15 Plant Genetic Diversity

Statistical Methods for Analyzing Distribution and Diversity of Species

Murari Singh, Ardeshir B. Damania, and Yogendra P. Chaubey

CONTENTS

Availability of plant genetic diversity is fundamental to the existence of the living planet. Conservation of biodiversity has been a practice of all concerned professions, including farmers, since ancient times. However, the changes in environmental conditions and pressures from population and technological change have resulted in genetic modification, including replacement of landraces and erosion of genetic diversity.

The relationship between loss of diversity and climate change has been well recognized with unprecedented higher levels of species extinction (Hooper et al. 2012). In order to minimize genetic erosion and capture diversity, a need for collection and conservation of biodiversity has been stressed in the Convention on Biological Diversity (Articles 8 and 9, CBD 1992), Agenda 21 (Chapters 14 and 15, UNCED 1992), and the Global Biodiversity Strategy (WRI et al. 1992), using several mechanisms such as in-situ, ex-situ, and in-vitro conservation. A number of references dealing with various aspects can be found in Guarino et al. (1995). Technical guidelines for germplasm exploration and collection including planning, methods, and procedures illustrated with real germplasm collection missions are given in Engels et al. (1995), whereas examples of planning and execution of a genetic resource collection have been given in Bennett (1970), Chang (1985), Damania (1987), and Kameswara Rao and Bramel (2000). This chapter discusses, in brief, statistical features of collection and analysis of data in this context.

SAMPLING FOR SPECIES AND GENETIC DIVERSITY

One practical way for preserving the species and genetic diversity is to collect and conserve (and regenerate) samples with maximum diversity in species and genomic information in relation to the environment/region. The sampling strategy will depend on the population structure, distribution of the traits or genes, and the statistical measure of the diversity captured in the sample as well as the precision required. The sample (sampling fraction) should be as large as possible with maximum information, on one hand, to serve the principle, but should be small enough to be collected and maintained within the limited time and resources available in practice on the other. An optimum sampling strategy for genetic conservation of crop plants under threat of extinction has been discussed by Marshall and Brown (1975), Weir (1990), and Brown and Marshall (1995), among others. This depends on the genetic variation in the set of populations under investigation in terms of genotype and allele frequencies or allelic richness, gene diversity, heterozygosity levels, and disequilibrium coefficients, and so on. Often in practice, for a single population, allelic richness is measured as the average number of alleles for a large number of markers. In case of sampling from several populations, the sampling strategy depends on the extent of genetic divergence among populations (e.g., in terms of number of alleles that attain appreciable frequencies in individual populations) and the level of genetic variation (e.g., in the distribution of number of alleles per locus).

~~Size of Samples~~

From the neutral theory of Kimura and Crow (1964), the approximate number of neutral alleles (k) in a sample of S random gametes, from a population of size N in equilibrium, at a locus with mutation rate u is given as (Brown and Briggs 1991, Brown and Marshall 1995).

$$k \approx \theta \ln\left[\frac{(S+\theta)}{\theta}\right] + 0.6, \text{ where } \theta = 4\mathrm{Nu} > 0.1 \text{ and } S > 10$$

A basic sampling strategy should take into account the number and location of sampling units, the number of individual plants sampled at a site, the choice of individuals, and the number and type of propagules per plant. This strategy needs refinement in view of the information available on the genetic structure of the target populations. Thus, modifications are required to the basic sampling strategy for different species to address spatio-temporal distribution, life history, genetic system (mating structure), and mixture of the populations. Modifications in sampling strategy are also required when sampling is for specific goals (e.g., collecting for additional genes for a resistance to a specific disease). Sedcole (1977) gave expressions for the sample size (S) to recover, with 95% confidence, a minimum number r of plants with a trait that occurs in population with frequency p:

$$S \approx \frac{\left[r + 1.645r^{0.5} + 0.5\right]}{p}$$

In cases where partial information is available from a previous collection on the target species, re-sampling from the region could be done to improve the information content on the genetic diversity. A discussion of the advantages and disadvantages from various angles of covering unstudied areas, returning to the areas with high genetic diversity, or collecting for specific ecotypes are presented in Nabhan (1990). The gain due to re-sampling could be obtained in terms of change in diversity measure. A statistical test for significance of additional information from recollection is available in Rao (1973).

Brown (1989, 1992) made recommendations on the sample sizes: sample about 50 populations in an eco-geographical area or on a specific mission; the size of a sample field should be 50 individuals per site to capture locally common alleles with $p > 0.95$; the total collection size should be 30,000 individuals per species to include widespread rare alleles present at mutation rate; the core collection size should be 3,000 individuals per species to give the expected number of alleles equal to number of alleles in the species with a frequency $>10^{-4}$; the minimum number of sites for endangered species should be five (collecting 10 individuals per site) to ensure survival of worthwhile genotypes.

Methods of Sampling

General sampling procedures have been covered in standard texts (Cochran 1977, Sukhatme et al. 1984, among others). Application of any sampling technique requires preparation of a sampling frame (the list of all sampling units) of the population; of all the subpopulations (in the case of stratified sampling, where the population under study is divided into strata or subpopulations). In a simple random sampling, the sampling units are selected with equal probability. In systematic sampling, the sampling units are aligned on a rectangular grid (rows and columns). One of the rows (or columns, also called *clusters*) is selected using simple random sampling. It may be noted that while systematic sampling may be operationally very convenient, it does not provide an unbiased estimate of the population variance. Applications of various sampling methods in the context of wheat germplasm collection in the West Asia and Northern Africa region are discussed in Damania (1987), Valkoun and Damania (1990), and Porceddu and Damania (1991). When the samples are collected with geo-reference coordinates, spatial models should be used to analyze such data (Cressie 1993). The distance-sampling approach, used to collect the information on the spatial pattern of species or genetic diversity, comprises data on distances from a randomly placed line or point to the object of interest (Buckland et al. 1993).

DATA ANALYSIS METHODS

Several statistical methods are available to analyze data collected for various objectives relevant to plant genetic resources. We discuss a few of these methods used for specific purposes.

Measurement of Diversity

Diversity is widely used to judge the suitability of a habitat for conservation (Magurran 1988). It has two components: (1) variety/richness in terms of entities such as alleles, genes, varieties, populations, species, and genus and (2) relative abundance of the entities. The diversity measures have been defined by combining these two components in various ways. The most popular indices are Margalef's diversity index (Clifford and Stephenson 1975): $D_{\mathrm{Mg}} = (S-1)/\ln N$ and Menhinick's index (Whittaker 1977): $D_{\mathrm{Mn}} = S/\sqrt{N}$, where S is the number of species (groups/clusters) recorded and N is the total number of individuals (samples) summed over all the species (groups/clusters).

Shannon and Simpson Indices

Shannon and Simpson indices are heterogeneity measures, and they include both richness and abundance in one single value. The Shannon index is $H' = -\sum p_i \times \ln(p_i)$ and the Simpson index is $D = E(p_i)^2$, where p_i is estimated as n_i/N, where n_i is the species record (specimens, records from the flora, and germplasm data) and N is the total number of individual records of all the species.

The Shannon index has been corrected for bias and is given as

AQ 1

$$H' = -\sum p_i \times \ln(p_i) - \left[\frac{(S-1)}{N}\right] + \left[\frac{\left(1 - \sum p_i^{-1}\right)}{\left(12N^2\right)}\right] + \sum\left[\frac{\left(p_i - 1 - p_i^{-2}\right)}{\left(12N^3\right)}\right]$$

with variance

AQ 2

$$\mathrm{Var}(H_1) = \left(\sum p_i \times \ln(p_i)\right)^2 - \left(\sum p_i \times \ln(p_i)\right)^2 / N - (S-1)/\left(2N^2\right)$$

See Hutcheson (1970) and Bowman et al. (1971).

A study of plant diversity may throw light on describing abundance of species, estimating the number of species in a given region, species diversity via indices, comparing diversities across regions, association between species and geographical region, spatial modeling of abundance of species, and so on. Various indices, such as those given above, may be computed using a number of tools in Excel, R-package (R Development Core Team 2009), Genstat statistical software (Payne 2014), and others. To illustrate, consider a hypothetical dataset with observed abundances (53, 33, 26, 16, 8, 2, and 1) for seven species. The following Genstat procedure~~s~~

```
ECDIVERSITY [PRINT=index,estimate;
INDEX=hshannon,jshannon,simpson,isimpson; BMETHOD=bootstrap;\
CIPROBABILITY=0.95; NBOOT=100; SEED=12431]!(53,33,26,16,8,2,1)
```

would yield the output of Table 15.1.

The standard errors of the diversity indices have been evaluated using bootstraps of 100 replications. An open source package *BiodiversityR* in the R-software application (R Development Core Team 2009) may also be used to compute these indices (Kindt and Coe 2005).

Distributional Behavior of Abundance of Species

A number of theoretical models have been studied for their goodness of fit to the observed abundances of species (frequencies) (Engen and Taillie 1979). Let us denote the probability or relative abundance of the ith species by p_i, where $i = 1,\ldots,$ s, and s is the number of species in the population (in the region under consideration). The most frequently used models include the following:

1. Uniform or completely even model

$$p_i = \frac{1}{s}, \quad \left(i = 1,\ldots,s\right)$$

TABLE 15.1
Values of Diversity Indices, Bootstrap Diversity Statistics, and Confidence Intervals Produced by Genstat from a Hypothetical Dataset (See Text)

	Shannon-Wiener H	Shannon-Wiener J	Simpson 1-D	Simpson 1/D
Index	1.532	0.7874	0.7519	4.030
Bootstrap estimate	1.510	0.8099	0.7469	3.970
Bootstrap s.e.	0.052	0.0399	0.0177	0.271
95% confidence interval	(1.404, 1.607)	(0.7333, 0.9088)	(0.7059, 0.7777)	(3.400, 4.499)

s.e. is the standard error.

2. Broken stick model

$$p_i = \left(\frac{1}{s}\right)\left(\frac{1}{i} + \frac{1}{(i+1)} + \cdots + \frac{1}{s}\right), \ (i = 1,\ldots,s)$$

3. Geometric series model

   A function to describe the distribution of such frequencies is the geometric frequency distribution function (May 1975), which is given by

$$n_i = NC(k)k(1-k)^{i-1}, \quad i = 1,\ldots,s, \ 0 < k < 1$$

   where:

   k is the unknown parameter representing the proportion of available niche space or resource that each species occupies
   n_i is the number of individuals in the ith species
   N is the total number of individuals
   s is the total number of species
   $C(k) = [(1 - (1 - k)^s]^{i-1}$

   An example of the geometric distribution fitted satisfactorily to data on wild wheat (species of *Aegilops*) with the estimates of the parameter $k = 0.1966$ with a standard error of 0.0213 is provided by Bari and Singh (1998).

4. Infinite geometric model

$$n_i = Nk(1-k)^{i-1}, \quad i = 1,\ldots,s, \ 0 < k < 1$$

   As an example, to fit a geometric model to the hypothetical data used above—observed abundances (53, 33, 26, 16, 8, 2, and 1) for seven species—we may use the following command in the Genstat software:

```
ECFIT [PRINT=summary,estimates; MODELTYPE=geometric]
!(53,33,26,16,8,2,1)
```

   This results in the following output (Table 15.2) and the plot indicating the goodness of fit (Figure 15.1).

5. Gamma model

In order to account for variation in abundances over time and space, a number of continuous probability models have been found suitable. The gamma model, with index k and mean k/θ, has the following probability density function:

AQ 3

$$f(p) = \left(\frac{\theta^k p^{k-1} \mathrm{e}^{-\theta p}}{\Gamma(k)}\right), \ \theta > 0$$

TABLE 15.2
Summary of Genestat Output for the Geometric Model with a Hypothetical Dataset

Parameter	Value
Deviance	5.89 on 6 d.f.
Number of individuals	139
Number of species	7
Estimate of *k*	0.4006
Standard error of estimate	0.03487

d.f. is the degrees of freedom.

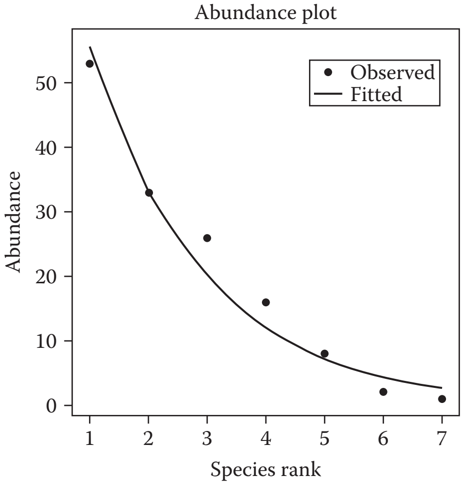

FIGURE 15.1 Geometric model fitted to the hypothetical data on abundances.

Expected Number of Species in a Region

Based on an observed abundance data, one may be interested in estimating the species richness in terms of the number of species that may be expected from a given sample size. Across various regions of collection, sample sizes are not always equal. In such cases, rarefaction is a way to counter this problem. This can be done by using the rarefaction technique of Sanders and modified by Hurlbert (1971):

$$E(S)=\Sigma\left[1-\frac{\binom{N-Ni}{n}}{\binom{N}{n}}\right]$$

where:

$E(S)$ is the expected number of species
n is a standardized sample size (number of individual in the smallest sample)
N is the total number of individuals recorded
N_i is the number of individual in the ith species

In order to compute a confidence interval for the number of species in the region, it would be worthwhile to obtain an expression for variance of S. In its absence, one may use either a bootstrap method (Efron and Tibshirani 1993) or standard error from a number of independent estimates, for example, by dividing the region randomly into a number of groups (Bari and Singh 1998).

Core Collections

The idea of a core collection was put forth by Frankel (1984). Frankel and Brown AQ 4 (1984) and Brown (1989) described the essential features: “A core collection consists of a limited set of accessions derived from an existing germplasm collection, chosen to represent as much as possible the genetic diversity present in the whole spectrum. The remaining accessions in the collection are called the reserve collection.” The advantages of having a core collection includes serving as a standard for including new accessions, efficiency of conservation, characterization, evaluation, enhancement, and distribution (Brown 1995).

The basic issues in forming a core collection include determining the (optimum) size and quality of the accessions selected in terms of representing the diversity (Brown 1995). The first step is to begin with the passport and characterization data for the available collection, followed by grouping of the accessions, selection of entries from each group, and evaluating the core accessions thus selected. AQ 5 Yonezawa et al. (1995) concluded that an optimal stratification sampling strategy is a proportional allocation to the number of accessions in the group and they gave procedures to retain both the pattern and the range of genetic diversity in the whole collection.

M- and H-Strategies

Schoen and Brown (1995) presented two core-collection strategies, termed *M-strategy* and *H-strategy*, for maximizing genetic diversity in core collection, using stratified sampling and marker-gene data for selecting allele-rich accessions from different regions. The H-strategy (after Nei’s diversity index) allots the number of accessions in proportion to the sum, over all loci, of the θ coefficient ($=h/(1-h)$), a function of Nei’s diversity index (h). The M-strategy (maximization strategy) pinpoints the individual accessions from each geographic group to be selected in the core collection to maximize the allelic diversity (using genetic-marker loci) expected to maximize the target allelic diversity. M-strategy uses a linear programming approach for minimizing the loss of marker alleles with respect to those present in the entire collection, subject to the conditions based on the number of accession in each group and on the total number of accessions in the core.

Coefficient of Variation

The coefficient of variation (CV) of a quantitative character used in defining the core collection can indicate the genetic diversity/variability of the collection and allow an estimate of how the inclusion or deletion of an accession can influence the diversity of the core.

Multivariate Methods

We discuss a number of multivariate analyses that have been used in diversity studies from a number of perspectives (Digby and Kempton 1987). The data at the DNA level are most appropriate for quantifying genetic diversity, while morphological and agronomic traits indicate the influence of environmental factors.

Classification Methods

Multivariate information has generally been used to show similarity of the entries using grouping/classification and ordination methods. Genetic diversity can be depicted by classifying/grouping the accessions into genetic diversity trees or dendrograms. The structure of genetic diversity in the form of the tree can then guide selection of accessions to form the core. The classification methods are of two types: hierarchical (often known as *cluster analyses*) and nonhierarchical (known as *K-means cluster analysis*).

In hierarchical analysis, the accessions will be arranged into groups with similar properties and the number of groups, at a given level of similarity, is unknown in advance. This method thus reflects a more natural interrelation, or closeness, or diversity in the accessions. An algorithm for a hierarchical method requires the following:

1. A measure of similarity between two items (e.g., two accessions). Depending on the nature of the traits observed, several measures have been suggested: Euclidean distance, cityblock (also known as *Manhattan*), simple matching, ecological distance, and Jaccard’s measure.
2. A concept of treating initially all the accessions as separate clusters and fusing the two most nearby clusters into one at each of the subsequent stages (agglomerative method) or treating initially all the accessions as a single cluster and dividing it into two groups at subsequent stages (divisive methods).
3. A method of deriving similarity between two groups or clusters of items. Often-used procedures are single linkage or nearest neighbor, complete linkage or furthest neighbor, average linkage, centroid method, or a group average (Cormack 1971). The Genstat routine HCLUSTER can be used to produce the required information.

In a nonhierarchical method, accessions are divided into a given number (determined in advance) of disjoint groups such that the groups are reasonably homogeneous within and different between. Again, a number of criteria influence the algorithm for grouping. The commonly used ones are based on maximizing the between-group sum of squares; minimizing the determinant of the pooled within-class dispersion matrix; maximizing the total Mahalanobis squared distance between the groups; and choosing the maximal predictive classification (for which the Genstat routine CLUSTER could be used).

Comparison of Classifications

When there are several methods of classification, then one may wonder whether there is any natural order in the items. One may either work out some other independent grouping based on environmental and or morphological traits, or try examining whether two or more classifications give the same groupings. Comparison of two grouping methods could be done using chi-square contingency tests for independence of the two classifications produced by the two methods.

Ordination Techniques

Ordination techniques are used to order a group of objects, for example, accessions, populations in multi-dimensions (e.g., based on the response from analyses with multivariables or multimarker information) with a view to finding if there is any interrelationship among the set of objects and their association with other factors such as site or environmental factors. The ordination can be done using direct gradients of environmental factors (e.g., abundance versus soil pH) and weighted scores for objects over environments. In absence of environmental factor information, indirect methods can be used, such as principal component analysis, bi-plots, correspondence analysis, canonical variate analysis, and principal coordinate analysis, which are based on data in the form of two-way tables or matrices of objects and environments.

Principal Component Analysis

Let the $X = (x_{ij})$ be the data ~~point~~ (e.g., abundance~~)~~ for the ith object (species) from the jth variate (site), where $i = 1,\ldots, n$ and $j = 1,\ldots, p$~~.~~ The values of the variates are standardized into matrix Y to have mean of zero and unit variance. The singular value decomposition of matrix Y is given as follows:

$$Y = USV'$$

where:

U is the $n \times p$ orthonormal matrix
S is the diagonal matrix of order p
V is the transpose of the orthogonal matrix V $(p \times p)$

The diagonal elements of S are arranged in descending order to give an order of approximation to Y in terms of the matrix $A = US$, called *scores*. Another principal component analysis (PCA) approach is to find a new set of transformed variates (linear combinations of observed variates), which account more effectively for the variation among the individuals. For this, one finds the use of spectral decompositions of the sum of squares, then the product matrix in Y, that is, $\Sigma = Y'Y$, yields the eigenvalues and eigenvectors (loadings) of Σ, arranged in descending order. Such an ordering assigns the associated vectors (scores as linear combinations of Y values and loadings) as principal components (PCs) and the associated eigenvalue gives the variation attributable to that PC. PCs can be used to display the objects in two dimensions (e.g., the first PC versus the average abundance of species). The Genstat ~~routing~~ for PCA is PCP.

Bi-Plot

PCs can be worked out for environment or sites as well, following the above procedure. The two sets of PCs, one for the species and the other for the environments, will give bi-plots. Such bi-plots can be used to determine if a specific set of species are associated with certain sites. The Genstat library procedure for this is BIPLOT.

Correspondence Analysis

Considering the species-by-site data as row-columns, one can obtain iteratively the scores for species and sites using the direct gradient method. The transformed data $Y\,(= y_{ij})$ are obtained from the observed frequency data $X\,(= x_{ij})$ after correcting for the proportional model. Thus, $y_{ij} = x_{ij} - x_{i.}x_{.j}/x_{..}$ (where the dots indicate totals over that suffix set), which can be used to calculate the scores for species and for sites:

$$a_i = \left(\frac{\rho^{-1} \sum y_{ij} b_j}{x_{i.}} \right) \text{and } b_j = \left(\frac{\rho^{-1} \sum y_{ij} a_i}{x_{.j}} \right)$$

AQ 6

where ρ is a constant to keep the score within range.

This is called *reciprocal averaging* (Digby and Kempton 1987). The correspondence analysis generalizes the reciprocal averaging approach in two dimensions, using matrix algebra tools and the results of spectral decomposition. The correspondence analysis was also approached using concepts in mechanics (Benzecri 1973, Greenacre 1984) leading to similar results. The Genstat library procedure is CORRESP.

Principal Coordinate Analysis

Principal coordinate analysis (PCO), unlike PCA and other methods, uses an $n \times n$ (e.g., number of species) symmetric matrix of associations, similarity, or distances (Gower 1966). Then PCO is employed on such a matrix to give principal coordinate scores. The Genstat ~~routing~~ is PCO.

Canonical Variate Analysis

When the species or sites are grouped (or have some structure, perhaps indicating similarity), canonical variate analysis (CVA) could be used to validate the existing groupings or to assign the membership to a new species in one of the groups. Using the $n \times p$ data from n species and p sites (variates), one can obtain within-group and between-group sums of squares and product matrices. CVA, also known as *linear discriminant analysis*, provides linear functions to maximize the ratio of between-group to within-group variation. The Genstat ~~routing~~ for this analysis is CVA.

Canonical Correlation Analysis

Consider the case where a set of variables can be divided into two groups (e.g., abundance of species as one set of variables and environmental factors as the second set, observed over sites). Canonical correlation analysis provides a linear combination of variables in the first set (species abundance) and another linear combination of variables in the second set (environmental factors) such that the correlation between the two variables generated by the two linear combinations is maximal. Similarly, a second pair of linear combinations could be generated to give the next maximal correlation. Thus, the linear combinations of the variables could be used for prediction. The GENSTAT library procedure CANCOR can be used to generate these linear combinations.

Methods ~~of~~ Comparing Ordinations

Procrustes rotation was named after the innkeeper in Greek mythology who used to match the guest to the bed by adjusting the limbs of the guests. If there are more than one ordination of the same set of objects, the Procrustes rotation is used to determine the consistency between them. For example, one may be interested in comparing two ordinations for the same set of sites, perhaps one based on species abundance and the other based on environmental factors. Here, one of the two sets of coordinates of the *n* points in *r*-dimension is treated as a fixed configuration (the *X*-matrix), while the other configuration (the *Y*-matrix) is shifted and rotated to best match with *X*-matrix. A measure of goodness of fit is produced as the residual sum of squares. Its generalization to more than two ordinations is called the generalized Procrustes rotation (Gower 1975, 1985). The Genstat ~~routine~~ ROTATE and library procedure GENPROC can be used.

Spatial Pattern of the Species

A feature of species may be that their abundance is associated with specific locations. Spatial distribution of the abundance of species with site references can be used to predict the abundance at a site where data were not collected. Spatial models use the stochastic behavior of the variable over space (Cressie 1993), in contrast to the classical approaches. The observations of the variable being modeled are assumed to be independent and randomly distributed, and predictions are made on means irrespective of the location. The modeling requires the concept of variation with distance, a variogram, and its parameters such as nugget (micro-scale variation), sill (the maximum variation between any two points), and range (distance within which there is variation in the variogram) beyond which variance does not depend on distance. Geo-statistical programs are used. Genstat ~~routing~~ FVARIOGRAM and library procedures MVARIOGRAM and KRIGE could be used.

GERMPLASM DATABASES, BIOINFORMATICS, AND SOFTWARE

Information on collection missions, environment, sites of collection, and GIS maps/spatial tables of environmental variables, accession-specific data, and morphological and molecular data are valuable resources (e.g., the genetic resource database at

ICARDA). With advances in biotechnology, the data at the molecular levels are being electronically stored with volume of the database growing with time. Bioinformatics, a system to store, retrieve, manipulate, and interpret genomic data, is now being recognized as a very useful and active discipline.

REFERENCES

AQ 7

Bari, A. and M. Singh. 1998. Analysis of plant genetic resources data. Lecture notes for a training course on *Documentation and information management of plant genetic resources, 22 Nov–3 Dec 1998*. Aleppo, Syria: ICARDA.

Bennett, E. 1970. Tactics of plant exploration. In *Genetic resources in plants—Their exploration and conservation*, ed. O.H. Frankel and E. Bennett, 157–179. Oxford: Blackwell.

Benzecri, P.J. 1973. *L'analyse des donnees*, *Vol 2. L'analyse des correspondances*. Paris, France: Dunod.

Bowman, K.O., K. Hutcheson, E.P. Odum, and L.R. Shenton. 1971. Comments on the distribution of indices of diversity. In *Statistical ecology, Vol. 3. Many species populations, ecosystems, and systems analysis*, ed. G.P. Patil, E.C. Pielou, and W.E. Walters, 315–336. University Park, PA: Pennsylvania State University Press.

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

Brown, A.H.D. 1992. Human impact on plant gene pools and sampling for their conservation. *Oikos* 63:109–118.

Brown, A.H.D. 1995. The core collection at the crossroads. In *Core collections of plant genetic resources*, ed. T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and E.A.V. Morales, 3–19. New York: John Wiley & Sons.

Brown, A.H.D. and J.D. Briggs. 1991. Sampling strategies for genetic variation in ex situ collections of endangered plant species. In *Genetics and conservation of rare plants*, ed. D.A. Falk and K.E. Holsinger, 99–122. New York: Oxford University Press.

Brown, A.H.D. and D.R. Marshall 1995. A basic sampling strategy: Theory and practice. In *Collecting plant genetic diversity: Technical guidelines*, ed. L. Guarino, V. Ramanatha Rao, and R. Reid, 76–91. Wallingford: CAB International.

Buckland, S.T., D.R. Anderson, K.P. Burnham, and J.L. Laake. 1993. *Distance sampling: Estimating abundance of biological populations*. London: Chapman & Hall.

CBD. 1992. *Convention on biological diversity*. https://www.cbd.int/doc/legal/cbd-en.pdf

Chang, T.T. 1985. Collection of crop germplasm. *Iowa State J Res* 59:349–364.

Clifford, H.T. and W. Stephenson. 1975. An introduction to numerical classification. New York: Academic Press.

Cochran, W.G. 1977. *Sampling techniques*, 3rd ed. New York: John Wiley & Sons.

Cormack, R.M. 1971. A review of classification. *J Roy Stat Soc A Sta* 134:321–367.

Cressie, N.A.C. 1993. *Statistics for spatial data*, Rev. ed. New York: John Wiley & Sons.

Damania, A.B. 1987. Sampling cereal diversity in Morocco. *Plant Genetic Resources Newsletter* 72:29–30.

Digby, P.G.N. and R.A. Kempton. 1987. *Multivariate analysis of ecological communities*. London: Chapman & Hall.

Efron, B. and R.J. Tibshirani. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

Engels, J.M.M., R.K. Arora, and L. Guarino. 1995. An introduction to plant germplasm exploration and collecting: Planning, methods and procedures, follow-up. In *Collecting plant genetic diversity: Technical guidelines*, ed. L. Guarino, V. Ramanatha Rao, and R. Reid, 31–63. Wallingford: CAB International.

Engen, S. and C. Taillie. 1979. A basic development of abundance models: Community description. In *Statistical distributions in ecological work*, ed. J.K. Ord, G.P. Patil, and C. Taillie, 289–311. Fairland, MD: International Co-operative Publishing House.

Frankel, O.H. 1984. Genetic perspectives of germplasm conservation. In *Genetic manipulation: Impact on man and society*, ed. W. Arber, K. Illmensee, W.J. Peacock, and P. Starlinger, 161–169. Cambridge: Cambridge University Press.

Frankel, O.H. and A.H.D. Brown. 1984. Current plant genetic resources: A critical appraisal. In *Genetics: New frontiers, Vol. 4. Applied genetics. Proceedings of the International Congress of Genetics, 15th, New Delhi, India*. December 12–21, 1983, ed. V.L. Chopra, B.C. Joshi, R.P. Sharma, and H.C. Bansal, 1–11. New Delhi, India: Oxford and IBH.

Guarino, L., V. Ramanatha Rao, and R. Reid (eds.). 1995. *Collecting plant genetic diversity: Technical guidelines*. Wallingford: CAB International.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53:325–328.

Gower, J.C. 1975. General procrustes analysis. *Psychometrika* 40:33–51.

Gower, J.C. 1985. Measures of similarity, dissimilarity and distance. In *Encyclopaedia of Statistics, Vol. 5*, ed. N.L. Johnson, S. Kotz, and C.B. Read, 397–405. New York: John Wiley & Sons.

Greenacre, M.J. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Hooper, D.U., E.C. Adair, B.J. Cardinale, et al. 2012. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature* 486:105–108.

Hurlbert, S.H. 1971. The nonconcept of species diversity: A critique and alternative parameters. *Ecology* 52(4):577–586.

Hutcheson, K. 1970. A test for comparing diversities based on the Shannon formula. *J Theor Biol* 29:151–154.

Kameswara Rao, N. and P. J. Bramel (eds.). 2000. *Manual of genebank operations and procedures*. Technical Manual no. 6. India: ICRISAT.

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

Kindt, R. and R. Coe. 2005. *Tree diversity analysis: A manual and software for common statistical methods for ecological and biodiversity studies*. http://www.worldagroforestry.org/resources/databases/tree-diversity-analysis.

Magurran, A.E. 1988. *Ecological diversity and its measurement*. Kent: Croom Helm.

Marshall, D.R. and A.H.D. Brown. 1975. Optimum sampling strategies in genetic conservation. In *Crop genetic resources for today and tomorrow*, ed. O.H. Frankel and J.G. Hawkes, 53–80. Cambridge: Cambridge University Press.

May, R.M. 1975. Patterns of species abundance and diversity. In *Ecology and evolution of communities*, ed. M.L. Cody and J.M. Diamond. 81–120. Cambridge, MA: Harvard University Press.

Nabhan, G.P. 1990. Wild *Phaseolus* Ecogeography in the Sierra Madre Occidental, Mexico. In *Systematic and Ecogeographic Studies on Crop Genepools 5*. Rome, Italy: IBPGR.

Payne, R.W. (ed.). 2014. *The guide to GenStat® Release 17. Part 2: Statistics*. Hemel Hempstead: VSN International.

Porceddu, E. and A.B. Damania. 1991. *Sampling strategies for conserving variability of crop genetic resources in seed crops*. Technical Manual No. 17. Aleppo, Syria: ICARDA.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0 http://www.R-project.org.

Rao, C.R. 1973. *Linear statistical inference and its applications*. New York: John Wiley & Sons.

Schoen, D.J. and A.H.D. Brown. 1995. Maximizing genetic diversity in core collections of wild relatives of crop species. In *Core collections of plant genetic resources*, ed. T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and E.A.V. Morales, 55–76. New York: John Wiley & Sons.

Sedcole, J.R. 1977. Number of plants necessary to recover a trait. *Crop Sci* 17:667–668.

Sukhatme, P.V., B.V. Sukhatme, S. Sukhatme, and C. Asok. 1984. *Sampling theory of surveys with applications*, 3rd ed. Ames, IA: Iowa State University Press.

UNCED. 1992. *Agenda 21*. United Nations Conference on Environment and Development, June 3–14, 1992. http://www.unep.org/Documents.Multilingual/Default.asp?documentid=52.

Valkoun, J. and A.B. Damania. 1990. *Report on the germplasm collecting mission to Tibet (Autonomous Region of China)* (unpublished). Aleppo, Syria: ICARDA.

Weir, B.S. 1990. *Genetic data analysis*. Sunderland, MA: Sinauer Associate.

Whittaker, R.H. 1977. Evolution of species diversity in land communities. *Evol Biol* 10:1–67.

WRI, IUCN, and UNEP. 1992. *Global biodiversity strategy: Guidelines for action to save, study, and use Earth's biotic wealth sustainably and equitably*. Washington, DC: WRI, IUCN and UNEP. http://pdf.wri.org/globalbiodiversitystrategy_bw.pdf.

Yonezawa, K., T. Nomura, and H. Morishima. 1995. Sampling strategies for use in stratified germplasm collections. In *Core collections of plant genetic resources*, ed. T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and E.A.V. Morales, 35–53. New York: John Wiley & Sons.

Author Query Sheet

Chapter No.: 15

Query No.	Queries	Response
AQ 1	Please confirm whether the equation as set is OK.	
AQ 2	Please confirm whether the equation could be set as follows: $\mathrm{Var}\left(H_1\right)=\left\{\frac{\left[\sum p_i\times\ln\left(p_i\right)\right]^2-\left[\sum p_i\times\ln\left(p_i\right)\right]^2}{N}\right\}-\left[\frac{(S-1)}{\left(2N^2\right)}\right]$	
AQ 3	Please confirm whether the equation is OK as edited.	
AQ 4	If available, please provide a page number for the quote.	
AQ 5	The name of citation “Yonezama et al. (1995)” has been changed to “Yonezawa” to match with reference list. Please check and confirm if this is OK.	
AQ 6	Please confirm whether the equation as edited is OK.	
AQ 7	Heading “Literature cited” has been changed to “References”. Please check if this is OK.	
AQ 8	Please provide missing editor name for reference “Nabhan (1990)”.	