REVIEW & INTERPRETATION

Analysis of Genetic Diversity in Crop Plants—Salient Statistical Tools and Considerations

S. A. Mohammadi and B. M. Prasanna*

ABSTRACT

Knowledge about germplasm diversity and genetic relationships among breeding materials could be an invaluable aid in crop improvement strategies. A number of methods are currently available for analysis of genetic diversity in germplasm accessions, breeding lines, and populations. These methods have relied on pedigree data, morphological data, agronomic performance data, biochemical data, and more recently molecular (DNA-based) data. For reasonably accurate and unbiased estimates of genetic diversity, adequate attention has to be devoted to (i) sampling strategies; (ii) utilization of various data sets on the basis of the understanding of their strengths and constraints; (iii) choice of genetic distance measure(s), clustering procedures, and other multivariate methods in analyses of data; and (iv) objective determination of genetic relationships. Judicious combination and utilization of statistical tools and techniques, such as bootstrapping, is vital for addressing complex issues related to data analysis and interpretation of results from different types of data sets, particularly through clustering procedures. This review focuses on application of statistical tools and techniques in analysis of genetic diversity at the intraspecific level in crop plants.

Analysis of genetic relationships in crop species is an important component of crop improvement programs, as it serves to provide information about genetic diversity, and is a platform for stratified sampling of breeding populations. Accurate assessment of the levels and patterns of genetic diversity can be invaluable in crop breeding for diverse applications including (i) analysis of genetic variability in cultivars (Smith, 1984; Cox et al., 1986), (ii) identifying diverse parental combinations to create segregating progenies with maximum genetic variability for further selection (Barrett and Kidwell, 1998), and (iii) introgressing desirable genes from diverse germplasm into the available genetic base (Thompson et al., 1998). An understanding of genetic relationships among inbred lines or pure lines can be particularly useful in planning crosses, in assigning lines to specific heterotic groups, and for precise identification with respect to plant varietal protection (Hallauer and Miranda, 1988). Analysis of genetic diversity in germplasm collections can facilitate reliable classification of accessions, and identification of subsets of core accessions with possible utility for specific breeding purposes. Significant emphasis is being paid to comprehensive analysis of genetic diversity in numerous crops, including major field crops such as wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), maize (*Zea mays* L.), barley (*Hordeum vulgare* L.), and soybean [*Glycine max* (L.) Merr.].

Study of genetic diversity is the process by which variation among individuals or groups of individuals or populations is analyzed by a specific method or a combination of methods. The data often involve numerical measurements and in many cases, combinations of different types of variables. Diverse data sets have been used by researchers to analyze genetic diversity in crop plants; most important among such data sets are pedigree data (Bernardo, 1993; Messmer et al., 1993; van Hintum and Haalman, 1994), passport data–morphological data (Smith and Smith, 1992; Bar-Hen et al., 1995), biochemical data obtained by analysis of isozymes (Hamrick and Godt, 1997) and storage proteins (Smith et al., 1987), and, recently, DNA-based marker data that allow more reliable differentiation of genotypes. Since each of these data sets provide different types of information, the choice of analytical method(s) depends on the objective(s) of the experiment, the level of resolution required, the resources and technological infrastructure available, and the operational and time constraints, if any (see Karp et al., 1997, for detailed review).

Sampling Strategies

Genetic diversity in crop plants may be analyzed at different levels: individual genotypes such as inbred lines or pure lines or clones, populations, germplasm accessions, and species. Sampling strategies in each of the above cases would vary, primarily because of the differences in the nature of genetic materials. In contrast to inbred lines or pure lines, sampling strategies for genetic diversity analysis at population level are complicated because of various factors including linkage, inbreeding, migration, and subpopulation differentiation. Genotypes in a population may not be distributed in Hardy-Weinberg frequencies. With most measures of genetic diversity, the form of their underlying sampling distributions is largely unknown. However, on the basis of statistical genetics theories, analytical formulae have been developed for estimating the sampling variance of some genetic diversity measures (Brown and Weir, 1983; Weir, 1990). In general, sampling variances of diversity

S.A. Mohammadi, Department of Agronomy & Plant Breeding, Faculty of Agriculture, Tabriz University, Tabriz 51664, Iran. B.M. Prasanna, Division of Genetics, Indian Agricultural Research Institute, New Delhi–110012, India; Received 18 April 2001. *Corresponding author (prasanna@ndf.vsnl.net.in).

Published in Crop Sci. 43:1235–1248 (2003).

Abbreviations: PIC, polymorphism information content; GD, genetic distance; GS, genetic similarity; SE, standard error; PCA, principal component analysis; PC, principal component; PCoA, principal coordinate analysis; MDS, multidimensional scaling.

measures depend particularly on the number of individuals sampled per population, the number of loci sampled, genotypic and allelic compositions of population, mating system, and effective population size (Nei and Chesser, 1983; Namkoong, 1988; Weir, 1990). A large portion of the sampling variance of diversity estimates is due to the variation of diversity levels among loci across the genome (Nei, 1987; Weir, 1990). The sampling error associated with the sampling of loci shall be considerably reduced if the same set of loci is monitored in each population of a species.

Frankel et al. (1995) suggested that two distinct concepts of genetic variation are applicable at the population level: (i) "richness" of any population or sample from it, corresponding to the total number of genotypes or alleles present in the population, and (ii) "evenness" or the frequency of different types or alleles in the population or samples analyzed. "Allele richness" is estimated by taking into account the mean number of alleles per locus and percent polymorphic loci. This estimate is sensitive to the presence or absence of distinct or rare alleles (5% or lower in frequency) in a population (or sample), as a high degree of sampling error could be associated with detection of such alleles (Nei, 1987; Namkoong, 1988; Sjogren and Wyone, 1994). Therefore, in addition to the total number, it would be useful to monitor the number of alleles in the sample above a frequency threshold (say 5%). The percentage of polymorphic loci in a population is a crude estimation of genetic variation, as it is subject to a large genomic sampling error; this estimate is reliable only when a large number of loci are sampled (Brown and Weir, 1983). The evenness of allele or genotype frequencies is accounted for by the measures of average observed heterozygosity, expected heterozygosity, and effective number of alleles. None of these measures are sensitive to the sampling error associated with rare alleles. Sampling strategy, sample size, and distribution of a sample over population subdivision (occurrence of subpopulations within a population with differences in allelic frequencies) affects the probability of sampling rare alleles. Clustered sampling with sufficient samples per subpopulation or groups can alleviate the complexity associated with sampling of rare alleles.

In terms of sampling for analysis of genetic diversity, the law of diminishing marginal returns holds true. While the cost of sampling new individuals, particularly by means of molecular markers, is directly proportional to the size of the sample, the probability of detecting an additional allele with each added individual sample decreases rapidly with increasing sample size (Marshall and Brown, 1975; Brown, 1989; Frankel et al., 1995). In studies aimed at analysis of population structure, it is necessary to balance the need to collect as large a sample size as possible, against the need to screen as many populations as possible, and the need to get allele frequencies from as many loci as possible. There is no simple recommendation for the ideal sample size, number of samples, or number of loci. However, whether or not the aim is to describe genetic variation in taxonomic units (populations or species), it will be necessary to estimate the variation within the taxonomic unit, to be able to determine the degree of differentiation among taxonomic units.

Marshall and Brown (1975) recommended that a sample size of 59 or more unrelated gametes (assured by a random sample of 50 diploid individuals) is sufficiently large to have a 95% probability of detecting all alleles of 5% or greater in frequency. Crossa et al. (1993) showed that the sample size (n) required to retain, with probability (P), at least one copy of each of k allelic classes in each of m loci could be calculated as follows:

$$n > \frac{\log[1 - (P)^{1/m} - \log(k - 1)]}{\log(1 - P)}$$

With 48 individuals, for $m = 5$ loci with $k = 5$ alleles per locus, there is a 95% probability of detecting all alleles with $P = 0.05$ or greater (Warburton et al., 2002). Baverstock and Moritz (1996) presented a table of sample size (diploid individuals) needed to detect given differences in allele frequency in populations for a given statistical power. Sampling considerations in relation to analysis of intraspecific differentiation were discussed in detail by Baverstock and Moritz (1996) and Weir (1996).

Estimation of Genetic Distance

Genetic distance is "that difference between two entities that can be described by allelic variation." (Nei, 1973). This definition was later elaborated by Nei (1987) as "the extent of gene differences… between populations or species that is measured by some numerical quantity." A more comprehensive definition of genetic distance is "any quantitative measure of genetic difference, be it at the sequence level or the allele frequency level, that is calculated between individuals, populations or species" (Beaumont et al., 1998).

Measures of Genetic Distance–Similarity

Genetic distance–similarity between two genotypes, populations, or individuals may be calculated by various statistical measures depending on the data set. Discussions on various distance measures are available in the literature (Felsenstein, 1984; Nei, 1987; Weir, 1990, 1996; Beaumont et al., 1998).

Euclidean or straight-line measure of distance is the most commonly used statistic for estimating genetic distance (GD) between individuals (genotypes or populations) by morphological data. Euclidean distance between two individuals i and j, having observations on morphological characters (p) denoted by $x_1, x_2, \ldots, x_p$ and $y_1, y_2, \ldots, y_p$ for i and j, respectively, can be calculated by the following formula:

$$d_{(i,j)} = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_p - y_p)^2]^{1/2}$$

On the basis of data obtained by measurement of quantitative traits in inbred lines, Smith et al. (1991) applied another measure of genetic distance as follows:

$$d_{(i,j)} = \sum[(T_{1(i)} - T_{2(i)})^2/varT_{(i)}]^{1/2}$$

where T_1 and T_2 are the values of the ith trait for inbred lines 1 and 2, respectively, and the $varT_{(i)}$ is the variance for the ith trait over all inbreds.

Gower (1971) described a general coefficient for measuring genetic distance between individuals on the basis of various types of characters, such as dichotomous, qualitative, and quantitative. For qualitative characters, the distance between two individuals is scored as 0 (wherever there is a match) and 1 (wherever there is a mismatch). For quantitative characters, the distance between two individuals is calculated as the difference in the trait values divided by the overall range for the trait. This method converts the distance for quantitative characters to a specific value on a scale of 0 to 1; this, in turn, allows simultaneous use of both quantitative and qualitative data in generating a distance matrix. For this purpose, the individual character distances for each pair of individuals are summed and then divided by the number of characters scored in both individuals. Gower's measure of distance between individuals (i and j) is defined as $DG_{ij} = 1/p\ \Sigma w_k\ d_{ijk}$ where p is number of characters, d_{ijk} is the contribution of the kth character to the total distance between two individuals; $d_{ijk} = |x_{ik}–x_{jk}|$, where x_{ik}, x_{jk} are the values of the kth character on the individuals i and j, respectively, and $w_k = 1/R_k$, where R_k is the range of the kth character in the sample (Franco et al., 1997).

Various genetic distance measures have been proposed for analysis of molecular marker data for the purpose of genetic diversity analysis. For molecular marker data where the amplification products may be equated to alleles, as in case of simple sequence repeats (SSRs) and restriction fragment length polymorphisms (RFLPs), allele frequencies can be calculated. The genetic distance between individual i and j can be estimated using the formula,

$$d(i,j) = Constant\ (\sum_{a=1}^{n} |X_{ai} - X_{aj}|^r)^{1/r}$$

where X_{ai} is the frequency of the allele a for individual i, n is number of alleles per locus, and r is constant based on the coefficient used. In its simple form (that is, when $r = 1$), genetic distance can be calculated as

$$d_{1ij} = \tfrac{1}{2} \sum_{a=1}^{n} |X_{ai} - X_{aj}|$$

When $r = 2$, d_{ij} is referred to as Rogers' (1972) measure of distance (RD), where

$$RD_{ij} = \tfrac{1}{2}[\sum (X_{ai} - X_{aj})^2]^{1/2}$$

Although allele frequencies can be calculated for some of the molecular markers, the data is most widely employed to generate a binary matrix for statistical analysis. The commonly used measures of genetic distance or genetic similarity (GS) using such binary data are (i) Nei and Li's (1979) coefficient (GD_{NL}), (ii) Jaccard's (1908) coefficient (GD_J), (iii) simple matching coefficient (GD_{SM}) (Sokal and Michener, 1958), and (iv) Modified Rogers' distance (GD_{MR}). Genetic distances determined by these measures can be estimated as follows:

$$GD_{NL} = 1 - [2N_{11}/(2N_{11} + N_{10} + N_{01})]$$

$$GD_J = 1 - [N_{11}/(N_{11} + N_{10} + N_{01})]$$

$$GD_{SM} = 1 - [(N_{11} + N_{00})/(N_{11} + N_{10} + N_{01} + N_{00})]$$

$$GD_{MR} = [(N_{10} + N_{01})/2N]^{0.5}$$

where N_{11} is the number of bands–alleles present in both individuals; N_{00} is number of bands–alleles absent in both individuals; N_{10} is the number of bands–alleles present only in the individual i; N_{01} is the number of bands–alleles present only in the individual j; and N represents the total number of bands–alleles.

GD_J takes into consideration only matches between bands–alleles that are present and ignores pairs in which a band–allele is absent in both individuals. In contrast, GD_{NL} measures the proportion of bands–alleles shared as the result of being inherited from a common ancestor, and represents the proportion of bands–alleles present and shared in both individuals divided by the average of proportion of bands–alleles present in each individual. GD_{SM}, a Euclidean measure of distance, takes into account mismatches and matches, and gives equal weight to both in estimating genetic distance. GD_{MR}, another Euclidean distance measure, considers each locus scored as an orthogonal dimension (Link et al., 1995; Johns et al., 1997).

One specific problem often encountered during analysis of genetic diversity in crop plants by molecular markers, particularly with the microsatellite or SSR markers, is the failure of some genotypes to show amplification for some SSR primer pairs. It is often difficult to ascertain whether such lack of amplification is due to "null alleles" (Robinson and Harris, 1999). Unless the researcher is confident about the null status of a genotype for a specific SSR locus, such data might be considered missing data during computation of genetic similarity–distance matrix (Warburton and Crossa, 2000), to minimize the possibility of erroneous interpretation.

Choice of a Distance Measure

Appropriate choice of a genetic distance measure, on the basis of the type of the variable and the scale of measurement, is an important component in analysis of genetic diversity among a set of genotypes. GD_{NL} and GD_J differ in the weighting of dominant and codominant polymorphic markers. While both measures lead to identical rankings of GD among pairs of inbred lines, the GD estimates may differ when one analyzes heterozygous loci in hybrids (Link et al., 1995) or in case of populations where heterozygous genotypes are expected to occur commonly. For codominant markers (such as RFLPs and SSRs), the expected GD_{NL} of related pairs of lines is a linear function of their coancestry coefficient (Melchinger, 1993). For dominant markers, this property applies to GD_J but not to GD_{NL} (Link et al., 1995). In case of codominant markers, both GD_J and GD_{NL} may lead to identical ranking of GD estimates among inbred lines. In general, GD_J and GD_{NL} suffer from unknown statistical distributions resulting from the denominator, which is a random variable. The distribution of any statistic is indispensable for calculating sampling variance and confidence interval. To overcome this problem, the "bootstrap" technique (discussed later) can be effectively used to empirically estimate sampling variance (Brown, 1994).

Euclidean distances, such as GD_{SM} or GD_{MR}, can be

modeled as binomial variables with known statistical properties when the genome is randomly sampled (Tivang et al., 1994; Lombard et al., 2000). Among the various genetic distance measures, GD_{MR} is widely preferred because of its excellent genetical and statistical properties. GD_{SM} has Euclidean metric properties that allows its use in hierarchial clustering strategies (described later) such as the minimum variance method within a group, proposed by Ward (1963), and the analysis of molecular variance, AMOVA (Excoffier et al., 1992), which can be used for the estimation of the variance components among and within groups. However, many researchers do not prefer using GD_{SM} as it gives equal weight to both 0-0 and 1-1 matches in case of binary data. The 1-1 matches in reality indicate more similarity than the 0-0 matches because there are many reasons for lack of amplification or absence of bands, and a 0-0 match may not reflect identity by descent, but rather identity in state.

In case the researcher is interested to make use of more than one measure of genetic distance to analyze a given data set or different data sets, it is important to ascertain the correspondence between matrices derived from different distance measures. The test of matrix correspondence, popularly known as Mantel test (Mantel, 1967), analyzes matrix correspondence on the basis of the assumption of asymptotic normality for a particular test criterion. Mantel test is a regression in which the variables are themselves distances or dissimilarity matrices summarizing pair-wise similarities–dissimilarities between units of study. It is based on a simple cross-product term, $Z = \Sigma X_{ij} Y_{ij}$, and is normalized by means of the following formula:

$$r = 1/(n - 1)\sum[(X_{ij} - \overline{X})/S_X][(Y_{ij} - \overline{Y})/S_Y]$$

where X_{ij} and Y_{ij} are the off-diagonal elements of matrices X and Y, n is the number of elements in the distance matrices, and the S_X and S_Y are standard deviations for variables X and Y, respectively. This standardized equation allows one to consider variables of different measurement units within the same framework, rescaling the statistic to the range of a conventional correlation coefficient bounded on -1 to 1. Because the elements of a distance matrix are not independent, Mantel's test of significance is evaluated via permutation procedures to overcome the problem of dependent elements (Manly, 1991). Note that the Mantel test is based on linear correlation, and hence, is subject to the same set of assumptions that beset a common Pearson correlation. However, the test of spatial dependence is averaged over all distances in the simple Mantel test, and so this test cannot discover changes in the pattern of correlation at different distances (scales).

Because the Mantel test proceeds from a dissimilarity–similarity matrix, it can be applied to different types of variables (categorical, rank, or interval-scale data). This is especially important in analysis of genetic diversity, where various data sets may be used to assess the relationships among different individuals or populations. Mantel test has been used in analysis of genetic diversity in crop plants, particularly in ascertaining correspondence of matrices derived by means of different marker systems over the same set of genotypes (van Bueningen and Busch, 1997; Bohn et al., 1999; Lombard et al., 2000; Lübberstedt et al., 2000; Virk et al., 2000; Vuylsteke et al., 2000).

Genetic Differentiation of Populations

Several approaches have been proposed to estimate the amount of genetic differentiation between populations and in subdivisions of a population. χ^2 Tests using frequency-based statistics have greater power for detecting differences between populations or population subdivision when mutation rate (and thus allelic diversity) is low. It is also possible to quantify the extent of between–within population differentiation by the F statistics of Wright (1951) or the analogous measures of Cockerham (1969, 1973). Wright's approach consists of three different F coefficients that estimate (i) correlation of genes within individuals over all populations (F_{IT}), (ii) correlation of genes of different individuals in the same population (F_{ST}), and (iii) correlation of genes within individuals within populations (F_{IS}). F_{ST}, F_{IT}, and F_{IS} are interrelated so that

$$1 - F_{IT} = (1 - F_{ST})(1 - F_{IS})$$

$$F_{ST} = (F_{IT} - F_{IS})/(1 - F_{IS})$$

F_{ST} equals 0 when the subpopulations are identical in allele frequencies, and 1 when they are fixed for different alleles. F_{ST} is a measure of genetic differentiation over subpopulations and is always positive. F_{IS} and F_{IT} are measures of deviation from Hardy-Weinberg proportions within subpopulations and in the total population, respectively, where positive values indicate a deficiency of heterozygotes, and negative values indicate an excess of heterozygotes. Estimation of the F statistics, and the inferences from such estimates, were discussed in detail by Weir (1996).

Parameters analogous to F_{ST} have been defined by several authors on the basis of alternative assumptions about the evolutionary model and consequent modifications to the algorithm. Nei (1973) suggested another statistic, G_{ST}, which utilizes information from several loci simultaneously. G_{ST} is calculated from allele frequencies rather than genotype frequencies (assuming Hardy-Weinberg equilibria in all subpopulations). G_{ST} measures the proportional amount of variation within subpopulations as compared with the total population and does not specify the identity of alleles involved. When subpopulations appear similar, G_{ST} is biased and results in an overestimate of the degree of substructuring. Because of the dependence of G_{ST} on the level of diversity, Nei (1973) proposed an absolute measure of gene differentiation called the "minimum genetic distance" (D) which is independent of gene diversity within populations. The F_{ST} value is not expected to be affected by the type of genetic marker used (i.e., markers with different evolutionary rates, e.g., isozymes and microsatellites). In contrast, measures of absolute genetic differences, like Nei's genetic distance D, are expected to give different results depending on the evolutionary rate (muta-

tion rate) of the actual marker. For example, mini- and microsatellites give larger *D*-values than isozymes.

Some subpopulations may have further levels of obvious structure, enabling grouping on the basis of regions or colonies. Assuming that there is a local regional level into which subpopulations can be placed, additional measures such as F_{SR} and F_{RT} that partition variation into the diversity among subpopulations within a region and that among regions for a total population, respectively, may be adopted (Weir and Cockerham, 1984; Nei, 1987). Data related to molecular differences between alleles can be highly useful in determining hierarchical array of groups in a population by sequence-based statistics (e.g., N_{ST}; Lynch and Crease, 1990). Nucleotide diversity data can be inferred from differences in allele size (microsatellites) or single nucleotide polymorphisms (SNPs).

Multivariate Methods

With increases in the sample sizes of breeding materials and germplasm accessions used in crop improvement programs, methods to classify and order genetic variability are assuming considerable significance. The use of established multivariate statistical algorithms is an important strategy for classifying germplasm, ordering variability for a large number of accessions, or analyzing genetic relationships among breeding materials. Multivariate analytical techniques, which simultaneously analyze multiple measurements on each individual under investigation, are widely used in analysis of genetic diversity irrespective of the dataset (morphological, biochemical, or molecular marker data). Among these algorithms, cluster analysis, principal component analysis (PCA), principal coordinate analysis (PCoA), and multidimensional scaling (MDS) are, at present, most commonly employed and appear particularly useful (Melchinger, 1993; Johns et al., 1997; Thompson et al., 1998; Brown-Guedira et al., 2000). We shall focus here only on the salient features of statistical methodologies and some important considerations, specifically in relation to genetic diversity in crop plants at the intraspecific level.

Cluster Analysis

"Cluster analysis" refers to "a group of multivariate techniques whose primary purpose is to group individuals or objects based on the characteristics they possess, so that individuals with similar descriptions are mathematically gathered into the same cluster" (Hair et al., 1995). The resulting clusters of individuals should then exhibit high internal (within cluster) homogeneity and high external (between cluster) heterogeneity. Thus, if the classification is successful, individuals within a cluster shall be closer when plotted geometrically and different clusters shall be farther apart (Hair et al., 1995).

There are broadly two types of clustering methods: (i) distance-based methods, in which a pair-wise distance matrix is used as an input for analysis by a specific clustering algorithm (Johnson and Wichern, 1992), leading to a graphical representation (such a tree or dendrogram) in which clusters may be visually identified; and (ii) model-based methods, in which observations from each cluster are assumed to be random draws from some parametric model, and inferences about parameters corresponding to each cluster and cluster membership of each individual are performed jointly using standard statistical methods such as maximum-likelihood or Bayesian methods. Pritchard et al. (2000) discussed some of the constraints of distance-based methods, and described an innovative model-based clustering method based on Bayesian statistics for inferring population structure using multilocus genotypic data consisting of unlinked markers. The strength of this structured association approach lies in effective analysis of population structure, accurate clustering and assignment of individuals into their appropriate populations, even using a modest number of unlinked markers, and identification of migrants and admixed individuals. Using this approach, one can estimate the proportion of an individual's genome contributed by a specific subpopulation, referred to as "genetic background matrix" (*Q*). By suitably modifying the test statistic to deal with quantitative traits, Thornsberry et al. (2001) provided the first empirical demonstration of the utility of the "structured association" method in plant genetics, identifying a gene associated with variation for flowering time in maize. At present, distance-based methods are most frequently applied.

Distance-based clustering methods can be categorized into two groups: hierarchical and nonhierarchical. Hierarchical clustering methods are more commonly employed in analysis of genetic diversity in crop species. These methods proceed either by a series of successive mergers or by a series of successive divisions of group of individuals. The former, known as "agglomerative hierarchical" methods, start with a single individual. Thus, there are initially as many clusters as individuals. The most similar individuals are first grouped and these initial groups are merged according to their similarities. Among various agglomerative hierarchical methods, the UPGMA (Unweighted Paired Group Method using Arithmetic averages) (Sneath and Sokal, 1973; Panchen, 1992) is the most commonly adopted clustering algorithm, followed by the Ward's minimum variance method (Ward, 1963).

The nonhierarchical clustering procedures do not involve construction of dendrograms or trees. These procedures, also frequently referred to as "K-means clustering," are based on "sequential threshold," "parallel threshold," or "optimizing" approaches for assigning individuals to specific clusters, once the number of clusters to be formed is specified (Everitt, 1980). Options for performing nonhierarchical clustering are available in statistical packages such as SAS [FASTCLUS] and SPSS [QUICK CLUSTER]. Nonhierarchical clustering methods are rarely used for analysis of intraspecific genetic diversity in crop plants. The primary reason could be the lack of prior information about the optimal number of clusters that are required for accurate assignment of individuals. Thompson et al. (1998) and Thompson and Nelson (1998) also reported that the

FASTCLUS procedure did not separate three of the predominant and well-defined ancestral groups in soybean germplasm, unlike hierarchical clustering.

Choice of a Clustering Method

UPGMA dendrograms have tended to predominate in past literature. Although some studies indicated the relative advantages of UPGMA clustering algorithm in terms of consistency in grouping biological materials with relationships computed from different types of data (Ajmone-Marsan et al., 1992; Mohna et al., 1992; Mumm et al., 1994), a single clustering method might not be always optimal or effective in revealing genetic associations. Despite some favorable attributes in UPGMA, the underlying assumptions are rarely met. Also, very few studies have analyzed the congruence between results obtained by application of different clustering procedures and the relative strengths and constraints of each. We shall cite some of the studies performed in recent years, particularly those that have attempted to utilize different datasets and multivariate methods in analysis of genetic diversity in crop plants.

Five clustering methods, namely UPGMA, UPGMC (Unweighted Paired Group Method using Centroids), Single Linkage, Complete Linkage, and Median, were compared for their utility in revealing genotype associations in barley germplasm collections (Peeters and Martinelli, 1989). UPGMA and UPGMC were found to be almost comparable with a relatively high level of accuracy, in accordance with pedigrees, compared to other methods. Single Linkage and Median clustering methods led to "chaining effect," which gave poor resolution of individual groups and complicated the interpretation of results. UPGMA, Single Linkage, Complete Linkage, UPGMC, Ward's method, and Principal Component Analysis (PCA), were compared in assessing genetic diversity in dent and popcorn maize inbred lines based on intersimple sequence repeat polymorphism (Kantety et al., 1995). UPGMA provided results most consistent with known heterotic groups and pedigree information, while PCA clearly separated the dent corn lines from the popcorn germplasm.

One way of comparing the efficiency of different clustering algorithms is through estimation of the "cophenetic correlation coefficient," which is a product-moment correlation coefficient measuring agreement between the dissimilarity–similarity indicated by a phenogram–dendrogram as output of analysis and the distance–similarity matrix as input of cluster analysis. A method yielding a high cophenetic correlation coefficient can be considered as an appropriate method for a particular analysis (Romesburg, 1984). The degree of fit can be interpreted subjectively as: $0.9 \leq r$, very good fit; $0.8 \leq r < 0.9$, good fit; $0.7 \leq r < 0.8$, poor fit; $r < 0.7$, very poor fit (Rohlf, 1992). However, a low cophenetic correlation coefficient does not mean that the dendrogram has no utility, but only indicates that some distortion might have occurred. There is no statistical test for the correlation coefficient because of the lack of independence of the individual coefficient in the dissimilarity matrices (Rincon et al., 1996). With distance matrices as input of clustering, the magnitude of cophenetic correlation coefficient decreases if the number of individuals increases to about 50, but no changes may result over 50 (Rohlf and Fisher, 1968). For a large sample of individuals, the cophenetic correlation coefficients have similar values and are not affected by the number of characters.

Several clustering methods were compared in grouping maize accessions on the basis of agronomic and morphological characters; UPGMA method was generally consistent with regard to the allocation of clusters, when different types and number of characters were used (Rincon et al., 1996). UPGMA also revealed higher cophenetic correlation coefficient in comparison to UPGMC, Single Linkage, and Ward's method. Genetic relationships in rapeseed (*Brassica* spp.) cultivars were analyzed on the basis of amplified fragment length polymorphisms (AFLP) by means of UPGMA and Ward's method in combination with Jaccard, Simple Matching, and Modified Simple Matching coefficients (Lombard et al., 2000). Despite very high correlations between distance matrices obtained through use of different coefficients, and derivation of the same patterns with both clustering methods, Ward's method was found more suitable as it avoided the chaining effects that are often observed with UPGMA. Similar observations were made in analysis of genetic diversity among maize inbred lines based on RFLP data (Dubreuil et al., 1996).

Apart from cophenetic correlation, another alternative and simple way of comparison is possible when there is a prior idea about the structure of groups according to geographical or germplasm origin of individuals. Here, the best method is that which recovers much of the expected structure. By simulating different hierarchical cluster methods and measures of distance on data with various levels of noise, Milligan and Cooper (1985) found that the single linkage cluster method was found to be the worst cluster strategy to recover the true structure, while Ward's and UPGMA were the best for similar and different group sizes, respectively (Milligan and Cooper, 1985). Mahalanobis distance (D^2) between centroids (vectors of means) of the groups can be used to identify the best clustering algorithm (Franco et al., 1997). The best clustering method produces the largest distance, D^2, among groups or clusters; this method may be particularly appropriate for quantitative data.

Cluster analysis based on algorithms such as UPGMA, UPGMC, Ward's, Single Linkage, and Complete Linkage has drawbacks. For instance, these algorithms do not provide an objective definition of what constitutes an optimal tree or dendrogram, and systemic errors are likely to be introduced during cluster analysis reconstructions. Such constraints may possibly be overcome by employing alternative methods, such as neighbor joining or Fitch–Margoliash, that remove the assumption that the data are ultrametric (Swofford et al., 1996). Methods such as neighbor joining have been more commonly used for phylogenetic studies; but very few researchers (Liu et al., 2000) have applied this method for intraspecific differentiation in crop plants. To our

knowledge, no in-depth analysis has been made to objectively ascertain the efficiency of neighbor-joining method or other related algorithms over commonly used clustering algorithms such as UPGMA and Ward's. Whatever algorithm is used for generating the dendrogram, it is useful to carry out bootstrapping of the allele frequencies (followed by calculation of genetic distances, etc.) to assess the reliability of the nodes.

Multivariate Analysis of Genetic Diversity– Some Important Considerations

Researchers should carefully consider the following points (Franco et al., 1997) while applying diversity measures and multivariate methods for analysis of genetic diversity: (i) judicious and effective use of different types of variables like continuous, discrete, ordinal, multistate, binomial etc.; (ii) application of multiple data sets (morphological, biochemical and molecular marker data); (iii) proper choice of a genetic distance measure and clustering algorithm(s) (discussed earlier); and (iv) determination of optimal number of clusters. Strategies required to address the above issues vary depending on the genetic materials being analyzed and the objectives of the experiment. Nevertheless, some of the ways for deriving objective solutions to these issues are presented below.

Using Diverse Data Sets

Multivariate methods such as cluster analysis can be performed on morphological (qualitative and quantitative), biochemical, and molecular marker data or combinations of such data. Different types of morphological variables, their associations, and implications for cluster analysis, were analyzed by Anderberg (1973) who discussed the properties of mean, range, and standard deviation as alternatives for removing measurement scale of different types of variables, and equalizing their effects in the final output of clustering. When characters with different scales such as field evaluation data and per se performance are used as inputs for cluster analysis, scale differences can be eliminated by standardizing each variable by means of either its standard deviation or its range to give equal weightage and contribution of all the characters in the final output. However, standardization of variables by range is a better option than standard deviation (Milligan and Cooper, 1985). When binary data such as morphological (qualitative) data, and molecular marker data (scored as 1 or 0) are used, standardization is not warranted since the distribution is binomial and not normal.

Principal components can be used as input for clustering, rather than directly applying data from quantitative characters, particularly when the correlations among the characters are significant (Goodman, 1972; Everitt, 1980). Principal component analysis provides variable independence and balanced weighting of traits, which leads to an effective contribution of different characters on the basis of respective variation. On the basis of quantitative morphological traits, van Bueningen and Busch (1997) applied such a procedure for analysis of genetic diversity among North American spring wheat cultivars.

Individual versus Combined Analyses of Data Sets

Two questions assume considerable significance: (i) whether analysis and interpretation should be based on individual or combined data sets when multiple data sets are available; and (ii) how to combine different data sets effectively. Hillis et al. (1996) provided an excellent discussion on these two questions.

The most important point to consider before combining different data sets is the congruence or correspondence among the results derived from individual data sets. Several studies in recent years have analyzed correlations among genetic distance–similarity matrices derived from application of different DNA-based marker systems, such as RFLP, random amplified polymorphic DNA (RAPD), SSR, and AFLP, in diverse crop species (Powell et al., 1996). However, very few studies have attempted to compare results derived from individual versus combined data sets (even for molecular marker data) in relation to the study of genetic diversity (Russell et al., 1997; Franco et al., 1997; Ajmone-Marsan et al., 1998). Also, comprehensive analyses of data sets of different nature (combination of qualitative and quantitative morphological data, or biochemical and molecular marker data, or morphological data with either biochemical or molecular marker data) to ascertain first, whether the total evidence is within the confidence limits of evidence from individual data sets, and second, whether such combinations provide a better estimate of genetic diversity, are highly scarce. In limited studies, biochemical data and morphological data were combined for deriving common distance measure (for example, Wrigley et al., 1982). Seberg et al. (1996) studied phylogenetic relationships among a small number of Triticeae species using individual and combined analysis of five data sets (one morphological and four molecular).

There are divergent opinions about the utility of combining data sets from different types of variables for the purpose of analyzing genetic diversity. Caution is required in combining data from qualitative and quantitative measures because of possible biases in distances estimated on the basis of quantitative characters and the high correlation of qualitative characters; instead, combining parentage and genetic marker information can lead to a better estimate of the genetic relationships (Souza and Sorrells, 1991). Assigning differential weight to the characters is often advocated to take effectively into account the possible incongruence among characters in terms of their genetic nature and contributions to genetic diversity in individuals or populations (Hillis, 1987; Chippindale and Weins, 1994). Such a procedure is difficult to adopt since there are no fool-proof criteria for determining appropriate weight to each character under analysis.

The Modified Location Model (MLM) combines all the categorical variables into one multinomial variable, W, which can be then used with the available continuous

variables (Franco et al., 1998). Initial grouping can be performed by Ward's minimum variance method, and then improved by the MLM. This strategy was successfully employed to classify maize accessions from most of the Latin American and U.S. gene banks (Taba et al., 1999). When simultaneously using genetic markers and phenotypic attributes to classify genotypes, this method could be further extended to obtain a relevant minimum subset of marker fragments that can be used in conjunction with morphoagronomic data to classify genotypes better rather than can be done with classifications based on individual data sets (Franco et al., 2001). It is not only reasonable to analyze data sets separately on the basis of different modes of inheritance, but also to "… analyze your data in as many ways as possible and sensible, then draw your conclusions." (Pedersen and Seberg, 1998).

Determining Optimal Number of Clusters

Another important aspect in cluster analysis is determining the optimal number of clusters or number of acceptable clusters. In essence, this involves deciding where to "cut" a dendrogram to find the true or natural groups. An "acceptable cluster" is defined as "a group of two or more genotypes with a within-cluster genetic distance less than the overall mean genetic distance and between cluster distances greater than their within cluster distance of the two clusters involved" (Brown-Guedira et al., 2000).

Some relatively simple ways of finding optimal number of clusters are the D^2 and the "upper tail approach" (Wishart, 1987). On the basis of D^2, the best point for cutting a dendrogram is the one that shows the largest D^2 between centroids of the groups created at that point (Franco et al., 1997). The upper tail approach is a simple procedure in which the mean and the standard deviation of distance values at the fusion points are used to calculate the optimal number of clusters.

Use of statistical techniques such as bootstrap, MANOVA (Multivariate Analysis of Variance), or discriminant analysis can facilitate determination of optimal number of clusters. In MANOVA, clusters or groups obtained in each cutting point are considered as treatments and individuals falling within that group are considered as replications for that treatment. The analysis is performed individually for each cut point with all characters or variables selected for cluster analysis. The optimal number of clusters or groups will be at that specific point which reveals the highest F value. This is based on the principle that at a proper cut point, within-group variance (error variance) shall be less than between-group variance (between-treatment variance), leading to a higher F value. Similarly, discriminant analysis can be effectively utilized to determine the best possible grouping on the basis of discrimination among groups achieved by different cut points.

Principal Component Analysis (PCA) and Principal Coordinate Analysis (PCoA)

PCA and PCoA can be utilized to derive a 2- or 3-dimensional scatter plot of individuals, such that the geometrical distances among individuals in the plot reflect the genetic distances among them with minimal distortion. Aggregations of individuals in such a plot will reveal sets of genetically similar individuals (Melchinger, 1993; Karp et al., 1997; Warburton and Crossa, 2000).

PCA is defined as "a method of data reduction to clarify the relationships between two or more characters and to divide the total variance of the original characters into a limited number of uncorrelated new variables" (Wiley, 1981). This will allow visualization of the differences among the individuals and identify possible groups. The reduction is achieved by linear transformation of the original variables into a new set of uncorrelated variables known as principal components (PCs). The first step in PCA is to calculate eigenvalues, which define the amount of total variation that is displayed on the PC axes. The first PC summarizes most of the variability present in the original data relative to all remaining PCs. The second PC explains most of the variability not summarized by the first PC and uncorrelated with the first, and so on (Jolliffe, 1986). Because PCs are orthogonal and independent of each other, each PC reveals different properties of the original data and may be interpreted independently. In this way, the total variation in the original data set may be broken down into components that are cumulative. The proportion of variation accounted for by each PC is expressed as the eigenvalue divided by the sum of the eigenvalues. The eigenvector defines the relation of the PC axes to the original data axes.

When using PCA on molecular marker data, it is preferable not to include negative eigenvalues or any with very low (<1) eigenvalues. To eliminate negative eigenvalues, the similarity matrix may be transformed by the following formula,

$$S'_{ij} = S_{ij} - S_{i.} - S_{.j} + S_{..}$$

where S_{ij} is the coefficient of similarity between individuals i and j, $S_{i.}$ is the mean of the values for the ith row in the similarity matrix, $S_{.j}$ is the mean of the values for the jth column and $S_{..}$ is the overall mean of similarity coefficients. This transformation renders the similarity matrix to have zero root but preserves the distance properties on which the methodology is based (Hayes et al., 1997).

PCA can be performed on two types of data matrices: a variance–covariance matrix and a correlation matrix. With characters of difference scales, a correlation matrix standardizing the original data set is preferred. If the characters are of the same scale, a variance–covariance matrix can be used. In the use of these two types of matrices, one has to consider that with the variance–covariance matrix, absolute changes among individuals can be studied. But, with the correlation matrix, only differences relative to the standardized data can be interpreted (Wiley, 1981).

PCA can also be used to determine the optimum number of clusters in a study. In this case, the objective is to maximize the variation explained by the first PC of each cluster. It begins with all individuals in a single cluster and splits them until the second eigenvalue of

all clusters is less than a level specified by the user. The second eigenvalue may be set at 0.75 to be certain that most of the variation is explained by the first PC (Thompson et al., 1998).

PCoA is a scaling or ordination method that starts with a matrix of similarities or dissimilarities between a set of individuals and aims to produce a low-dimensional graphical plot of the data in such a way that distances between points in the plot are close to original dissimilarities. Thus, the starting point matrix of similarities or dissimilarities for PCoA is different from that of PCA, which starts with the initial data matrix (e.g., presence versus absence of alleles in molecular marker data).

When there are relatively few characters and no missing data, the output of PCA and PCoA will be similar. However, Rohlf (1972) found that in PCoA, the treatment of missing data is more satisfactory than that in PCA. In PCA, each missing value is simply replaced by the mean value for the corresponding character or marker when computing the input matrix for analysis. Thus, one might expect that individuals with lots of missing data may group more closely to the centroid of the group when using PCA compared to PCoA. To overcome the problem of missing data in PCA, the coefficient between two individuals should be independently computed by only using those characters that have been recorded for both the individuals. PCoA is recommended over PCA when there are lots of missing data, and when there are fewer individuals than characters (Rohlf, 1972).

When the first two or three PCs explain most of the variation, PCA and PCoA become useful techniques for grouping individuals by a scatter plot presentation. In PCA or PCoA, when the original data are not highly correlated, the first few PCs do not usually explain much of the original variation. In such a case, assessment of genetic relationships on the basis of the first two or three PCs could lead to misleading interpretations. To avoid such distortion, analysis of genetic relationships among individuals should be based on optimal number of PCs that explain maximum amount of original data variation. The eigenvalue of PCs can be used as a criterion to determine how many PCs should be utilized. The PCs with eigenvalue >1.0 are considered as inherently more informative than any single original variable alone (Iezzoni and Pritts, 1991).

Multidimensional Scaling

Multidimensional scaling (MDS), also referred to as "perceptual mapping," is a procedure that "represents a set of individuals or genotypes (n) in a few dimensions (m) using a similarity/distance matrix between them such that the inter-individual proximities in the map nearly match the original similarities/distances" (Johnson and Wichern, 1992). The technique, thus, attempts to find configurations in $m \geq n - 1$ dimensions, such that the match is as close as possible. It is possible to arrange the n individuals in a low-dimensional coordinate system on the basis of only the rank order of $n(n - 1)/2$ original similarities–distances and not their magnitude. This type of geometric representation is called "non-metric" MDS. If the actual magnitudes of original similarities–distances are used to obtain a geometric representation in m dimensions, the process is called "metric" MDS (Johnson and Wichern, 1992).

The closeness between original similarities–distances and interindividual proximities in the map can be tested by different methods. The most commonly used test is a numerical measure of closeness called "stress." Stress indicates the proportion of the variance of the disparities not accounted for by the MDS model, and is measured as follows: Stress $= [(d_{ij} - \hat{d}_{ij})^2/(d_{ij} - \overline{d})^2]^{1/2}$, where $\overline{d}$ is the average distance ($\Sigma d_{ij}/n$) on the map. The stress value becomes smaller as the estimated map distance approaches the original distance. The interpretation of stress in terms of goodness-of-fit is as follows: a stress level of 0.05 provides excellent fit; with 0.1 a good fit; 0.2 a fair fit; and 0.4 a poor fit (Kruskal, 1964). However, a problem often encountered with the use of stress is analogous to that of R^2 in multiple regression, in that stress always improves with increased dimensions.

For the purpose of visualizing genetic relationships, the distance matrix can be converted into two or more dimensional coordinates by means of MDS (Schiffman et al., 1981; Beebe et al., 1995). In MDS, one can effectively employ the distance matrix obtained among a set of genotypes with data sets such as morphological, biochemical, or molecular marker data as input, to generate a spatial representation of these genotypes in a geometric configuration as output (Thompson et al., 1998; Skroch et al., 1998). The resulting multidimensional distance matrices, reflecting the relationships among a set of genotypes, can be presented as a 2- or 3-dimensional representation that can be more easily interpreted. The pattern obtained from MDS can also be used to estimate the actual number of groups that may be obtained by cluster analysis.

The actual configurations of individuals resulting from PCA, PCoA and MDS are usually similar (Rohlf, 1972). However, results based on MDS might differ in comparison with PCA and PCoA since (i) differences between close individuals are, in general, reflected better by MDS, and (ii) the smaller or greater distances between individuals are not necessarily represented by MDS to the same scale. MDS is preferable over PCA and PCoA when the number of individuals is very large (Rohlf, 1972). Only if there are no missing data or many more individuals than characters, should PCA be employed.

Comparison of Efficiencies of Cluster Analysis, PCA, and PCoA

An increasing number of researchers are employing PCA or PCoA as a "pattern-finding method" to complement cluster analysis (for instance, Kantety et al., 1995; Rincon et al., 1996; Schut et al., 1997; Russell et al., 1997; Johns et al., 1997; Dubreuil and Charcosset, 1998; Lanza et al., 1997; Thompson et al., 1998; Barrett and Kidwell, 1998; Lombard et al., 2000). When there are nonhierarchical and reticular patters of diversity, the

hierarchical algorithms are somewhat limited in their usefulness to investigate pattern of genetic diversity (Lessa, 1990). In such a case, ordination methods such as PCA and PCoA, and particularly MDS, which does not assume linearity, might be more useful (Rendine et al., 1986; Derish and Sokal, 1988).

Using molecular marker data, Melchinger (1993) compared PCA, PCoA, and cluster analysis with respect to their efficiency in analyzing genetic diversity in crop plants. By analyzing a set of five studies in maize and barley, in general, PCA or PCoA provided faithful portrayal of the relationships between major groups of lines, but distances between close neighbors were often distorted when a small proportion (<25%) of the total variation was explained by the first two or three PCs or principal coordinates. Cluster analysis proved to be more sensitive and reliable for detecting pedigree relationships among genotypes than PCA or PCoA when the first two or three PCs explained <25% of the total variation. To extract maximum information from the molecular marker data, PCA or PCoA can be used in combination with cluster analysis, particularly when the first two or three PCs explain >25% of the original variation (Messmer et al., 1992).

The major advantage of ordination methods over cluster analysis is that these methods facilitate the detection of individuals or populations that show some intermediacy between two groups (Lessa, 1990). However, ordination methods such as PCA or PCoA become impractical when more than a few dimensions are needed to present the relationships among genotypes. Also, PCA and PCoA may yield distorted picture of genetic relationships among genotypes or populations when variables are nonlinearly related (Wartenberg et al., 1987). This could be a common problem for frequency data, such as allele frequencies, particularly if the data is heterogeneous. Linkage disequilibrium may also lead to unreliable and unstable patterns (Lessa, 1990).

Utility of Resampling Techniques

Resampling techniques such as "Bootstrap" and "Jackknife" are attracting considerable attention, particularly in relation to application of molecular marker data for analysis of genetic diversity and for finding the smallest set of markers that can provide an accurate assessment of genetic relationships among a set of genotypes or groups or populations (Tivang et al., 1994). The bootstrap technique is a general resampling procedure for estimating the distribution of a statistic on the basis of independent observations (Efron, 1979). The technique resamples the actual data to reveal some its subtler patterns. The basic notion is that the data themselves, viewed as a frequency distribution, represent the best available image of the frequency distribution from which they were drawn. Thus, the bootstrap metaphor refers to the sense in which the data itself is effectively used to assess its own utility in statistical analysis (Crowley, 1992). Bootstrap methods have been mostly employed to estimate standard errors, confidence intervals, and other measures of accuracy for statistical parameters for which analytical methods are not available or are difficult to calculate. The measures of statistical accuracy in a bootstrap analysis are generated from sampling. The parameter of interest is first estimated from the original sample. A vast number of bootstrap samples of size equal to the original sample are then generated by repeatedly sampling the entire original data with replacement. The statistic of interest is then calculated for each bootstrap sample produced (Efron and Tibshirani, 1986, 1993).

An important issue in application of molecular marker data for analysis of genetic diversity concerns the number of markers that can provide a precise estimate of genetic relationships. It is clear that use of large numbers of polymorphic markers or bands which are uniformly distributed over the genome will provide an increasingly more precise estimate of genetic relationships and will reduce the variance estimation of genetic relationship due to over or under sampling of certain regions of the genome (Tivang et al., 1994). Because assaying a large number of polymorphic markers is often prohibitively expensive, it may be desirable to estimate genetic relationships using the smallest set of polymorphic markers with minimum sampling variance. Bootstrap analysis may be used to determine the effective number of molecular markers in analysis of genetic diversity through empirical estimation of sampling variance of genetic distances or similarities calculated from different marker data sets (for instance, Pejic et al., 1998; Vuylsteke et al., 2000). The relationship between the number of bands and sampling variance of genetic similarity or distance among all pairs of genotypes can be used to identify a suitable number of markers providing adequate information, provided that an adequate number of markers were sampled in the first place. The effective number of markers is one where the standard deviation of the estimates is not significantly affected by reducing or increasing the number of loci–bands analyzed. Because molecular markers are capable of generating a large amount of data, they provide an excellent opportunity for bootstrap sampling using whole data sets as well as with smaller partitions of the data set. If N markers are randomly sampled over the genome, the standard error (SE) of Rogers' distance (RD) between homozygous inbreds can be calculated as $SE = RD(1 - RD)/N$ (Dubreuil et al., 1996), which is identical to Jackknife estimate of SE (Melchinger et al., 1991). Alternatively, the SE of GD estimates can be determined by the bootstrap procedure (Tivang et al., 1994).

Bootstrapping can be effectively utilized for estimating the statistical support to the internal branches in a tree (Felsenstein, 1985). For instance, if a specific branching pattern is observed 80% of the time, this branching pattern is said to have 80% bootstrap support. The exact statistical interpretation of bootstrap results is still an active subject of study, but the rule of thumb is that internal tree branches that have >70% bootstrap are likely to be correct at the 95% level (Hillis and Bull, 1993). Some recent studies have utilized such a strategy in indicating bootstrap proportions for internal branches in a tree (for example, Barrett and Kidwell, 1998; Lom-

bard et al., 2000). However, a high bootstrap percentage, indicated by this nonparametric bootstrapping strategy, still does not guarantee that long branch attractions have not biased the results. Also, in many cases, the overall tree structure provides better information than a particular branch (Hillis et al., 1996). Wherever clear formulation of a priori hypotheses regarding genetic relationships is possible, it is preferable to apply parametric rather than nonparametric bootstrapping. Hillis et al. (1996) discussed in detail the strengths and limitations of parametric and nonparametric bootstrapping approaches in this regard.

Another numerical resampling technique is the Jackknife technique, where resampling is performed without replacement (Efron, 1979). Although this is the simplest resampling technique that provides estimates of bias and variance for genetic parameter estimates (for instance, Dje et al., 2000), it imposes a limitation on the number of resampling units and provides little information for the distribution of the estimates. Resampling techniques such as Bootstrap and Jackknife are now increasingly used to analyze the variance of any parameter without any prior information about calculation methods and type of distribution. However, the Bootstrap technique, unlike the Jackknife, is not limited by the number of resampling units, and can thus provide as many new samples as needed to give a reasonable approximation of the distribution of the original estimator.

Concluding Remarks

Many software packages are available for analyzing genetic diversity (Labate, 2000). There are two primary considerations when choosing between the various software packages for analysis of genetic data: (i) statistical packages that offer analyses on the basis of relevant evolutionary models and, (ii) user-friendliness of the packages. It is clear from many reports and publications that the latter consideration appears to rank highest. Feature-packed, menu-driven statistical packages, such as NTSYS-pc (F.J. Rohlf, State University of New York, Stony Brook, USA) and PHYLIP (J. Felsenstein, University of Washington, Seattle, USA), are providing useful means of analyzing diverse data sets for assessment of genetic diversity in plants and animals, and new packages are constantly being developed. Many of the recently developed packages include a range of possible options for analyzing (i) level of polymorphism (ii) allele and genotype frequencies, (iii) homozygosity and heterozygosity, (iv) conformance with Hardy-Weinberg expected proportions, (v) heterogeneity, (vi) cluster patterns, and (vii) numerical resampling such as bootstrapping or jackknifing.

Each data set (morphological, biochemical, or molecular) has its own strengths and constraints, and there is no single or simple strategy to address effectively various complex issues related to choice of distance measure(s), clustering methods, determination of optimal number of clusters or analysis of individual, and combined data sets by means of various statistical tools. However, empirical data generated in recent years by different strategies has provided an enhanced understanding of the above issues, and reasonably effective means of analyzing genetic diversity at various levels (individuals, populations, or species). With the recent development and use of model-based clustering methods based on Bayesian statistics, the possibilities of carrying out association studies in crop plants for identifying genes for agronomically important but complex traits have been enhanced (Pritchard, 2001; Thornsberry et al., 2001). There is still a distinct need for developing comprehensive and user-friendly statistical packages that facilitate an integrated analysis of different data sets for generating reliable information about genetic relationships, germplasm diversity, and favorable allele variation. Equally important, and perhaps more challenging, is the concerted and planned utilization of germplasm in crop breeding programs on the basis of knowledge accrued from studies on genetic diversity.

ACKNOWLEDGMENTS

We thank Marilyn Warburton and Gael Pressoir (CIMMYT, Mexico) for a critical reading of the manuscript and for their valuable comments. We appreciate the suggestions and helpful remarks made by the anonymous referees.

REFERENCES

Ajmone-Marsan, P., C. Livini, M.M. Messmer, A.E. Melchinger, and M. Motto. 1992. Cluster analysis of RFLP data from related maize inbred lines of the BSSS and LSC heterotic groups and comparison with pedigree data. Euphytica 60:139–148.

Ajmone-Marsan, P., P. Castiglioni, F. Fusari, M. Kuiper, and M. Motto. 1998. Genetic diversity and its relationship to hybrid performance in maize as revealed by RFLP and AFLP markers. Theor. Appl. Genet. 96:219–227.

Anderberg, M.R. 1973. Cluster Analysis for Applications. Academic Press, New York.

Bar-Hen, A., A. Charcosset, M. Bourgoin, and J. Cuiard. 1995. Relationships between genetic markers and morphological traits in a maize inbred lines collection. Euphytica 84:145–154.

Barrett, B.A., and K.K. Kidwell. 1998. AFLP-based genetic diversity assessment among wheat cultivars from the Pacific Northwest. Crop Sci. 38:1261–1271.

Baverstock, P.R., and C. Moritz. 1996. Project design. p. 17–27. *In* D.M. Hillis et al. (ed.) Molecular systematics. Sinauer Associates, Sunderland, MA.

Beaumont, M.A., K.M. Ibrahim, P. Boursot, and M.W. Bruford. 1998. Measuring genetic distance. p. 315–325. *In* A. Karp et al. (ed.) Molecular tools for screening biodiversity. Chapman and Hall, London.

Beebe, S.E., I. Ochoa, P. Skroch, J. Nienhuis, and J. Tivang. 1995. Genetic diversity among common bean breeding lines developed for Central America. Crop Sci. 35:1178–1183.

Bernardo, R. 1993. Estimation of coefficient of coancestry using molecular markers in maize. Theor. Appl. Genet. 85:1055–1062.

Bohn, M., H.F. Utz, and A.E. Melchinger. 1999. Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs and SSRs and their use in predicting progeny variance. Crop Sci. 39:228–237.

Brown, A.H.D. 1989. The case for core collection. p. 136–156. *In* A.H.D. Brown et al. (ed.) The use of plant genetic resources. Cambridge Univ. Press, Cambridge, England.

Brown, A.H.D., and B.S. Weir. 1983. Measuring genetic variability in plant populations. p. 219–229. *In* S.D. Tanskley and T.J. Orton (ed.). Isozymes in plant genetics and breeding. Part A. Elsevier, Amsterdam.

Brown, J.K.M. 1994. Bootstrap hypothesis tests for evolutionary trees

and other dendrograms. Proc. Natl. Acad. Sci. (USA) 91:12293–12297.
Brown-Guedira, G.L., J.A. Thompson, R.L. Nelson, and M.L. Warburton. 2000. Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. Crop Sci. 40:815–823.
Chippindale, P.T., and J.I. Weins. 1994. Weighting, partitioning and combining characters in phylogenetic analysis. Syst. Biol. 43:273–287.
Cockerham, C.C. 1969. Variance of gene frequencies. Evolution 23: 72–84.
Cockerham, C.C. 1973. Analysis of gene frequencies. Genetics 74: 679–700.
Cox, T.S., J.P. Murphy, and D.M. Rodgers. 1986. Changes in genetic diversity in the red winter wheat regions of the United States. Proc. Natl. Acad. Sci. (USA) 83:5583–5586.
Crossa, J., C.M. Hernandez, P. Bretting, S.A. Eberhart, and S. Taba. 1993. Statistical genetic considerations for maintaining germplasm collections. Theor. Appl. Genet. 86:673–678.
Crowley, P.H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. Annu. Rev. Ecol. Syst. 23: 405–447.
Derish, P.A., and R.R. Sokal. 1988. A classification of European populations based on gene frequencies and cranial measurements: A map-quadrant approach. Hum. Biol. 60:801–824.
Dje, Y., M. Heuretz, C. Lefebvre, and X. Vekemans. 2000. Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. Theor. Appl. Genet. 100:918–925.
Dubreuil, P., and A. Charcosset. 1998. Genetic diversity within and among maize populations: A comparison between isozymes and nuclear RFLP loci. Theor. Appl. Genet. 96:577–587.
Dubreuil, P., P. Dufour, E. Krejci, M. Causse, D. de Vienne, A. Gallais, and A. Charcosset. 1996. Organization of RFLP diversity among inbred lines of maize representing the most significant heterotic groups. Crop Sci. 36:790–799.
Efron, B. 1979. Bootstrap methods: Another look at the jacknife. Ann. Statist. 7:1–26.
Efron, B., and R.J. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Statist. Sci. 1:54–77.
Efron, B., and R.J. Tibshirani. 1993. An introduction to bootstrap. Chapman and Hall, London.
Everitt, B. 1980. Cluster analysis. 2nd edition, Halstead Press, New York.
Excoffier, L., P. Smouse, and J. Quattro. 1992. Analysis of molecular variance inferred for metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.
Felsenstein, J. 1984. Distance methods for inferring phylogenies: A justification. Evolution 38:16–24.
Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:789–791.
Franco, J., J. Crossa, J. Villasenor, S. Taba, and S.A Eberhart. 1997. Classifying Mexican maize accessions using hierarchical and density search methods. Crop Sci. 37:972–980.
Franco, J., J. Crossa, J. Villasenor, S. Taba, and S.A. Eberhart. 1998. Classifying genetic resources by categorical and continuous variables. Crop Sci. 38:1688–1696.
Franco, J., J. Crossa, J.M. Ribaut, J. Betran, M.L. Warburton, and M. Khairallah. 2001. A method for combining molecular markers and phenotypic attributes for classifying plant genotypes. Theor. Appl. Genet. 103:944–952.
Frankel, O.H., A.H.D. Brown, and J.J. Burdon. 1995. The conservation of plant biodiversity. Cambridge Univ Press, Cambridge, England.
Goodman, M.N. 1972. Distance analysis in biology. Syst. Zool. 21: 174–186.
Gower, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–874.
Hair, J.R., R.E. Anderson, R.L. Tatham, and W.C. Black. 1995. Multivariate data analysis with readings. 4th edition, Prentice-Hall, Englewood Cliffs, NJ.
Hallauer, A.R., and J.B. Miranda. 1988. Quantitative genetics in maize breeding. 2nd edition, Iowa State University Press, Ames, IA.
Hamrick, J.L., and M.J.W. Godt. 1997. Allozyme diversity in cultivated crops. Crop Sci. 37:26–30.
Hayes, P.M., J. Cerono, H. Witsenboer, M. Kuiper, M. Zabeau, K. Sato, A. Kleinhofs, D. Kudrna, A. Kilian, M.A. Saghai-Maroof, and D. Hoffman and The North American Barley Genome Mapping Project. 1997. Characterizing and exploiting genetic diversity and quantitative traits in barley (*Hordeum vulgare*) using AFLP markers. J. Quant. Trait Loci (no longer available) (see http://www.ncgr.org/jag/papers97/paper297/jqtl1997-02.html; verified 10 February 2003).
Hillis, D.M. 1987. Molecular versus morphological approaches to systematics. Annu. Rev. Ecol. Syst. 18:23–42.
Hillis, D.M., and J.J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.
Hillis, D.M., B.K. Mable, and C. Moritz. 1996. Applications of molecular systematics: The state of the field and a look into the future. p. 515–543. *In* D.M. Hillis et al. (ed.) Molecular systematics. 2nd edition, Sinauer Associates, Sunderland, MA.
Iezzoni, A.F., and M.P. Pritts. 1991. Applications of principal component analysis to horticultural research. HortScience 26:334–338.
Jaccard, P. 1908. Nouvelles researches sur la distribution florale. Bull. Soc. Vaudoise Sci. Natl. 44:223–270.
Johns, M.A., P.W. Skrotch, J. Neinhuis, P. Hinrichsen, G. Bascur, and C. Munoz-Schick. 1997. Gene pool classification of common bean landraces from Chile based on RAPD and morphological data. Crop Sci. 37:605–613.
Johnson, A.R., and D.W. Wichern. 1992. Applied multivariate statistical analysis. 3rd edition, Prentice-Hall, Englewood Cliffs, NJ.
Jolliffe, I.T. 1986. Principal component analysis. Springer-Verlag, Berlin.
Kantety, R.V., X. Zeng, L.B. Jeffrey, and B.E. Zehr. 1995. Assessment of genetic diversity in dent popcorn (*Zea mays* L.) inbred lines using inter-simple sequence repeat (ISSR) amplification. Mol. Breed. 1:365–373.
Karp, A., S. Kresovich, K.V. Bhat, W.G. Ayad, and T. Hodgkin. 1997. Molecular tools in plant genetic resources conservation: A guide to the technologies. IPGRI, Rome.
Kruskal, J.B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27.
Labate, J.A. 2000. Sofware for population genetic analysis of molecular marker data. Crop Sci. 40:1521–1528.
Lanza, L.L.B., C.L. Souza, Jr., L.M.M. Ottobani, M.L.C. Vieira, and A.P. de Souza. 1997. Genetic distance of inbred lines and prediction of maize single cross performance using RAPD markers. Theor. Appl. Genet. 94:1023–1030.
Lessa, E.P. 1990. Multidimensional analysis of geographic genetic structure. Syst. Zool. 39:242–252.
Link, W., C. Dickens, M. Singh, M. Schwall, and A.E. Melchinger. 1995. Genetic diversity in European and Mediterranean faba bean germplasm revealed by RAPD markers. Theor. Appl. Genet. 90: 27–32.
Liu, S., R.G. Cantrell, J.C. McCarty, and M.D. Stewart. 2000. Simple sequence repeat-based assessment of genetic diversity in cotton race accessions. Crop Sci. 40:1459–1469.
Lombard, V., C.P. Baril, P. Dubreuil, F. Blouet, and D. Zhang. 2000. Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: Consequences for varietal registration. Crop Sci. 40:1417–1425.
Lübberstedt, T., A.E. Melchinger, C. Duβle, M. Vuylsteke, and M. Kuiper. 2000. Relationship among early European maize inbreds: IV. Genetic diversity revealed with AFLP markers and comparison with RFLP, RAPD and pedigree data. Crop Sci. 40:783–791.
Lynch, M., and T.J. Crease. 1990. The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. 7:377–394.
Manly, F.F.J. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, New York.
Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27:209–220.
Marshall, D.R., and A.H.D. Brown. 1975. Optimum sampling strategies in genetic conservation. p. 53–80. *In* O.H. Frankel and J.G.

Hawkes (ed.) Crop Genetic resources for today and tomorrow. Cambridge Univ. Press, Cambridge, England.
Melchinger, A.E. 1993. Use of RFLP markers for analyses of genetic relationships among breeding materials and prediction of hybrid performance. p. 621–628. *In* D.R. Buxton (ed.) Proceedings of the International Crop Science Congress, 1st, Ames, IA. July 1992. CSSA, Madison, WI.
Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman, and K.R. Lamkey. 1991. Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci. 31:669–678.
Messmer, M.M., A.E. Melchinger, J. Boppenmaier, R.G. Herrmann, and E. Brunklaus-Jung. 1992. RFLP analyses of early-maturing European maize germplasm: I. Genetic diversity among flint and dent inbreds. Theor. Appl. Genet. 83:1003–1012.
Messmer, M.M., A.E. Melchinger, R.G. Herrmann, and J. Boppenmaier. 1993. Relationships among early European maize inbreds: II. Comparison of pedigree and RFLP data. Crop Sci. 33:944–950.
Milligan, G.W., and M. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159–179.
Mohna, F.I., P. Shen, S.C. Jong, and K. Orikono. 1992. Molecular evidence supports the separation of *Lentinula edodes* from *Lentinus* and related genera. Can. J. Bot. 70:2446–2452.
Mumm, R.H., J. Hubert, and J.W. Dudley. 1994. A classification of 148 U.S. maize inbreds: II. Validation of cluster analysis based on RFLPs. Crop Sci. 34:852–865.
Namkoong, G. 1988. Sampling for germplasm collections. HortScience 23:79–81.
Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. (USA) 70:3321–3323.
Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.
Nei, M., and W. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. (USA) 76:5269–5273.
Nei, M., and R.K. Chesser. 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47:253–259.
Panchen, A.L. 1992. Classification, evolution and the nature of biology. Cambridge Univ. Press, Cambridge, England.
Pedersen, G., and O. Seberg. 1998. Molecules vs morphology. p. 359–365. *In* A. Karp et al. (ed.) Molecular tools for screening biodiversity. Chapman and Hall, London.
Peeters, J.P., and J.A. Martinelli. 1989. Hierarchical cluster analysis as a tool to manage variation in germplasm collections. Theor. Appl. Genet. 78:42–48.
Pejic, I., P. Ajmone-Marsan, M. Morgante, V. Kozumplick, P. Castiglioni, G. Taramino, and M. Motto. 1998. Comparative analysis of genetic dissimilarity among maize inbred lines detected by RFLPs, RAPDs, SSRs and AFLPs. Theor. Appl. Genet. 97:1248–1255.
Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol. Breed. 2:225–238.
Pritchard, J.K. 2001. Deconstructing maize population structure. Nat. Genet. 28:203–204.
Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Rendine, S., A. Piazza, and L.L. Cavalli-Sforza. 1986. Simulation and separation by principal components of multiple demic expansions in Europe. Am. Nat. 128:681–706.
Rincon, F., B. Johnson, J. Crossa, and S. Taba. 1996. Cluster analysis, an approach to sampling variability in maize accessions. Maydica 41:307–316.
Robinson, J.P., and S.A. Harris. 1999. Amplified fragment length polymorphisms and microsatellites: a phylogenetic perspective. *In* E.M. Gillet (ed.) Which DNA marker for which purpose? Final Compendium of the Research Project Development, Optimisation and Validation of Molecular Tools for Assessment of Biodiversity in Forest Trees in the European Union (see http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm; verified 10 February 2003).
Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. Studies in genetics. VII. Univ. Tex. Publ. 2713:145–153.
Rohlf, F.J. 1972. An empirical comparison of three ordination techniques in numerical taxonomy. Syst. Zool. 21:271–280.
Rohlf, F.J. 1992. NTSYS-pc (Numerical Taxonomy and Multivariate Analysis System). Version 1.70. Exeter, Setauket, NY.
Rohlf, F.J., and D.R. Fisher. 1968. Tools for hierarchical structure in random sets. Syst. Zool. 17:407–412.
Romesburg, H.C. 1984. Cluster analysis for researchers. Lifetime Learning Publications, Belmont, CA.
Russell, J.R., J.D. Fuller, M. Macaulay, B.G. Hatz, A. Jahoor, W. Powell, and R. Waugh. 1997. Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. Theor. Appl. Genet. 95:714–722.
Seberg, O., G. Pedersen, and C. Baden. 1996. The phylogeny of *Psathyrostachys* (Triticeae, Poaceae)—Are we able to see the wood for the trees? p. 247–253. *In* R.R-C. Wang et al. (ed.) Proc. 2nd Int. Triticeae Conf. Utah State Univ. Press, Logan, UT.
Schiffman, S.S., M.L. Reynold, and F.W. Young. 1981. Introduction to multidimensional scaling: Theory, methods and applications. Academic Press, New York.
Schut, J.W., X. Qi, and P. Stam. 1997. Association between relationship measures based on AFLP markers, pedigree data and morphological traits in barley. Theor. Appl. Genet. 95:1161–1168.
Sjogren, P., and P.I. Wyone. 1994. Conservation genetics and detection of rare alleles in finite populations. Conserv. Biol. 8:267–270.
Skroch, P.W., J. Nienhuis, S. Beebe, J. Tohm, and F. Pedraza. 1998. Comparison of Mexican common bean (*Phaseolus vulgaris L.*) core and reserve germplasm collections. Crop Sci. 38:488–496.
Smith, J.S.C. 1984. Genetic variability within U.S. hybrid maize: Multivariate analysis of isozyme data. Crop Sci. 24:1041–1046.
Smith, O.S., and J.S.C. Smith. 1992. Measurement of genetic diversity among maize hybrids; A comparison of isozymic, RFLP, pedigree, and heterosis data. Maydica 37:53–60.
Smith, J.S.C., O.S. Smith, S.L. Boven, R.A. Tenborg, and S.J. Wall. 1991. The description and assessment of distances between inbred lines of maize. III: A revised scheme for the testing of distinctiveness between inbred lines utilizing DNA RFLPs. Maydica 36:213–226.
Smith, J.S.C., S. Paszkiewics, O.S. Smith, and J. Schaeffer. 1987. Electrophoretic, chromatograhic and genetic techniques for identifying associations and measuring genetic diversity among corn hybrids. p. 187–203. *In* Proc. 42nd Annu. Corn Sorghum Res. Conf., Chicago, IL. Am. Seed Trade Assoc., Washington, DC.
Sneath, P.H.A., and R.R. Sokal. 1973. Numerical taxonomy. Freeman, San Francisco.
Sokal, R.R., and C.D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38:1409–1438.
Souza, E., and M.E. Sorrells. 1991. Relationships among 70 North American oat germplasms: I. Cluster analysis using quantitative characters. Crop Sci. 31:605–612.
Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic inference. p. 407–514. *In* D.M. Hillis et al. (ed.) Molecular systematics. 2nd edition, Sinauer Associates, Sunderland, MA.
Taba, S., J. Diaz, J. Franco, J. Crossa, and S.A. Eberhart. 1999. A core subset of LAMP, from the Latin American Maize Project. CD-ROM, CIMMYT, Mexico, D.F., Mexico.
Thompson, J.A., and R.L. Nelson. 1998. Core set of primers to evaluate genetic diversity in soybean. Crop Sci. 38:1356–1362.
Thompson, J.A., R.L. Nelson, and L.O. Vodkin. 1998. Identification of diverse soybean germplasm using RAPD markers. Crop Sci. 38:1348–1355.
Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. Nat. Genet. 28:286–289.
Tivang, G., J. Nienhuis, and O.S. Smith. 1994. Estimation of sampling variance of molecular marker data using the bootstrap procedure. Theor. Appl. Genet. 89:259–264.
van Bueningen, L.T., and R.H. Busch. 1997. Genetic diversity among North American spring wheat cultivars: I. Analysis of the coefficient of parentage matrix. Crop Sci. 37:570–579.
van Hintum, Th.J.L., and D. Haalman. 1994. Pedigree analysis for composing a core collection of modern cultivars, with examples from barley (*Hordeum vulgare*). Theor. Appl. Genet. 88:70–74.

Virk, P.S., H.J. Newbury, M.T. Jackson, and B.V. Ford-Lloyd. 2000. Are mapped markers more useful for assessing genetic diversity? Theor. Appl. Genet. 100:607–613.

Vuylsteke, M., R. Mank, B. Brugmans, P. Stam, and M. Kuiper. 2000. Further characterization of AFLP data as a tool in genetic diversity assessments among maize (*Zea mays* L.) inbred lines. Mol. Breed. 6:265–276.

Warburton, M., and J. Crossa. 2000. Data analysis in the CIMMYT Applied Biotechnology Center for Fingerprinting and Genetic Diversity Studies. CIMMYT, Mexico.

Warburton, M.L., X. Xianchun, J. Crossa, J. Franco, A.E. Melchinger, M. Frisch, M. Bohn, and D. Hoisington. 2002. Genetic characterization of CIMMYT inbred maize lines and open pollinated populations using large scale fingerprinting methods. Crop Sci. 42:1832–1840.

Ward, J.H., Jr. 1963. Hierarchical grouping to optimize an objective function. J. Am. Statist. Assoc. 58:236–244.

Wartenberg, D., S. Ferson, and F.J. Rohlf. 1987. Putting things in order: A critique of detrended correspondence analysis. Am. Nat. 129:434–448.

Weir, B.S. 1990. Genetic data analysis. Sinauer Associates, Sunderland.

Weir, B.S. 1996. Intraspecific differentiation. p. 385–403. *In* D.M. Hillis et al. (ed.) Molecular systematics. 2nd edition, Sinauer Associates, Sunderland, MA.

Weir, B.S., and C.C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370.

Wiley, E.O. 1981. Phylogenetics: The theory and practice of phylogenetics and systematics. John Wiley, New York.

Wishart, D. 1987. CLUSTAN user manual. 3rd edition, Program Library Unit, Univ. of Edinburgh, Edinburgh.

Wrigley, C.W., J.C. Autran, and W. Bushuk. 1982. Identification of cereal varieties by gel electrophoresis of the grain proteins. Adv. Cereal Sci. Technol. 5:211–259.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. 15:323–354.