Hierarchical Multiple-Factor Analysis for Classifying Genotypes Based on Phenotypic and Genetic Data

Jorge Franco, José Crossa,* and Santosh Desphande

ABSTRACT

A numerical classification problem encountered by breeders and gene-bank curators is how to partition the original heterogeneous population of genotypes into non-overlapping homogeneous subpopulations. The measure of distance that may be defined depends on the type of variables measured (i.e., continuous and/or discrete). The key points are whether and how a distance may be defined using all types of variables to achieve effective classification. The objective of this research was to propose an approach that combines the use of hierarchical multiple-factor analysis (HMFA) and the two-stage Ward Modified Location Model (Ward-MLM) classification strategy that allows (i) combining different types of phenotypic and genetic data simultaneously; (ii) balancing out the effects of the different phenotypic, genetic, continuous, and discrete variables; and (iii) measuring the contribution of each original variable to the new principal axes (PAs). Of the two strategies applied for developing PA scores to be used for clustering genotypes, the strategy that used the first few PA scores to which phenotypic and genetic variables each contributed 50% (i.e., a balanced contribution) formed better groups than those formed by the strategy that used a large number of PA scores explaining 95% of total variability. Phenotypic variables account for much variability in the initial PA; then their contributions decrease. The importance of genetic variables increases in later PAs. Results showed that various phenotypic and genetic variables made important contributions to the new PA. The HMFA uses all phenotypic and genetic variables simultaneously and, in conjunction with the Ward-MLM method, it offers an effective unifying approach for the classification of breeding genotypes into homogeneous groups and for the formation of core subsets for genetic resource conservation.

J. Franco, Facultad de Agronomía, Universidad de la República del Uruguay, Av. Garzón 780 CP 12900, Montevideo, Uruguay (present address: International Institute for Tropical Agriculture [IITA], Oyo Rd., PMB 5320, Ibadan, Nigeria); J. Crossa, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, Mexico D.F., Mexico; S. Desphande, International Crop Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, 502324, India. Received 30 Jan. 2009. *Corresponding author (j.crossa@cgiar.org).

Abbreviations: CA, correspondence analysis; DArT, diversity array technology; HMFA, hierarchical multiple-factor analysis; MCA, multiple correspondence analysis; MFA, multiple factor analysis; MLM, Modified Location Model; PA, principal axis; PCA, principal component analysis; QTL, quantitative trait locus; SNP, single nucleotide polymorphism; SSR, simple sequence repeat.

Breeders and gene-bank curators face a typical multivariate problem when classifying genotypes or gene-bank accessions for forming core subsets, studying genetic diversity, classifying landraces, and grouping genotypes for specific environmental conditions, among other tasks. In these cases, the researcher typically has a set of n genotypes (or gene-bank accessions) on which p attributes (traits or variables) have been measured. A numerical classification problem arises when the researcher attempts to partition the genotypes (or gene-bank accessions) into homogeneous, non-overlapping groups of different sizes and use all the available information (i.e., phenotypic and genetic data) with the aim of grouping breeding genotypes into, for example, different maturity and grain yield clusters and/or when attempting to form core subsets for genetic resource conservation. The problem is how to partition the original heterogeneous population into homogeneous subpopulations. A set of n observations, each

Published in Crop Sci. 50:105–117 (2010).
doi: 10.2135/cropsci2009.01.0053

677 S. Segoe Rd., Madison, WI 53711 USA

one with p variables, may be represented by a $\mathbf{Y}_{n\times p}$ matrix formed by n row vectors of size $1 \times p$. This representation is related to a p-dimensional Euclidean space in which each individual is represented by a point. If the set is structured into groups, the cloud of points will display high- and low-density areas; the "natural groups" may then be defined as high-density areas, separated by others having low density.

Because the aim of classification is to identify homogeneous groups of individuals, it is necessary to define and calculate the distance (dissimilarity) between individuals or between groups of individuals. The aim of ordination is the same as that of classification, except that a similarity measure (one minus dissimilarity) between individuals or groups is used. The measure of distance that may be defined between any two individuals or groups depends on the type of variables measured (the word "variable" is used interchangeably with "trait" or "attribute"; all refer to phenotypic or genotypic characteristics measured on different genotypes, possibly under different environmental conditions). The important point is to define a distance using all types of variables to achieve an effective classification. Researchers usually measure several different types of attributes on each individual, and variables may be classified as phenotypic (i.e., agro-morphological) or genetic (i.e., molecular markers). Phenotypic variables can be classified into continuous and discrete variables, whereas genetic variables can be classified into molecular markers that are characterized solely by whether they are present or absent, or into markers characterized by their allele frequencies. The type of variable determines the kind of distance measure used, and the numerical classification and ordination that can be performed. When all variables are continuous, the most commonly used distances between two individuals (observations) are the Euclidean distance and the Manhattan distance. The issue is that classification and ordination multivariate techniques should be based on a mixture of different types of variables and therefore an appropriate distance measure is required.

When a mixture of variables is used for classification, Gower (1971) proposed balancing the effect of continuous and categorical variables for calculating the distance between two observations that have continuous and discrete variables measured simultaneously. Wishart (1986) generalized Gower's distance to be used with geometric clustering methods. Podani (1999) extended Gower's distance to ordinal variables. However, for a mixture of phenotypic and genetic data, different types of distance measures must be employed, and various classification and ordination methods can be applied. For example, for a set of quantitative variables, principal component analysis (PCA) is used as ordination. Correspondence analysis (CA) is used for frequency variables and multiple correspondence analysis (MCA) is used for a set of categorical variables. In general, results have shown that groups formed based on continuous and categorical variables separately achieve very poor consensus for clustering genotypes.

An important issue when classifying individuals is how different phenotypic and genetic variables influence classification and/or ordination. Furthermore, how much do these variables contribute to classification? When using a mixture of phenotypic and genetic data in plant breeding and genetic resource conservation, different variables have different effects on the classification and ordination of individuals. Therefore, there is a need for constructing global multiple tables that place all the different types of variables in a common background and balance their influence on the classification and/or ordination of individuals.

Multiple factor analysis (Escofier and Pagès, 1994, 1988–1998; Pagès, 2002) standardizes the results of PCA and MCA for continuous and categorical variables, and balances their influence on classification. When a nested structure of groups and subgroups of variables is present, it permits the hierarchical multiple-factor analysis (HMFA) proposed by Le Dien and Pagès (2003) as an extension of MFA. The HMFA is useful for combining multiple tables of quantitative and categorical variables and for finding common ground for balancing the different effects of all the different types of variables by generating a common, nonstandardized principal axes analysis as a step before clustering.

The underlying idea of applying the two-stage clustering Ward Modified Location Model (MLM) is that the initial groups are formed based on a geometric technique (such as Ward minimum variance within groups) that includes all continuous and discrete variables. A mixture of distribution models, the MLM then acts on the previous cluster (Franco et al., 1998, 2002; Franco and Crossa, 2002). Franco et al. (2001) applied the Ward-MLM approach for classifying genotypes using phenotypic and genetic information and found relevant marker information with available morpho-agronomic attributes that form compact and well-differentiated groups.

The main objective of this research was to propose an approach that combines the use of HMFA and the two-stage Ward-MLM classification strategy for classifying genotypes and/or gene-bank accessions while balancing the effects of many different types of continuous and categorical phenotypic variables, as well as different types of molecular marker information. The HMFA results given by the principal axes contributed by the different types of variables are then used as input data in the two-stage Ward-MLM strategy for classifying observations and forming core subsets. Four data sets are used to illustrate the use of this approach.

MATERIALS AND METHODS

Data Sets

To illustrate the use of the proposed method for grouping breeding genotypes and/or classifying gene-bank accessions for genetic

resource conservation and forming core subsets, four different data sets were employed with the purpose of covering most scenarios a researcher would encounter. We included data sets with phenotypic variables that comprise continuous and categorical attributes, and genetic data with molecular markers measured for their presence or absence, or their frequencies. The four data sets contain morpho-agronomic field information with continuous and categorical variables as well as three different kinds of molecular markers: diversity array technology (DArT), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs). The DArT and SNP markers are typically binary variables (0 = absence of the marker and 1 = presence of the marker), whereas SSR markers for a diploid species may have three values, 0 (absence of the marker, mm), 0.5 (marker with a frequency of 0.5, Mm) and 1 (marker with a frequency of 1, MM).

Wheat-DArT Data Set

This wheat (*Triticum aestivum* L.) data set contains 46 entries, 12 continuous phenotypic attributes, and 75 DArT molecular markers. The 12 continuous variables were percentage of leaf rust (LR; caused by *Puccinia triticina* Eriks.), stem rust (SR; caused by *P. graminis* Pers.:Pers. f. sp. *tritici* Eriks. and E. Henn.), and 1000-grain weight (GW), measured in four different environments. The DArT markers have two values: 0 = absence and 1 = presence.

Wheat-SSR Data Set

This data set contains the same morphological data as the Wheat-DArT data set, with continuous variables LR, SR, and GW measured in four environments, but the genetic variables correspond to frequencies of 12 SSR markers, for 49 alleles. Because the genotypes are wheat lines, the SSRs only have values of 0 and 1; however, alleles are grouped per marker.

Sorghum-SSR Data Set

This data set for sorghum [*Sorghum bicolor* (L.) Moench] contains 90 entries involving six morphological variables, three continuous variables (time to 50% flowering counted in days, DF; plant height, PH; and percent glume cover, PGC), and three discrete, nominal scale variables (panicle type, PT; glume color, GlC; and grain color, GrC). Marker data include the frequencies of 46 SSR markers totaling 336 alleles. The SSR allele frequencies have only three values (0, 0.5, and 1).

Simulated Maize Data Set

A simulated maize (*Zea mays* L.) data set containing 200 entries and five morphological, continuous variables (days to silking, DS; days to anthesis, DA; ear height, EH; plant height, PH; and grain yield, GY) were measured in three environments. The five traits and three environments were combined into 15 variables. In addition, 257 SNP markers covering 10 chromosomes were simulated. The original purpose of this simulation was to generate a doubled haploid maize population that was phenotyped for various traits in different environments with the objective of assessing the performance of different selection indices that include phenotypic and genotypic information (J. Jesús Cerón-Rojas, personal communication, 2008). The simulator system used on the maize data was developed by Wang et al. (2004) and has two engines, QU-GENE and QuCim, which require different input data. To simulate a population, the input file for QU-GENE should contain the genetic structure of the genotypes for each specific trait, for example, number of genes (or quantitative trait loci [QTLs]); gene effect for each trait, including additivity, dominance, and epistasis; linkage among the genes in one chromosome; and trait heritability. Components of QU-GENE can generate genotypes making up populations of cross-pollinated or self-pollinated species, or create different environmental conditions in which the simulated genotypes will be evaluated. On the other hand, the input file for QuCim must have the type of crosses and selection method to be used in each breeding strategy.

Original data on the five traits mentioned above represented an actual doubled-haploid, maize QTL mapping population made up of 236 genotypes; QTLs for all five traits were mapped (Cerón-Rojas et al., 2008). This data was used to generate the 200 doubled-haploid genotypes that form the population used in this study. A total of 257 (0,1) SNP markers was evenly distributed across the 10 chromosomes.

Conceptual Framework for Grouping Variables and Balancing their Influence When Classifying Genotypes

Groups and subgroups of variables are created based on their biological and/or agronomic characteristics (e.g., phenotypic and genetic or continuous and discrete variables measured in different environments or at different plant parts). Note that groups and subgroups of variables can be formed based on different biological rationalities, as will be explained later for the four data sets used in this study. Because the different variables used to classify the individuals are measured in different units of scale, the proposed method needs to simultaneously convert all available variables at the group and subgroup levels to a common unit of measurement by means of transformation. Because of transformation, the influence of different variables at the different hierarchies was better balanced.

The proposed method is applied in two steps: (i) after defining a hierarchy of groups and subgroups of variables, HMFA is used to balance the influence of the different groups and subgroups of variables across the hierarchy on the classification of individuals (next step); this step can be regarded as data transformation in the sense that a new set of variables are derived from the original ones; (ii) transformed data from the previous step are used to classify individuals; in this step, we propose using the Ward-MLM approach of Franco et al. (1998) since this method has been proven to be useful and efficient for forming well-defined and cohesive clusters; nevertheless, any classification method could be employed at this stage.

First Step: Balancing the Effects of Different Groups and Subgroups of Variables

The idea underlying this first step is to balance the importance of the different groups and subgroups of attributes based on their internal variability, by employing different PA geometrical methods depending on the type of variables. For example, for continuous variables the PCA is used, whereas for frequency variables and categorical variables the CA and the MCA are used, respectively.

When classifying individuals by use of phenotypic and genetic information simultaneously, phenotypic data may

comprise continuous and/or discrete variables (sometimes measured in different environments), whereas genetic data may include discrete and/or frequency information. Therefore, different hierarchies of variables can be defined. For example, a high level of hierarchy may involve defining two main groups of variables—phenotypic (morphological and/or agronomic) and genetic (molecular markers)—and a low level of hierarchy can then be achieved by creating subgroups of phenotypic variables based on biological criteria (e.g., agronomic traits measured in different environments). Subgroups of genetic variables can also be defined (e.g., alleles from a polymorphic SSR could be a subgroup within the main group of genetic information). Different criteria for grouping variables and defining hierarchies correspond to different biological hypotheses and refer to different types of inquiries; thus the researcher should use a hierarchical structure of the variables that is in close agreement with the natural structure of the biological data at hand. Therefore, although the hierarchical grouping of variables may seem somewhat arbitrary, in reality it is not because it is done following natural biological, agronomic, and genetic patterns that will facilitate testing different hypotheses.

In HMFA, MFA is performed at different stages to account for the different hierarchies (i.e., groups and subgroups of variables); the aim is to achieve a balance among groups and subgroups of variables in the final classification result. Mathematical and geometrical details of MFA and HMFA can be found in Escofier and Pagès (1988–1998), Pagès (2002), Le Dien and Pagès (2003), and Bécue-Bertaut and Pagès (2008). Here we give a few details of both the methods.

A Geometrical Approach for Balancing the Influence of Groups of Variables in a Mixture: Multiple Factor Analysis

Following the ideas and notation of Bécue-Bertaut and Pagès (2008), MFA can be used to solve a classification problem in a data set matrix $\mathbf{X}$ that contains a mixture of continuous and categorical variables. The matrix $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_{J_q}|\mathbf{Z}_1|\mathbf{Z}_2|\ldots|\mathbf{Z}_{J_c}]$ is formed by I rows (entries) and K columns (variables). Each $\mathbf{X}_j$ matrix ($j = 1,2,\ldots,J_q$) contains a set of K_j quantitative (continuous) variables and is composed of K_j columns (each column represents one variable). Each $\mathbf{Z}_j$ matrix ($j = 1,2,\ldots,J_c$) represents a set of categorical variables, each of which is expressed as a group of indicator (0,1) variables (a binary variable has two columns, a nominal three-level variable has three columns, etc., each column representing one category from one variable). Thus, we have J tables formed by J_q sets of continuous (or quantitative) variable tables and J_c sets of categorical variable tables; the total number of columns is $K = \sum_{j=1}^{j=J_q+J_c} K_j$.

The problem can be addressed using a PA geometrical frame in which different methods are proposed for different kinds of variables: PCA is used when only continuous variables are studied, whereas MCA is the appropriate method if variables are all categorical; CA is used for frequency variables. In this research, we have a mixture of variables: phenotypic (continuous and/or categorical) and genetic (categorical {0, 1} or {0, 0.5, 1}).

Originally, MFA was proposed for combining data sets with continuous variables only. The idea of bringing a mixture of continuous and categorical variables into the MFA framework was first proposed by Escofier and Pagès (1988–1998) and Pagès (2002). These authors demonstrated that results obtained from MCA could also be obtained by a nonstandardized weighted PCA on the indicator matrices $\mathbf{Z}_j = \{z_{ikj}\}$ transformed into $\mathbf{Y}_j = \{y_{ijk}\}$ by means of $y_{ikj} = \dfrac{(z_{ikj} - w_{ikj})}{w_{kj}}$, where $w_{kj} = \sum_{i\in I} p_i z_{ikj}$ is the proportion of entries belonging to the column k_j ($k_j = 1,2,\ldots,K_j$, number of columns in the $\mathbf{Y}_j$ matrix), p_i is the weight associated with each entry (in our case, $p_i = 1/I$ for all $i\hat{I}I$; i.e., equal weight is given to each entry), and w_{kj}/Q_j (where $Q_j = \sum_{k\in K_j} w_{kj}$) is used as column weight. The steps for doing an MFA can be summarized as follows:

Step 1. The MFA performs an individual PCA analysis on each $\mathbf{X}_j$ (a group of continuous, standardized, or nonstandardized variables) and a weighted nonstandardized PCA on each $\mathbf{Y}_j$ ($j = 1,2,\ldots,J_c$) matrix, thus obtaining the $J_q + J_c$ first eigenvalues λ_1^j, $j = 1,2,\ldots, J_q + J_c$, corresponding to the directions of maximum variance (or maximum inertia, where inertia is defined as the variability of a set of points in a Euclidean space) within each group of variables.

Step 2. A nonstandardized weighted PCA is then performed on the $\mathbf{X} = [\mathbf{X}_1|\ldots|\mathbf{X}_{J_q}|\mathbf{Y}_1|\ldots|\mathbf{Y}_{J_c}]$ matrix, where $(1/\lambda_1^j)$ from step 1 is used as the column weight for the $\mathbf{X}_j$ matrices, and $w_{kj}/Q_j\lambda_1^j$ is used as the column weight for the $\mathbf{Y}_j$ matrices; the proportion (1/I) is used as the weight for every row. In this manner, the maximum inertia explained for each group of variables is equal to one, and all groups of variables are equally important (because they were standardized to have the maximum inertia equal 1).

Step 3. The scores for each entry and the contribution to the total inertia of each entry, variable, and group of variables are the results obtained from the MFA.

A Geometrical Approach for Balancing the Influence of Groups of Variables in a Nested Structure: Hierarchical Multiple-Factor Analysis

The HMFA method (Le Dien and Pagès, 2003) extends the ideas of the MFA with the objective of accounting for any nested structure among variables. The HMFA uses MFA analyses in a sequential fashion to obtain a set of column weights to be used in a weighted and nonstandardized PCA global analysis that will balance the effects of different groups of variables at every level of the hierarchy and within hierarchies. The steps for doing HMFA analyses are as follows:

Step 1. At the lowest level of the hierarchy, HMFA performs step 1 of MFA. The first eigenvalues at this step are named $\lambda_1^{h(j)}$, where $h = 1$ and $j = 1,2,\ldots,g_1$ (where $g_1 = J_q + J_c$ is the number of groups of variables at this level).

Step 2. At the next higher level of the hierarchy, HMFA performs step 1 of MFA again within each of the high level groups, obtaining a new set of g_2 (number of groups at the high level) eigenvalues $\lambda_1^{h(j)}$, where $h = 2$ and $j = 1, 2,\ldots, g_2$, g_2 being the number of groups at the second level. If the hierarchy includes more than two levels (for example, p levels), this step is repeated to obtain p sets of eigenvalues according to the number of groups at each level.

Step 3. A global weighted and nonstandardized PCA on the whole $\mathbf{X} = [\mathbf{X}_1|\ldots|\mathbf{X}_{J_q}|\mathbf{Y}_1|\ldots|\mathbf{Y}_{J_c}]$ matrix is then performed using 1/I as every row weight (each entry has equal weight) and the product of calculated column weights across the hierarchy as the weight column:

$$\prod_{h=1}^{p}\frac{1}{\lambda_1^{h(j)}}\text{, for columns in the } \mathbf{X}_j \text{ matrices} \quad [1]$$

$$\frac{w_{kj}}{Q_j}\prod_{h=1}^{p}\frac{1}{\lambda_1^{h(j)}} \text{ for columns in the } \mathbf{Y}_j \text{ matrices} \quad [2]$$

We used HMFA to produce new coordinates (scores from the global PCA) for each entry by considering different levels of the groups and subgroups of variables. This is equivalent to transforming variables (phenotypic-discrete, phenotypic-continuous, and genetic) into PA scores that can be treated as continuous variables at the classification stage.

The Contribution of Each Variable to the New Principal Axis in HMFA

Pagès (2004) proposed a procedure for measuring the contribution of one original variable (j_q if continuous, j_c if categorical) to the variability of a new axis υ. This author showed that the total variability explained by one variable (from the mixture of continuous [j_q] and categorical [j_c] variables) on the new axis υ could be expressed as

$$\sum_{j\in J_q} r^2(j,\upsilon) + \sum_{j\in J_c}\eta^2(j,\upsilon) = 1 \quad [3]$$

where r is the correlation coefficient between each original variable and the new axis, and η is the correlation coefficient between the set of k_j indicator variables associated with each categorical variable and the new axis. Dagnièle (1998) (vol 1, page 133; referenced by Pagès, 2004) demonstrated that the coefficient of correlation η between the set of transformed indicator variables γ_{ikj} and the principal axis (or factor) F_s could be written as

$$\eta(\gamma_{ikj}, F_s) = \sqrt{\frac{w_{kj}}{1-w_{kj}}}\frac{F_s(j_k)}{\sqrt{\lambda_s}} \quad [4]$$

where $F_s(j_k)$ is the projection on the axis of range s of the center of gravity of individuals with the kth modality of the jth categorical variable, λ_s is the eigenvalue associated with F_s, and w_{kj} was defined above. Using these concepts, HMFA allows measuring the contribution of each variable and each group of variables to each of the new PAs obtained in the final result. We will use this approach for measuring the contribution of the various original phenotypic and genetic variables to the PA obtained using the MFA and HMFA methods.

Second Step: The Ward-MLM Classification Method under Two Strategies for HMFA Scores

Classifications using transformed data obtained from HMFA were done in all cases using the two-stage Ward-MLM method (Franco et al., 1998). The method starts by applying a hierarchical Ward (1963) minimum variance-clustering algorithm using the scores from the PA previously obtained by HMFA using the Gower (1971) distance. PseudoT^2 and pseudo F statistics are then used to select a set of possible optimal number of groups. Then the mixture of multinormal and multinomial variables model (MLM) is applied to the Ward grouping at each of the possible number of groups, and the maximum likelihood is obtained at convergence for each possible number of groups. The "optimal" number of groups is selected as the smallest number showing the highest increment in likelihood. Once a convenient number of groups is defined, the MLM method attempts to improve the composition of the original Ward groups by maximizing the likelihood of the sample.

When the method is applied to HMFA scores, the PA coordinates can be treated as normal variables; therefore, the MLM method works on a mixture of multinormal distributions assumed to have homogeneity within group variance-covariance matrices. We compared two classification strategies:

1. Ward-MLM using PA scores that explain 95% of total variability (total inertia), called STR-95, and
2. Ward-MLM using PA scores to which the contribution of the original phenotypic and genetic variables is approximately 50% for each; it is called STR-50.

The first classification strategy uses more information (a higher number of PAs) than the second, because almost all the original information (95%) is expressed in PA scores used in classification. However, the second strategy implies a better balance between the effects contributed by both types of variables: phenotypic and genetic.

Distance Measures

The two-stage Ward-MLM cluster strategy is used with the Gower (1971) distance on the Ward within-cluster, minimum variance algorithm and the Mahalanobis distance for the MLM stage. When using Ward-MLM on the continuous scores of the axes obtained from HMFA, the distance for the first stage (Ward) is the Manhattan distance (which is the Gower distance for continuous variables), and the distance for the second stage (MLM) is the Mahalanobis distance. The Mahalanobis distance is a Euclidean distance weighted by the variance-covariance within-group matrix, so we are working in a Euclidean metric space.

When calculating the Euclidean distance between entries, HMFA induces a distance corresponding to a weighted sum of the separate distances from every group of variables. Following the Bécue-Bertaut and Pagès (2008) notation, the squared distance between any two entries i and l is

$$d_{il}^2 = \sum_{j\in J_q}\frac{1}{\lambda_1^j}\sum_{k\in K_j}\left[\frac{x_{ikj}-x_{lkj}}{s_{kj}}\right]^2 + \sum_{j\in J_c}\frac{1}{\lambda_1^j}\sum_{k\in K_j}\frac{1}{Q_j w_{kj}}[z_{ikj}-z_{lkj}]^2 \quad [5]$$

where all terms are defined for previous equations. This distance can be replaced by the Euclidean distance between i and l individuals using all the PAs, or approximated if only some of the PAs are used.

Criteria Used for Comparing Strategies

To evaluate and compare the different classification strategies, we used six criteria: (i) the average distances between entries within a cluster using the Euclidean distance based on HMFA scores; (ii) the average distances between observations within a group using the Gower (1971) distance based on the original phenotypic and genetic variables; (iii) and (iv) the reduction in the Euclidean and

Gower distance, respectively, obtained by comparing the average distance between observations within a group with the average distance between observations for the entire unclassified data set (population), expressed as a percentage of the population average distance; (v) the Mahalanobis distance between groups using only continuous (phenotypic in all cases) variables; and (vi) reduction in the genetic diversity or expected heterozygosity index (*He*) within a group with respect to the *He* of the entire data set.

When forming well-differentiated and homogeneous groups, a strategy is regarded as better than another if it shows smaller values of within-group average distance, higher values of average distance reduction with respect to the unclassified population, higher values for the Mahalanobis distance, and greater reduction in within-group genetic diversity. When selecting a core subset, a strategy is better if it generates a core subset with higher values for all distance and diversity measures.

Creating the Variables Hierarchy for HMFA for Each Data Set

For the four data sets, we defined two hierarchical levels of variables; we will call them subgroups (or low-level groups) and groups (or high-level groups). For each data set, the hierarchy of variables was defined in a particular way, as explained in the following sections.

Wheat-DArT

We defined four low-level subgroups of variables by joining attributes measured in different environments: subgroup 1 has four columns corresponding to values of continuous attribute LR measured in four environments; subgroup 2 contains values of continuous variable SR measured in four environments; subgroup 3 includes values of continuous variable GW measured in four environments; and subgroup 4 contains 75 (0,1) columns corresponding to DArT markers.

At the higher level, we formed two groups of variables by joining previously defined low-group variables: group 1 has three subgroups (1–3) of continuous variables and group 2 comprises 75 DArT columns. In this case, the first HMFA step (the PCA) within each of the three subgroups of continuous variables was performed with no standardization, so that the expression of each genotype in each environment within each variable would reflect genotype × environment interaction.

Wheat-SSR

At the low hierarchical level, we defined 15 subgroups of variables: subgroups 1 to 3 are similar to the Wheat-DArT data set, and subgroups 4 to 15 contain a set of alleles corresponding to each of the 12 SSR markers. At the high level, we defined two groups of variables: group 1 containing three low-level subgroups of continuous variables and group 2 containing 47 SSR columns grouped into 12 SSR markers at the lower level. In this case, PCA within each of the first three low-level subgroups of continuous variables was done with no standardization, allowing genotypic expression in each environment within each subgroup of variables.

Sorghum-SSR

At the low level, we defined 46 subgroups of variables: subgroup 1 containing three continuous phenotypic variables (DF, PH, and PGC); subgroup 2 containing the three discrete nominal variables (PT, GIC, and GrC); and subgroups 3 to 46 comprising 336 SSR columns grouped into 44 SSR markers. At the high hierarchical level, we defined two groups: group 1 formed by two subgroups of phenotypic variables and group 2 formed by 44 SSR markers. In this case, PCA of the low-level subgroup of continuous variables was done using standardization to correct for differences in scale effects.

Maize-SNP Simulation

At the low level, we defined 15 subgroups of variables: subgroups 1 to 5 containing the response of the continuous traits (DS, DA, EH, PH, and GY) in three environments; subgroups 6 to 15 comprising 257 (0,1)-SNP columns grouped by chromosome (10 chromosomes). At the high level, we used two groups: group 1 formed by five low subgroups of phenotypic variables, and group 2 formed by the 10 chromosomes. In this case, PCA within each of the five low-level continuous subgroups was performed with no standardization, allowing the expression of each environment within each group of variables.

Selecting Core Subsets

The process used to select the core subset (of a size equal to 20% of the population) for each data set (Wheat-DArT, Wheat-SSR, Sorghum-SSR, Maize-SNP) and each classification strategy (PA scores that explain 95% of total inertia and PA scores with 50% contribution from phenotypic variables and 50% contribution from genetic variables) involved (i) calculating the Euclidean distance between genotypes within a group defined by the Ward-MLM strategy when using PA scores from the HMFA for each strategy; (ii) defining the number of genotypes to be selected from each group using the D-method (Franco et al., 2005) with the Euclidean distance; (iii) forming 100 independent core subsets; and (iv) from those 100 core subsets, selecting the one showing the maximum Gower distance between genotypes.

Software

The HMFA analysis was performed step by step using multivariate procedures from SAS (SAS Institute, 2006) and R Development Core Team (2008) software. In addition, HMFA was run using the package FactoMineR from Lê et al. (2008). Cluster analysis was done using a code written by Franco et al. (1998) using the IML procedure from SAS (SAS Institute, 2006). Genetic diversity analysis was performed using PowerMarker software (Liu and Muse, 2005).

RESULTS

Hierarchical Multiple Factor Analysis using Two Strategies

The two proposed strategies were (i) classification of entries by the Ward-MLM clustering method using PA scores that explain 95% of total inertia in HMFA (STR-95); and (ii) classification of entries by the Ward-MLM clustering method using PA scores to which the contribution of the original phenotypic and genetic variables is approximately 50% each (STR-50).

There were no important differences in the optimal number of groups using the two classification strategies (g95 vs. g50, Table 1); on the other hand, there were strong differences in the number of principal axes used at the clustering stage (PA95 vs. PA50). Figures 1a to 1d reveal the importance (accumulated percentage of contribution to total inertia) of the two types of variables in each of the first 25, 25, 49, and 77 PAs for the Wheat-DArT (d25, Fig. 1a), Wheat-SSR (d25, Fig. 1b), Sorghum-SSR (d49, Fig. 1c), and Maize-SNP (d77, Fig. 1d) data sets, respectively, and the cumulative inertia explained by those axes.

For the four data sets, phenotypic variables made a greater contribution to the first PAs, reaching a 50 to 50% equilibrium with the contribution of genotypic variables at dimensions d7, d2, d7, and d4 PAs for Wheat-DArT (Fig. 1a), Wheat-SSR (Fig. 1b), Sorghum-SSR (Fig. 1c), and Maize-SNP (Fig. 1d), respectively. Because the ratio between the number of phenotypic traits to the number of genetic traits (columns) was 12:75, 12:49, 6:336, and 15:257 for Wheat-DArT, Wheat-SSR, Sorghum-SSR, and Maize-SNP, respectively, the effect of each column on the first PAs depends on its own discriminatory ability. The genetic variables increased their contributions to later PAs, whereas the phenotypic variables had greater influence on earlier PAs. This is a very important characteristic of MFA and HMFA by which the set of variables formed by a greater number of components (columns) will explain more of the total inertia, but the contribution will be distributed along all axes, and will not necessarily be concentrated in the first PA.

Table 1. Data set name, number of entries (accessions in the collection, *N*), number of phenotypic variables (nph), number of molecular markers (nm), number of alleles (na), total number of columns in the analysis (ncols), number of principal axes explaining 95% of the inertia (PA95), number of groups obtained using PA95 (g95), number of principal axes for which phenotypic and genetic variables contribute 50% each (PA50), and number of groups obtained using PA50 (g50).

Data set†	*N*	nph	nm	na	ncols	PA95	g95	PA50	g50
Wheat-DArT	46	12	75	75	87	25	4	7	4
Wheat-SSR	46	12	12	49	61	25	5	3	4
Sorghum-SSR	90	6	46	336	342	50	5	7	5
Maize-SNP	200	15	257	257	272	78	13	4	13

†DArT, diversity array technology; SSR, simple sequence repeat; SNP, single nucleotide polymorphism.

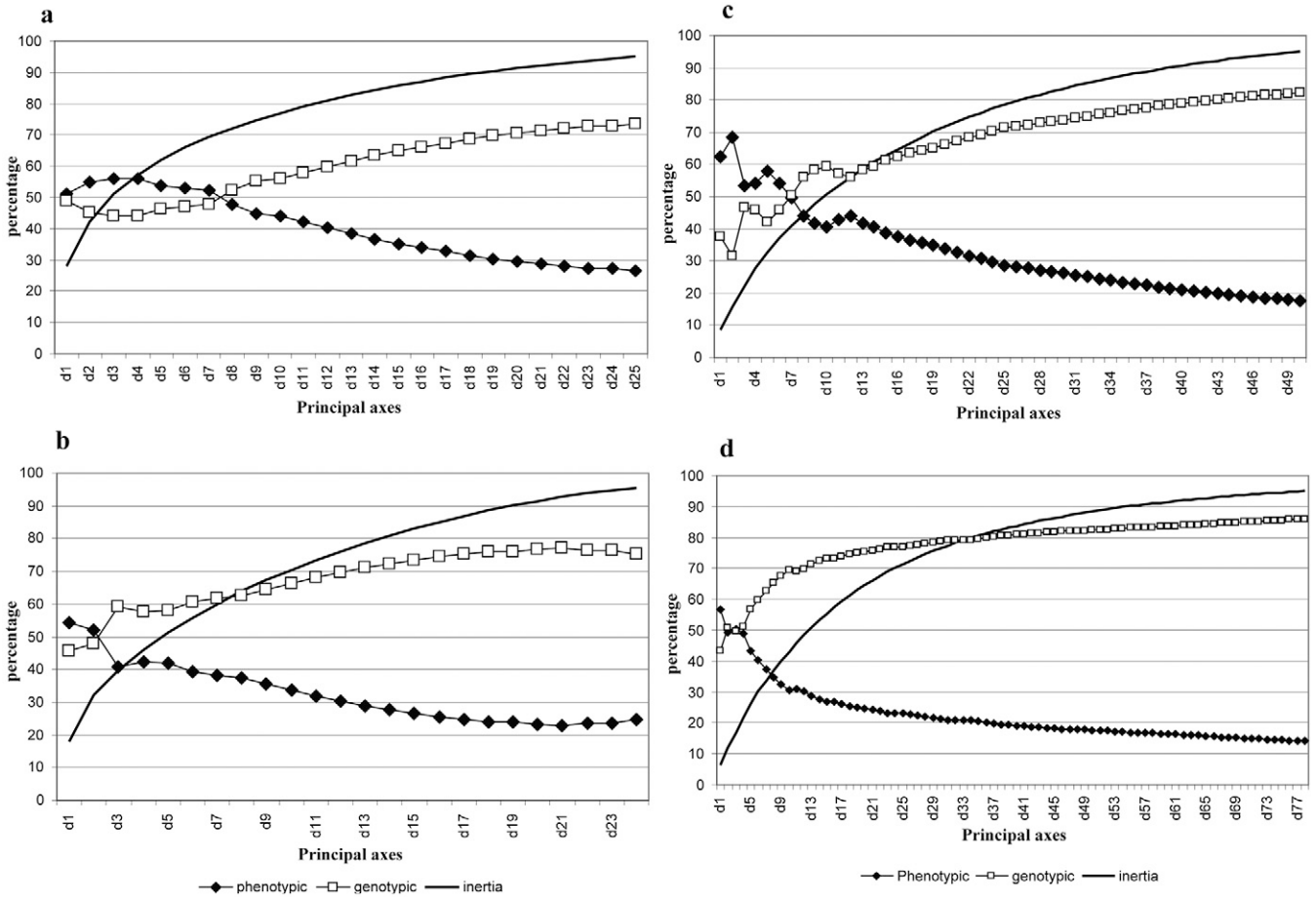

Figure 1. Contribution (%) of phenotypic and genetic variables to the principal axes scores used in the classification of genotypes and percentage of total inertia explained by the principal axes of the global principal components analysis. The data sets are (a) Wheat-DArT, (b) Wheat-SSR, (c) Sorghum-SSR, and (d) Maize-SNP. The number of principal axes on the *x* axis is represented by the letter 'd' referring to dimension and a number denoting the number of axes. DArT, diversity array technology; SSR, simple sequence repeat; SNP, single nucleotide polymorphism.

Table 2. Average Euclidean (E-mean) and Gower (G-mean) distances within groups; standard error of the means (SE); reduction in Euclidean (E-gain) and Gower (G-gain) distances with respect to the entire population, and Mahalanobis distance (MD) between groups for continuous variables. The Ward Modified Location Model used two types of principal axes scores: (i) those from the principal axes that explain 95% of the inertia (STR-95); and (ii) those from the principal axes to which phenotypic and genetic variables contributed 50% each (STR-50).

Strategy	E-mean	SE	E-gain (%)	G-mean	SE	G-gain (%)	MD
Wheat-DArT							
STR-95	2.88	0.035	13.6	0.252	0.005	15.4	4.2
POP†	3.33	0.034		0.298	0.005		
STR-50	1.71	0.029	37.3	0.198	0.004	33.4	39.9
POP	2.73	0.039		0.298	0.005		
Wheat-SSR							
STR-95	3.18	0.024	16.7	0.209	0.002	13.0	9.1
POP	3.82	0.033		0.240	0.003		
STR-50	1.13	0.021	49.3	0.166	0.002	31.1	22.8
POP	2.23	0.039		0.240	0.003		
Sorghum-SSR							
STR-95	4.66	0.024	12.5	0.135	0.001	22.0	3.9
POP	5.32	0.027		0.174	0.001		
STR-50	2.07	0.018	36.6	0.149	0.001	14.3	3.9
POP	3.26	0.026		0.174	0.001		
Maize-SNP							
STR-95	6.56	0.019	2.7	0.471	0.001	2.87	5.4
POP	6.74	0.004		0.485	0.001		
STR-50	1.75	0.005	42.9	0.447	0.001	7.73	11.4
POP	3.06	0.007		0.485	0.001		

†POP, distances and standard errors using the entire (unclustered) data sets.

For all data sets, the average Euclidean distance between entries within a group (E-mean) was smaller than the average Euclidean distance between entries of the entire population (POP) using both strategies (Table 2); similar results were obtained using the Gower distance (G-mean). The reduction (gain) in average Euclidean distance between entries within a group with respect to the average distance between entries of the entire population showed values within the 2.7 to 16.7% interval for the STR-95 strategy, and values within the 36.6 to 49.3% interval for the STR-50 strategy.

For all data sets and both distance measures (except for Gower's distance in the Sorghum-SSR data set), the STR-50 strategy produced greater reductions in the distance between genotypes within a group (%E-gain and %G-gain) than the STR-95 strategy; similar behavior was observed for the Mahalanobis distance (MD) used only on continuous variables (Table 2). Furthermore, Table 3 indicates that for all data sets (except the Sorghum-SSR data set), the groups formed using the STR-50 strategy had lower *He* values within groups, higher *He* values between groups, and higher values for the proportion of genetic diversity among groups (G_{ST}) (Nei, 1973) than the STR-95 strategy. These results indicate that by using the scores of the few first axes, to which the contribution of phenotypic and genetic variables is similar (50 to 50%), we can obtain more compact, better defined, and more cohesive groups of genotypes than those obtained using a large number of axes accounting for 95% of total variability.

Distances between genotypes measured using the Euclidean distance based on PA scores from HMFA show high correlation coefficients with distances measured using Gower's distance on the original variables (data not shown). All correlation coefficients were high and significant ($P < 0.001$) using Fisher's transformation and the *t* test, which is the appropriate test for large sample sizes (Bhattacharyya and Johnson, 1977).

Contribution of Phenotypic and Genetic Variables to New Principal Axes

The contributions of phenotypic and genetic variables to the PAs obtained from HFMA for the four data sets are shown in Table 4 (for Wheat-SSR, Wheat-DArT, and Maize-SNP data sets) and Table 5 (for the Sorghum-SSR data set). The contribution of the groups of phenotypic and genetic variables to the PA explaining 95% of the inertia showed values from 14 to 86% in the Maize-SNP data set (15 phenotypic columns, 257 genetic columns) to 25 to 75% in the Wheat-SSR data set (12 phenotypic columns, 49 genetic columns). The rank of the contribution of each phenotypic variable or group of phenotypic variables to the final PA was different in each data set and for each strategy within each data set, but the most important variable (or group of variables) was the same for both strategies (STR-50 and STR-95) within each data set: (i) for the Wheat-DArT and Wheat-SSR data sets, the most important phenotypic variable was LR in both strategies; (ii) for the Sorghum-SSR data set, the

Table 3. Genetic diversity (*He*) of the entire population (total), averaged within groups (within), and between groups (between); G_{ST} statistic, and level of differentiation between groups following Wright (1951) for four data sets: Wheat-DArT, Wheat-SSR, Sorghum-SSR, and Maize-SNP. The Ward Modified Location Model used two types of principal axes scores: (i) those from the principal axes that explain 95% of the inertia (STR-95); and (ii) those from the principal axes to which phenotypic and genetic variables contributed 50% each (STR-50).†

Source	Wheat-DArT		Wheat-SSR		Sorghum-SSR		Maize-SNP	
	STR-50	STR-95	STR-50	STR-95	STR-50	STR-95	STR-50	STR-95
Total	0.309	0.309	0.474	0.474	0.593	0.593	0.497	0.497
Within	0.199	0.254	0.292	0.341	0.492	0.432	0.425	0.448
Between	0.110	0.055	0.182	0.133	0.101	0.161	0.072	0.049
G_{ST}	0.356	0.178	0.384	0.281	0.170	0.272	0.145	0.099
Level‡	Very large	Large	Very large	Very large	Large	Very large	Moderate	Moderate

†DArT, diversity array technology; SSR, simple sequence repeat; SNP, single nucleotide polymorphism.

‡Genetic differentiation between groups, following Wright (1951).

Table 4. Results of Wheat-SSR, Wheat-DArT, and Maize-SNP data sets: contribution (%) of each variable (Var) using the first principal axes (three for Wheat-SSR, seven for Wheat-DArT, four for Maize-SNP) for the case in which phenotypic and genetic variables contribute 50% each to the principal axes (STR-50), and the contribution of each variable to the principal axes (25 for Wheat-SSR, 25 for Wheat-DArT, 78 for Maize-SNP) explaining 95% of the inertia (STR-95).

Wheat-SSR†				Wheat-DArT‡				Maize-SNP§			
STR-50		STR-95		STR-50		STR-95		STR-50		STR-95	
Var	%	Var	%	Var	%	Var	%	Var	%	Var	%
LR	16.4	LR	9.5	LR	23.4	LR	10.3	PH	13.9	PH	3.5
SR	12.8	GW	8.4	GW	15.1	SR	8.7	GY	11.8	EH	3.1
GW	11.6	SR	6.7	SR	13.6	GW	7.3	DA	8.4	GY	2.9
SSR11	11.3	SSR12	11.1	D15	1.23	D22	2.01	DS	7.3	DA	2.5
SSR3	7.7	SSR10	10.5	D39	1.21	D03	1.89	EH	7.3	DS	2.2
SSR6	6.3	SSR4	9.6	D73	1.2	D63	1.85	CHR4	9.3	CHR1	10.9
SSR10	5.7	SSR7	6.7	D12	1.17	D73	1.69	CHR2	8.9	CHR7	9. 8
SSR12	5.2	SSR5	6.5	D40	1.15	D11	1.6	CHR3	6.8	CHR3	9
SSR1	4.8	SSR1	6.5	D59	1.09	D16	1.49	CHR8	5.4	CHR5	8.6
SSR7	4.3	SSR9	5.4	D45	1.05	D35	1.43	CHR1	5	CHR9	8.5
SSR4	3.9	SSR11	5.4	D13	1.02	D34	1.42	CHR10	4.6	CHR8	8.5
SSR2	3.7	SSR3	4.6	D41	1.01	D56	1.38	CHR5	3.2	CHR4	8
SSR5	2.5	SSR8	4.6	D69	0.99	D70	1.38	CHR7	3	CHR10	7.9
SSR9	2.4	SSR6	2.9	D10	0.95	D60	1.37	CHR9	2.9	CHR2	7.5
SSR8	1.4	SSR2	1.7	D42	0.91	D44	1.33	CHR6	2	CHR6	7.3
Total phenotypic	40.8	–	24.6	–	52.2	–	26.3	–	48.7	–	14.1
Total Genetic	59.2	–	75.4	–	47.8	–	73.7	–	51.3	–	85.9

†LR, leaf rust, SR, stem rust, GW, 1000-grain weight; SSR1-SSR12, microsatellite molecular markers.

‡DArT, Diversity Arrays Technology markers D1-D75.

§PH, plant height; GY, grain yield; DA, days to anthesis; DS, days to silking; EH, ear height; CHR1 to CHR10, chromosomes 1–10.

group of categorical phenotypic variables (and inside this group, the glume color variable, GlC) was more important than the group of continuous phenotypic variables for both strategies (Table 5); and (iii) for the simulated Maize-SNP data set, the most important phenotypic variable was plant height for both strategies; the other variables showed different rankings (Table 4).

The contribution of molecular marker variables to the PA did not show a pattern similar to that of phenotypic variables, but some interesting observations were noted: in the Wheat-SSR data set, markers SSR10 and SSR12 were within the five most important markers in both strategies (Table 4); in the Sorghum-SSR data set, there were 15 molecular markers in the set of the 20 most important molecular markers in both strategies that contributed to the PA (Table 5); in the Maize-SNP data set, chromosomes CHR3 and CHR1 were among the five variables that contributed most in both strategies; in the Wheat-DArT data set, only 1 of the 12 most important markers (out of 75) was shared by both strategies (Table 4).

The above results indicate that some phenotypic and genetic variables that turned out to be important under both strategies are in fact the most relevant variables in each data set, for they contributed the most to genotypes classification. These variables are most important for forming the different groups at the clustering stage.

Further Examination and Interpretation of the Results of the Wheat-SSR Data Set

In this section, we examine and interpret in more detail the groups formed in the Wheat-SSR data set. For the Wheat-SSR data set, the STR-50 strategy formed better groups than the SRT-95 strategy relative to E-mean (1.13 in STR-50 vs. 3.18 in STR-95), G-mean (0.166 in STR-50 vs. 0.209 in STR-95), E-gain (%) (49.3% in STR-50 vs. 16.7% in STR-95), G-gain (%) (31.1% in STR-50 vs. 13.0% in STR-95), and MD (22.8 in STR-50 vs. in 9.1 STR-95) (Table 2). Concerning *He*, STR-50 formed better groups than SRT-95 within groups (0.292 in STR-50 vs. 0.341 in STR-95) and between groups (0.182 in STR-50 vs. 0.133 in STR-95), and also produced higher G_{ST} values (0.384 in STR-50 vs. 0.281 in STR-95) (Table 3).

As for continuous variables of the Wheat-SSR data set, reductions in average variances within groups with respect to variances in the entire data set for STR-50 and STR-95 were 11.2% and 6.1% for LR, 37.7 and 34.9% for GW, and 51.8 and 8.2% for SR, respectively (data not shown). These results show that groups of genotypes formed by STR-50 are less variable and more compact and cohesive than those formed by the STR-95 strategy.

The effect of the genetic variables was studied using the genetic diversity or *He* (Weir, 1996) calculated for each marker in the entire Wheat-SSR data set and compared with the weighted average within-group *He* calculated for

Table 5. Results of Sorghum-SSR data set for the case where phenotypic and genetic variables contribute 50% each to the principal axes (STR-50), and the contribution of each variable to the principal axes explaining 95% of the inertia (STR-95).

	STR-50				STR-95		
		Variable	%			Variable	%
			Phenotypic†				
1	Categorical		32.9	1	Categorical		13.4
	1	GlC	16.1		1	GlC	4.8
	2	GrC	11.8		2	PT	4.7
	3	PT	5.0		3	GrC	3.9
2	Continuous		16.7	2	Continuous		4.3
	1	PH	6.1		1	PGC	1.6
	2	PGC	5.7		2	PH	1.5
	3	DF	4.9		3	DF	1.2
			Genetic‡				
	1	SSR39	2.5		1	SSR44	4.9
	2	SSR33	2.4		2	SSR36	4.4
	3	SSR44	2.3		3	SSR40	4.3
	4	SSR31	2.2		4	SSR39	3.6
	5	SSR25	2.0		5	SSR6	3.4
	6	SSR38	1.9		6	SSR35	3.4
	7	SSR6	1.8		7	SSR33	3.4
	8	SSR9	1.7		8	SSR31	3.3
	9	SSR36	1.6		9	SSR37	2.9
	10	SSR42	1.6		10	SSR25	2.6
	11	SSR27	1.6		11	SSR13	2.5
	12	SSR43	1.5		12	SSR1	2.3
	13	SSR41	1.4		13	SSR27	2.3
	14	SSR35	1.4		14	SSR41	2.3
	15	SSR28	1.4		15	SSR9	2.3
	16	SSR1	1.4		16	SSR29	2.2
	17	SSR10	1.3		17	SSR7	2.2
	18	SSR29	1.3		18	SSR38	2.2
	19	SSR17	1.3		19	SSR30	2.1
	20	SSR37	1.2		20	SSR22	1.9
Total phenotypic			49.7				17.7
Total genetic			50.3				82.3

†Continuous: three continuous variables (DF, time to flowering in days; PH, plant height; PGC, percentage of glume coverage); categorical: three categorical, nominal variables (PT, panicle type; GlC, glume color; GrC, grain color); SSR1 to SSR44: microsatellite molecular markers.

‡Highest 20 (out of 46) contributions.

the groups formed by the STR-50 and STR-95 strategies (using the number of genotypes per group as a weight). For all markers and both strategies, the *He* index was smaller within groups than for the entire population (Table 6). Percentages of *He* reductions were within the interval (9.6, 69.3) for the STR-50 strategy and within the interval (16.0, 49.7) for the STR-95 strategy, the mean reduction in *He* being 38.3 and 28.0% for the STR-50 and STR-95 strategies, respectively. Again, STR-50 proved to be more efficient than STR-95 for forming final groups of genotypes that are more homogeneous, compact, and cohesive with respect to genetic variables.

Molecular marker SSR11 showed the highest contribution with the STR-50 strategy (11.3%, Table 4) and was the marker showing the smallest value of *He* for the entire data set (0.0841, Table 6). This marker has three alleles with frequencies of 0.95, 0.02, and 0.02 (representing 44, 1, and 1 genotypes, respectively). On the other hand, marker SSR12 showed the highest contribution with the STR-95 strategy (11.1%, Table 4), and the largest *He* in the entire data set (0.6904, Table 6). This marker has six alleles with frequencies of 0.044, 0.174, 0.022, 0.022, 0.413, and 0.304 (representing 2, 8, 1, 1, 19, 14 genotypes and a missing value). Marker SSR12 is more polymorphic and has a more balanced allelic distribution across genotypes than SSR11.

In summary, the most diverse marker showed the greatest effect on the groups when using the 25 PAs explaining 95% of total inertia, and the least diverse marker showed the greatest effect on the groups when using the first three PAs explaining phenotypic and genetic variables (50% each). This result was to be expected because, as mentioned before, the weights of the columns (markers) of rare alleles were greater than the weights of markers with more evenly distribution alleles. On the other hand, more diverse markers showed more columns (alleles), and their effect on the groups was increased as more PAs were included in the classification (STR-95 case).

A researcher could expect groups separated mainly by markers with rare alleles (lower diversity values) when using a few PAs, and groups separated mainly by markers with a greater number of more evenly distributed alleles when using a larger number of principal axes at the cluster analysis stage.

Comparing Results of the Wheat-SSR Data Set with Those of the Sorghum-SSR Data Set

It is interesting to note that in contrast to results for the Wheat-SSR data set, for the Sorghum-SSR data set, the STR-95 strategy formed more compact and cohesive groups than the STR-50 strategy for Gower distance within group (0.135 for STR-95 vs. 0.149 for STR-50) (Table 2), gain in Gower distance within group (22.0 for STR-95 vs. 14.3 for STR-50) (Table 2), and genetic diversity (0.432 STR-95 vs. 0.492 STR-0.50 for within group; 0.161 for STR-95 vs. 0.101 STR-50 for between groups; 0.272 G_{ST} for STR-95 vs. 0.170 G_{ST} for STR-50) (Table 3).

The effects of different groups of variables and the two strategies on the formation of the final groups can be studied by calculating the distance among genotypes using all the variables together (phenotypic and genetic), or each one of them separately, and separating the phenotypic-continuous from the phenotypic-discrete variables (in the Sorghum-SSR data set). For the Wheat-SSR data set when using all variables, the gain in distance reduction within groups was greater with the STR-50 strategy than with the STR-95 strategy (31 vs. 13%, respectively) (Table 7); the result was similar when using only the phenotypic or only

the genetic groups of variables. On the other hand, for the Sorghum-SSR data set, STR-95 was better than STR-50 (gain of 22% vs. 14%, respectively), but this advantage was based only on genetic variables (Table 7). Based only on phenotypic variables, the gains for STR-50 were 37 vs. 26% given by the STR-95 strategy. In general, groups generated by the STR-50 strategy were better than those generated by the STR-95 strategy in both data sets with respect to phenotypic variables and with respect to genetic and phenotypic variables in the wheat data set. The groups were better for STR-95 only with respect to genetic variables in the sorghum data set. Thus, the difference in performance between the strategies applied to both data sets could be attributed to the different numbers of columns defined by phenotypic and genetic variables in both data sets: the wheat data set had 12 phenotypic and 49 genetic columns, whereas the sorghum data set had only 6 phenotypic but 336 genetic columns. This difference produced an important genetic columns effect when using a greater number of PAs in the classification step, that is, when using the STR-95 strategy.

Table 6. Ranking of each marker in the Wheat-SSR data set contributing to the first principal axis of each of the classification strategies (STR-50 and STR-95).† Genetic diversity in the entire population (*He*-all), average *He* within group (*He*-mean), and percentage of *He* reduction obtained by grouping the genotypes based on both strategies (STR-50 and STR-95) with respect to *He* of the entire population.

Marker	Ranking			STR-50		STR-95	
	STR-50	STR-95	*He*-all	*He*-mean	*He* reduction (%)	*He*-mean	*He* reduction(%)
SSR12	5	1	0.6904	0.529	23.4	0.498	27.8
SSR9	11	7	0.6900	0.392	43.3	0.485	29.7
SSR7	7	4	0.6607	0.465	29.6	0.515	22.0
SSR10	4	2	0.6021	0.185	69.3	0.367	39.0
SSR1	6	6	0.5424	0.409	24.7	0.430	20.7
SSR6	3	11	0.4679	0.300	36.0	0.235	49.7
SSR2	9	12	0.4537	0.196	56.7	0.325	28.4
SSR8	8	10	0.4357	0.344	21.0	0.283	35.1
SSR3	2	9	0.3563	0.118	66.9	0.294	17.5
SSR5	10	5	0.3563	0.241	32.3	0.298	16.4
SSR4	8	3	0.3512	0.255	27.4	0.294	16.4
SSR11	1	8	0.0841	0.076	9.6	0.071	16.0
Mean			0.4742	0.292	38.3	0.341	28.0

†Phenotypic and genetic variables contribute 50% each to the principal axes (STR-50), and principal axes explaining 95% of the inertia (STR-95).

Characteristics of the core subsets of the four data sets and for each strategy (STR-95 and STR-50) for genetic diversity measured by *He* and the Gower distance (*Gd*) are shown in Table 8. For the four data sets, the *He*-core and the *Gd*-core in the core subsets were greater or equal to their respective distances between genotypes of the unclustered population (*He*-POP and *Gd*-POP). In the Maize-SNP data set, *He*-core and *Gd*-core values were smaller than *He*-POP and *Gd*-POP values, but only in 1% of the entire population. Therefore, the core subsets satisfied the requirement of reducing the number of genotypes while maintaining the diversity of the entire population. In addition, core subsets obtained using a smaller number of PAs with STR-50 strategy were equal to or better than those obtained with the STR-95 strategy, which uses more PAs and does not balance phenotypic and genotypic effects. This result indicated that the balance between phenotypic and genetic variables captured in the first few PAs produced better cores than those obtained using a large number of PAs.

DISCUSSION

The HMFA method was useful for classifying breeding genotypes and/or gene-bank accessions using all available information simultaneously. There was a natural imbalance in the traits used to classify the genotypes; whereas the number of phenotypic columns was much smaller than the number of genetic columns and the phenotypic traits were much more variable than the molecular markers (which usually have two or three different values). The HMFA method allowed balancing out this situation by using classification strategies that consider different contributions of phenotypic and genetic variables to the scores of the new PAs. The STR-50 strategy is an approach to the problem of balancing the

Table 7. Average and standard errors for the Gower distances (*Gd*) between individuals in the entire population (POP), and between individuals within group in the groups formed by applying the two strategies (STR-50 and STR-95) using all variables, only phenotypic variables, and only genetic variables for Wheat-SSR and Sorghum-SSR data sets. In the sorghum data set, the phenotypic distance is separated into two components: distance using only the continuous variables and distance using only the categorical variables.†

Data set	Variables	POP		STR-50			STR-95		
		Gd	SE	*Gd*	SE	Gain	*Gd*	SE	Gain
Wheat-SSR	All	0.240	0.09	0.166	0.06	31	0.209	0.07	13
	Phenotypic	0.252	0.004	0.238	0.003	5	0.306	0.004	–21‡
	Genetic	0.237	0.003	0.140	0.002	41	0.180	0.002	24
Sorghum-SSR	All	0.174	0.05	0.149	0.05	14	0.135	0.05	22
	Phenotypic	0.448	0.003	0.283	0.002	37	0.329	0.003	26
	Continuous	0.225	0.002	0.254	0.002	–13†	0.209	0.002	7
	Categorical	0.671	0.005	0.278	0.003	58	0.367	0.004	45
	Genetic	0.169	0.001	0.138	0.001	18	0.131	0.001	23

†Phenotypic and genetic variables contribute 50% each to the principal axes (STR-50), and principal axes explain 95% of the inertia (STR-95). SSR, simple sequence repeat.

‡Only in those cases where the distance between individuals within groups is greater than the respective distance in the whole data set.

Table 8. Genetic diversity (*He*); Gower distance (*Gd*) in the entire population (*He*-POP, *Gd*-POP) and in the core subset (*He*-core, *Gd*-core); and percentage of *He* and *Gd* increase (%) with respect to the entire population for four data sets: Wheat-DArT, Wheat-SSR, Sorghum-SSR, and Maize-SNP.†

	Wheat-DArT		Wheat-SSR		Sorghum-SSR		Maize-SNP	
	STR-50	STR-95	STR-50	STR-95	STR-50	STR-95	STR-50	STR-95
He-POP	0.309		0.474		0.593		0.497	
He-core	0.363	0.359	0.565	0.565	0.629	0.629	0.492	0.492
%	17.4	16.2	19.2	19.2	4.4	4.4	–1.0	–1.0
Gd-POP	0.298		0.240		0.174		0.485	
Gd-core	0.406	0.402	0.323	0.322	0.219	0.208	0.492	0.492
%	36.2	34.9	34.6	34.2	25.9	19.5	1.4	1.4

†Phenotypic and genetic variables contribute 50% each to the principal axes (STR-50), and principal axes explain 95% of the inertia (STR-95). DArT, diversity array technology; SSR, simple sequence repeat; SNP, single nucleotide polymorphism.

effects between two kinds of variables (phenotypic and genotypic); this problem has not been tackled for forming core groups within the context of genetic resources conservation. Although 50:50 is only one of several possibilities (e.g., we could use 30:70 or any other ratio), (i) this strategy selects a small set of markers and phenotypic variables with which the biologist can work when interpreting the resulting classification; this avoids having to look at irrelevant markers or phenotypic variables; and (ii) by balancing the effects, the classification concentrates on the most discriminatory variables (phenotypic and genetic) and does not allow any of them to dominate others. It has been clearly shown that when more columns of genetic variables are included in the analysis, their contribution to the new PA starts to appear in the later axes.

Second, the HMFA allows the researcher to decide whether to standardize a continuous variable or a set of continuous variables. For example, on the Wheat-DArT data, the variable hierarchy for HMFA defined four low-level subgroups of variables by joining traits measured in different environments (e.g., subgroup 1 has four columns corresponding to values of leaf rust measured in four environments). Because different environments had different potentials and, therefore, different variabilities for different variables, the HMFA was performed on nonstandardized data within each subgroup so that traits could be expressed differently in each environment. The HMFA gives the researcher freedom to perform the classification without masking possible environmental effects that may affect the traits differently; this facilitates the natural expression of traits under different environmental conditions.

Third, the HMFA gives the researcher the opportunity to use PA scores to which the contributions of the original phenotypic and genetic variables may vary. As shown in this study, a well-balanced contribution produces the most cohesive and best defined groups of genotypes. Finally, the consistency of the HMFA across different data sets with different types of variables and under different circumstances, such as using data from multienvironment trials, different types of markers, and different numbers of phenotypic and genetic traits.

The HMFA is useful for combining multiple tables of quantitative and categorical variables, and for finding common ground for balancing the different effects of all the different types of variables. A general guideline for the proper use of the introduced method is to form groups and subgroups of variables based on biological rationalities, such as joining attributes measured in different environments and/or defining hierarchy of variables based on continuous and discrete variables. Furthermore, if different types of markers are used, a logical grouping will be to join markers of the same type in one hierarchy.

Results of this study show that the HMFA in conjunction with the Ward-MLM method is useful for breeders who need to form homogeneous groups of breeding genotypes with similar performance across different response variables measured in different scale and in different environmental conditions. The approach presented in this article shows how all available information can be used to form homogeneous groups. Also, the HMFA allows balancing the effect of the different variables on the transformed dimensions represented by the new principal axes. Furthermore, the HMFA and the Ward-MLM can be useful for gene-bank managers that usually form core subsets based only on phenotypic traits or genetic (molecular marker) traits but never using both types of variables simultaneously.

Acknowledgments

The authors thank Drs. Rajan Sharma, R.P. Thakur, and C.T. Hash of ICRISAT for the sorghum data set. The sorghum data were generated in a research project sponsored by the Sehgal Foundation Endowment Fund.

References

Bécue-Bertaut, M., and J. Pagès. 2008. Multiple factor analysis and clustering of a mixture of quantitative, categorical, and frequency data. Comput. Stat. Data Anal. 52:3255–3268.

Bhattacharyya, G.K., and R.A. Johnson. 1977. Statistical concepts and methods. John Wiley and Sons, New York.

Cerón-Rojas, J.J., F. Castillo-González, J. Sahagún-Castellanos, A. Santacruz-Varela, I. Benítez-Riquelme, and J. Crossa. 2008. A molecular selection index method based on eigenanalysis. Genetics 180:547–557.

Dagnièle, P. 1998. Statistique théorique et appliqué. De Boeck Université Pubisher, Bruxelles, Belgium.

Escofier, B., and J. Pagès. 1988–1998. Analyses factorielles simples et multiples: Objectifs, méthodes, et interprétation. Dunod, Paris.

Escofier, B., and J. Pagès. 1994. Multiple factor analysis: AFMULT package. Comput. Stat. Data Anal. 18:121–140.

Franco, J., and J. Crossa. 2002. The modified location model for classifying genetic resources: I. Association between categorical and continuous variables. Crop Sci. 42:1719–1726.

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.

Franco, J., J. Crossa, S. Taba, and S.A. Eberhart. 2002. The modified location model for classifying genetic resources: II. Unrestricted variance–covariance matrices. Crop Sci. 42:1727–1736.

Franco, J., J. Crossa, S. Taba, and H.A. Shands. 2005. Sampling strategy for conserving genetic diversity when forming core subsets. Crop Sci. 45:1035–1044.

Franco, J., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1998. Classifying genetic resources by categorical and continuous variables. Crop Sci. 38:1688–1696.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–874.

Lê, S., J. Josse, and F. Husson. 2008. FactoMineR: An R package for multivariate analysis. J. Stat. Soft. 25:1–18.

Le Dien, S., and J. Pagès. 2003. Analyse factorielle multiple hiérarchique. Rev. Statistique Appliquée 51:47–73.

Liu, K., and S.V. Muse. 2005. PowerMarker: Integrated analysis environment for genetic marker data. Bioinformatics 21:2128–2129.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70:3321–3323.

Pagès, J. 2002. Analyse factorielle de données mixtes. Rev. Statistique Appliquée 50:5–37.

Pagès, J. 2004. Analyse factorielle multiple aux variables qualitatives et aux données mixtes. Rev. Statistique Appliquée 50:93–111.

Podani, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. Taxon 48:331–340.

R Development Core Team. 2008. R: A language and environment for statistical computing. Available at www.R-project.org. R Foundation for Statistical Computing, Vienna, Austria.

SAS Institute. 2006. SAS/STAT, Version 9.1. SAS Inst., Cary, NC.

SAS Institute. 2006. SAS/IML, Version 9.1. SAS Inst., Cary, NC.

Wang, J., M. van Ginkel, R. Trethowan, G. Ye, I. Delacy, D. Podlich, and M. Cooper. 2004. Simulating the effects of dominance and epistasis on selection response in the CIMMYT wheat breeding program using QuCim. Crop Sci. 44:2006–2018.

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

Weir, B. 1996. Genetic data analysis II. Sinauer Associates, Sunderland, MA.

Wishart, D. 1986. Hierarchical cluster analysis with messy data. p. 453–460. *In* W. Gaul and M. Schader (ed.) Classification as a tool of research. Elsevier Science, Amsterdam, the Netherlands.

Wright, S. 1951. The genetical structure of populations. Ann. Eugenetics 15:313–354.