Vol. 65 August, 1995 No. 8

Plant Breeding Abstracts

Core collections - accomplishments and challenges

S. Hamon[1], S. Dussert[1], M. Noirot[1], F. Anthony[3] and T. Hodgkin[2]
[1]ORSTOM, BP 5035, 34032 Montpellier, France,
[2]IPGRI, Via delle Sette Chiese 142, 00146 Rome, Italy,
[3]CATIE, Aparto 59, 7170 Turrialba, Costa Rica

ABSTRACT

The major issue facing gene bank managers is the need to improve the accessibility of their collections to a large panel of users, from plant breeders to fundamental geneticists. The core collection concept was proposed as one approach to this problem. A core collection has been described as a collection which contains, with a minimum of repetitiveness, the greatest possible genetic diversity of a crop species and its wild relatives. Such a collection is not intended to replace existing gene bank collections, but rather to render the variation within such collections more accessible to users. The theory was developed on the basis of neutral markers, and the objective of including over 70% of the total variation within 10% of the whole collection can certainly be achieved. A number of procedures were developed from the R strategy (random sampling), in order to take into account the hierarchical structure of the gene pool, or when possible the genetic structure of genotypes. It now seems clear that a hierarchical approach is a prerequisite, but the hierarchical basis is still contentious. Strategies must be adapted to knowledge of the gene pool (taxonomy, evolutionary history, domestication process etc.). When germplasm information is scarce, it is possible to use the country of origin as a good starting point. The screening of a large collection with molecular markers remains difficult and expensive, but these markers could help curators solve a great many problems. It is also clear that breeders, ultimately the most important users, understand and agree with the fundamental neutral character approach, but are interested in agronomic quantitative characters. The work carried out to date shows that it is possible to develop core collections using existing data on the accessions in a collection. Several strategies adapted to each application have been derived from the original theory. They provide a relatively simple and effective way of improving the accessibility of collections and the quality of information on variation in collections. This is a prerequisite for improving the use of plant genetic resources held in collections.

INTRODUCTION

Gene bank managers have a responsibility to optimise conservation of the total variation in a crop. The sheer size of many collections has frequently been cited as a barrier to their increased use (Holden, 1984).

Plant breeders are interested in having fairly small numbers of genotypes which possess, or are likely to possess, the characters needed in their breeding programmes. In contrast, molecular biologists require representative samples for more in-depth studies on neutral characters. Frankel and Brown (1984) proposed that one way of satisfying these conflicting needs lay in the development of core collections. These would represent, with minimum repetitiveness, the genetic diversity of a crop species and its relatives.That is, they would be limited sets of accessions derived from existing germplasm collections, chosen to represent the genetic spectrum in the whole collection and including as much as possible of its genetic diversity.

The idea of developing core collections aroused considerable interest and discussion among genetic resources workers. Brown (1989 a, b) explored a number of the issues involved in developing such collections and other research workers initiated programmes to develop core collections for particular crop germplasm collections.

Table 1. Comparison of molecular electrophoretic markers in evolutionary and breeding studies.

	Allozymes	RFLPs	Seed proteins	Minisatellite sequences	RAPDs
Polymorphism	low	low-high	high	very high	low-high
Environmental stability	moderate	high	high	high	high
Number of loci	moderate(<50 loci)	high	low (<10 loci)	moderate	high
Molecular basis of polymorphism	simple	intermediate	complex	complex	complex
Practicality	quick, cheap	slow, expensive	quick, cheap	intermediate	quick, expensive

Source : Gepts, 1994.

The purpose of this minireview is to assess the current status of the core collection concept. The paper is divided into four parts. The first is devoted to the choice of an adequate hierarchical structure for the target species or gene pool. The second describes the sampling strategies that have been developed both for neutral and non-neutral characters. In the third part we examine the divergence between the concept, the supposed limits and constraints. In the fourth part an overview is given of the current applications of the concept.

THE CHOICE OF AN HIERARCHICAL STRUCTURE

Brown (1989a) argued that, before sampling, the most effective strategy for developing a core collection would involve hierarchical stratification of the whole collection into groups of accessions that share common characters. The criteria used could be taxonomic, geographical or ecological, or could be based on shared, neutral or non-neutral characters. As in all aspects of the management of genetic resources, the first step is to determine the available data (Asley, 1987). The availability and use of these data depend on the history of the crop plant, the target species (cultivars, landraces, wild species), and the practical needs of plant breeders or other users.

The first stage in the development of any core collection is to assemble the available data on the whole collection. These will certainly include the passport data available to the gene bank and as much characterization data as have been collected. Evaluation data may also be included, and, where available, data from biochemical studies of isozymes, seed proteins and molecular markers. Practical experience suggests that there will always be problems at this stage as to the completeness and reliability of available data. Passport data may be minimal for many accessions, with only the country of origin recorded. Peeters and Williams (1984) estimated that for most collections the only passport data available for about 75% of the accessions was the country of origin. However, Hodgkin (1991) found that precise origin data (latitude and longitude) were available for 78% of the world's *ex situ* holdings of wild *Triticum* and *Aegilops* species. Even when detailed passport data are available, characterization and evaluation data may be lacking for a significant proportion.

Core collections should not contain any duplication and the identification of duplicates in a gene bank would be an important part of developing such a collection. The identification and management of possible duplicates in gene banks has been discussed by Van Hintum (1994).

1. Grouping by country of origin

There is considerable evidence to suggest that country of origin, which is commonly recorded in collections, could be a reliable unit for grouping (Peeters and Martinelli, 1989; Holbrook *et al.*, 1993). In addition, where there are reliable agronomic data, such an approach may be modified to take into account phenotypic similarity of accessions from different countries (Spagnoletti-Zeuli and Qualset 1987, 1993; Diwan *et al.* 1994; Hamon *et al.* 1995, Matthews and Ambrose, 1994). When possible, ecogeographical data or geographical patterns of variation can be used to complement grouping based on country of origin (Perry *et al.*, 1991; Peeters *et al.*, 1993).

2. Grouping on neutral markers

Gepts (1995) considered that genetic diversity was best quantified on the basis of data as close to DNA as possible. The heritable material comprises genomic and cytoplasmic DNA. It may differ in DNA sequences and alleles, but also in allele combinations and genotypes. A comparison of the properties of molecular and isozyme markers in evolutionary and breeding studies is given in Table 1.

Before the rise of PCR (polymerase chain reaction) technology, at the end of the 1980s, isozymes were widely used in studies of both interspecific and intraspecific diversity. They have been used for taxonomic studies, in the search for the centre of diversity of a species, to trace the route of domestication, to study the relation between environment and diversity, and to study a complete gene pool. Brown (1990) underlined the advantages of isozymes in the design of sampling strategies for in-depth diversity analysis. In particular, the relatively low technology level needed means that large populations can be studied.

In the past decade, the two primary sources of molecular variation exploited for phylogenetic purposes have been the choloroplast genome and the nuclear ribosomal DNA of repeat region. The chloroplast genome varies little in size, structure and gene content among angiosperms (135-165 kb). Olmstead and Palmer (1994) reviewed methods and data analysis used with chloroplast DNA. These authors indicated that in most reports the use of PCR-based strategy (i.e. RAPD [random amplified polymorphic DNA]) for various applications naturally paints a rosy picture of the successful results obtained. Yet when researchers get together off the record to discuss their results, stories of unanticipated and sometimes inexplicable results usually come up. Although molecular studies have clearly illuminated relationships within many groups of plants, especially at lower taxonomic levels, most of the germplasm accessions maintained in collections are uncharacterized at this level. The fact that PCR requires only minute amounts of DNA suggests that accessions will be studied more in the near future.

DNA fingerprinting with specific oligonucleotide probes for simple repetitive DNA sequences could be very useful for germplasm characterization (Beyermann *et al.*, 1992). In comparison to PCR, DNA fingerprinting allows good discrimination of genotypes but is awkward and time consuming. Describing their experience with gene banks, Lux and Hammer (1994) suggested that PCR-based assays should be used for discrimination of greater genetic differences between accessions (i.e. the definition of an hierarchical structure) and DNA fingerprinting to

distinguish between closely related forms (i.e. elimination of duplicates). Melchinger *et al.* (1994) characterized 48 accessions (24 winter, 24 spring) of European barley germplasm. They found that 7 probe-enzyme combinations were able to distinguish closely related lines.

Currently, for most crops, it is possible to develop phenograms or cladograms which could serve as a basis for a hierarchical structure using neutral markers. In addition, molecular tools are likely to be particularly useful for in-depth analysis and the identification of duplicates.

3. Grouping on neutral markers and correlation with neutral markers

Quantitative markers are dependent on both genotype and environment, but include useful morphological and agronomic characteristics (e.g. earliness, drought resistance) not all of which are always suitable for the study of genetic diversity.

The expression of morphological and agronomic characteristics indicates adaptation to environmental factors. Hintum (1995) suggested adding different major features of the crop such as 2-rowed and 6-rowed types of barley, or spring and winter wheat, or a mixture of these features as grouping criteria. When morphological variation is to be used for phylogenetic analysis, it is advisable to take note of characteristics (Stevens, 1991) and the clustering process (Williams, 1971). In particular, characters should be delimited by carefully analysing discontinuities.

A major problem is the degree of concordance between molecular markers and morpho-agronomic traits. Although country of origin is useful in establishing a core collection, there is not always a clear correlation between geographic distance and genetic distance (Lefort-Busson and Devienne, 1985). Gepts (1995) concluded that it was always possible to identify examples showing, or not showing, concordance between patterns of diversity identified either by molecular and biochemical markers or by morpho-agronomic traits. Perhaps Hillis (1987) provided a useful compromise: morphological work on large samples, molecular work on smaller samples, and studies that combine them to thereby maximise both information content and usefulness. Nevertheless, molecular and taxonomic classifications may well be incongruous as in the case of *Brassica* (Prahan *et al.*, 1992).

SAMPLING STRATEGIES

When the hierarchical structure of a core collection is adopted, the second step of the process is to sample accessions within each group. As for all genetic problems, some argue that neutral markers are essential, while others suggest that they only need adaptive or agronomically useful characters.

1. Models developed on the basis of neutral alleles

The first model on neutral alleles was developed by Brown (1989b). In this model, the collection was considered as a unit and four classes of alleles were identified in a given germplasm collection according to their frequency and distribution: (i) common and widespread, (ii) common and localized, (iii) rare and widespread, and (iv) rare and localized. The sampling procedure was developed for class iii. Other classes of alleles were removed for different reasons: class i will almost certainly be included; class iv does not lead to any guiding principle other than the impracticability of conserving anything; class ii will largely be included when groups have been successfully described. For class iii, on the basis of the neutral allele model developed by Kimura and Crow (1964) the expected number of neutral alleles with frequency greater than a definite number in the population is given in Table 2.

Table 2. Expected number of neutral alleles (n_a) of frequency greater than (p) in the population and the expected number of alleles (n_r) in a sample of size N_r

	Level of polymorphism		
	$\theta = 0.5$	$\theta = 1.0$	$\theta = 2.0$
Frequency class (p)	Number of alleles in the collection (n_a)		
> 0.10	1.82	2.30	2.8
> 0.01	2.99	4.61	7.2
> 0.001	4.15	6.91	11.8
> 0.0005	4.50	7.60	13.2
> 0.0001	5.30	9.21	16.4
Sample size (N_r)	Number of alleles in the sample (n_r)		
50	3.15	5.1	8.3
100	3.50	5.8	9.7
300	4.05	6.9	11.7
1.000	4.65	8.1	14.3
3.000	5.20	9.2	16.5
10.000	5.80	10.4	18.9

Source : Brown (1989a)

Brown (1989a) showed that a random sample of 10% from the whole collection (or at least 3000 accessions) can be expected to contain over 70% of the variation in a species.

A slightly different approach was developed by Crossa *et al.* (1993). Using probability models for outbreeding crops that incorporate a number of alleles at independent loci, these authors determined optimal sample sizes for regenerating germplasm accessions. Sample sizes with a 95% probability of including at least one copy of alleles with a given frequency are reported in Table 3. As an example, for 100 loci each with 10 alleles, 191 individuals are required to retain with 95% probability at least one copy of one allele with a frequency of 0.05.

Table 3. Sample sizes required to achieve a 95% probability of including at least one copy of alleles with p_o of 0.05, 0.03 and 0.01 from each allele class for several alleles at each locus.

Number of k alleles	Number of loci 1	2	5	10	50	100	150
				$p_o = 0.05$			
2	58	72	89	103	134	148	156
3	72	85	103	116	148	161	169
4	80	93	111	124	156	169	177
10	101	115	132	146	177	191	198
15	110	123	141	154	186	199	207
				$p_o = 0.03$			
2	98	121	151	173	226	249	262
3	121	143	173	196	249	271	285
4	134	157	187	209	262	285	298
10	170	193	223	245	298	321	334
15	185	207	237	260	313	335	349
				$p_o = 0.01$			
2	298	366	456	525	685	754	794
3	367	435	525	594	754	823	863
4	407	475	565	634	794	863	903
10	517	584	675	744	903	972	1013
15	561	628	719	787	947	1016	1057

Source : Crossa *et al.*, 1993

Brown (1989b) introduced hierarchical sampling and then three new sampling strategies. These involved sampling in a given group according to its size as follows: (i) a constant fraction (C strategy); (ii) a proportion of the number of accessions available per group (P strategy), and (iii) a proportion of the logarithm of the number available per group (L strategy). The L strategy is thought to be the most appropriate since the R, C, P and L strategies are only based on allele frequencies.

Schoen and Brown (1995) suggested strategies H and M, which refer to genetic index. H (heterozygosity) strategy refers to Nei's gene diversity index defined as one minus the sum of squared allelic frequencies at the ith locus in the jth group (Nei, 1973). The M (maximization) strategy differs from all other procedures because it refers to individual accessions and the variance and covariance measured at different loci. It is assumed that variation at a selected number of marker loci is representative of the variation at loci of interest in genetic conservation. Strategies H and M were tested by the authors and compared with those mentioned above. With real data sets, using estimated and target loci, the results showed that for the overall average the ranking of the six strategies from highest (rank 1) to lowest (rank 6) in terms of expected allele retention was M > H > P > L > C > R. Bataillon (1994) used computer simulation to confirm the superiority of the M strategy in maximizing allele inclusion in the core collection. In addition he found that (i) the M strategy could be helpful for maximizing non-neutral diversity for an autogamous species and/or a species subdivided into genetically isolated populations; (ii) only a limited number of markers are necessary (10 in this case).

2. The introduction of non-neutral characters

Spagnoletti-Zeuli and Qualset (1993) utilized an approach derived from Brown's (1989b) suggested strategies. To select a core collection from 3000 *Triticum durum* accessions, four qualitative and eight quantitative characters were used. Each of the following strategies generated about 500 accessions for the core sample: (i) random (R), random systematic according to chronology of entries into the collection, (ii) stratified by country of origin, (iii) stratified by log frequency (L) by country of origin, and (iv) stratified by canonical variables. The first three strategies produce samples representative of the whole collection, but the remaining two produce the desired effect of increasing the frequency from less well represented countries. In addition, the stratified canonical sample increased phenotypic variation (Table 4). These authors concluded that the multivariate approach is useful but requires considerable data from the whole collection.

The 10% sampling is given as a guideline but is modified to take account of the known features of the crop or the needs

Table 4. Means, standard errors, variances and coefficients of variation for quantitative characters of the spike in a world collection of *durum* wheat and for five sampling strategies. Means and variances were tested for significant differences from the world collection values.

Strategy	Spike length			
	Mean	SE	Var	CV
World collection	76.0	0.221	142.75	15.7
Random sample	75.6	0.516	125.69	14.7
Systematic	75.4	0.455	139.51	15.7
Stratified by origin	75.9	0.543	141.14	15.7
Stratified by log of origin	75.1	0.558	147.44	16.2
Stratified by canon var/origin	76.0	0.602	176.60**	17.5

** $P<0.01$

Source : Spagnoletti-Zeuli and Qualset, 1993.

Table 5. Means for six morphological variables for the entire groundnut germplasm collection and for the core collection and the range of these variables in the core collection expressed as a percentage of the range in the entire collection.

Entire collection	Core collection		
Variable	Mean	Mean	Range of entire collection %
Plant type	2.16	2.17	100
Pod type	3.96	3.87	100
Seed size	5.13	5.14	100
Testa color	2.95	3.05	100
Seed per pod	3.23	3.22	80
Seed weight	50.81	51.46	100

Source : Holbrook, 1994.

of the users [Crossa (1989), Crossa *et al.* (1993, 1994) on maize; Hamon *et al.* (1995) on coffee; Diwan *et al.* (1994) on medicago)]. In a study of groundnut (*Arachis hypogaea*) Holbrook *et al.* (1993) did not find differences between the core and entire collections in the mean or in the variation range of the quantitative markers used for developing the core collection (Table 5).

Radovic and Jevolac (1994) determined divergent populations within a pool (populations heterotic to an inbred line) in order to select a core collection. The differences between the pool and the core were determined according to the Goodness of Fit Test (G-statistics) for 18 morphological traits. Four groups of combining ability were defined from a pool of 902 populations tested with 4 divergent inbred testers. With 72 populations (7% selection), it could be assumed that the difference between the core and the pool is only found for one trait.

Noirot *et al.* (1995) described the mathematical principle of a general selection scheme in a given hierarchical group: Principal Component Score Strategy (PCSS). The diversity (i.e. variability) is determined by the between-individual differences for one or more traits. In order to give the same weight to each trait, the Euclidean distance is weighted by the reciprocal of the standard deviation. To avoid the common variable colinearity, a principal component analysis is applied. Each individual is then characterized by its relative contribution (CRi) to the Generalized Sum of Square (GSS). The first step of selection consists in keeping the farthest individual of the set centre as initial sub-set (i.e. the highest CRi). Then at every iteration the cumulative GSS of the sub-set is known and the procedure can be stopped. Sub-set size and GSS are simultaneously taken into account. An application to simulated data gave results that agree with previous findings noted from application to coffee trees (Hamon *et al.*, 1995).

FROM THE CORE COLLECTION CONCEPT TO ITS APPLICATIONS

When a new concept arises it is clear that some researchers will be enthusiastic and others sceptical. A number of doubts have been expressed about the development of core collections and were discussed by Brown (1995). The major concerns can be summarised as follows: (i) what is the future of accessions not included in the core? (ii) a good core needs so much information that it is impracticable to develop and, where it is possible, it will be inflexible with regard to content; (iii) the bias towards representing neutral diversity ignores usefulness.

1. The future of the whole collection

It has been argued that the accessions in large collections will be neglected and that curators will be less prepared to

provide resources for this maintenance if core collections are established. Breeders will express limited interest because they always have their own working collections. It is clear from Brown's definition that the core is not an entity apart, but is a guide and entry point for the whole collection. His main objective is to stimulate and improve the use of genetic resources. This is very difficult with a large base collection. The establishment of a core would increase the value of the whole collection. For this, links between the core collection and the whole collection are needed. The core collection provides a first step in a two-stage process of identifying the most desirable accession. When the core is a separate, small collection of each germplasm collection, the risk is evident. This risk is lower in cases such as European barley or American groundnut (see below). For barley, the core is the result of collaboration and is built up with samples from different sources. For groundnut, the allocation of an accession to the core is only an index in a computer file and the samples are increased only for exchange purposes.

The core is impracticable and inflexible

It has been argued that the creation of a reasonably representative core collection would require so much information that it is impracticable. When sufficient information has been collected to create such a collection, it would be unnecessary. The examples noted below show that core collections can be constructed on the basis of what is already known. Limited passport data and basic characterization data for major morphological characters can provide an effective grouping.

Some believe that the core collection is inflexible and that once developed it cannot be changed. This is not so. It is difficult to estimate the rate at which changes should be made in the composition of the core, but the whole collection changes so the core can be expected to change. For certain crops in some gene banks, it may be sufficient to mark the accessions in a database as belonging to the core and to identify the different groups of which they are the representatives (Holbrook *et al.*, 1993). The difficulties in handling large and growing numbers of accessions in gene banks during the decade prior to 1984 were largely responsible for the core collection. Mackay (1995) has argued that this limit on use is largely negated by modern database technology. This author discussed the need for one or many cores. One basis for considering many core collections is that germplasm users often request a set of accessions that is likely to contain a previously undescribed characteristic. The geographical distribution of boron tolerance is a good example (Table 6).

Mackay also considered that the core collection concept has the merit of selecting a set representative of diversity from a larger assemblage of germplasm, but different germplasm users will have various objectives in sampling germplasm. The core(s) collection(s) could be included in the base collection; accessions will only be amplified for diffusion. Where collections are extremely well characterized and of reasonable size, the techniques proposed by Mackay (1995) may be of considerable value, but there will remain many situations where a core collection is still the most rational basis for improving use of a collection and management of the whole collection.

3. What about the usefulness?

Does the representation of neutral diversity ignore usefulness? It is true that in many instances the core is unlikely to contain the single most useful source of agronomic characters. As pointed out by Brown *et al.* (1995), it provides a logical general strategy to identify the best source. One advantage for the breeder is the chance to become acquainted with the diversity of phenotypes in a crop and its related wild species. The strategy to be adopted depends on the nature of the data available and the extent to which genetic and agronomic considerations are balanced in the development of the core collection. In addition, it has always been recognised that core collections will not necessarily include extremely rare genes. A proportion will be included, but this will be accidental and many will not be present. In the examples developed below, it is clear that usefulness is predominant on population genetics, on which the core collection theory is largely based. A much higher rate of change occurs in genetic diversity during domestication than during the relatively slow process of natural evolution. Plant breeding techniques could influence the level of diversity. Diversity can clearly be increased by several techniques, such as induced mutations, hybridization between previously incompatible populations and introgression. In contrast, inbred lines, cultivars for high input agriculture and vegetative propagation lead to reduction in diversity.

AN OVERVIEW OF CURRENT CORE COLLECTIONS

We shall not attempt to review here all examples of existing core collections but rather to give an overview of practical applications of the core collection concept. None of these core collections strictly corresponds to the definition. Some include only cultivated species, others include wild species but with a very variable approach.

Table 6. The geographical distribution of boron-tolerant plants based on visual assessment of tolerance. VS indicates very sensitive, S = sensitive, MS = moderately sensitive, MT = moderately tolerant, T = tolerant, VT = very tolerant. (This table is reproduced by kind permission of Moody *et al.* 1988).

Region or Origin	Visual assessment of tolerance						Total	Frequency of MT-VT classes (%)
	VS	S	MS	MT	T	VT		
N. America	11	40	27	5	-	2	85	8
Mexico	6	34	31	9	3	1	84	15
S. America	3	12	21	12	8	-	56	36
Europe	18	98	98	26	7	3	250	14
Asia-Asia Minor	22	82	89	105	49	10	357	47
Australia	5	31	45	8	1	-	90	10
Egypt	25	22	7	-	-	-	54	0
Kenya, NW Africa	5	10	9	5	2	-	31	23
Unknown origin	90	220	192	46	18	3	569	12
Total	185	549	519	216	88	19	1576	
Frequency (%)	12	35	33	14	5	1		

Source : Mackay, 1994 (Moody 1988).

1. Examples of cultivated species only

Core collections devoted only to cultivated species have been set up for cassava, phaseolus bean, groundnut, maize, ryegrass (*Lolium*) and coffee.

A Brazilian cassava core collection has been developed by Cordeiro *et al.* (1995). Brazil is an area of genetic diversification of the genus *Manihot*, but wild species were not taken into account and the study only concerns the cultivated species *Manihot esculenta*. Brazilian cassava accessions are distributed among a number of centres involved in maintenance or breeding of cassava, and the data collection phase established the origin of the existing accessions and allowed putative duplicates to be identified. Some 4132 accessions were identified from different collections in Brazil, of which 1200 were considered to be duplicates. This phase also led to the development of an effective hierarchical classification in which the first criterion was the category of the material (landraces 2035, improved selections 339, unknown status 558) and the second was the agroecological zone of origin, of which nine were defined. A final criterion for grouping was based on characterization and evaluation data. This has shown that the data assembly phase can be valuable in its own right to improve gene bank management, but we are very far from the initial model and a large amount of work remains to be done.

For the CIAT *Phaseolus* core collection, the same principles as for cassava were applied, but in a much more sophisticated manner (Tohme *et al.*, 1995). Following preliminary grouping according to the known history of bean cultivation, more weight was given to centres of high diversity by constructing an agroecological classification of Latin America in which each 10-minute grid was classified into one of 54 classes on the basis of soil type, altitude, available water and photoperiod. The origin of the Latin American landraces for which passport data were available was determined and they were grouped into the 54 identified agroecological groups. A second step involved a weighting process designed to ensure that variation in growth habit, seed colour and seed size were fully represented in the core collection.

The US germ plasm collection of peanut consists of 7432 accessions. Holbrook *et al.* (1993) used data from the US Germplasm Resources Information Network (GRIN) to select a core collection. Data on peanut in the GRIN included country of origin and plant type, pod type, seed size, testa colour, number of seeds per pod and average seed weight. The entire collection was stratified by country of origin and by the amount of available morphological data. Using available data, this simple procedure established 9 sets. Random sampling (10%) was then used to select from each group. The resulting 831 accessions form the core collection. Examination of the six phenotypic traits indicated that the genetic variation expressed in the entire collection is preserved in the core. For all peanut plant introductions, relationships to the clustering procedure described above were included in the GRIN, as well as the core number for the selected accessions. In this example the authors adopt a hierarchical approach based on the country of origin. The sampling strategy is derived from the global 10% defined in Brown's model, but no neutral characters are used as test loci.

Suggestions for a maize core collection were made by Crossa *et al.* (1993). The collection was subdivided into non-overlapping groups based on racial complex and/or selected ecogeographical criteria. Within each racial complex, sub-samples of 25-100 accessions were selected. Such subsets will preserve alleles with frequencies higher than 0.03 in each race collection. Within each race complex, accessions can be grouped by region or elevation. Several morphological and agronomic attributes should be measured in multi-location trials. Cluster analysis could then help the curator to identify similar accessions. Crossa *et al.* (1994) gave practical consideration to maintaining germplasm in maize, for which 23 races and 3 subraces were described in Mexico. These authors considered the example of the race Tuxpeno, for which 848 accessions are available. A set of 175 accessions was selected based on lodging and adaptation assessed in multiple location trials. The second criterion was ecogeographic (dry ecology vs. wet ecology), and the third was derived from cluster analyses and principal component analysis. A very different approach is suggested by Radovic and Jelovac (1994), who established the hierarchy by using the combining ability compared with testers. They selected 7% of 902 maize populations of the Yugoslav Maize Genebank for the core.

Balfourrier *et al.* (1994) mention that many studies have been carried out concerning the loss of alleles as a result of random drift or founder effect (bottleneck) caused by a reduction in size. Most studies are theoretical and few experimental investigations have been made, and then only on animals. Balfourrier and Charmet (1994) built a core collection from 547 natural populations of perennial ryegrass (*Lolium perenne*) studied for agronomic characteristics. Forty-two populations were chosen and regrouped in order to create 9 breeding or pool populations representative of the observed diversity. For this purpose, agronomic and ecogeographic criteria were taken into account and multivariate analyses were used. Balfourrier *et al.* (1994) found that the allelic multiplicity and genotypic frequencies, tested with six isozyme markers, were conserved in 3 experimental breeding populations. Only very rare alleles ($p<0.01$) were reduced in frequency or lost by bulking 4 or 5 natural populations represented, respectively, by 20 to 25 plants, in a large polycross design. This approach is intermediate between the maintenance of pre-breeding populations and the concept of the core collection.

A slightly different procedure was used with coffee (Hamon *et al.*, 1995). A hierarchical structure was given for the genus (see below), but a particular sampling method was developed for one cultivated species (*Coffea liberica*). The Principal Component Score was used to develop a core collection for quantitative data. It was shown that about half the inertia was obtained when 10% of the 338 genotypes were selected, and 90% of the total inertia was obtained with a sample of 50% of these genotypes. The advantage of this procedure is that existing evaluation data can be used to identify the core accessions and that the process can be adapted to breeders' needs in respect of the numbers selected (Noirot *et al.*, 1995). The selected number did not depend on group size but on the original and selected variability.

2. Examples including wild species

For some genera, such as annual *Medicago, Pisum, Coffea* and *Hordeum*, core collections have been developed using wild species.

The United States National Plant Germplasm System contains 3159 accessions from 36 species of annual *Medicago*. This germplasm is under-utilized but there is an increase in interest in the field of sustainable agriculture. Diwan *et al.* (1994) selected a core collection by evaluating 1240 accessions for 16 agronomic traits. Within species, accessions were grouped by cluster analysis utilizing an unweighted pairing group method with arithmetic averages. A minimum of one accession per cluster was selected for each species for the core. Accessions were chosen within a species to represent the greatest diversity in geographical regions. The selected core collection of 211 accessions was re-evaluated and found to represent the variability of the germplasm collection and to be stable between evaluation years. The level of selection was highly variable (Table 7). Selected samples per species ranged

Table 7. The number of accessions of studied annual *Medicago* species in the US National Plant Germplasm System collection, in the initial subset, and in the core collection.

Medicago Species	US Collection	Initial subset	Core collection
		Number	
arabica (L.) Huds.	71	35	2
blancheana Boiss.	18	18	8
ciliaris (L.) Krocker	73	31	6
constricta Durieu	48	30	3
coronata (L.) Bart.	23	3	2
disciformis DC.	50	30	4
doliata Carmign.	127	40	3
granadensis Willd.	14	13	4
heyniana Greuter	2	2	1
intertexta (L.) Miller	22	19	6
italica (Miller) Fiori	83	32	9
laciniata (L.) Miller	130	52	10
lanigera Winkl. & Fedtsch	1	1	1
lesinsii E. Small	5	2	2
littoralis Rohde ex Lois.	120	50	6
lupulina L.	170	63	14
minima (L.) Bart.	274	101	4
murex Willd.	73	34	6
muricoleptis Tin.	7	7	1
noeana Boiss.	19	14	3
orbicularis (L.) Bart.	251	86	8
platycarpa (L.) Trautv.	6	5	1
polymorpha L.	651	217	36
praecox DC.	21	20	2
radiata L.	12	11	4
rigidula (L.) All.	329	104	6
rotata Boiss.	21	20	8
rugosa Desr.	43	28	11
sauvagei Negre	5	5	2
scutellata (L.) Miller	60	37	18
secundiflora Durieu	2	2	1
shepardii Post	4	4	1
soleirolii Duby	10	10	3
tenoreana Ser.	6	5	1
truncatula Gaerth.	325	71	8
turbinata (L.) All.	83	38	6
Total	3159	1240	211

Source : Diwan *et al.*, 1994.

from 100% (*M. lanigera* 1/1) to 1% (*M. minima* 4/274, *M. rigidula* 6/329) or 5% *M. polymorpha* (36/651).

A subset of accessions of the John Innes *Pisum* collection (total 2500) has been used for screening purposes for several years. Accessions were sorted into groups according to (i) species and subspecies, (ii) ecotypes and landraces, (iii) cultivars representative of various end-uses and, (iv) genetic stocks. Matthews and Ambrose (1994) moved this reference collection closer to Brown's concept. They used: (i) passport data, (ii) cross-referencing with morphological data, and descriptive statistics of revised groupings, and (iv) selection of representative accessions. The core collection currently stands at 157 accessions, which corresponds to a 6% selection.

The procedure used with *Coffea* was slightly different (Hamon *et al.*, 1995). The genetic organization of the coffee gene pool was examined in terms of biogeography, genetic resources and available data, in order to determine the hierarchical structure of a theoretical core collection. A core collection for coffee should consist of 88 genetic diversity groups (i.e. hierarchical class). Different strategies could be applied for the groups. Some strategies for cultivated species are well documented, others with botanically undescribed wild species are only known by the site of the collection. An *in vitro* core collection has been constituted with 33 groups of diversity corresponding to 15 species and 580 genotypes (LRGAPT, 1994). The selection rate from the base collection was about 7%.

The Barley Core Collection Project is a collaborative international initiative (Knüpffer and Hintum, 1995). The accessions selected (approximately 2000) will cover the entire *Hordeum* gene pool with defined numbers of landraces, improved cultivars and wild relatives from the primary, secondary and tertiary gene pools. Genetic stocks will also be included. It is envisaged that the core collection will be held at several centres and, to ensure that it remains constant, it is intended that each accession shall be an homozygous line.

This raises an interesting new dimension in core collection work in that the accessions now become additional entities maintained separately from those from which they were derived. The cultivated barley species is *Hordeum vulgare*. The species is diploid and shares the primary gene pool with the *H. spontaneum* complex. The secondary gene pool comprises *H. bulbosum*. The tertiary gene pool includes about 30 species. For this complex, the core collection is not a selection from the germplasm collection of a single institution but from the entire gene pool of the crop. It is a part of an existing gene bank but maintained separately. The objectives are to facilitate the co-ordination of efforts and to share responsibilities, to increase knowledge about barley in general, to use the existing germplasm, and to provide standards for studies on genetic diversity. This implies the accumulation of a large amount of data for a limited standard set of accessions. The starting point for research is a small sample covering a considerable part of the whole collection, thus avoiding expensive screening of large collections with duplicated material. The size should not exceed 2000 accessions. It is as follows: Category 1 (C1) 500 cultivars (phylogenetic group, including oriental and occidental); C2, 800 landraces (ecogeographical data, agricultural system practised, type of use) ; C3, 150-200 accessions of *H. spontaneum* (including *Agriocrithon* and introgression products with cultivated barley) (ecogeographical data: 2/3 from central areas, 1/3 from marginal areas); C4, 60-100 accessions of other wild species (2 per species - ecogeographical and morphological data); C5, 200 genetic stocks and reference materials from barley genetics experts. To ensure continued integrity, accessions will be homozygous and homogeneous lines as far as possible. Heterogeneous samples are only accepted for the 2 outcrossing species. The advantages are two-fold: identical guarantees for multiplication over generations and locations, and correspondence between information and material. Disadvantages are that variation within landraces is not reflected, with considerable reduction in the number of alleles. It is presumed that a single line of the wild relatives contains the genetic background common to the material it represents.

CONCLUSIONS

For most crops, it is possible nowadays to gather data on the organization of the genetic diversity of cultivated species and wild relatives. The results of taxonomic studies, knowledge of the centre of origin and diversity of species, routes of domestication, and relations between environment and diversity, must be collated. This yields a general scheme which can be used to define a hierarchical structure. A study which takes into account neutral and

non-neutral characters maximizes both information content and usefulness.

The global 10% defined in Brown's model is accepted as a reference and is often cited in the literature. When the number of accessions varies greatly from one group to another, the L strategy can be used. In the near future, when molecular data become available in large data sets, the M strategy will probably be widely used. Applications of the sampling procedures defined with neutral markers are applicable to non-neutral markers. Multivariate analyses show that overall means are not changed. In addition, some sampling strategies could be of greater help in optimizing both neutral and non-neutral characters.

When a gene bank, a network, or a consortium decides to develop a core collection, the risk to the whole collection is low for major crops. The problem could arise with minor crops for which the search for funding is more difficult at every stage. The core collection in this case could still be more helpful than deleterious. We have seen that the theoretical concept is not a severe limitation for the development of cores. When research groups are interested, they develop their core collection using what they want and with their own objectives in mind.

Ten examples give a general overview of possible interpretations of a unique concept for a large panel of plants. Their common points are: (i) the need for a reference collection of reduced size which contains substantial genetic polymorphism; (ii) the acceptance of the hierarchical classification approach; (iii) the theoretical acceptance of the 5-10% selection rate in a given cluster; (iv) the absence of an extensive data set which allows the use and testing of strategies defined for neutral characters. The different points are: (i) some researchers initially consider only cultivated forms; (ii) some of those studying wild species accept representation by only one sample; (iii) priority for non-neutral characters depends on area of origin, major features of the crops, or the status of pre-breeding level. These collections reflect the first attempts to put a theoretical concept into practice. Further work will provide new insights into how this can best be done and increase our knowledge, and hence use, of germplasm collections.

REFERENCES

Asley, D. (1987) Genetic resource conservation. *Experimental Agriculture* **23**, 245-257.

Balfourrier, F.; Charmet, G. (1994) Etude méthodologique de la conservation des ressources génétiques de rye-grass anglais (graminée fourragère) par multiplication en pools de populations naturelles. *Genetics, Selection, Evolution* **26**, suppl. 1, 203-218.

Balfourrier, F.; Charmet, G.; Grand-Ravel, C. (1994) Conservation of allelic and genotypic frequency by pooling wild populations of perennial rye-grass. *Heredity* **73**, 386-396.

Bataillon, T. (1994) Comparaison de diverses stratégies d'échantillonage pour la constitution de core collections de ressources génétiques végétales (étude par simulation informatique). Mémoire de DEA de Ressources Génétiques et Amélioration des Plantes. Paris, France; INA-PG. 43 pp.

Beyermann, B.; Nürnberg, P.; Weihe, A.; Meixer, M.; Epplen, J. T.; Börner, T. (1992) Fingerprinting plant genomes with oligonucleotide probes specific for simple repetitive DNA sequences. *Theoretical & Applied Genetics* **83**, 691-694.

Brown, A. H. D. (1989a) The case for core collections. In: *The use of Plant Genetic Resources* Cambridge, UK; Cambridge University Press. pp. 136-156

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

Brown, A. H. D. (1990) The role of isoenzyme studies in molecular systematics. *Australian Systematic Botany* **3**, 39-46.

Brown, A. H. D. (1995) The core collection at the crossroads. In: *Core Collections of Plant Genetic Resources*. Chichester, UK; John Wiley & Sons. pp. 3-20.

Charmet, G.; Balfourrier, F.; Ravel, C.; Denis, J. B. (1993) Genotype \multiply\ environment interactions in a core collection of French perennial ryegrass populations. *Theoretical & Applied Genetics* **86**, 731-736.

Cordeiro, C. M. T.; Morales, E.; Ferreira, P.; Rocha, D. M. S.; Costa, I.; Valois, A.; Silva, S. (1995) Towards a Brazilian core collection of cassava. In: *Core Collections of Plant Genetic Resources*. Chicester, UK; John Wiley & Sons. pp. 155-168.

Crossa, J. (1989) Methodologies for estimating the sample size required for genetic conservation of outbreeding crops. *Theoretical & Applied Genetics* **77**, 153-161.

Crossa, J.; Hernandez, C. M.; Bretting, P.; Eberhart, S. A.; Taba, S. (1993) Statistical genetic considerations for maintaining germplasm collections. *Theoretical & Applied Genetics* **86**, 673-678.

Crossa, J.; Taba, S.; Eberhart, S. A.; Bretting, P. (1994) Practical considerations for maintaining germplasm in maize. *Theoretical & Applied Genetics* **89**, 89-95.

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

Frankel, O. H.; Brown, A. H. D. (1984) Current plant genetic resources - a critical appraisal. In: *Genetics: New Frontiers*, Vol IV. New Delhi, India; Oxford and IBH Publishing Company. pp. 1-11.

Gepts, P. (1995) Genetic markers and core collections. In: *Core Collections of Plant Genetic Resources*. Chichester, UK; John Wiley & Sons. pp. 127-146.

Hamon, S.; Noirot, M.; Anthony, F. (1995) Suggested procedures for selecting a coffee core collection. In: *Core Collections of Plant Genetic Resources*. Chichester, UK; John Wiley & Sons. pp. 117-126.

Hillis, D. M. (1987) Molecular versus morphological approaches to systematics. *Annual Review of Ecology and Systematics* **18**, 23-42.

Hintum (van), T. J. L. (1994) Duplication in germplasm collections. *Proceedings of the Eucarpia meeting: Evaluation and exploitation of genetic resources. Pre-breeding*. 15-18th March 1994, Clermont Ferrand, France. pp.131-139.

Hintum (van), T. J. L. (1995) Hierarchical approaches to the analysis of genetic diversity in crop plants. In: *Core collections of Plant Genetic Resources*. Chichester, UK; John Wiley & Sons. pp. 23-24.

Holden, J. H. W. (1984) The second ten years. In: *Crop Genetic Resources: Conservation and Evaluation*. London, UK; George Allen & Unwin. pp. 277-285.

Hodgkin , T. (1991) The core collection concept. In: *Crop network - new concepts for genetics resources management*. International Crop Network series 4. Rome, Italy; IBPGR.

Holbrook, C. C.; Anderson, W. F.; Pittman, R. N. (1993) Selection of a core collection from the US germplasm collection of peanut. *Crop Science* **33**, 859-861.

Kimura, M.; Crow, J. F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725-738.

Knüpffer, H.; Hintum (van), T. J. L. (1995) The barley core collection - an international effort. In: *Core Collections of Plant Genetic Resources*. Chichester, UK; John Wiley & Sons. pp. 171-178.

Lefort-Busson, M.; De Vienne, D. (1985) Les distances génétiques. Versailles, France; INRA.

LRGAPT (1994) Annual report 1993. Montpellier, France; ORSTOM. 39 pp.

Lux, H.; Hammer, K. (1994) Molecular markers and genetic diversity - some experience from the genebank. *Proceedings of the Eucarpia meeting : Evaluation and exploitation of genetic resources. Pre-breeding*. 15-18th March 1994, Clermont Ferrand, France. pp. 49-53.

MacKay, M. C. (1995) One core collection or many? In: *Core Collections of Plant Genetic Resources*. Chichester, UK; John Wiley & Sons. pp. 199-210.

Matthews, P., Ambrose, M. J. (1994) Development and use of a core collection for the John Innes *Pisum* collection. *Proceedings of the Eucarpia meeting : Evaluation and exploitation of genetic resources. Pre-breeding*. 15-18th March 1994, Clermont Ferrand, France. pp. 99-102.

Melchinger, A. E.; Graner, A.; Singh, M.; Messmer, M. (1994) Relationships among European barley germplasm: 1. Genetic diversity among winter and spring cultivars revealed by RFLPs. *Crop Science* **34**, 1191-1199.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the USA* **70**, 3312-3323.

Noirot, M.; Hamon, S.; Anthony, F. (1995) The principal component scoring: a new method of constituting a core collection using quantitative data. *Genetic Resources and Crop Evolution* 42 (2) (in press).

Olmstead, R. G.; Palmer, J. D. (1994) Chloroplast DNA systematics : a review of methods and data analysis. *American Journal of Botany* 81, 1205-1224.

Peeters, J. P.; Williams, J. T. (1984) Towards better use of genebanks with special reference to information. *Plant Genetic Resources Newsletter* 60, 22-32.

Peeters, J. P.; Martinelli, J. A. (1989) Hierarchical cluster analysis as a tool to manage variation in germplasm collections. *Theoretical & Applied Genetics* 78, 42-48.

Peeters, J. P.; Wilkes, H. G., Galwey, N. W. (1993) The use of ecogeographical data in the exploitation of variation from gene bank. *Theoretical & Applied Genetics* 80 110-112.

Perry, M. C.; McIntosh, M. S.; Stoner, A. K. (1991) Geographical patterns of variation in the USDA soybean germplasm collection. II. Allozyme frequencies. *Crop Science* 31, 1356-1360.

Praham, A. K.; Prakash, S.; Mukhopadhyay ; Pental, D. (1992) Phylogeny of *Brassica* and allied genera based on variation in chloroplast and mitochondrial DNA patterns: molecular and taxonomic classifications are incongruous. *Theoretical & Applied Genetics* 85, 331-340.

Radovic, G.; Jelovac, D. (1994) The possible approach in maize core collection development. *Proceedings of the Eucarpia meeting: Evaluation and exploitation of genetic resources. Pre-breeding.* 15-18th March 1994, Clermont Ferrand, France. pp. 109-115.

Schoen, D. J.; Brown, A. H. D. (1995) Maximizing allelic diversity in core collections of wild crop relatives: the role of genetic markers. In : *Core Collections of Plant Genetic Resources.* Chichester, UK; John Wiley & Sons. pp. 55-76.

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

Spagnoletti-Zeuli, P. L.; Qualset, C. O. (1993) Evaluation of five strategies for obtaining a core subset from large genetic resources collection of *Triticum durum*. *Theoretical & Applied Genetics* 87, 295-304.

Stevens, P. F. (1991) Character states, morphological variation, and phylogenetic analysis. *Systematic Botany* 16, 553-583.

Tohme, J.; Jones, P.; Beebe, S.; Iwanaga, M. (1995) The combined use of agroecological and characterization data to establish the CIAT *Phaseolus vulgaris* core collection. In: *Core Collections of Plant Genetic Resources.* Chichester, UK; John Wiley & Sons. pp. 95-107.

Williams, W. T. (1971) Principles of clustering. *Annual Review of Ecology and Systematics* 2, 303-326.