Original article

Effects of quantitative and qualitative principal component score strategies on the structure of coffee, rubber tree, rice and sorghum core collections

Serge Hamon[a]*, Stéphane Dussert[a], Monique Deu[b],

Perla Hamon[b], Marc Seguin[b], Jean-Chistophe Glaszmann[b],

Laurent Grivet[b], Jacques Chantereau[b],

Marie-Hélène Chevallier[b], Albert Flori[b],

Philippe Lashermes[a], Hyacinthe Legnate[c], Michel Noirot[a]

[a] ORSTOM, Institut français de recherche scientifique pour le développement
en coopération, BP 5045, 34032 Montpellier, France
[b] CIRAD, Centre international de recherche agronomique pour le développement,
BP 5035, 34032 Montpellier, France
[c] IDEFOR, Institut des forêts, 01 BP 1536 Abidjan 01, Côte d'Ivoire

Abstract – The principal component score strategy (PCSS) is a multivariate method which allows the identification of a core subset from a germ-plasm collection. Previously described for quantitative data, the method is extended here to qualitative data provided by molecular markers. Quantitative and qualitative PCSS were then applied to real data on four tropical crops: coffee, rice, rubber tree and sorghum. The results show, in all cases, that the increase in the cumulated relative contribution (CRC) is very rapid but may depend on the species. Ten percent of the entire collection yielded between 22 and 58 % of the CRC. As expected, the variability of the quantitative characters in the subsets was little or not modified by a qualitative selection but largely increased by a quantitative one, whereas qualitative PCSS was more efficient in preserving rare alleles and increased the global diversity with limited quantitative changes. The range of crop plants tested made it possible to compare the respective impacts of the two methods and highlighted the advantage of combining both types of characters. © Inra/Elsevier, Paris

core collection / PCSS strategy / coffee / rice / rubber tree / sorghum

---

* Correspondence and reprints

Résumé – Étude des impacts sur la structure génétique des « core collections » de caféiers, hévéas, riz et sorghos, de sélections de type PCSS basées sur des variables quantitatives et qualitatives. La stratégie PCSS (*Principal Component Score Strategy*) est une méthode de sélection, basée sur des analyses multivariées, proposée pour constituer des « core collection » à partir de collections importantes de ressources génétiques. La méthode décrite sur des données quantitatives est adaptée ici à des données qualitatives de type moléculaire. Ces deux méthodes ont été testées pour leurs impacts sur la structure de quatre plantes tropicales: caféier, hévéa, riz et sorgho. Les résultats montrent, dans tous les cas, que l'augmentation des contributions relatives cumulées (CRC) sont très rapides mais diffèrent d'une espèce à l'autre. Dix pour cent de la collection totale permet d'obtenir de 22 à 58 % de CRC. Comme prévu, la variabilité des caractères quantitatifs dans les échantillons, est peu ou pas modifiée lorsque la sélection est qualitative mais elle l'est fortement par une sélection quantitative. La sélection qualitative apparaît comme la plus efficace pour conserver les allèles rares et augmenter la diversité globale avec des effets limités au niveau quantitatif. L'utilisation d'espèces très différentes a permis de comparer les impacts respectifs des deux méthodes et de mettre en lumière les avantages d'une sélection combinée sur les deux types d'approches. © Inra/Elsevier, Paris

core collection / stratégie PCSS / caféier / riz / hévéa / sorgho

1. INTRODUCTION

One of the major issues that gene-bank managers (curators) face is the need to increase the accessibility of their collection to a large group of users. Indeed, the sheer size of many collections, the low degree of characterisation of the accessions and the poor efficiency of data management systems, often lead to collections that are difficult to use effectively. Frankel and Brown (1984) first proposed that one way of alleviating this problem lay through the development of core collections (CCs), defined as combining 'the genetic diversity of a crop species and its relatives with minimum repetitiveness'.

The first procedures for establishing CCS were based on neutral characters. Brown (1989b) suggested that a random (R) sampling strategy of 10 % of accessions from the base collection (BC) should yield, for rare but widespread alleles, over 70 % of the variation in the BC. Brown (1989a) introduced hierarchical sampling which assumes that the BC can be structured into groups. These groups, or clusters, could be based on different types of data such as genetics, ecogeography, country of origin. According to the hierarchical sampling strategies proposed by Brown, each subgroup size is related to the initial group size according to different strategies: 1) a constant number of accessions per group (C strategy); 2) a number of accessions in proportion to the group size (P strategy); or 3) a number of accessions in proportion to the logarithm of the group size (L strategy). Because these strategies were based on the number of accessions in a given group and did not utilise population genetics information, Schoen and Brown (1993) suggested two additional strategies, H and M. The H strategy (heterozygosity) maximises Nei's genetic diversity index, whereas the M strategy (maximisation) is based on maximising the allele diversity in the core collection. In an empirical examination, these authors ranked the strategies from the highest to the lowest expected allele retention, as follows: M > H > P > L > C > R. Bataillon et al. (1996), using computer simulation, confirmed that the M strategy works well in maximising

the non-neutral diversity of an autogamous species, or a species subdivided into genetically isolated populations. Reports of the practical applications of the H and M strategies have not yet been published. One major reason is the lack of base collections that are fully characterised at the molecular level.

Multivariate methods, based on quantitative data, have been developed for the purpose of forming CCs. The approach was first introduced by Spagnoletti-Zeuli and Qualset (1993) using a three-step procedure: 1) groups were defined using cluster analysis; 2) within each group factor scores of each accession were computed using discriminant analysis; and 3) accessions were randomly sampled in zones delimited by factorial scores. As a result, the variance of quantitative characters in the CC was maximised. Basigalup et al. (1995) compared height putative CC strategies, and found similar results. Zhong and Qualset (1995) suggested the use of a generalised coefficient of phenotypic variation (GCPV), calculated on the basis of the coefficient of variation within and between populations. Mahajan et al. (1996) tested the Shannon diversity index (SDI) which is adapted to morphological qualitative characters. The results suggest that given rather complete data, principal component and cluster analysis are useful tools for grouping and selecting accessions used to build CCs. In fact, random sampling within groups leads to increases in the global variance. A theoretically more efficient method, the principal component score strategy (PCSS), was described by Noirot et al. (1996). In the PCSS, groups are defined on the basis of possible gene flow and lack of reproductive barriers. In a given group, the diversity of the CC is maximised after elimination of colinearity between variables, then accessions are selected according to their cumulated relative contribution (CRC). The CC size can be defined by determining the number of accessions or by fixing a CRC value.

Whatever the strategy used, the creation of CCs is intended to improve the efficiency of both conservation and utilisation of genetic resources. Recent experience suggests that CCs do, in fact, aid the end-user in discovering useful traits with fewer accessions to screen. As an example, Bouton (1996) found nearly the same frequency of acid soil tolerance in both BCs and CCs of alfalfa (*Medicago sativa*). The white clover (*Trifolium repens*) CC was also representative of BC for total cyanogenesis (Pederson et al., 1996). In addition, a two-stage screening approach for resistance to late leafspot in the peanut (*Arachis hypogaea*) CC clearly demonstrated that this CC can be used to improve the efficiency of peanut germ-plasm evaluation (Holbrook and Anderson, 1995). Despite this, some authors have advocated the development of a few situation-specific CCs, or subsets of CCs rather than a single core from a base collection (Mackay, 1995; Rana and Kochhar, 1996).

In this report, we present 1) an adaptation of the PCSS strategy to qualitative data; 2) the CRC for quantitative and qualitative data; 3) the differences observed when used with four crops that have contrasting biology; 4) the impact of the selections on the final allelic composition; and 5) the modification induced (or not) for means and variances of the quantitative descriptors.

1.1. Genetic background of the four crops studied

The evolutionary relationships among cultivars of the four crops tested are summarised here.

Asian cultivated rice (*Oryza sativa* L.) is an autogamous diploid species. Two sub-specific groups have been recognised for centuries in China. They were found to reflect the species structure in most other regions of the world (Oka, 1958) and have been named indica and japonica. Isozyme diversity confirmed the existence of the two major types (Second, 1982), which might be related to a 2 to 3 million year differentiation between two populations of wild rice, followed by two independent domestications (Second, 1985). A more precise analysis of isozymic diversity among Asian cultivars (Glaszmann, 1987) showed that several other specific types coexist with the two major groups. Their evolutionary origin is still unclear. This structure of the species was largely confirmed by subsequent analyses with molecular markers, including random amplified polymorphic deoxyribonucleic acids (RAPDs) (Virk et al., 1994; Mackill, 1995) and restriction fragment length polymorphisms (RFLPs) (Wang and Tanksley, 1989; Second and Ghesquière, 1994).

Cultivated sorghum forms are all included in the African species *Sorghum bicolor* and constitute the *S. Bicolor* ssp. *bicolor* subspecies. They are monoecious, preferentially self-pollinating and exhibit great phenotypic diversity. A simpler classification than that of Snowden (1936) was proposed by Harlan and de Wet (1972) using two morphological criteria: spikelet structure and panicle shape. Five basic races, bicolor, caudatum, durra, kafir and guinea and ten intermediate races (representing intermediates between two races) have been defined. A quantitative study, involving morphological and physiological traits, led to a classification with three groups characterised by different cropping performances (Chantereau et al., 1989). Isozymic markers do not discriminate the races but highlight a geographical structuring (Morden et al., 1989; Ollitrault et al., 1989). The variation of sorghum cultivars grown throughout the world is included in that of African forms. Nuclear deoxyribonucleic acid (DNA) diversity revealed a racial differentiation and a subrace division within the guinea race (Deu et al., 1994, 1995; de Oliveira et al., 1996).

Cultivated rubber tree (*Hevea brasiliensis*) is an allogamous diploid species ($2n = 36$), originating from the Amazon basin. All the elite cultivars (grafted clones) were selected from the few seeds introduced in Southeast Asia at the end of the nineteenth century (Wycherley, 1979). Significant germ-plasm collections were constituted recently with *H. brasiliensis* accessions surveyed in three Brazilian states (Acre, Rondonia and Mato Grosso). The IDEFOR International Conservation Centre in the Ivory Coast encompasses 2 423 trees surveyed in 1981 in 16 districts of these three states IRRDB collection (Chapuset et al., 1995). Important agronomic evaluations were carried out on this collection (Chapuset et al., 1995) as well as genetic diversity studies using morphological traits (Nicolas et al., 1988), isozymes (Chevallier et al., 1988; Seguin et al., 1995), nuclear RFLP (Besse et al., 1994) and mitochondrial RFLP markers (Luo et al., 1995). Molecular markers revealed four differentiated genetic groups in accordance with the geographic origin of the accessions, despite the predominance of the genetic diversity at the intragroup level (Seguin et al., 1996). A slight agronomic difference is also observed between the four molecular genetic groups (Chapuset et al., 1995).

Coffee trees belong to the *Coffeae* tribe of family Rubiaceae. The genus *Coffea* L. is subdivided in two subgenera: *Coffea* and *Baracoffea*. Approximately 80 taxa have been identified so far in the subgenus *Coffea* (Berthaud and Charrier,

1988). All species are woody, ranging from small shrubs to robust trees, and originate from the intertropical forests of Africa and Madagascar. Phylogenic relationships among species of coffee are well studied (Lashermes et al., 1996, 1997). Commercial coffee production relies on two species only, *C. arabica* L. and *C. canephora* Pierre. Since 1975, more than 1 000 *C. canephora* genotypes have been collected by ORSTOM and CIRAD, in collaboration with IBPGR and FAO, in five African countries: Guinea and the Ivory Coast in West Africa, Congo, Central African Republic and Cameroon in Central Africa (Berthaud and Charrier, 1988). A base field collection was established in the Ivory Coast (IDEFOR-DCC, Divo) to conserve the germ-plasm collected. An isozymic evaluation of the diversity, connected with intercrossing behaviour studies and morphological descriptions, revealed evidence of two genetic groups, the guinean and congolian (Berthaud and Charrier, 1988).

The four crops used represent four contrasting cases that are relevant to testing of our methodologies: rice and sorghum are annual and autogamous, whereas coffee (*C. canephora*) and rubber tree are perennial allogamous species; rice and *C. canephora* display a strong structure in specific groups, whereas sorghum and rubber tree display only a weak structure.

2. MATERIALS AND METHODS

2.1. Quantitative PCSS

Within-population diversity is determined by the level of between-individual differences in one or more traits. Generally, quantitative traits are of heterogeneous type. In order to give the same contribution (weight) to each trait j, the Euclidian distance is weighted by the reciprocal of the standard deviation σ_φ. The distance d_{ik} between two individuals i and k for the J quantitative traits is defined by the following formula:

$$d_{ik} = \sqrt{\sum_{j=1}^{J} \left[(x_{ij} - x_{kj}) \cdot \sigma_j^{-1} \right]^2}$$

where x_{ij} and x_{kj} are the observed values of the trait j on the individuals i and k, respectively. The between-individual distance is directly related to the number of differences. If traits are highly correlated (positively or negatively), this may lead to overestimation of the distance between individuals. To avoid the effect of colinearity among traits, principal component analysis was applied to standardised data, to yield J statistically independent and centred variables, or factors. The distance between two individuals i and k for the J factors is computed using a similar formula:

$$d_{ik} = \sqrt{\sum_{j=1}^{J} \left[(z_{ij} - z_{ij}) \cdot \sqrt{\lambda_j^{-1}} \right]^2}$$

where the square root of the 1_j eigenvalue allows weighting, and where z_{ij} and z_{kj} are the scores of the individuals i and k, respectively, on the factor j. Such a

procedure takes into account all factors with the same weight, including residual components – the result of chance or notation errors – in distance estimation. Removal of factors for which the eigenvalue is below one is arbitrarily applied to eliminate this disadvantage.

The generalised sum of squares (GSS) of a set of N individuals in the factorial space of K standardised (mean $= 0$; variance $= 1$) and independent (correlation coefficient $= 0$) variables is equal to the product $N.K$ (Lebart et al., 1977). The contribution P_i of the individual i to the GSS is equal to the sum of the squares of its K new scores:

$$P_i = \sum_{j=1}^{K} x_{ij}^2$$

The relative contribution CR_i of the individual i to the GSS of the set is given by:

$$CR_i = P_i/(N.K)$$

Preserving the greatest variability is equivalent to maximising the score of the subset of sampled individuals using a GSS estimator. The first step consists of keeping the farthest individual of the set centre as initial subset, i.e. the individual with the highest relative contribution. Iterative selection of individuals that maximise subset variability increases subset size and provides a core collection. At each iteration, the cumulative GSS of the subset (expressed in percentage of the total GSS) is calculated. The procedure can thus be stopped according to either the subset size or the GSS expressed in %. The two criteria can be simultaneously taken into account. In this case, the first criterion to be reached defines the stopping point for sampling.

2.2. Qualitative PCSS

The method just described was adapted here to qualitative data. Changes concern only the first step of the PCSS. As for quantitative data, between-variate relationships can also exist; for example, two molecular markers that are highly linked on a chromosome. In order to avoid between-variate relationships and to give the same weight to independent markers, a multivariate method was used to transform initial data into factor scores. Factorial analysis of correspondence (Benzécri, 1972) was adopted here.

The method uses the χ^2 distance instead of the Euclidian distance. A complete disjunctive table has to be used in this case. In this table, the presence and absence of an allele are considered as two different variates taking 0 or 1 as values. With p molecular markers observed on N individuals, we obtain a $2p \times N$ table. Consequently, all individuals show the same margin frequencies equal to p. In addition, the term $p\lambda_l$ (λ_i is the eigenvalue of factor i) is equal to the sum of the correlation ratios of the factor with the p variates (Saporta, 1990). This term is equivalent to the eigenvalue observed in principal component analysis. The sum of $p\lambda_l$ is equal to the number of markers (for the principal component analysis on quantitative data, the sum of eigenvalues is equal to the number of variates). As for quantitative data, factor scores are weighted. For qualitative

data, weights are the square root of the respective $p\lambda_l$. Other steps are the same.

A software was designed using Visual Basic (Microsoft copyright). Data were recorded as an Excel sheet (Microsoft copyright), and both algorithms were made available in the 'tools' option of the main menu of Excel.

2.3. Evolution of the CRC

For the four species, the initial data were first used to perform multivariate analysis, respectively principal component analysis on the quantitative data and factorial analysis on the qualitative data. Then, for each accession, the multivariate scores were used to perform quantitative PCSS (Quant PCSS) or qualitative PCSS (Qual PCSS). For each species, the CRC during the selection process was recorded as a function of the relative size of the initial collection.

For each crop, we arbitrarily selected, using Quant PCSS and Qual PCSS, two CCs which were selected at the CRC level of 50 %: the quantitative CC (Core Quant) and qualitative data (Core Qual). With this constraint, the selected samples differed in size. They were coffee (21/15), rubber tree (30/29), rice (68/29) and sorghum (40/70), where the first number is the size of the Core Quant and the Core Qual.

2.4. Allelic retention and genetic diversity

The allelic frequencies in the BC were calculated and then five categories were defined as follows: ($f < 5$ %, 5 % $< f < 10$ %, $10 < f < 20$ %, 20 % $< f < 40$ %, $f > 40$ %). With the constraint of CRC = 50 %, the selected samples differed in size, so we defined the allelic retention index as the number of alleles found in at least one individual of the core subset for a given category. The various subsets were also compared to the initial sample on the basis of the global diversity (Nei's diversity index) Nei (1978).

2.5. Plant data used

For the four species, a subsample of the collection was extracted to best represent the genetic diversity and was considered in the study as the BC (base collection).

2.5.1. Rice

The selected BC consists of 270 accessions from the world collection maintained at the IRRI. Characters used to define the BC included geographic origin, the culture type and the position in the isozyme classification. Two hundred sixty-five accessions in the BC were characterised with isozymes at 15 loci, as described by Glaszmann and colleagues (1987, 1988). In all, 49 alleles were observed. Two hundred fifty-six accessions were described for 11 morphological traits: seedling height (SDHT), leaf length (LLT), leaf width (LWD), ligule length (LIGLT), number of days from seedling to 50 % headed (HDG), culm length (CULT), culm number (CUNO), culm diameter of basal internode (CUDI), panicle length (PLT), 100-grain weight (100), grain length (GRLT) and grain width (GRWD). These descriptions were extracted from the IRRI database.

2.5.2. Sorghum

The selected BC consists of 347 accessions from the CIRAD collection. Ten enzymatic systems corresponding to 14 polymorphic loci were revealed for 347 accessions by Ollitrault et al. (1989). Among the 347 accessions, 151 were characterised for 16 morphological traits: number of days from planting to 50 % sprouting (NDS), number of days from planting to 50 % heading (NDH), plant height (PHE), length of the panicle (LEP), number of internodes (NIN), length of the third leaf under the panicle (LEL), width of the third leaf under the panicle (WIL), diameter of the stem at the third internode (DIS), straw weight (SWE), panicle weight (PWE), grain production (GPR), weight of 500 grains (W5G), vitrosity (VIT), germinative rate (GRA), number of flowering axillary stems (FAS) and number of flowering basal stems (FBS) (Chantereau et al., 1989).

2.5.3. Rubber tree

The selected BC consists of 183 IRRDB accessions completely characterised for 13 isozyme loci (Chevallier et al., 1988; Seguin et al., 1995) (without missing data) and for 26 agronomic traits (Chapuset et al., 1995). The agronomic data were obtained between 1986 and 1995 (# symbolises the year which is coded as follows: 91 = 1991) (Chapuset et al., 1995). Observed characters were mainly latex production (eight variables, Pro#) and girth growth (nine variables, Circ#), but encompassed also notations of crown architecture (two variables, Branc#), bark thickness (two variables, Epec#), wood production (one variable, Grum94) and leaf disease (one variable, Res89). Isozyme data were obtained according to Lebrun and Chevallier (1991) and were scored as presence/absence of 51 alleles at 13 loci.

2.5.4. Coffee

The BC consists of 73 wild and 62 cultivated *Coffea canephora* accessions which are maintained in a field collection kept in the Ivory Coast (IDEFOR-DCC, Divo). Genomic DNA was isolated from lyophilised leaves through a nuclei isolation step. Restriction enzyme digestion, gel electrophoresis, alkaline transfer, nonradioactive digoxigenin-labelling of DNA probes and southern hybridisation were carried out as previously reported by Lashermes et al. (1995). Twelve single-copy nuclear genomic clones from a *C. arabica* (cultivar N39) Puc18-*Pst*1 library were used as probes following digestion by either *Eco*RI or *Dra* I. Ninety of the 135 accessions were evaluated for ten quantitative traits: one corresponds to the annual yield (YIEL), six variables are foliar morphology traits (leaf length LEAL, leaf width LEAW, leaf shape LEAS, leaf area LEAA, acumen length ACUM and petiole length PETL), fertility (two variables, Caracoli bean rate CARA, outturn OUTT) and bean technological characteristic, the 100-seed weight SEWE.
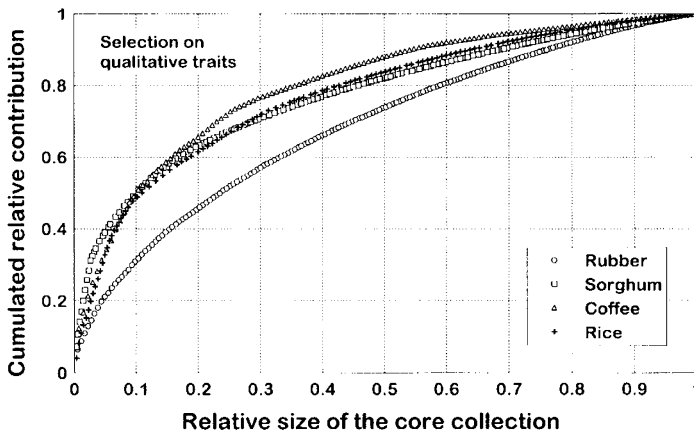
3. RESULTS

3.1. Cumulated relative contribution (CRC) and selection of the core collection (CC)

For both quantitative and qualitative data, the CRC of the CC is a function of its relative size (*figure 1*). With quantitative data (*figure 1a*) coffee and rice revealed similar CRC patterns. The CRC increased faster with rubber and much faster with sorghum. For coffee and rice, 50 % CRC was obtained with 30 % of the BC and 80 % CRC with 50 % of the BC. The sorghum case was different: 60 % CRC was selected with 10 % BC and 80 % CRC with 30 % of
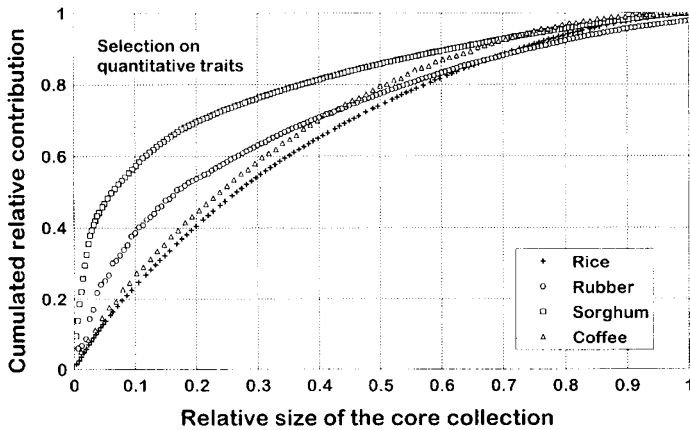
(a)

(b)

Relationships between the relative size of the core collection and the cumulated relative contribution: 1a, selection on quantitative traits and 1b, selection on qualitative traits.

Figure 1.

the BC. Altogether, these results show that in the initial phase the PCSS gave a rapid gain, but differences occurred among the crops.

The selection patterns using qualitative data (*figure 1b*) were similar for three of four crops. For coffee, rice and sorghum, 50 % CRC was selected with 10 % of the BC. Sixty percent of the CRC was selected with 20 % of the BC. The rubber tree curve was slightly different: only 30 % CRC was selected with 10 % BC and 45 % CRC with 20 % of the BC. These results show that, whatever the crop, as for the quantitative data, the PCSS gave a rapid gain in variability but the rate depends on the crop.

3.2. Allelic retention in the core subsets

For each crop, using Quant PCSS and Qual PCSS, two CC were selected at the CRC = 50 %: the Core Quant and Core Qual. With this constraint, the selected samples differed in size, and were coffee (21/15), rubber tree (30/29), rice (68/29) and sorghum (40/70), where the first number is the size of the Core Quant and the other the Core Qual.

Using the quantitative selection (*table Ia*), the number of alleles lost were 11 (33 %) for coffee, six (12 %) for rice, three (6 %) for rubber and four (11 %) for sorghum. All the alleles that were lost in the core set had an initial frequency lower than 0.05. Using the qualitative selection (*table Ib*), the number of alleles lost were two (6 %) for coffee, one (2 %) for rice, one (2 %) for rubber and one (2 %) for sorghum. Again, all of them had an initial frequency lower than 0.05.

The global diversity indices calculated using the allele frequencies at all loci gave another perspective. The allele frequency change (data not shown) was significant in several instances after selection based on quantitative data. For rice and sorghum, the average diversity was not affected (*table Ia*) because changes were in both directions; for rubber tree, the average diversity notably increased because most changes were in the same direction. After selection based on qualitative data, the allele frequency changes were more markedly affected. Many loci displayed significant changes, almost all towards an increased diversity; the frequency of the rare alleles increased. For all crops the average diversity index was thus higher than before the selection. Rubber tree had a distinct response as compared to other crops: the selection based on quantitative traits resulted in a limited loss of rare alleles and a global increase in molecular diversity.

3.3. Variance and mean differences between core subsets and the initial collection

For the four crops the variance homogeneity was tested (Levene test) between the BC. Core Qual and Core Quant. Means were compared (Bartlett test) when homogeneity of variances allowed the comparison.

For sorghum, the Core Qual compared to the BC showed only one slightly heterogeneous variance and one different mean (*table IIa*). In contrast, the comparison of the Core Quant and the BC variances gave seven heterogeneous variances, and three homogeneous with no mean differences. The three other crops, rubber tree (*table IIb*), rice (*table IIc*) and coffee (*table IId*), revealed similar situations. In most cases, as expected, variances were homogeneous when the BC was compared to the Core Qual and the variances were mostly heterogeneous when the Core Quant was compared to the BC. Nevertheless, this was not systematic and reciprocal situations were found.

Table Ia. Selection of core subset accessions for four different crops using the PCSS strategy on quantitative data. The selected number of alleles are reported according to their frequency in the initial collection and to their presence in the subset.

Alleles frequency (f)	Coffee total	$I_{50} = 21$ selected	Rice total	$I_{50} = 68$ selected	Rubber total	$I_{50} = 30$ selected	Sorghum total	$I_{50} = 40$ selected
$f < 5\%$	17	6	17	11	13	10	7	3
$5\% < f < 10\%$	1	1	4	4	8	8	3	3
$10\% < f < 20\%$	5	5	5	5	11	11	4	4
$20\% < f < 40\%$	8	8	5	5	6	6	8	8
$f > 40\%$	2	2	18	18	13	13	16	16
H	/	/	0.38	0.36	0.45	0.51	0.39	0.39

H: the heterozygosity of Nei, 1978.

Table Ib. Selection of core subset accessions for four different crops using the PCSS strategy on qualitative data. The selected number of alleles are reported according to their frequency in the initial collection and to their presence in the subset.

Alleles frequency (f)	Coffee total	$I_{50} = 15$ selected	Rice total	$I_{50} = 29$ selected	Rubber total	$I_{50} = 30$ selected	Sorghum total	$I_{50} = 70$ selected
$f < 5\ \%$	16	14	17	16	13	12	18	17
$5\ \% < f < 10\ \%$	2	2	4	4	8	8	3	3
$10\ \% < f < 20\ \%$	5	5	5	5	11	11	4	4
$20\ \% < f < 40\ \%$	8	8	5	5	6	6	8	8
$f > 40\ \%$	2	2	18	18	13	13	15	15
H	/	/	0.38	0.46	0.45	0.53	0.38	0.46

H : the heterozygosity of Nei, 1978.

Table IIa. Comparison, for the sorghum case, of ten trait distribution variables (Vardiff), means (MeanDiff) between the BC and the qualitative subset means (Core Qual) and between the BC and the quantitative subset means (Core Quant).

Variable	NDS	NDH	PHE	LEP	NIN	LEL	DIS	PWE	GPR	FBS
BC	4.58	63.08	204.54	272.82	10.45	68.65	18.51	85.74	58.38	0.061
Core Qual	4.79	64.62	226.71	300.40	10.96	70.45	18.55	84.27	55.67	0.053
VarDiff	NS	NS	NS	NS	NS	NS	NS	NS	NS	*
MeanDiff	**	NS	NS	NS	NS	NS	NS	NS	NS	/
Core Quant	4.71	63.52	230.16	281.58	10.68	68.47	17.27	83.85	55.03	0.112
VarDiff	***	NS	**	***	NS	***	NS	***	***	***
MeanDiff	/	NS	/	/	NS	/	NS	/	/	/

*, **, ***, significant at 0.05, 0.01, 0.001 levels respectively;
NS: not significant. See text for the definitions of abbreviations of variables.

Table IIb. Comparison, for the rubber tree case, of eight trait distribution variables (Vardiff), means (MeanDiff) between the BC and the qualitative subset means (Core Qual) and between the BC and the quantitative subset means (Core Quant).

Variable	Branc87	Circ87	Circ91	Circ94	Epec89	Epec96	Grum94	Pro87	Pro91	Pro95
BC	1.79	120.5	434.4	480.8	50.1	4.69	2.99	937	2 935	7 784
Core Qual	1.73	128.1	389.5	430.9	47.0	4.13	2.80	1 052	3138	6 075
VarDiff	*	*	NS	NS	NS	NS	NS	NS	NS	NS
MeanDiff	/	/	NS	NS	NS	*	NS	*	NS	NS
Core Quant	1.30	148.2	359.6	439.7	42.3	4.33	2.40	1 164	6 738	14 556
VarDiff	***	***	***	***	***	***	***	***	***	***
MeanDiff	/	/	/	/	/	/	/	/	/	/

*, **, ***, significant at 0.05, 0.01, 0.001 levels respectively;
NS: not significant. See text for definitions of abbreviations of variables.

Table IIc. Comparison, for the rice case, of ten trait distribution variables (Vardiff), means (MeanDiff) between the BC and the qualitative subset means (Core Qual) and between the BC and the quantitative subset means (Core Quant).

Variable	SDHT	LLT	LWD	LIGLT	HDG	CULT	CUNO	CUDI	100	GRWD
BC	1.47	3.33	1.35	18.57	93.7	4.65	2.10	4.74	2.43	3.09
Core Qual	1.70	3.44	1.46	21.7	95.5	5.07	1.92	4.62	2.62	3.10
VarDiff	NS	NS	NS	*	NS	NS	NS	NS	NS	*
MeanDiff	*	NS	NS	/	NS	NS	*	NS	*	/
Core Quant	1.42	3.39	1.45	18.5	93.7	4.52	2.10	4.60	2.43	3.12
VarDiff	NS	***	***	*	***	***	***	***	***	***
MeanDiff	NS	/	/	/	/	/	/	/	/	/

*, **, ***, significant at 0.05, 0.01, 0.001 levels respectively;
NS: not significant. See text for definitions of abbreviations of variables.

Table IId. Comparison, for the coffee tree case, of ten trait distribution variables (Vardiff), means (MeanDiff) between the BC and the qualitative subset means (Core Qual) and between the BC and the quantitative subset means (Core Quant).

Variable	CARA	SEWE	OUTT	YIEL	LEAL	LEAW	PETL	ACUM	LEAS	LEAA
BC	22.71	12.15	18.89	24.66	174.0	73.86	10.17	13.21	2.39	13 160
Core Qual	22.70	10.98	18.11	12.75	190.2	82.15	11.61	14.37	2.37	15 906
VarDiff	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
MeanDiff	NS	NS	NS	NS	NS	NS	NS	NS	NS	*
Core Quant	26.71	12.12	18.89	45.09	176.0	76.01	9.97	12.33	2.35	13 899
VarDiff	NS	***	NS	***	***	NS	NS	NS	NS	***
MeanDiff	**	/	NS	/	/	NS	NS	NS	NS	NS

*, **, ***; significant at 0.05, 0.01, 0.001 levels respectively;
NS: not significant. See text for definitions of abbreviations of variables.

3.4. Distribution differences in core subsets

For four variables, arbitrarily selected in the coffee example, the distributions of the entire collection, and both the Core Qual and Core Quant, are shown in *figure 2a–d*.

These distributions clearly show that, whatever the shape of the curve in the BC, the final profile is more or less regular along the x axes. Most of the redundancy is eliminated. In *table IId*, for CARA, there was no significant difference between the Core Qual, the Core Quant and the BC, and in *figure 2b* it is clear that Core Qual and Core Quant have similar distributions. Conversely, for LOFLA and SEWE for which the variances were different, it seems that these distributions are also different.

4. DISCUSSION

Tested on four different crops, the use of both qualitative and quantitative PCSS allowed rapid selection of a large CRC. The shapes of the CRC curves were alike except for sorghum (quantitative) and rubber tree (qualitative).

The explanation probably resides in the population genetic structure of the various species. Differentiation between several components will result in the occurrence of very distant individuals along the axes that bear this differentiation. Since the PCSS equalises the weight of all axes, it is the number of axes that bear this structure (i.e. the complexity of the structure rather than its intensity) that will determine the early gain in CRC.

In terms of morphological traits, sorghum is classified into five major races (Harlan and de Wet, 1972), whereas rice displays three types (indica, tropical japonica [Javanica] and temperate japonica; Glaszmann, 1987) and coffee, two (guinean and congolian; Berthaud and Charrier, 1988). Rubber tree has extensive but little structured diversity (Seguin et al., 1995). For qualitative data, the explanation is probably different. The CRC is comparable for all crops, only the rubber tree case is slightly different with a slower change in CRC. The cause may reside in its higher species homogeneity associated with allogamy and continuous geographic distribution in South and Central America.

The impact of the PCSS selection on the basis of quantitative data was as expected on the distribution of the corresponding traits. In most cases (30 of 40), the variances were significantly increased, highlighting a wider exploration of the diversity available. When the means could be compared, the rule was an absence of deviation. Conversely, when the qualitative selection was used, the variances of the characters were not greatly modified (5 of 40) and seven means were slightly different.

Frankel and Brown (1984) defined the core collection as a collection which is expected to reduce repetitiveness. Consequently, the core should not be a photocopy, in reduction, of the global collection, but a new organisation with a maximum variability in a minimum size. The diagram distributions for the quantitative traits were illustrative of this objective. The full amplitudes of the variation were represented and the top of the bell curves were eliminated.

The impact of the PCSS on the diversity for molecular markers was uneven among the crops. For three of them, some rare alleles were lost, in a proportion
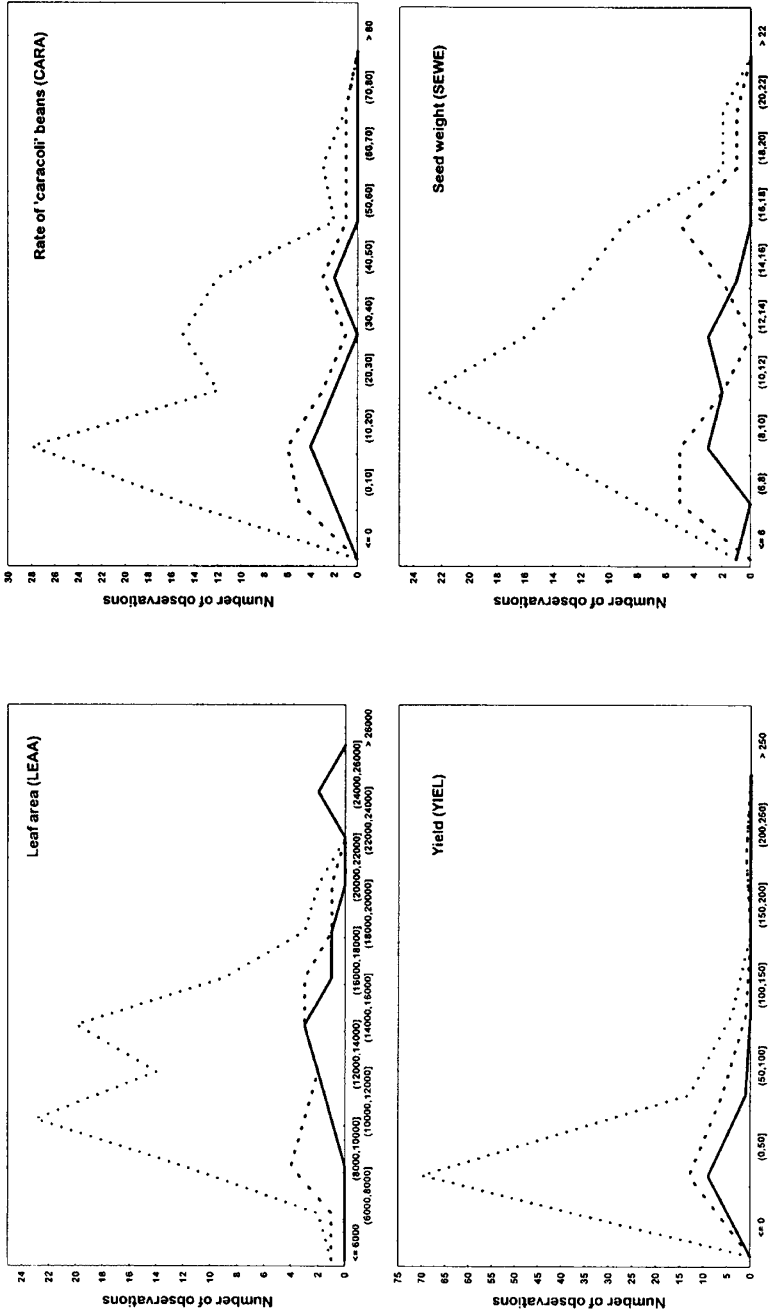
Figure 2. Coffee case. Distributions of leaf area, rate of caracoli beans, yield and seed weight in the base collection (dotted line), core collection selected on quantitative traits (hatched line) and in core collection selected with qualitative traits (solid line). Both core collections were defined by the 50 % cumulated contribution.

comparable to that expected through random selection; the global diversity, however, did not seem to be affected. For the fourth crop, rubber tree, most alleles were retained and the global diversity increased, suggesting that the two types of diversity were related. The PCSS selection on the basis of markers, in turn, had an obvious impact on the marker diversity, essentially by overestimating the rare alleles. The impact on quantitative traits was limited to slight deviations for a few characters.

The association between the two types of characters is responsible for cross-impacts of one PCSS basis on the other type of character. Two contrasting cases are worth examining in more detail.

In terms of the mating system, there is rice with autogamy and a marked structure of varietal groups. The PCSS on quantitative traits resulted in an overrepresentation of the japonica group, which hosts a high morphological diversity between tropical forms (so-called Javanica) and temperate forms. This in turn modified the marker frequencies in favour of the alleles predominant in this group, sometimes causing an increase in the diversity index and sometimes a decrease. The PCSS on qualitative traits resulted in an overrepresentation of minor varietal groups that have several rare alleles. Thus, both strategies led to a divergence between the core subsets.

The other extreme is rubber tree, with allogamy and continuous geographic distribution in the Amazon basin. In this case, both PCSS converged to an increase in marker diversity.

The above contrasts illustrate the potential differential impacts of one method or another, depending on the biology and the evolutionary history of the species.

The example of rice can be used to examine the status of the dilemma. A selection on the single basis of morphological traits will poorly sample types corresponding to the minor marker-based groups that were shown to hold alternative sources of factors for resistance to the major disease of rice (Glaszmann et al., 1996). Conversely, a selection on the single basis of molecular traits will leave little room for covering the wide ecogeographical adaptation of a group such as japonica, which is much more associated with morphological diversity than with markers; only the use of a large number of markers would reveal the differentiation within japonica.

On the other hand, the example of rubber tree illustrates the efficiency of the PCSS on quantitative data in enriching the diversity for both types of traits. When the variation is continuous, a selection on the basis of traits related to the use of the crop, which are of primary interest for the breeders, will be of little detriment to genetic diversity as a whole.

It has already been said that the PCSS should be applied after distinguishing clusters of materials that are separated by restrictions to recombination (Noirot et al., 1996). Molecular markers are best suited to reveal such restrictions, be they due to partial reproductive barriers or to factors such as geographic or seasonal isolation. As an example, an extensive AFLP analysis of the wild bean gene pool focused on insights into the genetic structure of the bean CC that were not possible by another approach (Thome et al., 1996). The difficulty lies in the necessity to split a collection into clusters when the variation is generally continuous between the clusters, when clusters are bridged by local

interfaces exhibiting gene flow and introgression and when the determination of the thresholds are essentially arbitrary.

Performing an appropriate classification requires a considerable amount of information that is seldom available. Therefore, the challenge for core-sampling strategies is to make the best use of the data available and to combine information of various kinds in a refined manner. This is obviously an avenue for future research, which will adapt the strategies to a wide array of biological situations.

REFERENCES

Basigalup D.H., Barnes D.K., Stucker R.E., Development of a core collection for perennial Medicago plant introductions, Crop Sci. 35 (1995) 1163–1168.

Bataillon T.M., David J.L., Schoen D.J., Neutral genetic markers and conservation genetics: simulated germplasm collections, Genetics 144 (1996) 409–417.

Benzécri J.P., Pratique de l'analyse des données : analyse des correspondances, Dunod, Paris, 1972.

Berthaud J., Charrier A., Genetic resources of *Coffea*, in: Clarke R.J., Macrae R. (Eds.), Coffee, vol. 4, Elsevier Applied Science, London, 1988, pp. 1–42.

Besse P., Seguin M., Lebrun P., Chevallier M.H., Nicolas D., Lanaud C., Genetic diversity among wild and cultivated populations of *Hevea brasiliensis* assessed by nuclear RFLP analysis, Theor. Appl. Genet. 88 (1994) 199–207.

Bouton J.H., Screening the alfalfa core collection for acid soil tolerance, Crop Sci. 36 (1996) 198–200.

Brown A.D.H., Core collections: a practical approach to genetic resources management, Genome 31 (1989a) 818–824.

Brown A.D.H., Size and structure of collection: the case for core collection, in: Hodgkin T., Brown A.D.H., Hintum T.J.L. van, Morales E.A.V. (Eds.), The Use of Plant Genetic Resources, John Wiley & Sons, Baffins Lane, Chichester, UK, 1989b, pp. 136–156.

Chantereau J., Arnaud M., Ollitrault P., Nabayago P., Noyer J.L., Étude de la diversité morphophysiologique et classification des sorghos cultivés, Agronomie Tropicale 44 (1989) 223–232.

Chapuset T., Legnate H., Doumbia A., Clément-Demange A., Nicolas D., Keli J., Agronomical characterisation of the 1981 germplasm in Côte-d'Ivoire: growth, production, architecture and leaf disease sensibility, in: IRRDB Symposium on the Physiological and Molecular Aspects of the Breeding of *Hevea brasiliensis*, 6–7 November 1995, IRRDB, UK, Penang, Malaysia, 1995, pp. 112–122.

Chevallier M.H., Lebrun P., Normand F., Approach of the genetic variability of germplasm using enzymatic markers, in: Jacob J.L., Prevôt J.C. (Eds.), Exploitation-physiologie et amélioration de l'Hévéa, IRCA-CIRAD, Paris, France, 1988, pp. 365–376.

Crossa J., Hernandez C.M., Bretting P., Eberhart S.A., Taba S., Statistical genetic considerations for maintaining germplasm collections, Theor. Appl. Genet. 86 (1993) 673–678.

Crossa J., Taba S., Eberhart S.A., Bretting P., Practical considerations for maintaining germplasm in maize, Theor. Appl. Genet. 89 (1994) 89–95.

De Oliveira A.C., Richter T., Bennetzen J.L., Regional and racial specificities in sorghum germplasm assessed with DNA markers, Genome 39 (1996) 579–587.

Deu M., Gonzalez-de-Leon D., Glaszmann J.C., Degremont I., Chantereau J., Lanaud C., Hamon P., RFLP diversity in cultivated sorghum in relation to racial differentiation, Theor. Appl. Genet. 88 (1994) 838–844.

Deu M., Hamon P., Chantereau J., Dufour P., D'Hont A., Lanaud C., Mitochondrial DNA diversity in wild and cultivated sorghum, Genome 38 (1995) 635–645.

Diwan N., Bauchan G.R., McIntosh M.S., A core collection for the United States annual *Medicago* germplasm collection, Crop Sci. 34 (1994) 279–285.

Diwan N., McIntosh M.S., Bauchan G.R., Methods for developing a core collection of annual *Medicago* species, Theor. Appl. Genet. 90 (1995) 755–761.

Dussert S., Chabrillange N., Anthony F., Engelmann F., Récalt C., Hamon S., Variability response within a coffee (*Coffea* spp.) core collection under slow growth conditions, Plant Cell Rep. 16 (1997) 344–348.

Frankel O.H., Brown A.D.H., Current plant genetic resources – a critical appraisal, in: Genetics New Frontiers, Proceedings of the 15th International Congress of Genetics, vol. 4, Oxford and IBH Publishing Co., 1984, pp. 3–13.

Glaszmann J.C., Isozymes and classification of Asian rice varieties, Theor. Appl. Genet. 74 (1987) 21–30.

Glaszmann J.C., de Los Reyes B.G., Khush G.S., Electrophoretic variation of isozymes in plumules of rice (*Oryza sativa* L.); a key to the identification of 76 alleles at 24 loci, IRRI Research Paper Series 134, IRRI, Manila, Philippines, 1988.

Glaszmann J.C., Mew T., Hibino H., Kim C.K., Vergel de Dios-Mew T.I., Vera Cruz C.M., Notteghem J.L., Bonman J.M., Molecular variation as a diverse source of disease resistance in cultivated rice, in: Proc. Rice Genetics III, IRRI, Manila, Philippines, 1996, pp. 460–465.

Hamon S., Hodgkin T., Dussert S., Noirot M., Core collection – accomplishments and challenges, Plant Breeding Abstr. 65 (1995) 1125–1133.

Hamon S., Noirot M., Anthony F., Suggested procedures for selecting a coffee core collection, in: Hodgkin T., Brown A.D.H., Hintum T.J.L. van, Morales E.A.V. (Eds.), The Use of Plant Genetic Resources, John Wiley & Sons, Baffins Lane, Chichester, UK, 1995, pp. 117–126.

Harlan J.R., de Wet J.M.J., A simplified classification of cultivated sorghum, Crop Sci. 12 (1972) 172–176.

Holbrook C.C., Anderson W.F., Evaluation of a core collection to identify resistance to late leafspot peanut, Crop Sci. 35 (1995) 1700–1702.

Holbrook C.C., Anderson W.F., Pittman R.N., Selection of a core collection from the U.S. germplasm collection of peanut, Crop Sci. 33 (1993) 859–861.

Lashermes P., Combes M.C., Cros J., Use of non-radioactive digoxigenin-labelled DNA probes for RFLP analysis in coffee, in: Bervillié A., Tressac M. (Eds.), Techniques et utilisations des marqueurs moléculaires, Les Colloques n° 72, Inra, Paris, 1995, pp. 21–25.

Lashermes P., Cros J., Combes M.C., Trouslot P., Anthony F., Hamon S., Charrier A., Phylogenic studies of coffee species using chloroplastic DNA, Theor. Appl. Genet. 93 (1996) 626–632.

Lashermes P., Combes M.C., Trouslot P., Charrier A., Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA, Theor. Appl. Genet. 94 (1997) 947–955.

Lebart L., Morineau A., Tabard N., Techniques de la description statistique. Méthodes et logiciels pour l'analyse des grands tableaux, Dunod, Paris, 1977.

Luo H., Van Coppenolle B., Seguin M., Boutry M., Mitochondrial DNA polymorphism and phylogenetic relationships in *Hevea brasiliensis*, Mol. Breeding 1 (1995) 51–63.

Mackay M.C., One core collection or many? In: Hodgkin T., Brown A.D.H., Hintum T.J.L. van, Morales E.A.V. (Eds.), The Use of Plant Genetic Resources, John Wiley & Sons, Baffins Lane, Chichester, UK, 1995, pp. 199–210.

Mackill D.J., Classifying japonica rice cultivars with RAPD markers, Crop Sci. 35 (1995) 889–894.

Mahajan R.K., Bisht I.S., Agraval R.C., Rana R.S., Studies of South Asian okra collection: methodology for establishing a representative core set using characterization data, Genet. Resources Crop Evol. 43 (1996) 249–255.

Morden C.W., Doebley J., Schertz K.F., Allozyme variation in old world races of *Sorghum bicolor* (Poaceae), Am. J. Bot. 76 (1989) 247–255.

Nicolas D., Chevallier M.H., Clément-Demange A., Contribution to the study and evaluation of new germplasm for use in Hevea genetic improvement, in: Jacob J.L., Prevôt J.C. (Eds.), C.R. IRRDB Hevea Meeting, Exploitation-Physiologie et Amélioration de l'Hévéa, IRCA-CIRAD, Paris, France, 1988, pp. 335–352.

Nei M., Estimation of average heterozygoty and genetic distance from a small number of individuals, Genetics 89 (1978) 583–590.

Noirot M., Hamon S., Anthony F., The principal component scoring: a new method of constituting a core collection using quantitative data, Genet. Resources Crop Evol. 43 (1996) 1–6.

Oka H.I., Intervarietal variation and classification of cultivated rice, Indian J. Genet. Plant Breed. 18 (1958) 79–89.

Ollitrault P., Arnaud M., Chantereau J., Polymorphisme enzymatique des sorghos. II. Organisation génétique et évolutive des sorghos cultivés, Agronomie Tropicale 44 (1989) 211–221.

Pederson G.A., Faibrother T.E., Greene S.L., Cyanogenesis and climatic relationships in the US white clover germplasm and core subset, Crop Sci. 36 (1996) 427–433.

Rana R.S., Kochhar S., Core subsets of base collections and priorities of national programmes: Indian perspectives, Genet. Resources Crop Evol. 43 (1996) 423–428.

Saporta G., Probabilités, analyse des données et statistiques, Technip, Paris, 1990.

Schoen D.J., Brown A.H.D., Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers, Proc. Natl. Acad. Sci. USA 90 (1993) 10623–10627.

Second G., Origin of the genetic diversity of cultivated rice (*Oryza sativa* L.): study of the polymorphism scored at 40 isozyme loci, Jpn. J. Genet. 57 (1982) 25–57.

Second G., Evolutionary relationships in the *Sativa* group of *Oryza* based on isozyme data, Genet. Sel. Evol. 17 (1985) 89–114.

Second G., Ghesquière A., Cartographie des introgressions réciproques entre les sous-espèces indica et japonica de riz cultivé (*Oryza sativa* L.), in : Colloque Techniques et Utilisations des Marqueurs Moléculaires, Les colloques de l'Inra 72, Inra, Paris, 1994, pp. 83–93.

Seguin M., Besse P., Lebrun P., Chevallier M.H., Hevea germplasm characterisation using isozymes and RFLP markers, in: Baradat P., Adams W.T., Müller-Starck G., (Ed.), Proceedings of the Symposium on Population Genetics and Genetic Conservation 24–28 August 1992, IUFRO Meetings, SPB Academic Publishing BV, Amsterdam, The Netherlands; France, 1995, pp. 129–134.

Seguin M., Besse P., Lespinasse D., Lebrun P., Rodier-Goud M., Nicolas D., Hevea molecular genetics, Plantation Recherche Développement 3 (1996) 77–88.

Snowden J.D., The Cultivated Races of Sorghum, Adlard, London, 1936.

Spagnoletti-Zeuli P.L., Qualset C.O., Geographical diversity for quantitative spike characters in a world collection of durum wheat, Crop Sci. 27 (1987) 235–241.

Spagnoletti-Zeuli P.L., Qualset C.O., Evaluation of five strategies for obtaining a core subset from large genetic resources collection of *Triticum durum*, Theor. Appl. Genet. 87 (1993) 295–304.

Tohme D.O., Gonzalez S., Beebe S., Duque M.C., AFLP analysis of gene pools of a wild bean core collection, Crop Sci. 36 (1996) 1375–1384.

Virk P.S., Ford-LLoyd B.V., Jackson M.T., Newbury H.J., Use of RAPD for the study of diversity within plant germplasm collections, Heredity 74 (1994) 170–179.

Wang Z.Y., Tanksley S.D., Restriction fragment length polymorphism in *Oryza sativa* L., Genome 32 (1989) 1113–1118.

Wycherley P.R., Rubber, in: Simmonds N.W. (Ed.), Evolution of Crop Plants, Longman Group Limited, Harlow, UK, 1979, pp. 77–80.

Zhong G.Y., Qualset C.O., Quantitative genetic diversity and conservation strategies for an allogamous annual species, *Dasypyrum villosum* (L.) Candargy (Poaceae), Theor. Appl. Genet. 91 (1995) 1064–1073.