A Method for Building Core Collections

Michel Noirot, François Anthony, Stephane Dussert and Serge Hamon

About 10,000 years ago, our ancestors discovered agriculture and thereby exerted new selection pressures on wild plants: the process of domestication began. Over the course of millennia, this process led to what are generally called 'primitive varieties'. From the 16th century onwards, with the development of intercontinental migrations, the cultivation of plants went far beyond their zone of origin and diversification. Finally, the 20th century has been marked by the rapid development of selection techniques and modes of cultivation and by the production of new idiotypes. Thus, a considerable diversity of species and forms has been created.

Vavilov (1935) was among the first to demonstrate the importance of gene banks in plant breeding. However, some decades passed before Harlan (1970), Frankel and Bennett (1970) and subsequently Pernes (1984) encouraged the sampling and evaluation of the diversity of natural populations.

The major cultivated plants, their related wild species, and the minor cultivated plants have been sampled on the initiative of the International Board for Plant Genetic Resources (IBPGR). Routine prospections rapidly led to difficulties in managing and conserving the collections thus assembled.

By the late 1980s, collections became enormous and difficult to regenerate and maintain. This considerable growth in collection size, as well as the inadequate documentation available for the samples, have often been cited as limitations on the effective use of genetic resources (Holden, 1984). Potential users require either populations representative of the diversity or accessions that present particular agronomic characters (e.g., disease resistance, drought resistance). In either case the managers of collections find it difficult to meet such needs.

Frankel and Brown (1984) were the first to emphasize the need to constitute small collections with maximal diversity: the core collections. For most users, a core collection helps them avoid the redundancy of genotypes (doubles), often linked to the mode of reproduction or overrepresentation of

cultivated varieties. This redundancy is rare in allogamous species, very frequent in autogamous species, and common in apomictic species or in plants with vegetative propagation.

In terms of practical use, the three major objectives of the core collection are to set up as wide a representation as possible of the genetic diversity, to be able to conduct intensive studies on a reduced set of genotypes, and to attempt to extrapolate the results thus obtained to facilitate research on appropriate genotypes in the base collections.

Presently, in the constitution of a core collection, most researchers agree on the need for a stratification prior to the sampling. In other words, the organization of the variability in groups and subgroups must be taken into account (Frankel and Brown, 1984; van Hintum, 1995; Yonezawa et al., 1995).

On the other hand, several processes are proposed for sampling within groups and subgroups. One such process is to take a random sampling in each previously defined group (or from the whole base collection if the organization into groups is unknown or does not exist). This type of sampling has the advantage of creating a core collection that is statistically representative of the base collection.

In this chapter, we propose a new method of sampling, principal component scoring (PCS), the aim of which is to maximize the diversity sampled. This diversity is measured by using quantitative or qualitative variables, the choice of which is discussed. We also examine the effects of PCS on the stratification of the sampling and on the size of the sample. Finally, we present some examples of presently developed core collections.

PRINCIPLES AND METHODS

The within-population diversity is determined by differences between individuals for one or several characters. These differences can be estimated by a distance that, for our purposes, must be a metric distance. The choice of this distance depends on the characters observed, quantitative or qualitative.

Quantitative Variables

THE CHOICE OF DISTANCE, COLINEARITY AND WEIGHTING

Quantitative characteristics are generally heterogeneous. They correspond to lengths (plant height, stem diameter), areas (leaf area, area of stigmata), weights (aerial biomass, reproductive biomass), or time (date of flowering, duration of fructification). They have, moreover, different forms of variability. In order to give the same weight to each character j, the Euclidean distance is weighted by the inverse of the standard deviation σ_j. The distance d_{ik} between two individuals i and k for J quantitative characters is defined by the following formula:

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

where x_{ij} is the value of the character j observed on the individual i and x_{kj} is the value of the character j for the individual k.

The distance between individuals is directly linked to the differences. If the differences come from characters that are strongly correlated, positively or negatively, the distance between certain individuals is greatly overestimated. Thus, if we measure the diameter of a tree trunk at different heights from the ground (1 m, 1.10 m, 1.20 m, etc.), the Euclidean distance between two trees will be greatly influenced by differences in diameter. This example is obvious, but such an effect, called the colinearity effect, is found for all the correlated characters.

To eliminate the colinearity effects, the principal components analysis has been applied to standardized variables to give J new centred variables, which are statistically independent: the factors. The distance between two individuals i and k for J factors is calculated by using a similar formula:

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

where the square root of the eigenvalue λ_j allows the weighting, and where z_{ij} and z_{kj} are respectively the coordinates of the individuals i and k on the factor j.

Such a procedure gives the same weight to all the factors in the estimation of distance, including the residual components resulting from noise or observation errors. Factors for which the eigenvalue is less than 1 (Kaiser criterion) are eliminated in order to prevent their intervention in the calculation of distance.

The choice of individuals maximizing diversity

The sum of generalized squares (SGS) of a lot of N individuals in the factorial space of K standardized variables (mean = 0, variance = 1) and independent variables (correlation coefficient = 0) is equal to the product NK (Lebart et al., 1977). The contribution P_i of the individual i to the SGS is equal to the sum of squares of these K new coordinates:

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

The relative contribution CR_i of the individual i to the SGS of the whole is given by:

$$CR_i = P_i(NK)$$

Conserving the greatest variability is equivalent to maximizing the score of the subgroup of individuals sampled by using an estimator SGS. The first step is to sample the individual furthest from the barycentre of the group, that is, the individual that makes the greatest relative contribution. The iterative selection of individuals maximizing the diversity of the core collection increases the core collection size. At each iteration, the cumulative SGS of the core collection, expressed as a percentage of the total SGS, is known. The procedure may be concluded either according to the size of the core collection or according to the percentage of diversity retained. The two criteria can be taken into account simultaneously. In that case, when the first criterion is met, the sampling comes to an end.

Qualitative Variables

The method described above is designed for quantitative data. The modifications required for qualitative data concern the first steps of the PCS. As with the quantitative data, there are relationships between the variables. For example, two molecular markers may be linked genetically. In order to eliminate the effects of this type of relationship on the distance and in order to give the same weight to independent variables, a method of multivariate analysis has been used to transform the initial data into factorial coordinates: this involves correspondence analysis (Benzecri, 1972).

The χ^2 distance is retained in place of the Euclidean distance and the analysis uses a binary table. In this table, the presence and absence of an allele are considered two different variables taking the values 1 and 0. With p molecular markers observed on N individuals, we obtain a table $2pN$. In consequence, all the individuals have the same marginal frequency equal to p. Moreover, the term $p\lambda_i$ ($p\lambda_i$ is the eigenvalue of the factor i) is equal to the sum of r^2 of this factor with p variables. This term is equivalent to the eigenvalue observed in the principal components analysis. The sum of $p\lambda_i$ is equal to the number of markers (for analysis of the principal components on quantitative data, the sum of eigenvalues is equal to the number of variables). As for quantitative data, the factorial coordinates are weighted. In our case, the weights are square roots of the corresponding $p\lambda_i$ values. The Kaiser criterion for the choice of number of factors is applied to the term $p\lambda_i$. The subsequent steps of the PCS are the same as for the quantitative characters.

DISCUSSION

Stratification: Conditions and Consequences

The simplest method for creating a core collection is random sampling over the entire base collection. When the genetic structure of the base collection is not known, such sampling is the best solution (Brown, 1989a). Nevertheless, it is less effective for alleles that are common locally but rare over the entire collection. This is why Brown (1989b) has suggested a stratification of the sampling.

Peeters and Martinelli (1989), Holbrook and Anderson (1995), and van Hintum (1995) are of a divided opinion and suggest that, as a basis of stratification, the country of origin of the plant material should be taken into account. Peeters et al. (1993) advise the use of precise ecogeographical data. These have been taken into consideration in establishing the soybean core collection (Perry et al., 1991). The major agronomic or biological characters—mode of reproduction, duration of cycle—are also used for the stratification (Spagnoletti-Zeulli and Qualset, 1987; Hamon and van Sloten, 1989; Diwan et al., 1994; Hamon et al., 1995). These characters are quantitative (height, diameter) as well as qualitative (colour, appearance), so they are difficult to treat simultaneously to obtain matrices of distances, except when the quantitative data are recorded so as to obtain classes of equal numbers or equal amplitudes. However, Cole-Rogers et al. (1997) propose an original method, the normed binary scale, which allows calculation of matrices of distances that integrate these two types of variables.

Molecular markers, used by Lux and Hammer (1994) and strongly recommended by Gepts (1995), have only begun to be taken into account in stratification. The percentage of accessions evaluated with the help of this type of marker is still low. Breeders are often not interested in it and the structuring of groups determined by the molecular markers does not always coincide with that which follows from morphoagronomic diversity.

However, random sampling, even within groups, does not enable us to reach the main aim of the core collection, which is to sample the maximum range of diversity. For example, the production of hydrocyanic acid in white clover (*Trifolium repens*) is controlled by two independent loci. This character confers a resistance to several species of insects and molluscs, and its expression is regulated by climatic parameters (temperature, day length, humidity). The base collection of the National Plant Germplasm System (NPGS) in the United States contains 602 accessions of white clover. A core collection of 91 accessions has moreover been established on the basis of a geographic stratification and according to a random selection within the groups. Pederson et al. (1996) determined the proportion of cyanogenic plants in the base collection and compared it to that of the core collection. No significant difference of frequency was found between the two collections,

which proves that the core collection did not 'maximize' the variability: it is simply a reduced version of the base collection.

Conversely, PCS modifies the mode of sampling so that it is no longer random, maximizes the diversity, and in most cases prevents doubles. It thus meets the objectives defined for a core collection.

By preferentially sampling the most distant individuals, the PCS method requires three conditions to be functional and effective. The first, and most important, is that all the individuals of the core collection can cross with each other and give intermediate types. This implies a prior knowledge of the genetic structure of the species complex (Pernes, 1984) and a good estimation of the level of reproductive barriers between compartments. The second condition concerns the efficiency of the PCS, which supposes a generalized additivity and high heritability (in the wide sense) for the quantitative characters. In cases where this hypothesis is not valid (dominance, superdominance, high plasticity, etc.), selection on phenotype diversity will not necessarily lead to selection on genetic diversity. The sampling can thus be considered random in relation to the hidden variability. The third condition is the absence of a polymodal structure of the base collection. Indeed, the existence of several groups may lead to a sampling of individuals alternatively in the most distant groups. In this case, maximizing the diversity increases redundancy. The stratification of the sampling is here a determining preliminary step, as in the case of random sampling.

Thus, for PCS, the stratification of sampling must depend on the genetic structure of populations and the limits of recombination. When such data are lost or absent, taxonomy must be taken into account. The bioclimatic and biogeographic information must then modify the structure by establishing subgroups corresponding to the genetic differentiation, in subspecies and in ecotypes. The PCS is thus applied within each group and subgroup.

Size of the Core Collection and its Strata

Discussion on the size of the core collection is always topical. Brown (1989a), using the theory of neutral alleles (Kimura and Crow, 1964) and that of sampling, demonstrated that a sample of 10% of the base collection contains at least 80% of the alleles, with a statistical risk of error of 5%. According to this author, the results are reliable in relation to the type of frequency distribution of alleles at each locus. This value of 10% is not modified by our mode of sampling.

The size of each subgroup within the core collection was studied by Brown (1989b). Three methods were compared to determine this size: choosing the same number of accessions per group, defining a number of accessions proportionate to the size of the group, or opting for a number proportionate to the logarithm of the group size. The author demonstrated that the third solution is a good compromise. Nevertheless, the choice of the size of the

subsample according to the size of the group supposes a relationship between the diversity of the group and its size, which is far from being always the case in the base collections. The ratio between the diversity and the size of the group depends, among other things, on the mode of reproduction and the economic importance of the plant (cultivated plant or related species). For example, in the base collections of the species complex of the genus *Coffea*, the cultivated and autogamous species *C. arabica* is overrepresented in relation to the wild and allogamous species *C. sessiliflora* from Africa and the East (Noirot et al., 1993).

The PCS method of sampling allows management of the diversity of the core collection. The sampling of individuals to increase the core collection can be halted according to the percentage of diversity already achieved. Taking into account the diversity already achieved is particularly useful in the case of species with high natural redundancy (agamic complexes, autogamous plants, plants that reproduce vegetatively).

The Primary Uses of Core Collections

Whatever the strategy used, the core collections are conceived to help managers to conserve and use genetic resources. The two examples recorded here show that users can find them helpful.

The US Department of Agriculture had in 1990 a base collection of perennial alfalfa of 2400 accessions. In order to extract a core collection of 200 accessions, Basigalup et al. (1995) decided on directed selection of genotypes after a geographic stratification, among the eight methods tested. Jung et al. (1997) subsequently used this core collection for research on proteic composition, biodegradation of leaves, digestibility, and lignin composition. Thus, this core collection was useful for characters other than those that determined its constitution.

Holbrook et al. (1993) established a core collection of 831 accessions of peanut from the base collection, which numbered 7432. This collection comprises 70% of samples evaluated on morphoagronomic characters and 30% of non-evaluated samples. For the evaluated accessions, multivariate analyses having revealed a structure in groups, 10% of accessions of each group are taken randomly. For the accessions that are not evaluated, 10% are taken randomly after stratification by country. Holbrook and Anderson (1995) tested the relevance of this core collection in relation to the base collection for resistance to cercosporiosis due to *Cercosporidium personatum*. This involved determining the number of resistant accessions that would allow us to detect the core collection in relation to the base collection. The process comprised two steps: In the first, the entire core collection was tested for this character, then the groups that seemed to indicate resistance in the core collection were examined in detail in the base collection. The rates of efficiency, in terms of proportion of resistant accessions identified, increased from 1/64 in the base

collection to 1/8 in the core collection. This result demonstrated the utility of a core collection not only in developing the genetic material, but also in improving the efficiency of the search for particular characteristics.

The PCS method of constituting the core collections enables us to achieve a greater efficiency. Of course, as with peanut, the implementation of this process depends on the availability of data required for the selections. Hamon et al. (1998) demonstrated on four plants—rice, coffee, sorghum and hevea—that the variability of quantitative characters in the core collections is only slightly or is not modified when the selection is qualitative. On the other hand, the means and variances of morphoagronomic characters are greatly modified by a quantitative selection. Qualitative selection seems the most effective to conserve the rare alleles and increase the overall diversity with limited numbers at the quantitative level. Quantitative selection leads to the loss of 6% of rare alleles (from initial frequency in the base collection of less than 5%) in hevea, 11% in sorghum, 12% in rice, and 33% in coffee. When qualitative selection is used, the losses are reduced to 2% for rice, sorghum, and hevea and 6% for coffee. In other words, this approach demonstrates that it is possible to maximize the allele richness—also known as the neutral variability—in the core collection while preserving the representativity of the morphoagronomic variability.

Core collections have now been set up for many plants. A consensus seems to have been reached on their size (around 10% of the base collection) and on the need for stratification. Random sampling leads to conservation of the variability contained in the base collection while retaining its faults (overrepresentation, redundancy, sampling bias). The main advantage of PCS is that it allows an increase in the neutral allelic diversity of the core collection without modifying the agronomic representativity or changing the relative intensity of the sampling (10% of the base collection). The present state of progress of molecular biology will certainly facilitate the use of molecular markers in the estimation and structuring of genetic diversity. The constitution of core collections must thus take into account the relations between different levels of variability.

REFERENCES

Basigalup, D.H., Barnes, D.K., and Stucker, R.E. 1995. Development of a core collection for perennial *Medicago* plant introductions. *Crop Science*, 35: 1163-1168.

Benzecri, J.P. 1972. Pratique de l'analyse des données: analyse des correspondances. Paris, Dunod, 424 p.

Brown, A.D.H. 1989a. Size and structure of collection: the case for core collection. In: *The Use of Plant Genetic Resources*. T. Hodgkin et al., eds., Chichester, Wiley, pp. 136-156.

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

Cole-Rogers, P., Smith, D.W., and Bosland, P.W. 1997. A novel statistical approach to analyze genetic evaluations using *Capsicum* as an example. *Crop Science*, 37: 1000-1002.

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

Frankel, O.H. and Bennett, E. 1970. Genetic resources. In: *Genetic Resources in Plants, their Exploration and Conservation*. O.H. Frankel and E. Bennett, eds., Oxford, Blackwell, 547 p.

Frankel, O.H. and Brown, A.H.D. 1984. Current plant genetic resources: a critical appraisal. In: *Genetics, New Frontiers* (vol. IV). New Delhi, Oxford and IBH.

Gepts, P. 1995. Genetic markers and core collections. In: *Core Collections of Plant Genetic Resources*. T. Hodgkin et al., eds., Chichester, Wiley, pp. 127-146.

Hamon, S., Dussert, J., Deu, M., Hamon, P., Seguin, M., Glaszmann, J.C., Grivet, L., Chantereau, J., Chevallier, M.H., Flori, A., Lashermes, P., Legnate, H., and Noirot, M. 1998. Effects of quantitative and qualitative principal component score strategies on the structure of coffee, rice, rubber tree and sorghum core collections. *Genetics, Selection, Evolution*, 30 (suppl. 1): 237-258.

Hamon, S., Noirot, M., and Anthony, F. 1995. Developing a coffee core collection using the principal components score strategy with quantitative data. In: *Core Collections of Plant Genetic Resources*. T. Hodgkin et al. eds., Chichester, Wiley, pp. 117-126.

Hamon, S. and van Sloten, D.H. 1989. Characterization and evaluation of okra. In: *The Use of Plant Genetic Resources*. A.D.H. Brown and O. Frankel, eds., Cambridge, Cambridge University Press, pp. 173-196.

Harlan, J.R. 1970. The evolution of cultivated plants. In: *Genetic Resources in Plants*, Their Exploration and Conservation. O.H. Frankel and E. Bennett, eds., Oxford, Blackwell, 547 p.

Holbrook, C.C. and Anderson, W.F. 1995. Evaluation of a core collection to identify resistance to late leafspot peanut. *Crop Science*, 35: 1700-1702.

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

Holden, J.H.W. 1984. The second ten years. In: *Crop Genetic Resources: Conservation and Evaluation*. J.H.W. Holden and J.T. Williams, eds. London, George Allen and Unwin, 296 p.

Jung, H.G., Sheaffer, C.C., Barnes, D.K., and Halgerson, J.L. 1997. Forage quality variation in the US alfalfa core collection. *Crop Science*, 37: 1361-1366.

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

Lebart, L., Morineau, A., and Tabard, N. 1977. *Techniques de la Description Statistique: Méthodes et Logiciels pour l'Analyse des Grands Tableaux*. Paris, Dunod.

Lux, H. and Hammer, K. 1994. Molecular markers and genetic diversity: some experience from the genebank. In: EUCARPIA meeting on evaluation and exploitation of genetic resources pre-breeding. Clermont-Ferrand, France, EUCARPIA, pp. 49-53.

Noirot, M., Hamon, S., and Anthony, F. 1993. L´obtention d´une core collection de caféiers: définition des groupes d´échantillonnage et méthodologie. In: XVI^{e} Colloque scientifique international sur le café. Paris, ASIC.

Pederson, G.A., Fairbrother, T.E., and Greene, S.L. 1996. Cyanogenesis and climatic relationships in the US white clover germplasm and core subset. *Crop Science*, 36: 427-433.

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

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

Pernes, J. 1984. Gestion des Ressources Génétiques des Plantes. Paris, ACCT, 212 p.

Perry, M.C., McIntosh, M.S., and Stoner, A.K. 1991. Geographical patterns of variation in the USDA soybean germplasm collection. 2. Allozyme frequencies. *Crop Science*, 31: 1356-1360.

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

van Hintum, T.J.L. 1995. Hierarchical approaches to the analysis of genetic diversity of crop plants. In: *Core Collections of Plant Genetic Resources*. T. Hodgkin et al. eds., Chichester, Wiley, pp. 23-34.

Vavilov, N.I. 1935. The origin, variation, immunity and breeding of cultivated plants. *Chronica Botanica*, 13 (6 volumes).

Yonezawa, K., Nomura, T., and Morishima, H.1995. Sampling strategies for use in stratified germplasm collections. In: *Core Collections of Plant Genetic Resources*. T. Hodgkin et al., eds., Chichester, Wiley, pp. 35-53.

Noirot Michel, Anthony François, Dussert Stéphane, Hamon Serge.

A method for building core collections.

In : Hamon P. (ed.), Seguin M. (ed.), Perrier X. (ed.), Glaszmann J.C. (ed.). Genetic diversity of cultivated tropical plants. Montpellier (FRA), Enfield : CIRAD, Science Publ., 2003, p. 67-75.

(Repères). ISSN 1251-7224