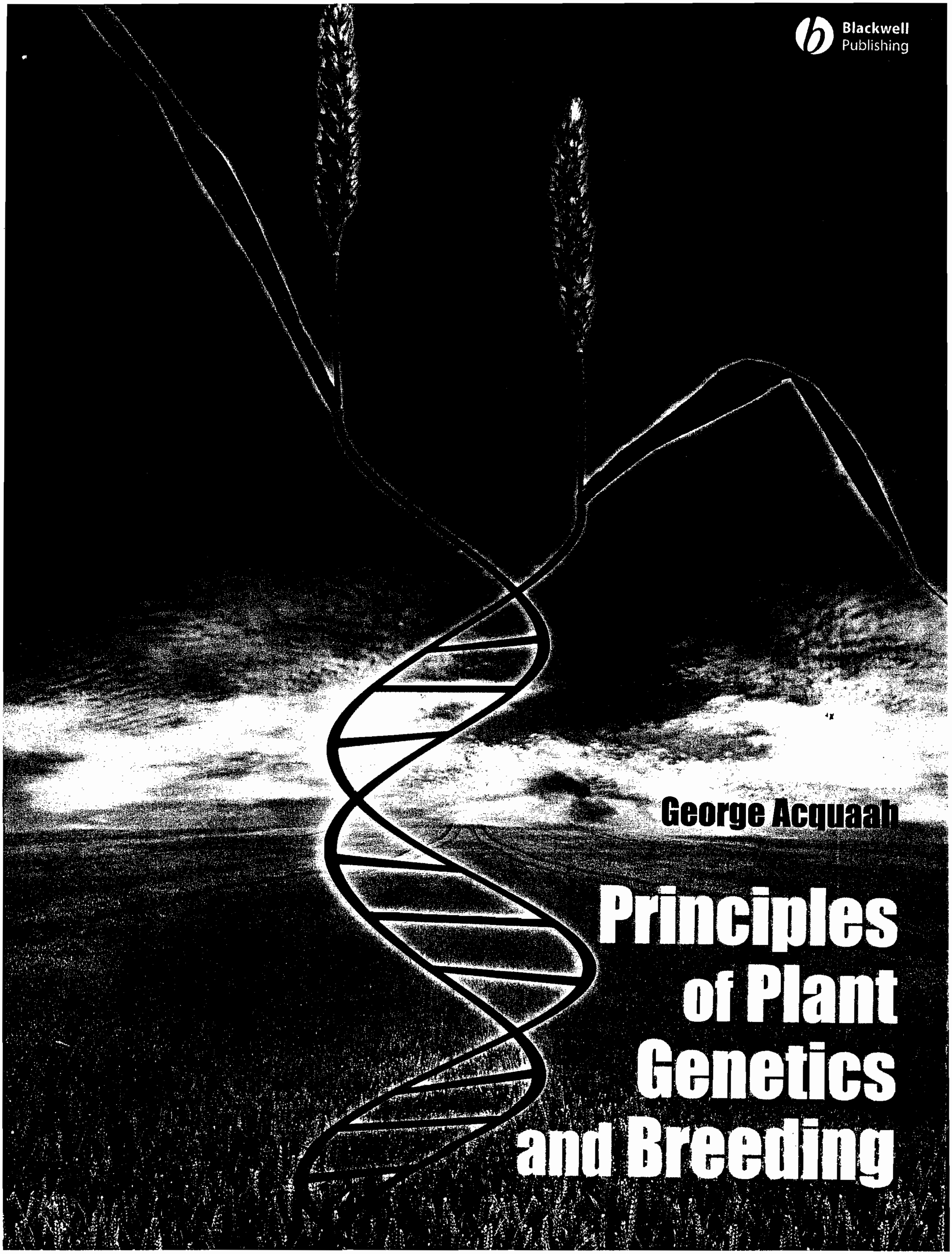


Blackwell
Publishing

George Acquaah

Principles of Plant Genetics and Breeding

Industry highlights boxes

Chapter 1
Normal Ernest Borlaug: the man and his passion
George Acquaah

Chapter 2
Introduction and adaptation of new crops
Jaime Prohens, Adrián Rodríguez-Burruezo, and Fernando Nuez

Chapter 3
No box

Chapter 4
Maize × *Tripsacum* hybridization and the transfer of apomixis: historical review
Bryan Kindiger

Chapter 5
No box

Chapter 6
Plant genetic resources for breeding
K. Hammer, F. Heuser, K. Khoshbakht, and Y. Teklu

Chapter 7
No box

Chapter 8
Recurrent selection with soybean
Joe W. Burton

Chapter 9
Multivariate analyses procedures: applications in plant breeding, genetics, and agronomy
A. A. Jaradat

Chapter 10
The use of the wild potato species, *Solanum etuberosum*, in developing virus- and insect-resistant potato varieties
Richard Novy

Chapter 11
Haploids and doubled haploids: their generation and application in plant breeding
Sergey Chalyk

Chapter 12
Current apple breeding programs to release apple scab-resistant scion cultivars
F. Laurens

Chapter 13
Application of tissue culture for tall wheatgrass improvement
Kanyand Matand and George Acquaah

Chapter 14
Bioinformatics for sequence and genomic data
Hugh B. Nicholas, Jr., David W. Deerfield II, and Alexander J. Ropelewski

Chapter 15
The intersection of science and policy in risk analysis of genetically engineered plants
David A. Lee and Laura E. Bartley

Chapter 16
Barley breeding in the United Kingdom
W. T. B. Thomas

Chapter 17
Developing a new cool-season perennial grass forage: interspecific hybrids of *Poa arachnifera* × *Poa secunda*
Bryan Kindiger

Chapter 18
Pioneer Hi-Bred International, Inc.: bringing seed value to the grower
Jerry Harrington

Chapter 19
Bringing Roundup Ready® technology to wheat
Sally Metz

Chapter 20
Genetic improvement of cassava through biotechnology
Nigel J. Taylor

Chapter 21
Discovering genes for drought adaptation in sorghum
Andrew Borrell, David Jordan, John Mullet, Patricia

Multivariate statistics in plant breeding

Multivariate analysis is the branch of statistics concerned with analyzing multiple measurements that have been made on one or several samples of individuals. Because these variables are interdependent among themselves, they are best considered together. Unfortunately, handling data with multicolinearity can be unwieldy and hence some meaningful summarization is needed.

The multivariate techniques in use may be divided into two groups:

1 **Interdependence models** – e.g., principal components analysis, factor analysis.
2 **Dependence models** – e.g., multivariate analysis of variance, classification functions, discriminant function analysis, cluster analysis, multiple correlation, canonical correlation.

W. W. Cooley and P. R. Jones further classified multivariate procedures into four categories according to the number of populations and the number of variables as follows:

1 One set of variables, one population – e.g., principal components analysis, factor analysis.
2 One set of variables, two or more populations – e.g., multivariate analysis of variance, discriminant functions, classification functions.
3 Two or more sets of variables, one population – e.g., polynomials fit, multiple correlation, canonical correlation, multiple partial correlation.
4 Two or more sets of variables, two or more sets of populations – e.g., multivariate covariance.

Multivariate analyses are done on computers because of their complexity. An overview of the common procedures is discussed next.

Factor analysis

A variable can be explained to the extent that its variance can be attributed to an identifiable source. **Factor analysis** may be used to find ways of identifying fundamental and meaningful dimensions of a multivariate domain. It is a decision-making model for extracting subsets of covarying variables. To do this, natural or observed intercorrelated variables are reformulated into a new set (usually fewer in number) of independent variables, such that the latter set has certain desired properties specified by the analyst. Naming factors is only a mnemonic convenience. It should be done thoughtfully so as to convey information to both the analyst and the audience. For example, a large set of morphological traits may be reduced to several conceptual factors such as "architectural factor" (loaded by variables such as internode length, number of internodes, etc.), whereas a "seed size factor" may be loaded by traits such as seed length and seed width.

Industry highlights

Multivariate analyses procedures: applications in plant breeding, genetics, and agronomy

A. A. Jaradat

USDA-ARS, Morris, 56267 MN, USA

Introduction

Plant breeders, geneticists, and agronomists are increasingly faced with theoretical and practical questions of multivariate nature. With increases in germplasm sizes, the number of plant and crop variables, and evaluation and characterization data on molecular, biochemical, morphological, and agronomic traits, multivariate statistical analysis (MVA) methods are receiving increasing interest and assuming considerable significance. Some MVAs (e.g., multivariate analysis of variance, MANOVA, and covariance, MANCOVA) are extensions of uni- and bivariate statistical methods appropriate for significance tests of statistical hypotheses. However, most MVAs are used for data exploration, the extraction of fundamental components of large data sets, the discovery of latent structural relationships, and the visualization and description of biological patterns. This review focuses on the salient features and applications of MVAs in multivariate data analyses of plant breeding, genetics, and agronomy data. These include MANOVA, MANCOVA, data reduction methods (factor, principal components, principal coordinates, perceptual mapping, and correspondence analyses), and data classification methods (discriminant analysis, clustering and additive trees).

Crop improvement programs – through breeding, selection, and agronomic evaluation – rely on available genetic diversity for specific trait(s) in the primary and, if needed, in the secondary gene pool of a particular crop species. Classic univariate analysis

Table 1 Summary of the significant effect ($P < 0.05$) for leaf area index (LAI) and dry weight of stems per plant in MANOVA and determination of the smallest set of variables.

Variable	Significant source of variation	Wilk's lambda	*F* approximation	Final set
LAI	Year	0.182	40.45***	y_{max}
	Genotype	0.175	4.73*	x_{max}
	Growth habit	0.479	26.85***	
Dry weight of stems	Year	0.552	7.30*	x_{inf}
	Genotype	0.067	13.99**	x_{inf}
	Growth habit	0.480	70.17***	
	Within winter types	0.105	13.12***	

y_{max}, maximum value of the variable; x_{max}, time in growing-degree days from sowing to y_{max}; x_{inf} time in growing degree-days from sowing to reach maximum rate of growth. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

procedures, limited to estimation and hypotheses testing, are not capable of detecting patterns and exploring multivariate data structures in genetic resources, breeding lines, or cultivars. Therefore, MVA methods to classify and order large numbers of breeding material, trait combinations, and genetic variation are gaining considerable importance and assuming considerable significance.

MANOVA and MANCOVA

MANOVA and MANCOVA perform a multivariate analysis of variance or covariance when multiple dependent variables are specified. MANOVA tests whether mean differences among groups for a combination of dependent variables are likely to have occurred by chance. A new dependent variable that maximizes group differences is created from the set of dependent variables. The new dependent variable is a linear combination of measured dependent variables, combined so as to separate the groups as much as possible. ANOVA is then performed on the newly created dependent variable. MANCOVA asks if there are statistically reliable mean differences among groups after adjusting the newly created dependent variable for differences on one or more covariates. In this case, variance associated with the covariate(s) is removed from error variance; smaller error variance provides a more powerful test of mean differences among groups.

MANOVA was used in the analysis of growth patterns and biomass partitioning of crop plants as a prerequisite for interpreting results of field experiments and in developing crop simulation models. Royo and Blanco (1999) utilized MANOVA to compare non-linear regression growth curves in spring and winter triticale and identified variables responsible for the differences between these curves. Results of these studies are partially presented in Table 1, along with the smallest set of variables required to characterize the growth curves. Wilk's lambda is the criteria for statistical inference and is estimated as the pooled ratio of error variance to effect variance plus error variance. In this example, all Wilk's lambda and *F*-approximation estimates are significant. For example, the differences within each growth habit (Table 1) were non-significant but differences between growth habits were significant. Thermal time needed to reach the maximum leaf area index was the variable responsible for these differences.

Variance components analysis (VCA)

Experimentation is sometimes mistakenly thought to involve only the manipulation of levels of the independent variables and the observation of subsequent responses on the dependent variables. Independent variables whose levels are determined or set by the experimenter are said to have fixed effects. A second class of effects, random effects, are classification effects where the levels of the effects are assumed to be randomly selected from an infinite population of possible levels. Many independent variables of research interest are not fully amenable to experimental manipulation, but nevertheless can be studied by considering them to have random effects.

Factor analysis (FA) and principal components analysis (PCA)

The primary purpose of FA and PCA is to define the underlying structure in a data matrix. As data reduction or exploratory methods, these procedures are used to reduce the number of variables and to detect structure in the relationships between these variables. FA reproduces the correlation matrix among variables with a few orthogonal factors; however, contrary to PCA, most forms of FA are not unique. PCA is a procedure for finding hypothetical variables (components) that account for as much of the variance in multidimensional data as possible. PCA is a unique mathematical solution; it performs simple reduction of the data set to a few components, for plotting and clustering purposes, and can be used to hypothesize that the most important components are correlated with some other underlying variables.

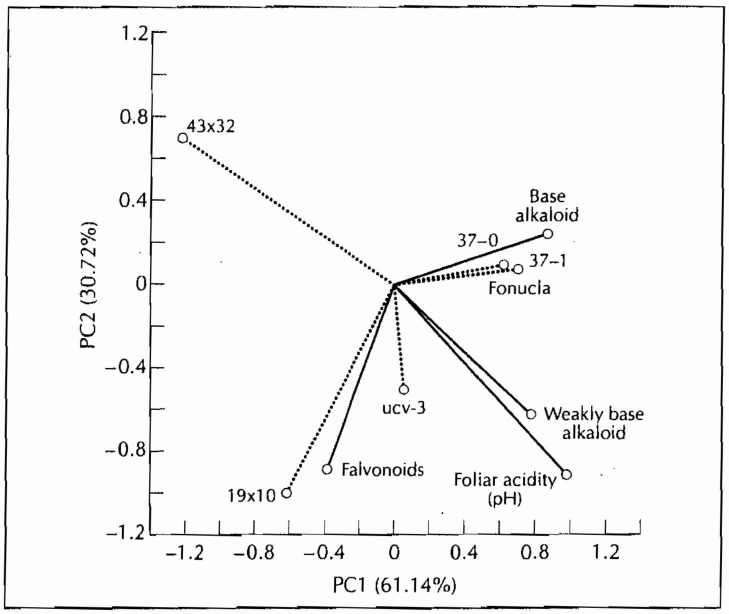


Figure 1 A graph based on PCA of five sesame genotypes as operational taxonomic units (dotted lines), and three secondary metabolites in leaves and foliar acidity as variables (solid lines).

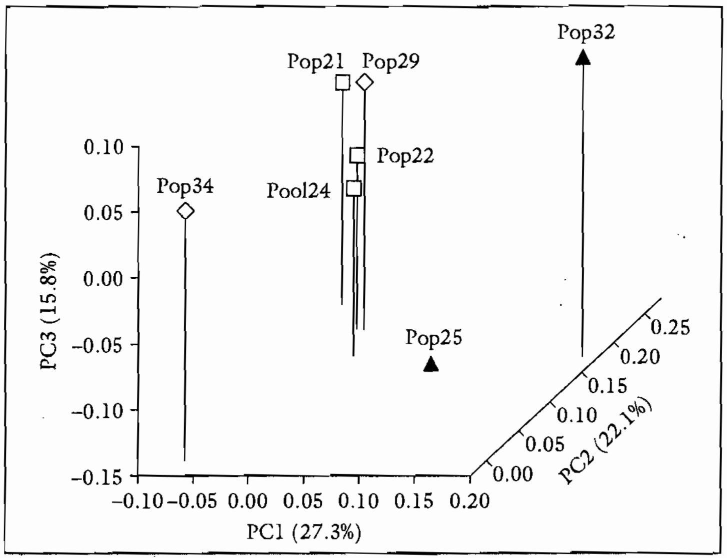


Figure 2 PCoA plot of seven tropical maize populations based on modified Roger's distance. PC1, PC2, and PC3 are the first, second, and third principal coordinates, respectively. Heterotic group A (Pop21, Pop22, and Pool24), heterotic group B (Pop25, Pop32), and populations not yet assigned to heterotic groups (Pop29, Pop34) are shown.

In PCA one can obtain a "biplot" in which the objects and the variables are superimposed on the same plot so that one can study their interrelationships (Figure 1). In PCA one judges proximities among the objects using Euclidean distances and among the variables using covariance or a correlation matrix. PCA was utilized in determining the phytochemical relationship of six sesame genotypes and their resistance to whitefly (Laurentin et al. 2003). Foliar acidity and flavonoids dominated PC1 and PC2, respectively. The five sesame genotypes were separated according to their phytochemical characteristics. A close relationship was found between secondary metabolites and foliar acidity, on the one hand, and incidence of whitefly on sesame, on the other, thus demonstrating the importancé of foliar acidity values of sesame genotypes as a resistance mechanism against whitefly.

Principal coordinates analysis (PCoA)

PCoA focuses on samples rather than variables and is based on a matrix containing the distances between all data points. A typical usage of PCoA is the reduction and interpretation of large multivariate data sets with some underlying linear structure. PCoA was instrumental in delineating relationships among tropical maize populations based on simple sequence repeats for breeding purposes (Reif et al. 2003). PCoA revealed very clear association among populations within certain heterotic groups (Figure 2). Reif et al. (2003) succeeded in identifying genetically similar germplasm based on molecular markers, and concluded that PCoA provides a more economic and solid approach for making important breeding decisions early in the breeding program.

Perceptual mapping (biplot and GE)

Success in evaluating germplasm, breeding lines, and cultivars in multiple environments and for complex traits to identify superior genotypes with specific or wide adaptation can be achieved if the genotypic (*G*) and environmental (*E*) effects and their interaction (*GE*) are precisely estimated (Yan et al. 2000). The *GE* biplot procedure has been used by breeders and agronomists for dissecting *GE* interactions and is being used to analyze data from genotype x trait, genotype x marker, environment x QTL, and diallel cross data. The biplot allows a readily visualized display of similarity and differences among environments in their differentiation of the genotypes, the similarity and differences among the genotypes in their response to locations, and the nature and magnitude of the interaction between any genotype and any location.

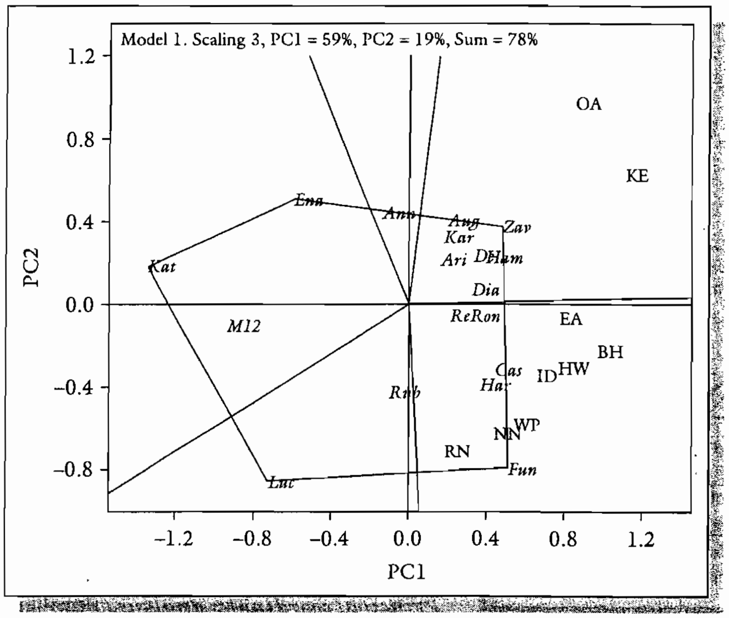


Figure 3 Biplot showing performance of different wheat cultivars (in italics) in different environments (in capital letters) as a selection method to identify superior cultivars for a target environment.

Biplot was used to compare the performance of wheat cultivars under several environments in the Ontario wheat performance trials (Figure 3) and to estimate relative variance components and their level of significance. Results of biplot analysis have several implications for future breeding and cultivar evaluation. A test for optimal adaptation can be achieved through the deployment of different cultivars for mega-environments, and the unpredictable genotype x location interaction can be avoided or minimized through cultivar evaluation and selection focusing on the main effects of genotype.

Multiple correspondence analysis (MCA)

MCA is a recently developed interdependence MVA procedure that facilitates both dimensional reduction of object ratings on a set of attributes and the perceptual mapping of objects relative to these attributes. MCA helps researchers quantify the qualitative data found in nominal variables and has the ability to accommodate both non-metric data and non-linear relationships. In order to facilitate the use of common bean landraces in genetic improvement, Beebe et al. (2000) used MCA to study the structure of genetic diversity, based on RAPD (random amplified polymorphic DNA), among common bean landraces of Middle American origin for breeding purposes. MCA results (Figure 4) indicated that the Middle American bean germplasm is more complex than previously thought with certain regions holding important genetic diversity that has yet to be properly explored for breeding purposes. The first dimension

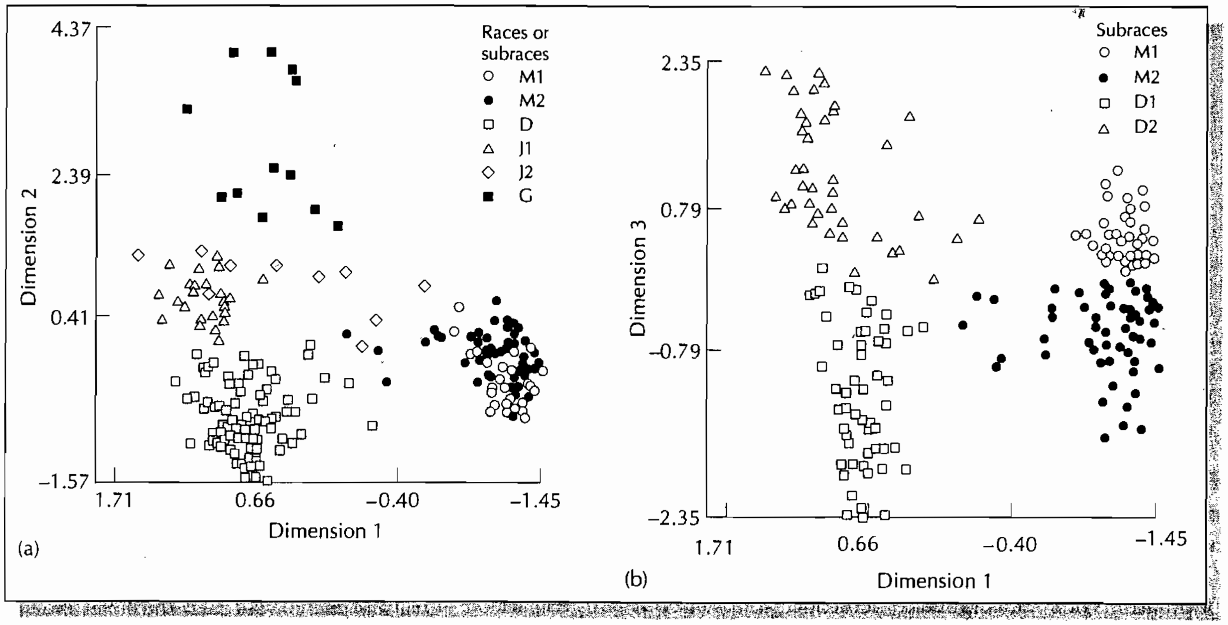


Figure 4 Plot of (a) 250 Middle American bean genotypes in dimension 1 and 2, and (b) 206 genotypes of two races in dimensions 1 and 3 of MCA based on RAPD data.

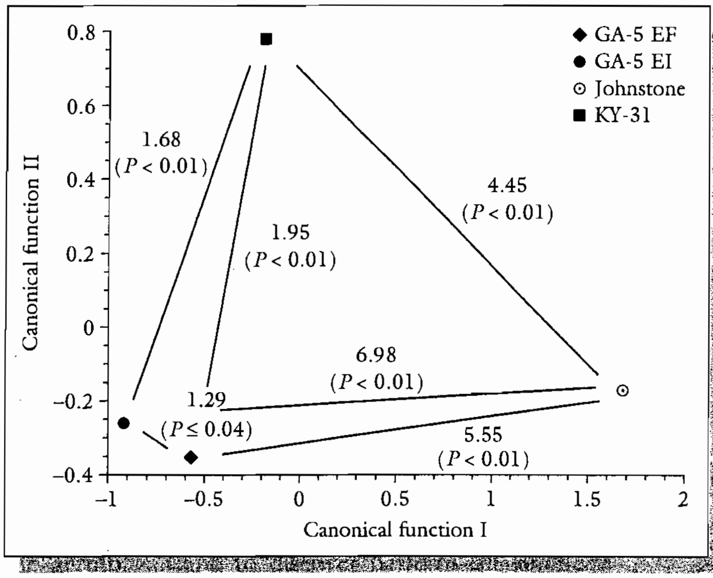


Figure 5 Scatterplot of centroid values of four tall fescue cultivars on two canonical discriminant functions. Mahalanobis distances and their probability values, in parentheses, measure the extent of genetic diversity between the four cultivars.

(Figure 4a) discriminated between lowland and highland races. The second dimension discriminated among highland races, whereas the third dimension (Figure 4b) divided the highland races according to their growth habit, geographic, distribution, and seed type. Results of MCA can be used to orient plant breeders in their search for distinct genes that can be recombined, thus contributing to higher genetic gain.

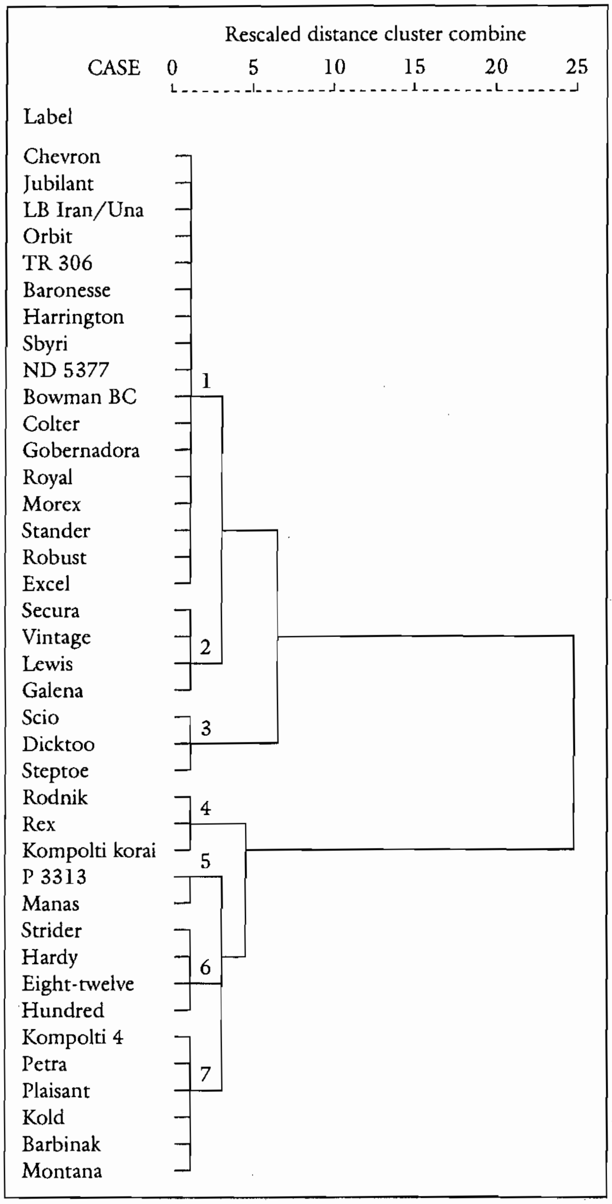
Canonical discriminant analysis (CDA)

CDA is used to study the variation among two or more groups (samples) of crop cultivars relative to the average variation found within the groups. Linear combinations of the original variables that account for as much as possible of total variation in the data set are constructed using PCA, then canonical correlation is used to determine a linear association between predictor variables identified in PCA and criterion measures. In CDA more distinct differentiation of cultivars is achieved as compared with univariate analysis, since all independent variables (e.g., traits) are considered simultaneously in the process. CDA can separate "among population" effects from "within population" effects thus maximizing the overall heritability estimates of canonical variates by placing very large weight on traits with low levels of environmental variability. CDA uses Mahalanobis distance to differentiate between cultivars or populations. The higher the canonical loadings (measures of the simple linear correlation between an original independent variable and the canonical variate) of traits of particular significance, the higher the genetic variation as compared with traits having low canonical loadings. Plant breeders can use this information to focus on particular trait(s) for genetic improvement of a particular crop. Vaylay and van Santen (2002) employed CDA in the assessment of genetic variation in tall fescue (Figure 5). They found that the genetic composition of four tall fescue cultivars differ mainly, in decreasing order, in maturity, cell wall content, flag leaf length, tiller number, and dry matter yield. Therefore, tall fescue breeders can concentrate on the most important traits of this perennial pasture crop knowing that the genetic composition of its cultivars changes with time.

Cluster analysis (CA)

CA is an analytical MVA procedure for developing meaningful subgroups of objects. It classifies a sample of objects into a small number of mutually exclusive groups based on the similarities among the objects. Stepwise clustering involves a combination or division of objects into clusters. Hierarchical CA starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. The divisive clustering method begins with all the objects in a single cluster, which is then divided at each step into two clusters that contain the most dissimilar objects. Additive trees, as an extension of clustering, are based on a dissimilarity distance matrix among all possible pairs of objects in order to retain the original distances among all pairs of these objects. Unlike other clustering algorithms that are based on the rigorous ultrametric relationships between objects, the additive tree precisely reflects distances among the objects.

Cluster analysis was used as a tool to optimize and accelerate barley breeding. Karsai et al. (2000) evaluated barley cultivars for five physiological and agronomic traits that have significant effects on heading date and winter hardiness. CA helped identify groups of cultivars representing different adaptational types. The wide level of diversity identified in the germplasm set was valuable in studying the genetics of adaptation to certain environments. It was possible to identify (numbered 1 through 7 in Figure 6) winter and spring groups, groups of cultivars with no vernalization response that had the lowest earliness *per se*, and other group of cultivars least sensitive to changes in photoperiod but with a strong vernalization response. A breeding scheme was designed on the basis of the clustering results (Figure 6) and was aimed at developing new cultivars better adapted to a given environment.



References

Beebe, S., P.W. Skroch, J. Tohme, M.C. Duque, F. Pedraza, and J. Nienhuis. 2000. Structure of genetic diversity among common bean landraces of Mesoamerican origin based on correspondence analysis of RAPD. Crop Sci. 40:264–273.

Karsai, I., K. Meszaros, L. Lang, P.M. Hayes, and Z. Bedo. 2000. Multivariate analysis of traits determining adaptation in cultivated barley. Plant Breed. 120:21–222.

Laurentin, H., C. Pereira, and M. Sanabria. 2003. Phytochemical characterization of six sesame (*Sesamum indicum* L.) genotypes and their relationships with resistance against the sweetpotato whitefly *Bemisia tabaci* Gennadius. Agron. J. 95:1577–1582.

Reif, J.C., A.E. Melchinger, X.C. Xia, et al. 2003. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43:1275–1282.

Royo, C., and R. Blanco. 1999. Growth analysis of five spring and five winter triticale genotypes. Agron. J. 91:305–311.

Vaylay, R., and E. van Santen. 2002. Application of canonical discriminant analysis for the assessment of genetic variation in tall fescue. Crop Sci. 42:534–539.

Yan, W., L.A. Hunt, Q. Sheng, and Z. Szlavincs. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Sci. 40:597–605.

Figure 6 Cluster analysis of 39 barley cultivars based on a matrix of vernalization response, photoperiod sensitivity, earliness *per se*, frost tolerance at −10 and −13°C, and heading dates under different photoperiod regimes. The dendrogram was created using the Ward minimum variance method. Groups (1–7) were characterized by having specific levels of one or more traits.

Principal components analysis

Principal components analysis (PCA) reduces the dimensions of multivariate data by removing intercorrelations among the traits being studied and thereby enabling multidimentional relationships to be plotted on two or three principal axes. PCA reduces the number of variables to be used for prediction and description. By examining a set of 15 quality traits, researchers at Michigan State University Bean Breeding Program were able to ascertain that certain quality traits (dry characteristics, soaking characteristics, cooking characteristics) of dry beans were independent. This prompted the researchers to suggest a tandem selection procedure to be followed by the construction of selection indices for their breeding program.