origins of *T. cacao* (**b**)

than from Criollo (D = 0.233). On the other hand, the Nacional type seems to be a specific group, distant from the other groups but more related to Forastero (D = 0.299) than to Trinitario (D = 0.381) and Criollo (D = 0.416). Thus, when the Nacional type was not considered in the analysis, the G_{ST} value of the morphological classification fell to 0.052.

According to the matrix of modified Rogers distances calculated between geographic origins, Venezuelan and Mexican accessions appear as the closest groups. Ecuadorian trees share more homologies to Costa Rican (D = 0.274) and Peruvian trees (D = 0.278) than to the others (D > 0.346). GU genotypes appeared highly distant from the others (D > 0.378). The dendrogram (Fig. 2b) shows that Venezuelan and Mexican trees clustered with those of Trinidad, whereas Peruvian and Costa Rican ones were grouped. The hybrid origin of some Costa Rican genotypes from Peruvian clones (as CC212 and CC231) could explain those results.

We could conclude that there is no clear concordance between the classification based on morphological groups or geographic origins and the amounts of differentiation at the molecular (RFLP) level. This could be the consequence of an intense genetic mixing in the complex species promoted by the allogamous habit of cocoa.

Figure 3 shows the representation of the principal components analysis (PCA) based on the RAPD study and the superposition of these molecular data with morphological classification and geographic origins. The variability observed on axis 2 was essentially due to the French Guyanese genotypes, probably because of specific amplified signals being detected for these plants. Therefore, representation based on axes 1 and 3 of the PCA is given here.

As observed on the dendrogram based on morphological classification, a genetic similarity was noted between Criollo and Trinitario. Forastero and Nacional formed heterogeneous groups. Overlapping areas confirmed that a clear distinction between groups is not obtained at molecular level. Axis 1 characterized most of the Ecuadorian genotypes and axis 3 most of the Peruvian ones. The Ecuadorian Forastero (EET332, EET333 and EET400) genotypes coming from the east side of the Andes seemed closer to the Peruvian SCA genotypes than to other Ecuadorian trees. Moreover, the majority of Venezuelan, Mexican and Trinidadian trees were grouped and shared therefore common RAPD variability.

The study showed that the cocoa trees labelled SA and BCH grouped together as well as with some EET genotypes. These trees were characterized by a low heterozygosity level compared to most of the other Nacional or Ecuadorian genotypes. Thus, when Nacional genotypes were grouped on the basis of RFLP heterozygosity level, the high-heterozygous genotypes (more than 20%) were more related to Forastero (D = 0.186) and Trinitario (D = 0.238) genotypes than to the low-heterozygous Nacional ones (D = 0.284) (data not shown). The low-heterozygous Ecuadorian trees constituted the most distant group, but it was closer to the high-heterozygous one than to other groups.

Globally, the use of either RFLP and RAPD markers resulted in the same conclusions even if each of the PCA revealed some specificities (data not shown). While RAPD could clearly provide evidence of the specificity of French Guyanese trees, RFLP could clearly differentiate the IMC and SCA Peruvian clones.

Discussion

Comparison between RFLP and RAPD

In this study, the genetic variability of cocoa accessions was assessed using RAPD and RFLP. The two types of DNA markers gave highly related estimates of genetic relationship between trees. In *Brassica*, Thormann et al. (1994) also observed, especially at the intraspecific level, a high correlation between RFLP and RAPD when comparing the genetic similarity matrices (R = 0.97). However, because of their codominant nature, RFLPs are more efficient than RAPDs for evaluating genetic parameters (Liu and Furnier 1993). The discrepancies between estimations done with RFLP and RAPD, when observed, may result from a difference in the nature of the DNA fragments. For example, in a study based on RAPDs in *Theobroma cacao*, repeated and single-copy sequences did not structure the variability in the same way (N'goran et al. 1994). Moreover each method had different properties and would identify preferably rearrangements or point mutations on the basis of the number of DNA bases analyzed and the

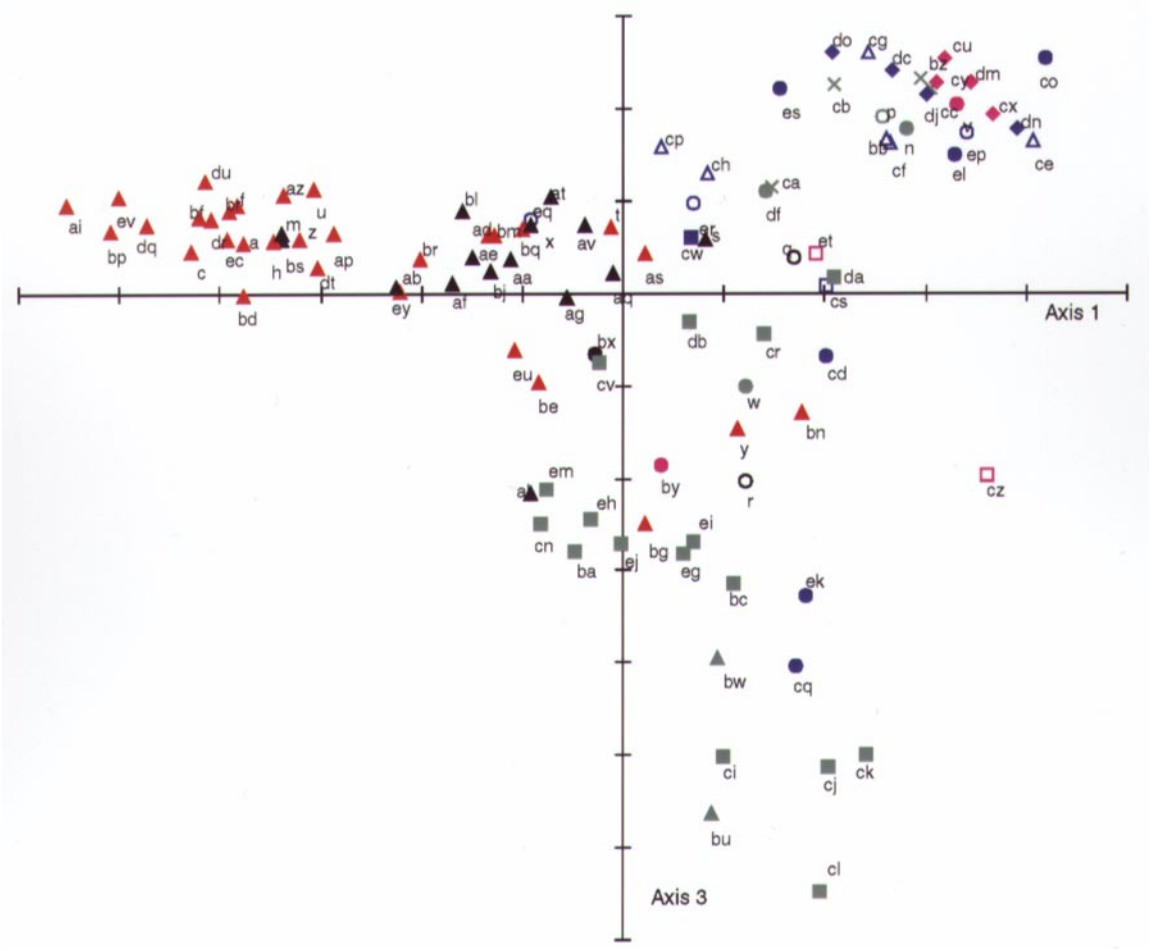

Fig. 3 Principal Component Analysis based on RAPD study, 155 clones and 40 RAPD loci. The letters plotted represent individual accessions and correspond to the ones listed in Table 1. The first and third principal components accounted for 15.6% and 9.8%, respectively, of the variance for the RAPD analysis. a = b = j = k = l = dw = dy = ea = ee, c = d = e = g = i = ds = dz, m = eb, n = o, x = au = aw = ay, aa = ac, ad = ai = am = an = ao, ae = eo, af = ah = ai = aj = ex, at = ax = en, bf = bh = bo = bv = ef, bj = bk, br = bt, bu = bv, ci = cm, ct = cy = dk, dc = dd = de = dg = dh = di = dp, dj = dl, dq = dx = dw, ec = ed. *Red* Nacional, *green* Forastero, *blue* Trinitario, *pink* Criollo, *black* undefined, ○ Costa Rica, ▲ Ecuador, × French Guyana, ◆ Mexico, ■ Peru, △ Trinidad, □ Venezuela, ● other countries

length of the fragments detected (Halldé et al. 1994). Heun and Helentjaris (1993) also postulated that DNA amplification could be selective and influenced by the other regions of the genome.

Organization of cocoa genetic diversity

Despite the reduced sampling of some groups, we observed genetic diversity within the three traditional botanical morphological groups of cocoa. However, the genetic distance between those groups was reduced with respect to that of Nacional, and thus the identification of these groups at a molecular level will be difficult. Genetic similarity between Trinitario and Criollo was observed by Laurent et al. (1994) who hypothesized a return of the hybrid Trinitario form to the parental type by successions of backcrosses, preferentially to the Criollo type in America and to the Lower Amazon Forastero type in Africa. A multiple origin of Trinitario is also suggested by the very high diversity of this group. From an another point of view, the values observed in our study for Ecuadorian and Peruvian clones did not provide evidence of a higher diversity compared to other countries even if a greater allelic richness was noted. Our results do not allow us to support either one of the two hypotheses developed on the differentiation of the *Theobroma cacao* species, the presence of specific allelic forms being found only in Forastero genotypes being in favor of the Cheesman (1944) hypothesis, and the high genetic diversity values of Criollo and Forastero supporting the Cuatrecasas (1964) assumption. Cheesman (1944) proposed a center of diversity in Ecuador and Peru (Cheesman 1944) followed by a dispersion of Forastero to the East and of Criollo to the North, whereas Cuatrecasas (1964) made the assumption of an independent differentiation of these two morphological groups on the two sides of the

Panama isthmus. The data of Laurent et al. (1994) supported the latter hypothesis. Actually, continuous hybridizations and introgressions between material have probably erased the origins of the cocoa clones. The presence of an autoincompatible system (S) in cocoa probably promotes outcrossing possibilities. Moreover, the situation is accentuated by the man-assisted movement of pods and seedlings from region to region, increasing the likelihood of genetic exchange between genotypes of different origins. Ronning and Schnell (1994) noted that historically cocoa germplasm has been extensively exchanged between breeding stations. For example, the UF genotypes come from a selection of the Atlantic Coast of Costa Rica, among cocoas from Trinidad and Venezuela or from derived forms of Nacional from Ecuador. Moreover, the Pleistocene climatic changes did probably influence the genetic evolution of the *Theobroma cacao* species in creating refuge areas (Simpson and Haffer 1978) and allopatric divergence. This phenomenon could explain the differentiation observed on the dendrogram between the French Guyanese genotypes and other groups. Lanaud (1987) and Laurent et al. (1994) noted the specificity of those GU trees. Other refuge areas have been observed, particularly in the Upper Amazon region (Prance 1973). Populations reexpanded from these isolates, sometimes giving rise to hybrid forms (Simpson and Haffer 1978). A recent molecular study revealed that some wild Yucatan plants may be relics of Mexican cocoa cultivation by the Maya population and would be a genetically distinct material which does not exist in the modern germplasm collections essentially constituted by South American origin cocoas (de la Cruz et al. 1995). If this is confirmed it would support the Cuatrecasas hypothesis.

Nacional specificity

Our data also show that Nacional has a lower diversity index than other morphological groups and must not be considered as an homogenous group. Its original status, which appeared here clearly, is inconsistent with the classifications of Soria (1970), Engels (1986), and Enriquez (1992) who integrated it into either the Forastero or the Criollo group. Previous molecular analyses on three Nacional accessions illustrated just how difficult it is to classify Nacional. Indeed, while data on the cytoplasmic DNA polymorphism have shown a similarity between Nacional and Criollo (Laurent et al. 1993a), results on nuclear DNA polymorphism have led to the inclusion of Nacional in the Upper Amazon or Trinitario groups (N'goran et al. 1994). Nacional has also been shown to have a specific nuclear rDNA polymorphism (Laurent et al. 1993b). The origin of the cultivated Nacional cocoa is unknown. This cocoa type is considered to be a native of Ecuador, either from the eastern slopes of the Andes from where some pods may have been brought by Indians centuries ago (Rorer, reported by Van Hall 1932) or from the banks of rivers of littoral zones where wild Nacional trees were discovered in the virgin forests in the 1930s (Pound 1938). At the end of the last century, the genotypes called Venezolano were introduced from Trinidad to Ecuador. Hybridizations and introgressions with Nacional may explain the heterogeneity of the Nacional group and the different levels of RFLP heterozygosity observed in this group. The low heterozygous SA and BCH clones issued from two old plantations and may represent a part of the original Nacional pool. Therefore, the low heterozygosity level could be a common trait of the pure Nacional type. Moreover, Ecuadorian trees may have differentiated on each side of Andes, since the genotypes of the most eastern part of Ecuador, Silecia trees (EET332, EET333 and EET400), clustered with the Peruvian genotypes and not with the Ecuadorian (Nacional) ones.

Gene differentiation and management of a cocoa germplasm collection

Cocoa germplam collections are located in Trinidad (University of West Indies) and Costa Rica (CATIE). However, the long-term conservation and the use of these genetic resources remain problematic. Therefore, efficient strategies have to be developed. A solution to this problem is more manageable collections with a reduced size, i.e. core collection, (Frankel 1984), which should enhance the interest of breeders in using the genetic diversity available in these collections. Several strategies for the constitution of a core collection with a minimum loss of diversity have been proposed (Brown 1989a) using information on passport data (species, variety and place of origins), morphological traits, and molecular polymorphisms. The strategy efficiency varies from species to species. Diwan et al. (1995) observed a better representation of the US annual *Medicago* species germplasm collection when the core collection was established from cluster analysis based on phenotypic or genetic diversity rather than from passport data, random selection of accessions, or by cluster analysis of geographical regions.

In the same way, our results bring useful information for the constitution of a cocoa-tree core collection. Because of the low values of G_{ST} and genetic distances observed, independent sampling within morphological groups or geographical origins cannot be an optimal strategy for reducing the size of the collection while eliminating redundancies. The Relative Diversity method (Diwan et al. 1994) could be applied to the cocoa collection. This method proposes to take into account the number of clusters in a dendrogram established from morphological or molecular data to select the number of accessions without taking into consideration an a priori classification.

Using isozymes in cocoa Ronning and Schnell (1994) found a higher gene differentiation among geographical origins than among morphological types even if the value was not significantly different. They suggested a "stratified random sampling" (Brown 1989b) over morphological types. However, as this sampling is generally preferred when groups are distinct, it does not seem fully satisfactory for cocoa. Bekele and Bekele (1996) detected a link between the geographic origins and accession grouping of 100 genotypes of the International Cocoa Gene bank of Trinidad and suggested collecting germplasm over a wide geographical range to capture as much diversity as possible. Lawrence et al. (1995) estimated that the size of a sample required in an outcrossing population to give a probability of 0.9999 of conserving an allele at a single locus, whose frequency is 0.01, is 450. For 43 loci the probability is 0.9957. However, a multifactor choice should be realized, and qualitative and agronomic traits as disease resistance should be considered.

Overall, on the basis of diversity indices, populations of outbreeding plants are much less differentiated than autogamous species because of the occurrence of gene flow. They contain low-frequency alleles which contribute little to diversity indices but which promote differentiation between them (Brown 1989a).

Despite this continuous genetic background, some cocoa types present interesting particularities. Studies on a larger sample might characterize other groups with more specificity.

Acknowledgements The authors are grateful to Dr. J. Tivang (University of Wisconsin, Madison) for the bootstrap analysis and for his helpful suggestions. They also want to thank Dr. A. Deshayes and Dr. T. Leroy for valuable comments on the manuscript and Dr. M. Paillard for her English correction. Estelle Lerceteau was supported by a grant from the Ministère de l'Enseignement supérieur et de la Recherche.

References

Bekele FL, Bekele I (1996) A sampling of the phenetic diversity of cacao in the international cocoa gene bank of Trinidad. Crop Sci 36 : 57–64

Brown AHD (1989a) The case for core collections. In: Brown AHD, Frankel OH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 136–156

Brown AHD (1989b) Core collections: a practical approach to genetic resources management. Genome 31 : 818–824

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

Cope FW (1976) Cacao, *Theobroma cacao* (Sterculiaceae). In: Simmonds NW (ed) Evolution of crop plants. Longman, London New York, pp 285–289

Crouzillat D, Lerceteau E, Pétiard 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 : 205–214

Cuatrecasas J (1964) Cacao and its allies: a taxonomic revision of the genus *Theobroma*. Bull US National Museum, Smithsonian Institution, Washington 35 : 379–614

de la Cruz M, Whitkus R, Gómez-Pompa A, Mota-Bravo L (1995) Origins of cacao cultivation. Nature 6530 : 542–543

Diwan N, Bauchan GR, McIntish MS (1994) A core collection for the United States annual *Medicago* germplasm collection. Crop Sci 34 : 279–285

Diwan N, McIntosh MS, Bauchan GR (1995) Methods of developing a core collection of annual *Medicago* species. Theor Appl Genet 90 : 755–761

Efron B, Tibishirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1 : 54–77

Engels JMM (1986) The identification of cacao cultivars. Acta Hortic 182 : 195–202

Engels JMM (1992) The use of botanical descriptors for cocoa characterisation: catie experiences. Edited by Haines Clark & Co. (UK). Int Workshop Conservation Characterisation Utilisation Cocoa Genetic Resources 21st Century. The Cocoa Research Unit, The University of the West Indies, Port-of-Spain, Trinidad, pp 69–76

Enriquez GA (1992) Characteristics of cacao "nacional" of Ecuador. Edited by Haines Clark & Co. (UK). Int Workshop Conservation Characterisation Utilisation Cocoa Genetic Resources 21st Century. The Cocoa Research Unit, The University of the West Indies, Port-of-Spain, Trinidad, pp 269–278

Figueira A, Janick J, Morris L, Goldsbrough P (1994) Re-examining the classification of *Theobroma cacao* L. using molecular markers. J Am Soc Hortic Sci 119 : 1073–1082

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber WK, Llimensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on Man and society. Cambridge University Press, Cambridge, pp 161–170

Gower JC (1985) Measures of similarity, dissimilarity, and distance. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 5. Wiley, New York, pp 397–405

Halldén C, Nilsson NO, Rading IM, Säll T (1994) Evaluation of RFLP and RAPD markers in a comparison of *Brassica napus* breeding lines. Theor Appl Genet 88 : 123–128

Heun M, Helentjaris T (1993) Inheritance of RAPDs in F_1 hybrids of corn. Theor Appl Genet 85 : 961–968

IBPGR Secretariat (1981) Report: IBPGR working group on genetic resources of cacao. ACP, IBPGR/80/56 March 1981, Rome

Lanaud C (1987) Nouvelles données sur la biologie du cacaoyer (*Theobroma cacao* L.): diversité des populations, systèmes d'incompatibilité, haploïdes spontanés. Leurs conséquences pour l'amélioration génétique de cette espèce. PhD thesis, Paris IX

Laurent V, Risterucci AM, Lanaud C (1993a) Chloroplast and mitochondrial DNA diversity in *Theobroma cacao*. Theor Appl Genet 87 : 81–88

Laurent V, Risterucci AM, Lanaud C (1993b) 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

Lawrence MJ, Marshall DF, Davies P (1995) Genetics of genetic conservation. I. Sample size when collecting germplasm. Euphytica 84 : 89–99

Liu Z, Furnier GR (1993) Comparison of allozyme, RFLP, and RAPD markers for revealing genetic variation within and between trembling aspen and bigtooth aspen. Theor Appl Genet 87 : 97–105

Lockwood G (1985) Genetic resources of the cocoa plant. Span 28 : 14–16

Marshall DR (1989) Limitations to the use of germplasm collections. In: Brown AHD, Frankel OH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 105–120

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

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New-York
Nei M, Chesser RK (1983) Estimation of fixation indices 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
Pound FJ (1938) Variety of cacao used for commercial purposes. In: Cacao and witchbroom disease of South America. Trinidad and Tobago, Port-of-Spain, pp 21–71
Prance GT (1973) Phytogeographic support for the theory of Pleistocene forest refuges in the Amazon Basin, based on evidence from distribution patterns in Caryocaraceae, Chrysobalanaceae, Dichapetalaceae and Lecythidaceae. Acta Amazon 3 : 5–28
Ronning CM, Schnell RJ (1994) Allozyme diversity in a germplasm collection of *Theobroma cacao* L J Hered 85 : 291–295
Simpson BB, Haffer J (1978) Speciation patterns in the Amazonian forest biota. Annu Rev Ecol Syst 9 : 497–518
Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco
Soria J (1970) Principal varieties of cocoa cultivated in tropical America. Cocoa Growers' Bulletin 19 : 12–21
Spencer ME, Hodge R (1992) Cloning and sequencing of the cDNA encoding the major storage proteins of *Theobroma cacao*. Identification of the proteins of the vicilin class of storage proteins. Planta 186 : 567–576
Swofford DL, Selander RB (1981) BIOSYS-1: A FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics. J Hered 72 : 281–283
Thormann CE, Ferreira ME, Camargo LEA, Tivang JG, Osborn TC (1994) Comparison of RFLP and RAPD markers to estimating genetic relationships within and among cruciferous species. Theor Appl Genet 88 : 973–980
Tivang JG (1992) Sampling variance of molecular marker data using the bootstrap procedure. MSc thesis, University of Wisconsin-Madison, Madison, Wis.
Van Hall CJJ (1932) Cacao, 2nd edn. Macmillan, London, pp 304–320
Warren JM (1994) Isozyme variation in a number of populations of *Theobroma cacao* L. obtained through various sampling regimes. Euphytica 72 : 121–126
Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18 : 6531–6535
Wright S (1978) Evolution and the genetics of populations, vol 4: variability within and among natural populations. University of Chicago Press, Chicago