THE UNIVERSITY OF THE WEST INDIES

COCOA RESEARCH UNIT

REPORT FOR 1996

THE UNIVERSITY OF THE WEST INDIES
ST. AUGUSTINE, TRINIDAD & TOBAGO

COCOA RESEARCH UNIT

REPORT FOR 1996

Genetic Diversity Assessment of *Theobroma Cacao* L. using Iso-enzyme and RAPD analyses

O. Sounigo, Y. Christopher and R. Umaharan

The ICG,T contains 88 cacao "populations" (term used here to designate accessions obtained from the same collecting expedition or selected at the same research centre), whose representation varies largely in size (from less than 5 to more than 200 accessions). Isozyme Electrophoresis and RAPD techniques were used to estimate the genetic diversity existing within and among 28 and 11 of these cacao "populations", respectively.

Material and Methods

For Isozyme Electrophoresis, flush leaves were extracted from 482 clones representing 28 populations present at UCRS and in the St. Augustine campus fields. In order to use the RAPD technique, mature leaves from trees representing 11 populations were collected from 127 clones present in Marper, the UCRS and the campus fields. Table 1 contains information on the populations studied.

Isozyme Electrophoresis

Five enzymatic systems were used (ACP, IDH, ADH, PGI and MDH) allowing studies on six polymorphic loci, according to Lanaud (1986).

RAPD analysis

DNA was extracted according to a modified version of the method of Edwards *et al.* (1991) (Johnson *et al.*, 1992). The DNA was quantitated and standardized based on fluorimeter measurements, and amplification performed according to the method of Christopher and Sounigo (1995). Amplifications were performed with 17 Operon primers, which revealed 49 stable polymorphic loci. The number of markers produced per primer varied from 1 to 9 giving rise to two to 24 molecular profiles per primer (Table 2).

Statistical analyses

Scoring of data

In the case of Isozyme Electrophoresis, the banding patterns obtained were interpreted as genotypes, according to the genetic model proposed by Lanaud (1986), while in the case of RAPD analysis, the markers were noted "present" or "absent".

Table 1 Characteristics of the studied populations.

Population	Origin	Group	Sample size	
			I.E	RAPD
ACT	Trinidad	mixed	7	
AM	Ecuador	Refractario	11	
B	Ecuador	Refractario	17	8
CC	Costa Rica	Trinitario	8	
CL/CLM	Ecuador	Refractario	20	9
DOM	Dominica	Trinitario	11	
EET	Ecuador	mixed	10	
GS	Grenada	Trinitario	15	
GU	French Guyana	Forastero	16	12
ICS	Trinidad	Trinitario & other	32	11
IMC	Peru	Forastero	24	12
JA	Ecuador	Refractario	21	11
LCT-EEN	Ecuador	Forastero	16	22
LP	Ecuador	Refractario	18	
LX	Ecuador	Refractario	12	
MAR	Martinique	Trinitario	11	
MO	Peru	Forastero	8	9
MOQ	Ecuador	Refractario	14	
NA	Peru	Forastero	37	12
P	Peru	Forastero	13	9
PA	Peru	Forastero	21	9
R	Mexico	Trinitario	10	
SC	Venezuela /Colombia	Trinitario?	9	
SCA	Peru	Forastero	9	11
SJ	Ecuador	Refractario	10	
SLA	Ecuador	Refractario	10	
TRD	Trinidad	Trinitario	16	
UF	Costa-Rica	Trinitario	12	

Table 2 Degree of polymorphism exhibited by the 17 Operon primers used for RAPD analyses.

Primer	Number of polymorphic markers	Number of molecular profiles
opA10	3	6
opB05	4	7
opB12	2	6
opB17	2	4
opC04	2	4
opD07	2	3
opF09	2	4
opG06	2	4
opG19	1	2
opG20	3	5
opH05	3	6
opM17	1	2
opO15	4	14
opP14	9	24
opR03	1	2
opR09	2	4
opR19	6	21

Measurement of genetic diversity

Isozyme Electrophoresis (I.E).

The diversity existing within populations was estimated from the 482 clones, for which data were available for at least four systems. Several indices were used to evaluate the diversity:

% of polymorphic loci;
% of heterozygosity (direct count);
average number of alleles per locus;
Shannon diversity index, based on genotypic frequencies (using BIODIV software); and
Nei diversity index, based on allelic frequencies (using BIOSYS-1 software).

The hypothesis of the Hardy-Weinberg equilibrium was tested for each population at each locus using BIOSYS software.

RAPD Analysis

The Shannon diversity index, calculated using band frequencies obtained for each primer, was used to evaluate the diversity existing within each population.

Partitioning of diversity

The total diversity was partitioned "within" and "among" populations, using the general formulae:

$$H_{within} = H_{pop} / H_{total} \quad \text{and} \quad H_{between} = (H_{total} - H_{pop}) / H_{total}$$

These formulae were applied to both the Shannon and Nei indices in the case of Isozyme Electrophoresis, and to the Shannon index in the case of RAPD analysis.

Description of genetic diversity

In the case of Isozyme Electrophoresis, cluster analysis was performed on Nei distances data using the UPGMA method for the 482 clones. For RAPD analysis, 89 clones completely characterized for 40 markers were used in a Principal Component Analysis, and a cluster analysis based on Balakrishnan and Sanghvi (1968) distances using the UPGMA method.

Fingerprinting

Different parameters were observed in order to evaluate the discriminative values of the two techniques: percentage of unique genotypes, number of groups of undifferentiated genotypes and the sizes of these groups, and the consequences of these values for the detection of mislabelled trees are described.

The efficiency of Isozyme Electrophoresis for fingerprinting was assessed in a sample of 215 accessions for which complete data were available for the 5 systems, while 40 RAPD markers were used to discriminate among 89 clones.

Results

Estimation of genetic diversity

Isozyme Electrophoresis

As shown in Table 3, striking differences were observed between populations for the level of genetic diversity. Some populations showed very limited diversity in terms of allelic richness as well as heterozygosity, such as: GU, MAR and DOM. On the other hand, other populations displayed a high level of diversity, for both allelic richness and heterozygosity, such as UF, MOQ, ICS, EET and AM. Some other populations contained a high level of heterozygosity but a low level of genotypic diversity or of allelic richness, as in the cases of R and SC populations. In contrast, some populations, such as LCT-EEN and SLA had a low level of heterozygosity and a high level of allelic richness.

RAPD

Table 4 shows that LCT-EEN, CL and IMC populations present a high level of diversity, while a low level of diversity is found within the B, NA, JA and GU populations.

Table 3 Diversity of the different populations, estimated by several indices calculated from I.E data and the Shannon diversity index for RAPD analysis. (The numbers in parentheses represent the standard errors).

Population	% polymorphic loci (I.E)	% heterozygosity (I.E)	Mean number of alleles per locus (I.E)	Shannon diversity index (I.E)	Nei diversity index (I.E.)	Shannon diversity index (RAPD)
ACT	66.7	23 (13)	*1.8 (0.3)*	0.5	0.29 (0.11)	
AM	83.3	40 (10)	2.2 (0.3)	**0.84**	**0.40 (0.11)**	
B	83.3	36 (11)	2.3 (0.3)	**0.88**	**0.41 (0.1)**	*0.29*
CC	83.3	46 (16)	2.2 (0.3)	0.6	**0. 40 (0.11)**	
CL	83.3	33 (11)	**2.5 (0.4)**	**0.92**	**0.39 (0.1)**	**0.71**
DOM	66.7	*16 (6)*	*1.8 (0.3)*	*0.38*	*0.16 (0.06)*	
EET	83.3	43 (13)	2.3 (0.4)	**0.87**	**0.41 (0.11)**	
GS	66.7	30 (10)	2.0 (0.3)	0.58	0.26 (0.09)	
GU	*33.3*	*13 (9)*	*1.5 (0.3)*	*0.34*	*0.14 (0.1)*	*0.5*
ICS	83.3	42 (11)	**2.5 (0.3)**	**0.91**	**0.41 (0.08)**	0.6
IMC	66.7	31 (10)	2.3 (0.4)	0.6	0.27 (0.08)	**0.72**
JA	83.3	35 (11)	2.3 (0.4)	**0.93**	**0.40 (0.1)**	*0.47*
LCT-EEN	**100**	29 (6)	**2.7 (0.3)**	**1.04**	**0.44 (0.07)**	**0.85**
LP	66.7	35 (12)	2.0 (0.4)	0.73	0.34 (0.12)	
LX	66.7	37 (14)	2.0 (0.4)	0.61	0.35 (0.11)	
MAR	*33.3*	*6 (4)*	*1.3 (0.2)*	*0.16*	*0.06 (0.04)*	
MO	83.3	31 (10)	2.2 (0.3)	0.71	0.36 (0.11)	0.55
MOQ	83.3	44 (13)	2.3 (0.4)	**0.89**	**0.42 (0.11)**	
NA	83.3	*19 (8)*	2.3 (0.2)	0.74	0.33 (0.07)	*0.42*
P	83.3	*18 (10)*	2.3 (0.3)	0.55	0.27 (0.05)	
PA	83.3	*18 (6)*	2.0 (0.3)	0.67	0.29 (0.08)	0.63
R	83.3	**67 (21)**	*1.8 (0.2)*	*0.08*	**0.40 (0.09)**	
SC	83.3	**66 (18)**	2.2 (0.3)	0.38	**0.44 (0.09)**	
SCA	83.3	30 (9)	2.2 (0.3)	0.79	0.36 (0.1)	0.6
SJ	66.7	41 (13)	*1.8 (0.3)*	0.62	0.35 (0.11)	
SLA	83.3	26 (10)	2.3 (0.4)	**0.84**	**0.43 (0.1)**	
TRD	83.3	22 (8)	2.0 (0.3)	0.60	0.29 (0.07)	
UF	**100**	**50 (12)**	**2.5 (0.3)**	0.77	**0.43 (0.06)**	

Partitioning of the diversity

The values obtained with the Nei indices obtained from I.E data were as follows:

H_{within} = 0.77 and $H_{between}$ = 0.23

The values calculated with the Shannon indices from RAPD data were as follows:

H_{within} = 0.55 and $H_{between}$ = 0.45

Hardy-Weinberg equilibrium

In 80% of the combinations of loci/population, the Hardy-Weinberg equilibrium was maintained allowing analyses based on allelic frequencies, in the case of I.E studies.

Cluster and multivariate analyses

Isozyme Electrophoresis

An Ascending Hierarchical Classification based on pairwise Nei distances using the UPGMA method produced the dendrogram presented in Figure 1. For d = 0.2, two major groups were observed:

- one grouping containing all of the Trinitario and Refractario populations, including also the IMC population;
- one grouping containing the GU and LCT-EEN populations and all the Forastero populations, except IMC.

For d = 0.15, the first major group was split into two sub-groups:

- the first contained all of the Trinitario populations (except ICS and CC);
- and the second consisted of all of the Refractario populations, and included the IMC, CC and ICS populations.

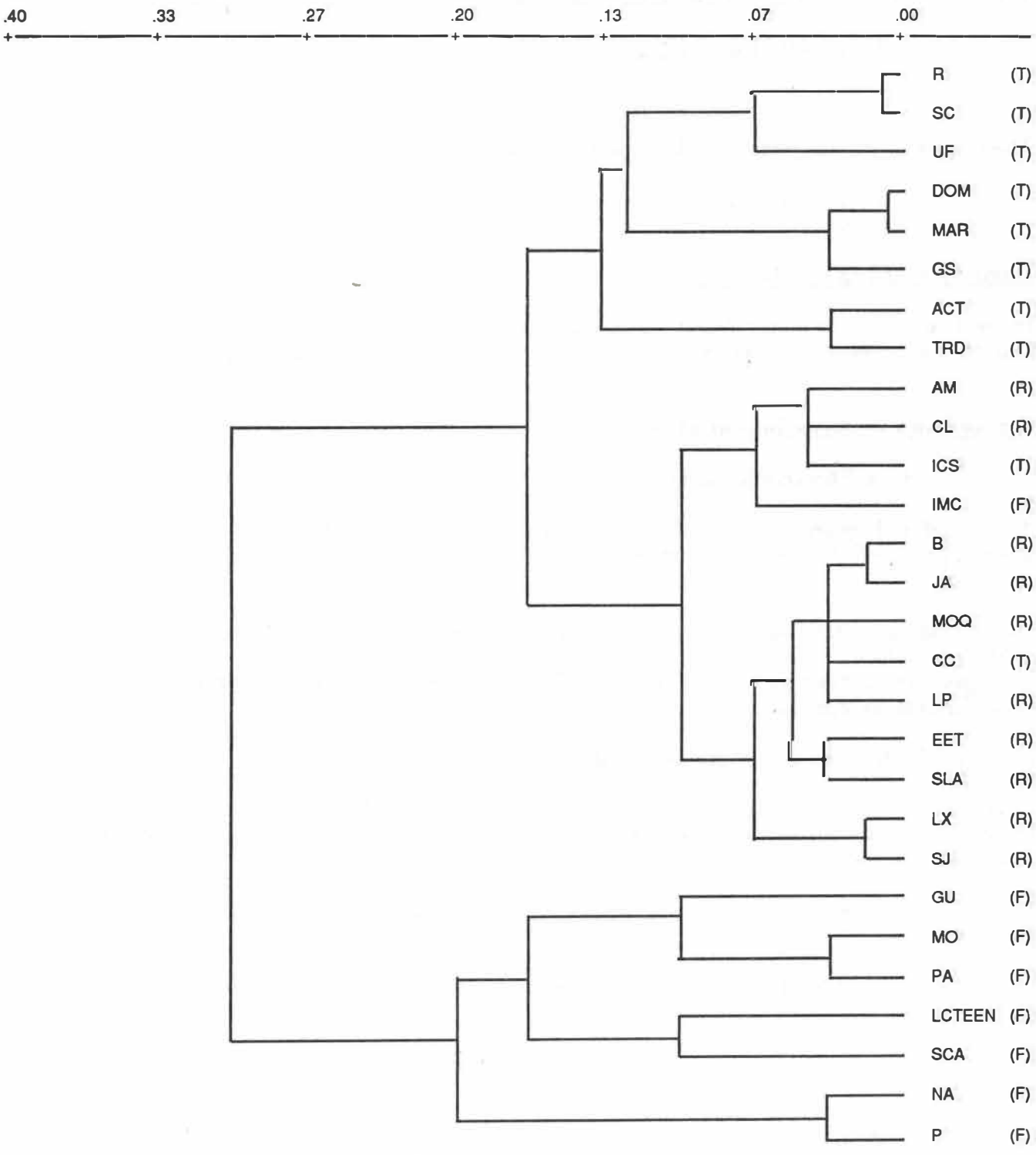
For d = 0.15, the second major group was split into three sub-groups:

- the first comprising of the GU, MO and PA populations;
- the second the LCT-EEN and SCA populations; and
- the third the NA and P populations.

For d = 0.1, the Trinitario group was split into three sub-groups, according to geographical origins:

- one grouping with the Central and South American Trinitarios (R, SC and UF);
- one grouping with the Caribbean Trinitarios (DOM, MAR and GS);
- one grouping with the Trinidadian populations (ACT and TRD).

Figure 1: Dendrogram based on Cluster Analysis (UPGMA Method) of the 28 cacao populations under study.



RAPD

An Ascending Hierarchical Classification based on Balakrishnan and Sanghvi distances using the UPGMA method produced the dendrogram presented in Figure 2.

At d = 4.5, a clustering of the three Refractario populations (B, CL and JA), on one hand, and of the five Forastero populations (IMC, NA, MO, PA and SCA) on the other hand, can be observed. The three other populations (ICS, LCT-EEN and GU) form three different groups.

At d = 4, the Forastero group was split into two sub-groups: one composed of IMC and NA populations and the other one by MO, PA and SCA populations.

The results of the Principal Component analysis are illustrated in Figures 3 and 4, representing the plots defined by the first two axes (28.9 % of the total diversity) and by the first and third axes (25.8% of the diversity), respectively. Although no complete discrimination is made between the groups defined by the cluster analysis, both plots show a clear separation between GU (f) and ICS (d) populations, both populations being confined to two clearly defined and different areas of the two plots. The first axis allows a fairly clear, but not perfect separation between Refractario (a) and Forastero (b) populations. The first two axes allow a fairly clear separation between ICS (d) and Refractario (a) populations. Both plots indicate a fairly clear separation between the two Forastero sub-groups (b and c), group b spreading much less than group c. The first and third axes show a clear separation between the LCT-EEN (e) and ICS (d) populations.

Fingerprinting

Isozyme Electrophoresis

From the 215 accessions used for this estimation, 120 different genotypes could be distinguished. Only 54 (25.15%) accessions were uniquely fingerprinted. For each of the other genotypes, fingerprints were shared by a number of accessions ranging from 2 to 11. Considering these data, the probability of having two randomly chosen accessions sharing the same genotype was calculated and found to be very low: P = 0.01. This means that mislabelling that may have occurred randomly would be detected in 99% of the cases. On the other hand, if this mislabelling problem occurred between trees belonging to the same population, the power of detection would decrease dramatically in some populations, the lowest values being for R (0%!) and MAR (42%) populations. Conversely, this power of detection would still remain high in the case of other populations such as LCT-EEN (98%), ICS (98.5%), TRD (97%), NA (97.4%) and SJ (93%).

Figure 2: Dendrogram obtained from a cluster analysis (using the UPGMA method) performed on Balakrishnan and Sanghvi distances, calculated from frequencies of RAPD markers.

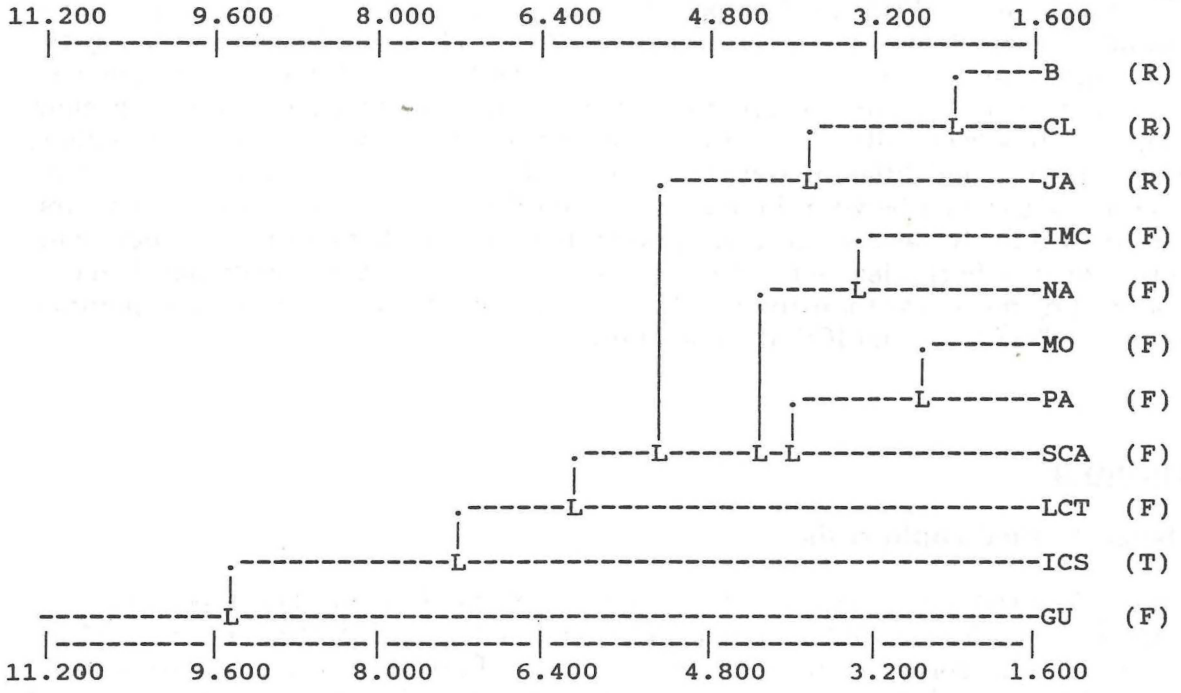


Figure 3: Plan defined by the first two axes of a Principal Component Analysis, performed on RAPD data obtained on 89 clones, using 49 markers. a = Refractario (B,CL and JA), b = IMC and NA, c = MO,PA and SCA, d = ICS, e = LCTEEN, f = GU.

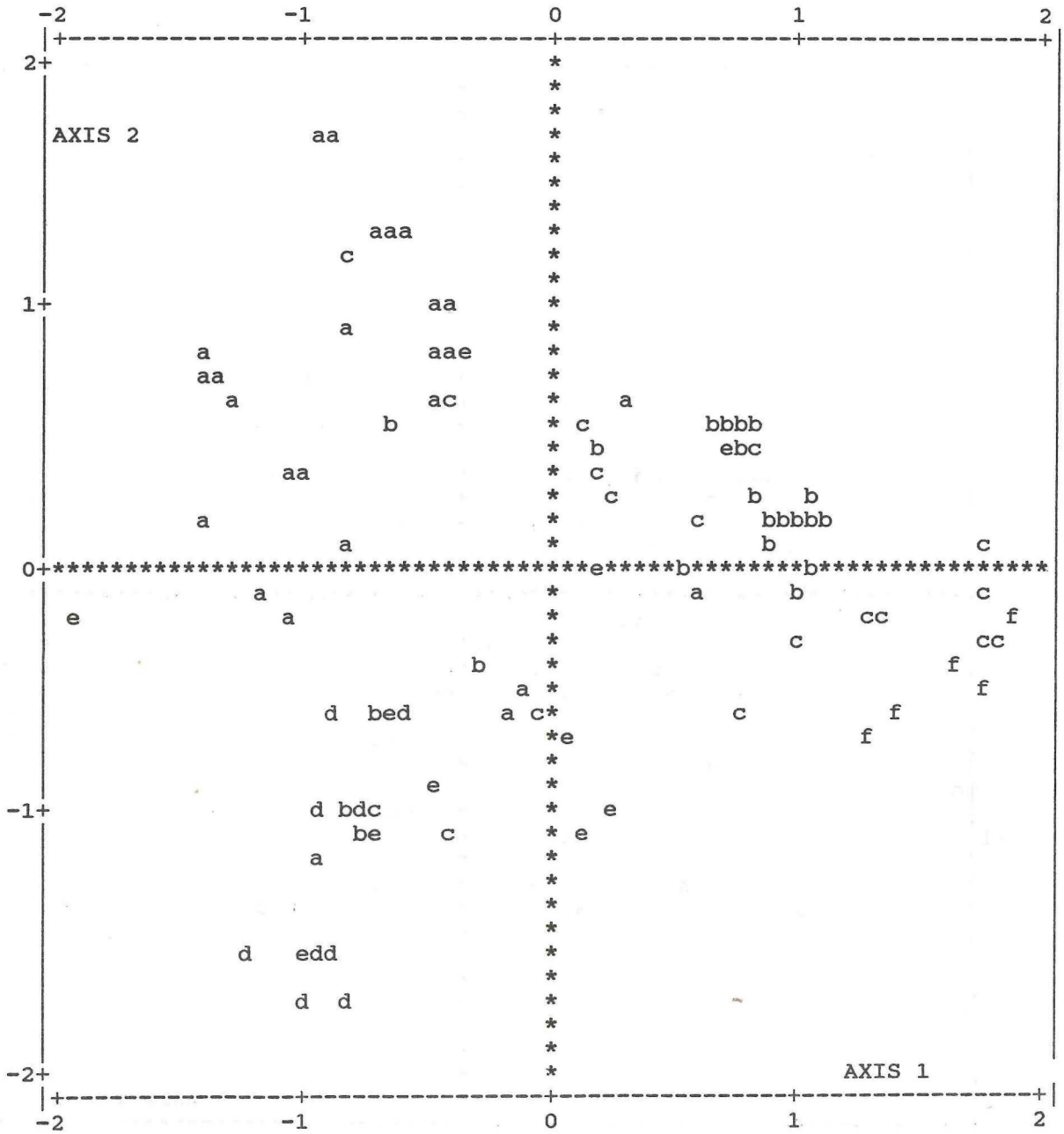
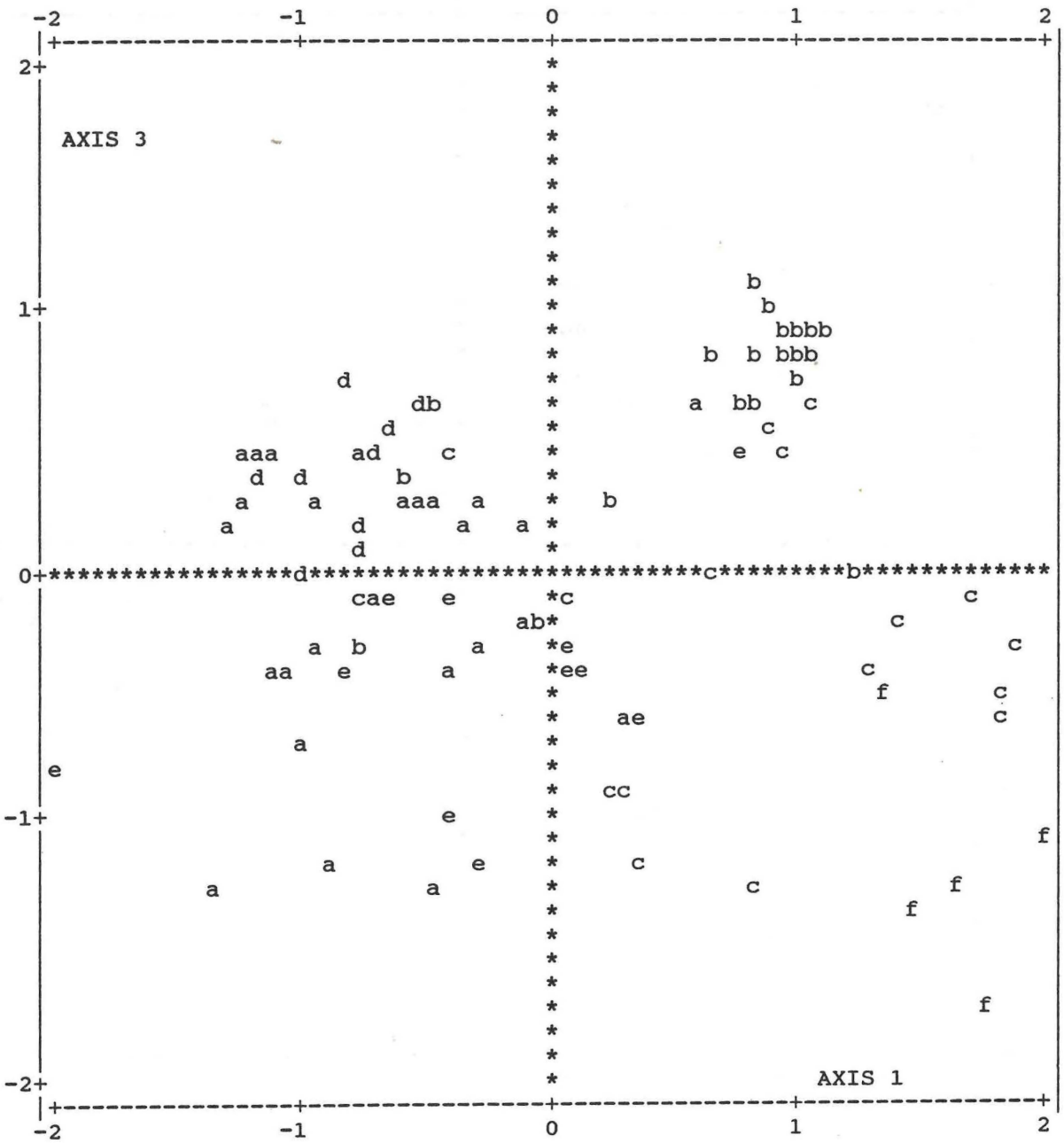


Figure 4: Plan defined by the first and third axes of a Principal Component Analysis, performed on RAPD data obtained on 89 clones, using 49 markers. a = Refractario (B,CL and JA), b = IMC and NA, c = MO,PA and SCA, d = ICS, e = LCTEEN, f = GU.



RAPD

The discriminative power of the RAPD technique has been evaluated using a subsample of 89 clones belonging to the 11 studied populations. Figure 5 shows a plot indicating the evolution of the discriminative power with the number of markers used. With 40 markers, 82 clones (92%) are uniquely fingerprinted, and 85 different phenotypes are obtained. There are 2 groups of 2 undistinguished clones, and 1 group of 3 undistinguished clones. Clones sharing the same phenotype always belong to the same population. These figures correspond to 99.9% chances of detecting mislabelled trees if the mislabelling problem occurs randomly, but this percentage decreases (91.7% in the case of JA and 96.4% in the case of NA) if the mislabelling problem occurs within the same population. The plot shows a lower slot between 30 and 40 markers, corresponding to a lower increase in the discriminative power. Indeed, with 30 markers, 75 clones (84.3%) are uniquely fingerprinted, and 81 different phenotypes are obtained, these figures corresponded to a 99.8% chance of detecting random mislabelling, this value falling to 88.1% in case of a mislabelling problem occurring within the NA population.

Some markers have been found to be characteristic of one or two clones, by their presence or absence, as in the case of NA 79 and SCA 23 with opP14 (Figure 6). Others have been found characteristic of one population, as in the case of GU with opP14 (Figure 6). Other markers appear only in one population, even if not in all accessions of this population, as in the case of LCT-EEN with opP14 (Figure 6).

Discussion

The data obtained by the use of both I.E and RAPD techniques show that striking differences exist between populations in their levels of diversity and heterozygosity. These differences reflect the heterogeneity in the origin, and in the collecting and sampling methods adopted for the collection of these populations. For example, the LCT-EEN population, which is in reality a group of populations, was collected over a very large area (about 160,000 km^2) (Allen and Lass, 1983), resulting in a large level of diversity in the sample present in the ICG,T. In contrast, accessions analysed from the GU population were collected over a much more restricted area (about 80 km^2) (Lachenaud and Sallée, 1993), resulting in much less diversity.

Among the clones obtained from collections in estates, striking differences are also noted. Indeed, a very low level of diversity is observed within the samples of Caribbean populations such as MAR and DOM. This could have been due to biased collecting or to high homogeneity of material in the estates visited. Nonetheless, the close relationship between these two populations, shown by the cluster analysis, seems to favour the second hypothesis. On the other hand, the high degree of diversity in the CL population probably reflects the fact that this Refractario population originated from three distinct genetic origins, as suggested by Pound (1943). These three proposed origins are:

- Trinitario from Trinidad;
- Local trees (Nacional);
- Trees introduced from lower Amazon regions by Amerindian immigrants.

It was stated that the Refractario material introduced into Trinidad would have probably been exclusively from the third origin, since this material was selected for its resistance to Witches' Broom disease, a characteristic found neither in Trinitario nor in Nacional trees. Nevertheless, there are Refractario trees bearing red pods at Marper farm, Trinidad, where these plants were first introduced, showing that this supposition was not exact.

Figure 5: Plot showing the % of fingerprinting in relation to the number of RAPD markers used.

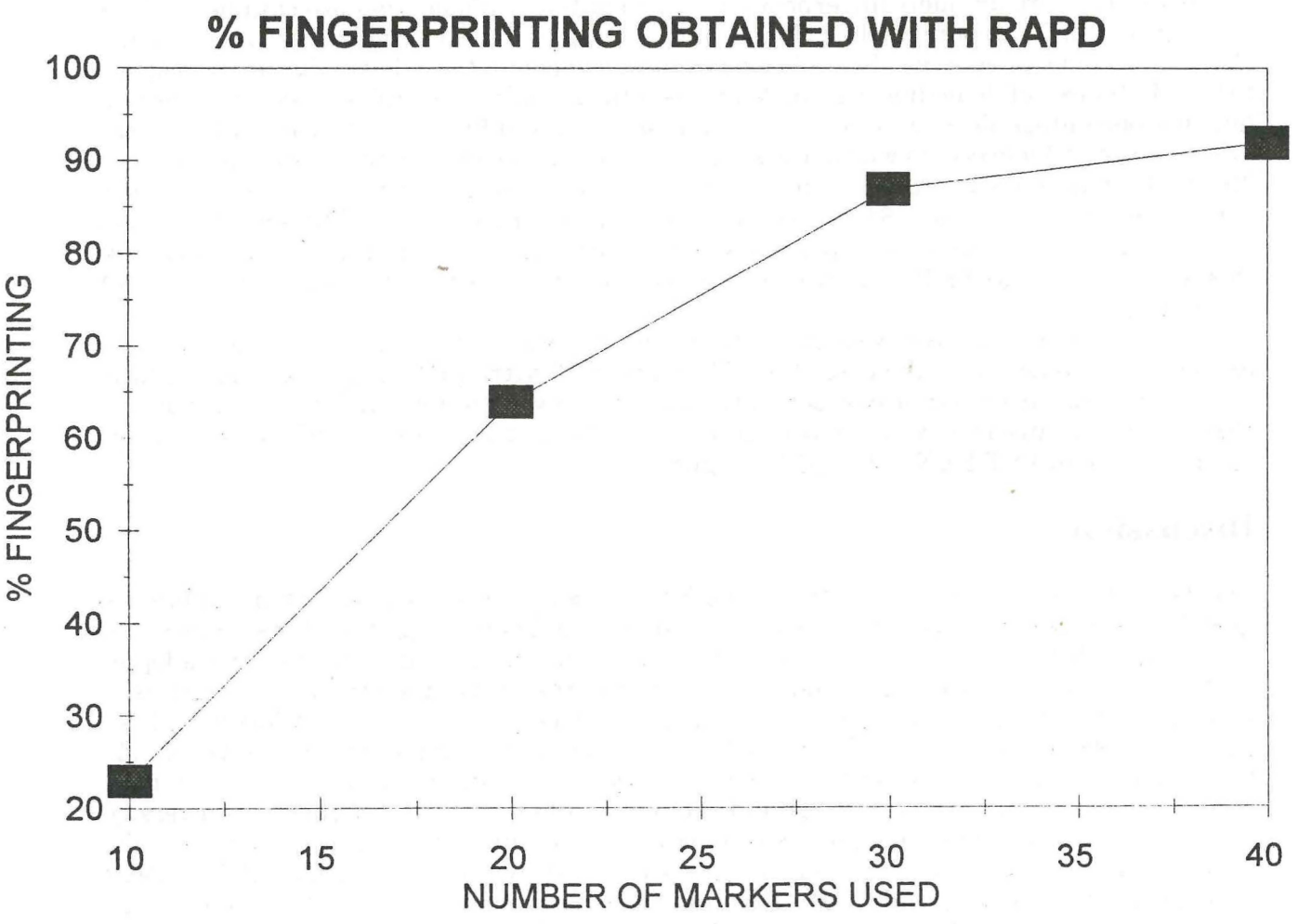
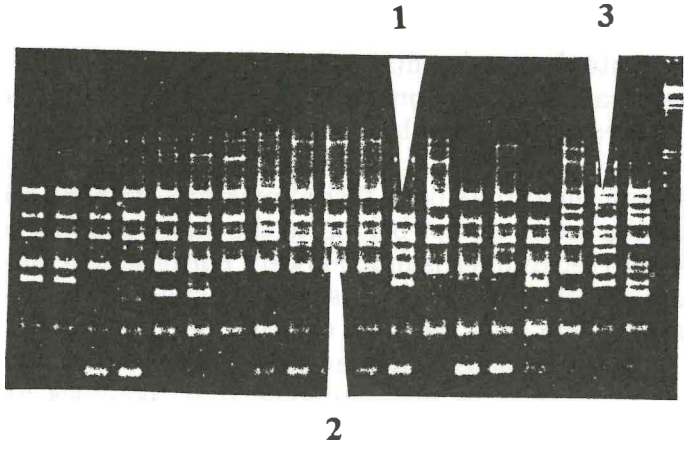


Figure 6: Examples of some specific bands observed with primer opP14. 1 = marker absent only in the case of Na79 and Sca23, 2 = marker specific to the GU population, 3 = marker specific to some LCTEEN clones.



In the future, the analysis of Refractario accessions should also be conducted in connection with their morphological types, rather than only based on the estate in which they were collected.

The level of heterozygosity was found to be low in the case of the two wild populations: LCT-EEN and GU, while some Trinitario populations such as UF, SC and CC show a high level of heterozygosity, which is not surprising due to the hybrid nature of the Trinitario group.

Partitioning of diversity in both I.E and RAPD studies shows that as much as half of the total diversity is due to variability within the populations, which would be expected in a highly out-crossing perennial species such as *Theobroma cacao* L. (Hamrick *et al.*, 1992). This confirms the findings of other researchers (Russell *et al.*, 1993, Ronning and Schnell, 1994).

These results demonstrate the need to maintain well represented populations in cacao genebanks, although the necessary level of representation varies greatly among populations.

RAPD data allowed a clear separation between Trinitario, Forastero and Refractario populations.

Fingerprinting analyses show that both techniques are efficient for discriminating among clones provided that the latter do not belong to the same population, in which case the discriminative power of both techniques is much lower. For practical purposes, under our conditions, different sources of "off-type" problems can be suggested. When budwood is collected from grafts on Marper farm or in the UWI campus fields in order to relocate clones, it can be very easily taken from the rootstock instead of the scion. In this case, the detection of the problem should be easy since it is most probable that rootstock and scion are not closely related. Mistakes could also occur where a rodent has attacked a pod, dropping a seed which is allowed to germinate at the base of the tree from which the seed was taken. The tree originating from this seed can grow and survive close to the mother tree, and if budwood is collected from the progeny instead of the mother tree, the detection of the mistake would be much more difficult because of the direct genetic link between the original tree and the younger one. Other mistakes could have occurred because of mislabelling of trees. These errors are likely to occur at the level of the number of the clone (for example confusion among B721, B7.21 and B72.1), in which case the distinction can be difficult. At the level of the name of the clone (for example, confusion among SCA, SPA and SLA), errors should be easy to detect. In conclusion, I.E is useful for detecting the most obvious mistakes while some more RAPD markers would be needed in case of confusion between closely related genotypes.

References

Allen, J.B., and Lass, R.A. (1983). London Cocoa Trade Amazon Project. Final Report. Phase 1, *Cocoa Grower's Bulletin* **34**

Christopher, Y. and Sounigo, O. (1995). The use of RAPD for characterization and genetic assessment of cacao, in: Annual Report of the Cocoa Research Unit for 1995, p. 38-51

Hamrick, J.L., M.J.W. Godt,and Sherman-Broyles, S.L. (1992). Factors influencing levels of genetic diversity in woody plant species, *New Forests* **6** 95-124

Johnson, E., Russel, J.R, Hosein, F., Powell, W., and Waugh, R. (1992). A laboratory manual for RAPD analyses in cocoa, Internal C.R.U Report, 17pp (unpublished).

Lachenaud, P., and Sallée, B. (1993). Les cacaoyers spontanés de Guyane. Localisation, écologie et morphologie, *Café Cacao Thé* vol XXXVII n°**2** 101-114

Lanaud, C. (1986). Utilisation des marqueurs enzymatiques pour l'étude génétique du cacaoyer: *Theobroma cacao* L. 1. Contrôle génétique et "linkage" de neuf marqueurs enzymatiques. *Café Cacao Thé* vol XXX n° **4** 259-270

Pound, F.J. (1943). Paper presented to the Agric. Society. of Trinidad and Tobago. Paper **846**.

Ronning, C.M., and Schnell, R.J. (1994). Allozyme diversity in a germplasm collection of *Theobroma cacao* L, *Journal of Heredity* **85** (4) 291-295

Russel, J.R., Hosein, F., Johnson, E., Waugh, R. and Powell, W. (1993). Genetic differentiation of cocoa (*Theobroma cacao* L.) populations revealed by RAPD analysis, *Molecular Ecology* **2** 89-97