J. Franco · J. Crossa · J.M. Ribaut · J. Betran
M.L. Warburton · M. Khairallah

A method for combining molecular markers and phenotypic attributes for classifying plant genotypes

Received: 15 January 2001 / Accepted: 13 March 2001

Abstract Classifying genotypes into clusters based on DNA fingerprinting, and/or agronomic attributes, for studying genetic and phenotypic diversity is a common practice. Researchers are interested in knowing the minimum number of fragments (and markers) needed for finding the underlying structural patterns of diversity in a population of interest, and using this information in conjunction with the phenotypic attributes to obtain more precise clusters of genotypes. The objectives of this study are to present: (1) a retrospective method of analysis for selecting a minimum number of fragments (and markers) from a study needed to produce the same classification of genotypes as that obtained using all the fragments (and markers), and (2) a classification strategy for genotypes that allows the combination of the minimum set of fragments with available phenotypic attributes. Results obtained on seven experimental data sets made up of different plant species, number of individuals per species' and number of markers, showed that the retrospective analysis did indeed find few relevant fragments (and markers) that best discriminated the genotypes. In two data sets, the classification strategy of combining the information on the relevant minimum fragments with the available morpho-agronomic attributes produced compact and well-differentiated groups of genotypes.

Keywords Molecular markers · Fragments · Cluster analysis · Simple matching coefficients · Analysis of molecular variance · Mixture models

Communicated by P.M.A. Tigerstedt

J. Franco (✉)
Facultad de Agronomía, Universidad de la República,
Ave. Garzón 780, CP 12900, Montevideo, Uruguay

J. Crossa · J.M. Ribaut · M.L. Warburton · M. Khairallah
International Maize and Wheat Improvement Center (CIMMYT),
Apdo. Postal 6-641, 06600 Mexico D.F., Mexico

J. Bertran
Department of Crop Sciences, Texas A&M University,
College Station, Texas 77843-2474, USA

Introduction

The study of phenotypic and genetic diversity to identify groups with similar genotypes is important for conserving, evaluating and utilizing genetic resources, for studying the diversity of pre-breeding and breeding germplasm, and for determining the uniqueness and distinctness of the phenotypic and genetic constitution of genotypes with the purpose of protecting the breeder's intellectual property rights.

To pursue these objectives, various types of attributes are commonly measured in each genotype: (1) continuous phenotypic variables such as morpho-agronomic traits (maturity, height, phenology, etc.); (2) discrete phenotypic variables such as grain color and texture, resistance to diseases and insects, etc. (these are usually multi-state variables); and, (3) discrete genetic marker characteristics using RFLPs and AFLPs that are binary (absence/presence). Therefore, for a total of *p* attributes, each genotype can be visualized as being located in a *p* multi-dimensional space in which each dimension is represented by one attribute. The true underlying homogeneous sub-populations (groups) of genotypes and their shape and structure are unknown. What is known is that the association between attributes affects the shape of the sub-populations and that their structure depends on the composition of the sub-populations. Hierarchical and non-hierarchical classification methods attempt to recover the true shape and structure of the underlying sub-populations.

Hierarchical clustering methods, such as those of Ward (1963) and UPGMA, are geometrical techniques widely used for finding homogeneous sub-populations where continuous or discrete variables can be combined and used by means of the Gower (1971) distance. Statistical classification methods use the concept of mixture models in which, beginning with an a priori classification of the genotypes into *g* sub-populations and considering each sub-population as one of the distributions in the mixture, the vector of mean attribute values ($\boldsymbol{\mu}$) and the variance-covariance matrix ($\boldsymbol{\Sigma}$) for each of the *g* groups are estimated by the maximum-likelihood method.

In general, when morpho-agronomic and genetic marker data are available on a set of genotypes for studying their diversity and the formation of homogeneous groups, two types of hierarchical classifications are independently performed. One is obtained based on the morpho-agronomic traits in which a standard metric distance (such as the Squared Euclidean) is computed and a clustering strategy, such as Ward or UPGMA, is applied. The other classification is obtained based on the genetic marker attributes when genetic similarities (or dissimilarities) of *n* individuals are determined with molecular markers such as RFLPs, AFLPs, or SSRs. Using each fragment as an attribute (with values of 0 and 1 denoting the presence or absence of the fragment in that genotype, respectively), and applying any clustering strategy (such as single or complete linkage, UPGMA, the centroid method, the Ward method, etc.), genotypes can be clustered into groups that are as homogeneous as possible and heterogeneous among groups. In general, results showed that groups formed based on both continuous and categorical classifications had a low to medium consensus. A unified statistical classification approach using both continuous and categorical variables seems useful for forming homogeneous groups.

Recently, Franco et al. (1998) proposed the Modified Location Model (MLM), which combines all the categorical variables into one multinomial variable, *W*, that is then used with the available continuous variables. Initial groups are defined by the Ward method and then improved by the MLM. This strategy is called the two-stage Ward-MLM method and has been extensively used to classify maize accessions from most of the Latin American and United States gene banks (Taba et al. 1999) Later, Franco et al. (1999) extended the two-stage Ward-MLM method to the case of three-way data when attributes of genotypes are measured in various environments (attributes×genotypes×environments), following the Basford and McLachlan (1985) approach.

An important issue that arises in the context of using genetic markers for classifying individuals relates to the number of fragments and markers needed to provide an adequate clustering. Researchers are interested in knowing if the clusters formed using all the marker-fragment combinations can be obtained by using a reduced number of marker-fragments. In other words, what is the minimum number of fragments (and markers) needed to find the underlying structural pattern of the population? Certainly, this depends on many factors such as the genetic and phenotypic diversity of the individuals, the number of individuals, the type of genetic markers used to characterize the individuals, and the similarity coefficient, etc. The precise answer to this question can only be obtained by performing a retrospective (a posteriori) analysis.

In genetic-resources conservation, in pre-breeding and breeding activities, and in the identification of unique and distinct genotypes, finding the minimum number of fragments (and markers) that differentiate the individuals of a given sample is useful when combined with available phenotypic attributes. Clusters can be formed using the mixture of categorical (marker-fragment) and continuous (morpho-agronomic) attributes by means of the MLM proposed by Franco et. al (1998). The main problem is that, with a large number of marker-fragment attributes combined with several phenotypic variables, the large number of parameters that need to be estimated ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for each of the *g* groups) can be a serious impediment for using the MLM or any other statistical model. Thus, in this context, reducing the number of marker-fragments is necessary. The approach presented in this study addresses this issue.

This study has two objectives: (1) to present a retrospective method of analysis for selecting a minimum number of fragments (and markers) that will produce the same (or nearly the same) classification of the individuals as that obtained using all the fragments in the study, and (2) to show that the relevant minimum set of marker-fragments obtained using the retrospective analysis can be combined with the continuous variables to obtain compact groups of genotypes that can be well-characterized based on the two kinds of attributes included in the analysis.

Materials and methods

A retrospective analysis for selecting a minimum set of marker-fragments

When genetic similarities (or dissimilarities) of *n* individuals are measured with DNA marker methods, such as RFLPs or AFLPs using *m* markers, the data can be arranged in a matrix of *n* observations (genotypes) and *k* variables (fragments). The *k* columns are called fragments or morphs, which may correspond to specific markers (in this study we will use the terms fragments and markers). The *k* fragments are grouped into *m* markers, and each marker group has a size k_i (each fragment belongs to one and only one marker), such that $\sum_{i=1}^{m} k_i = k$.

The proposed strategy for finding subsets of fragments consists of the following steps: (1) classifying the individuals using all the available *k* fragments, (2) estimating the optimum number of groups (*g* clusters), (3) ranking the fragments according to their ability to discriminate the clusters found in step 2, (4) using the ranking of all the fragments (found in step 3) to search for the minimum number of fragments that will form equal (or very similar) clusters to those obtained when all the fragments were included, and (5) identifying the markers that correspond to the selected fragments.

Classifying the individuals

Distance between individuals

For each genotype, the fragments are binary attributes (taking values of 0 or 1). Two similarity (S) measures (or their complements, distances D=1-S) are commonly used: the Jaccard coefficient and the Simple Matching coefficient (Kaufman and Rousseeuw 1990). Expressed as distance, the Simple Matching coefficient is $d_{ij}=(b+c)/k$; and the Jaccard coefficient is $d_{ij}=(b+c)/(k–d)$, where d_{ij} is the distance between individuals i and j, *k* is the total number of fragments, (*b*+*c*) is the number of fragments in which individuals i and j disagree (*b*=1,0 and *c*=0,1), and *d* is the number of fragments in which individuals i and j agree at zero (0,0). The difference between the respective distances is that Jaccard assumes that the agreement (0,0) is non-informative.

If the binary attributes are treated as continuous variables and the squared Euclidian distance is computed, the result is the Simple Matching distance (Wishart 1987; Kaufman and Rousseeuw 1990) and thus it has Euclidean metric properties, whereas Jaccard does not. This property of the Simple Matching distance allows its use in hierarchical clustering strategies such as the minimum variance within a group, proposed by Ward (1963), and the analysis of molecular variance, AMOVA (Excoffier et al. 1992), used for estimation of the variance components among and within groups.

The AMOVA is an algorithm for the analysis of the genetic structure of populations based on molecular data. For a specific structure, the algorithm estimates the within-groups sum of squares as the sum of the squared Euclidean distances between each genotype and its group centroid; the total sum of squares is obtained as the sum of squared Euclidean distances between each genotype and the centroid of the whole sample; and the among-groups sum of squares is obtained as the difference between the 'total' minus the 'within'. Because the Simple Matching distances are squared Euclidean distances, the AMOVA sum of squares has an interpretation that is straightforward. For the AMOVA analysis, we used the ARLEQUIN software (Schneider, et al. 2000) with the matrices of distances used as input.

The strategy for searching and selecting the minimum subset of fragments uses the Ward method for clustering individuals and the AMOVA procedure for defining the appropriate number of groups by examining the variance components among and within groups. Therefore, the distance measure used in this study is the Simple Matching coefficient.

NTSYSpc (Rohlf 1997) software was used to compute Simple Matching similarity coefficients (S), which were then transformed into distances D=(1-S).

Clustering method

The clustering method used in this study was the minimum variance within groups proposed by Ward (1963). Ward's hierarchical method combines, in each step, the two clusters whose fusion yields the least increase in the Euclidean sum of squares within groups; this sum of squares is defined as the sum of the squared distances from each individual to the centroid of the cluster to which it belongs. Therefore, the variance between groups is maximized while the variance within groups is minimized. Due to the Simple Matching distance properties, the Euclidean sum of squares is the sum of the Simple Matching distances among all the individuals belonging to a cluster. As expressed by Wishart (1987), "Ward's method is only meaningfully defined for Squared Euclidean distance" (or the Simple Matching distance). The classification was carried out using the HIERARCHY routine of the CLUSTAN software (Wishart 1987) applied to the matrices of Simple Matching distances.

Defining the optimum number of clusters

A preliminary number of clusters was determined using the fusion values obtained from the Ward method and the upper-tail rule (Mojena 1977). This procedure was employed by Franco et al. (1997, 1998) in a sequential clustering strategy using phenotypic attributes. Then, AMOVA (Excoffier et al. 1992) was used to estimate the "among" and "within" variance components for different numbers of clusters around the number determined by the upper-tail approach. The final number of clusters was defined as those producing the largest increment in the among-groups variance (largest reduction in the within-group variance) and/or the maximum average distance among groups.

The F value (F_{st}) from the AMOVA analysis is the fixation index (or Wright's F statistic) that measures the genetic differentiation of the sub-populations (or sub-groups). Values of F_{st} between 0.05 and 0.15 indicate moderate genetic differentiation among groups, whereas values of F_{st} between 0.15 to 0.25 and above 0.25 indicate great and very great genetic differentiation, respectively (Hartl and Clark 1997).

Searching for the minimum subset of fragments

A generalized linear model (McCullagh and Nelder 1983) was used for modelling the dependent variable, which is the proportion of genotypes that take the value of 1 in each of the clusters defined previously (1 being the presence of the fragment). The independent variable is the group (cluster), as in the usual ANOVA model, and the response variable is the proportion of the j^th^ fragment in the i^th^ cluster. This model is repeated *k* times with the objective of estimating the differences among groups for each of the fragments.

Assume a classification that gives rise to *g* clusters (i=1,2, … ,*g*), each of them with n_i genotypes $\left(\sum_{i=1}^{g} n_i = n\right)$. The j^th^ fragment will take the value of 1 in x_{1j}, x_{2j}, … , x_{gj}, genotypes, where x_{ij} is the observed value of a random variable X_{ij}: the number of successes (genotypes taking the value 1) in n_i trials (the n_i genotypes belonging to the i^th^ group); X_{ij} is assumed to take a Binomial distribution with parameters(n_i, θ_{ij}). The proportion $p_{ij}=x_{ij}/n_i$ estimates the parameter θ_{ij} of the *g* binomial distributions. For each fragment the linear model is

$$E(p_{ij}) = \theta_{ij} = \alpha_j + \tau_{ij},$$

where θ_{ij} is the expected proportion of the j^th^ fragment in the i^th^ cluster [$E(p_{ij})$], α_j is the intercept term, and τ_{ij} is the effect of the i^th^ group or cluster. The above model can be fitted to test the null hypothesis that the proportion of genotypes in which the j^th^ fragment takes the value of 1 is the same in all the clusters, i.e., Ho: $\theta_{1j} = \theta_{2j} = \ldots = \theta_{gj}$, "the groups are equal with respect to the j^th^ fragment." This model is very simple, but presents some estimation problems such as the possibility of obtaining values of p_{ij} that can be greater than 1. To overcome this problem, it is recommended that one uses the link function $\text{logit}(p_{ij}) = \log[p_{ij}/(1-p_{ij})]$ within the Generalized Linear Model framework.

Using the GENMOD procedure of SAS (1993), it is possible to compute the chi-square values and the probability of Type-I errors ($\hat{\alpha}$) for the likelihood ratio test for testing the null hypothesis. The probability values of each fragment are ordered so that each fragment is ranked by its ability to discriminate between clusters.

Selecting the minimum number of fragments

In order to search for the minimum number of fragments that can produce the same (or at least a very similar) classification as that obtained with all of them, the individuals are again clustered, but in this instance using a reduced number of fragments selected by their significance level in the model previously described.

To measure the degree of coincidence between the classification using all fragments and that based on a reduced number, three indices were used: Rand, Jaccard and Corrected Rand (Milligan et al. 1983). These consensus indices between classifications are based on the number of agreements and disagreements found in both classifications. Let *a* and *d* be the number of pairs of individuals that agree in both classifications; they agree in the sense that both classifications locate that pair of individuals in the same cluster (*a*) or in different clusters (*d*). Further, let *b* and *c* be the number of pairs of individuals that are classified differently by each method.

The consensus indices are defined, based on the proportion of agreements, as follows:

$$\text{Rand} = \frac{a+d}{a+b+c+d}, \text{ Jaccard} = \frac{a}{a+b+c},$$

$$\text{and C-Rand} = \frac{a+d-\text{N}_c}{a+b+c+d-\text{N}_c},$$

where N_c is a correction factor due to chance and is defined as the expected value of the Rand index [E(Rand)], assuming that each table of agreements and disagreements is the realization of a random variable with a Hypergeometric distribution subject to a true classification for which the value of the index is zero.

The Rand index is simply the proportion of agreements over the total number of pairs (similar to the Simple Matching similarity), whereas the Jaccard index is the proportion of agreements of the type "both in the same cluster" over the total number of pairs, but excluding the agreements of the type "both in different group" (similar to the Jaccard similarity). The Corrected Rand (C-Rand) is the Rand index with the correction due to chance. Milligan et al. (1983) gave the formulas to compute these indexes based on frequencies.

In addition, the correlation coefficients between distance matrices were calculated. One distance matrix is obtained when all fragments are used, and the other distance matrix is obtained when a reduced number of fragments are employed. These correlation coefficients were calculated using the MXCOMP routine of NTSYSpc (Rohlf 1997). The correlation coefficients between pairs of distance matrices are the Pearson's correlation coefficients between distances computed for each pair of individuals using all fragments or a subset of them.

Missing values

Missing values are common and can arise for several reasons, such as the failure of an amplification or restriction reaction, low-quality DNA or old reagents, poor electrophoresis conditions causing blank lanes, or inability to score a lane with confidence. In this study, the distances were calculated using only complete information for each pair of individuals, without any correction.

It should be pointed out that the presence of a large number of missing values could cause distortions and inconsistencies in the process of reducing the number of fragments. Furthermore, missing values can be a problem for estimating the 'between'-and 'within'-group variances in the AMOVA. In this study, AMOVA was applied to the distance matrices, and although the missing values did not directly affect the results, they did influence the distance calculations.

Table 1 shows the percentage of missing values out of the total number of values and the percentage of genotypes with at least one missing value, for each experimental data set used in this paper.

Experimental data

Seven data sets from three crops (maize, wheat and tomato) and two types of genetic markers (RFLP and AFLP) were used. The data sets contained different numbers of genotypes, markers and fragments. Two data sets were characterized with two different types of markers, allowing straight comparisons between types of markers. Table 1 describes the seven data sets in terms of the number of genotypes, the number of fragments, the type of markers, and other characteristics.

Two maize data sets were included: maize diallel 1 (MD1) and maize diallel 2 (MD2). RFLP and AFLP markers were used to evaluate the genetic diversity of these diallel experiments. The MD1 data set included 16 lines evaluated for their level of drought tolerance: five Tuxpeno Sequia (TS) lines, four La Posta Sequia (LP) lines, and seven elite lines from different maize sources (CML) presenting different performance under drought conditions. The MD2 data set comprised 15 diverse elite maize lines used as testers in the different CIMMYT maize subprograms.

The tomato genetic diversity (TGD) contained a set of tomato genotypes, landrace accessions, and one wild species. The genotypes and landraces were from the United States, Europe, and Central and South America, including the Galapagos Islands, and the sample as a whole showed a considerable amount of genetic diversity.

Two wheat data sets were included. The bread wheat diversity set (WBD) is a set of sister lines and a few related controls. Although these are all highly related lines, they segregate for the presence of a translocated segment of chromosome 1 (1B/1R), so lines with the translocation contain part of a rye chromosome. The genetic diversity in this set is therefore very high, but it is a direct function of this chromosomal region only. This is probably why this data set shows a high coefficient of variation (Table 1). The WGD data set is formed using the AFLP genotyping of 72 Spring bread wheats. These included 32 CIMMYT varieties released between the 1960s and 1990s, and 40 genotypes from different countries, among which 13 are landraces. The WGD data represent a highly diverse group of genotypes, although it includes a few highly related sister lines.

The Modified Location Model for combining the reduced set of fragments with the phenotypic data

Non-hierarchical statistical methods for classifying individuals include the mixture models, such as the Gaussian Model (GM) (which only deals with continuous variables) (Wolfe 1970). For the GM and other mixture models, the initial groups (or subpopulations) must be defined a priori, and then the GM attempts to improve them by a maximum-likelihood iterative process that results in a solution that corresponds to a global or local maximum of the likelihood function. The initial groups can be defined in different ways. Franco et al. (1997) proposed, with a hierarchical method such as Ward (or UPGMA), using Gower's distance. Defining the initial groups by this strategy allows the use of all available information, including categorical and continuous variables, for defining the a priori groups. This sequential (two-stage) clustering strategy, called Ward-GM, was extensively used by Franco et al. (1998) in 29 data sets and proved to form more-compact and separated groups than the initial groups formed by the Ward method per se. However, for improving the groups using the GM, only continuous attributes can be utilized, and information contained in the discrete attributes cannot be incorporated into the GM model.

Lawrence and Kraznowski (1996) extended the GM for the mixture of categorical and continuous variables in a model called the Location Model (LM). They proposed combining all of the categorical variables into one multinomial variable, *W*, and combining the *W* variable with the continuous variables. The LM must have observations in every combination of the initial groups and all of the levels of the multinomial variable *W*. In practical applications for classifying genotypes it is very likely that this assumption is not fulfilled so usually the LM cannot be applied.

The Modified Location Model (MLM) of Franco et al. (1998) was used for combining the information on categorical and continuous variables, and overcomes the problem of empty cells for some combinations of initial groups and levels of the multinomial variable. The initial groups are defined by the Ward method and then improved by the MLM. This two-stage strategy is called the Ward-MLM. Details on the theory and applications of the Ward-GM and Ward-MLM methods are given in Franco et al. (1998, 1999).

The Ward-MLM strategy was applied to data from diallel 1 (MD1) and diallel 2 (MD2) using, as discrete variables, the relevant fragments that produced the same classification of lines as that obtained when all fragments were used. The three continuous variables included in MD1 were the grain yield of the 16 lines under severe drought stress (SS), intermediate drought stress (IS), and well-watered conditions (WW). In MD2, five phenotypic variables measured in each of the 15 lines included in this study were days to anthesis, days to silking, plant height, ear height, and grain weight.

Table 1 Type of marker (RFLP, AFLP), data set name (MD1, MD2, TGD, WBD, WGD), number of genotypes (nG), number of fragments (nF), and number of markers (nM). Diversity was measured by the total variance (V_{tot}), average distances between all pairs of genotypes ($\bar{d}_n$), and the coefficient of variation for the distances (CV_d). Percentage of missing values (%mis) of the total nG×nF values, and percentage of genotypes with at least one missing value (%Gmis)

Item	RFLP			AFLP			
	MD1	MD2	TGD	MD1	MD2	WBD	WGD
	Crop						
	Maize	Maize	Tomato	Maize	Maize	Wheat	Wheat
nG	16	15	110	18	17	96	72
nF	434	490	131	175	175	45	366
nM	54	48	39	6	6	4	8
V_{tot}	0.159	0.145	0.120	0.195	0.184	0.151	0.102
$\bar{d}_n$	0.294	0.284	0.228	0.355	0.358	0.268	0.194
CV_d	12.6	7.8	45.5	16.2	11.5	50.4	25.8
%mis	1.4	0.4	1.9	0.5	0.0	4.4	0.6
%Gmis	67	12	65	17	0	59	14

Results and discussion

Searching for a minimum set of genetic markers

Genetic diversity and the number of final clusters

According to the values of total variance (V_{tot}) and the average distances between pairs of genotypes ($\bar{d}_n$), the data sets with the greatest genetic diversity are maize MD1 and MD2, when studied with AFLPs (Table 1). The data sets with the smallest genetic diversity were tomato, TGD, when studied with RFLPs, and one wheat set, WGD, when studied with AFLPs. The data sets with intermediate genetic diversity were MD1 and MD2 with RFLPs, and WBD with AFLPs.

The data sets with a minimum coefficient of variation (CV_d) among distances were MD1 and MD2 with RFLPs and AFLPs (Table 1). These results indicate that, for these data sets, all pair-wise distances are close to the mean, $\bar{d}_n$. On the other hand, the high coefficient of variation among distances for WBD and TGD indicates the presence of some very different pair-wise genotypes. These results are logical because in WBD data some lines contained the 1B/1R translocation from rye and most of the polymorphisms found in the markers is found on this region, causing a very large differentiation in lines carrying the translocation from those lacking it. Furthermore, the TGD data set contained some wild species. In contrast, in data sets MD1 and MD2, all the pair-wise genotypes were, on average, highly dissimilar and thus produced a small CV_d but a large $\bar{d}_n$.

The consensus indices and correlations

Values close to 1 for any of the consensus indices used (Jaccard, Rand and C-Rand) correspond to a total agreement between the classification of the genotypes obtained with all the fragments, and the classification obtained with the fragments selected using the proposed strategy.

Results obtained from the consensus indices are consistent with those obtained with the correlation matrices (*r*), i.e., both increased as the number of fragments included in the classification increased (Table 2). The exception is the WGD data set, in which the indices decreased when the fragments included increased from 132 to 202.

Finding the minimum number of marker-fragment combinations

For the MD1 and MD2 with RFLP data sets, using only 2% and 3% of the total fragments, respectively, it was possible to obtain exactly the same classification of genotypes as that obtained using all the fragments with a Rand index, Jaccard and C-Rand indices of 100% (Table 2). For MD1 with RFLPs, there were nine fragments (four markers with two fragments, and one marker with one fragment). For MD2 with RFLPs, there were 15 fragments (four with two fragments, and seven with one fragment). For TGD with RFLP data, 5% of the fragments will recover 79%, 93% and 87% of the classification obtained with all the markers according to the Jaccard, Rand and C-Rand indices, respectively. Thus, a consensus of at least 90% measured by the Rand index is achieved with 2–5% of the RFLP fragments (Table 2). For these three RFLP data sets, the proposed strategy found that only a very low percentage of the total RFLP fragments give rise to classifications similar to those obtained with all fragments. Furthermore, for MD1 and MD2 studied with AFLP markers, the fragment selection strategy showed that with 3% and 7% of the total fragments, a 100% consensus can be achieved. However, this may not always be the case. Indeed, for WBD and WGD, 56% and 36%, respectively, of the AFLP fragments were required to obtain a consensus of 90%.

Table 2 Number of significant fragments (sF), percent of significant fragments (%sF), number of associated markers (sM), percent of associated markers (%sM), significance level at the given number of selected clusters (Sig.), Jaccard, Rand, and Corrected Rand indices of consensus between the classification obtained with all fragments and using only the significant fragments. Correlation coefficients between pairs of distance matrices(*r*)

Marker	Data	sF	%sF	sM	%sM	Sig.	Jaccard	Rand	C-Rand	*r*
RFLP	MD1	9	2	5	9	0.002	1.000	1.000	1.000	0.351
RFLP	MD2	15	3	11	23	0.010	1.000	1.000	1.000	0.485
RFLP	TGD	106[a]	–	–	–	–	1.000	1.000	1.000	0.999
		14[b]	–	7	–	<0.001	0.868	0.960	0.924	0.876
		7[b]	5	4	10	<0.001	0.791	0.930	0.872	0.868
AFLP	MD1[b]	5	3	3	50	<0.001	1.000	1.000	1.000	0.467
		4	–	3	–	<0.001	0.508	0.810	0.680	0.443
AFLP	MD2[b]	13	7	5	83	0.010	1.000	1.000	1.000	0.444
		12	–	5	–	0.007	0.561	0.868	0.759	0.447
AFLP	WBD	34	–	4	–	0.050	0.936	0.984	0.968	0.981
		25	56	4	100	<0.001	0.747	0.928	0.843	0.938
AFLP	WGD	202	–	8	–	0.160	0.463	0.844	0.643	0.930
		132	36	8	100	0.050	0.573	0.899	0.743	0.622
		98	–	8	–	0.010	0.446	0.853	0.609	0.522

[a] Only 106 markers of a total of 131 presented variation (polymorphism)
[b] These rows are related to the classification based on 106 fragments

In a data set of 218 maize inbred lines (made up of about half of the major tropical, subtropical, and temperate CIMMYT maize germplasm) studied with 34 RFLP markers that identified 314 fragments, the fragment-selection strategy found that at least 64% of fragments were required in order to achieve a 90% consensus measured by the Rand index (data not shown). This indicates that when the genetic diversity of the included germplasm is large, or the number of individuals is high, the strategy still finds the required key fragments (and markers), although there may be additional key fragments in these cases.

It is interesting to note that of the five and 11 selected RFLP markers for MD1 and MD2, respectively, three were the same for both data sets. These results indicate that some markers may be better discriminators of genotypes, regardless of the types of genotypes included, and also that this strategy for selecting the relevant markers is capable of detecting them. The different gemplasm included in the study (maize, tomato and wheat) may have also played a role in the minimum set of fragments required for having the same classification of genotypes as that obtained by using all of the fragments.

Combining the information of the reduced set of fragments of diallel 1 and diallel 2 and the phenotypic attributes

The nine and 15 RFLP genetic fragments selected for MD1 and MD2, respectively, together with their corresponding continuous variables, were combined using the Ward-MLM strategy. To compare results, two other classifications were performed in each diallel data set. One classification was based only on the relevant marker-fragments using the Ward method per se. The other classification was based on the Ward-GM strategy, but the GM was applied based only on the continuous variables (without including the relevant marker-fragment data).

For diallel 1 (MD1), the Ward (D) identified three groups, whereas the Ward-GM (C) and the Ward-MLM (M) methods found four groups (Table 3). In Table 3, fragment 1 belongs to one marker and fragments 2–3, 4–5, 6–7 and 8–9 belong to four different RFLP markers. Based on the Ward-MLM (M) strategy, group 1 clustered the four TS lines and five CML lines with low grain yield performance on severe drought stress (SS) and intermediate drought stress (IS) as well as well-watered (WW) environmental conditions. With respect to the nine relevant RFLP fragments, lines of group 1 seem to have a fairly consistent response in terms of the absence/presence of the fragments; most of them were scored 0 for the fragments, except for fragment 5 with which most lines were scored for presence (1). Group 2 of the Ward-MLM strategy included two CML lines with high grain yield values in the three environments, but no clear pattern of response with regard to the nine fragments. Group 3 of the Ward-MLM strategy had one LP line that was the highest yielding line in IS and WW. Group 4 of the Ward-MLM strategy comprised four LP lines with intermediate grain yield in SS, IS and WW, and a highly consistent pattern of absence/presence on the nine fragments; contrary to group-1 lines, these lines showed the presence of most of the fragments, except fragment 5.

The average Mahalanobis distance between groups using the continuous variables showed that when the groups are formed based on only the continuous variables, they are well separated(D^2=61). When the groups are formed based on only the discrete variables, they do not show a clear separation, (D^2=3); but when the groups

Table 3 Values of grain yield in severe drought stress (SS), intermediate drought stress (IS) and well-watered (WW) environments, and nine relevant RFLP marker-fragments (1–9), for groups of 16 maize lines of diallel 1 (MD1) obtained using the Ward-MLM strategy on the mixture of variables (M), using the Ward-GM strategy for the three continuous variables (C), and using the Ward method for the nine relevant RFLP marker-fragments (D)

Entry	M	C	D	SS	IS	WW	Relevant fragment								
				– t/ha –			1	2	3	4	5	6	7	8	9
TS2	1	1	1	0.016	1.223	0.630	0	0	0	0	1	0	0	0	0
TS5	1	1	1	0.002	0.437	0.774	0	0	0	0	1	0	0	1	0
CML247	1	1	1	0.008	0.303	0.525	0	0	0	0	1	0	0	0	0
CML264	1	1	1	0.012	0.433	1.589	1	0	0	0	1	0	0	0	0
CML268	1	1	1	0.007	0.167	1.858	0	0	0	0	0	0	1	0	0
TS1	1	3	1	0.055	1.730	2.805	0	0	0	0	1	0	0	0	0
TS4	1	3	1	0.004	1.207	3.007	0	0	0	0	1	0	0	0	0
CML258	1	3	1	0.055	1.170	2.805	0	0	0	0	1	0	1	0	0
CML273	1	1	2	0.183	0.910	1.980	0	0	0	1	1	0	1	1	0
CML274	2	2	2	0.527	1.940	5.282	0	0	0	1	1	0	1	1	0
CML254	2	2	1	0.676	1.883	4.121	0	0	0	0	0	0	0	0	0
LP1	3	4	3	0.293	4.713	6.171	1	0	1	1	0	0	1	1	1
LP2	4	3	3	0.076	2.043	5.333	1	1	0	0	0	1	1	1	1
LP3	4	3	3	0.286	1.440	3.131	1	1	1	0	0	1	1	1	1
LP4	4	3	3	0.313	1.467	4.271	1	1	1	1	0	1	1	1	1
LP5	4	3	3	0.072	2.047	3.261	1	1	1	1	0	1	1	1	1
D^{2}[a]	45	61	3												
D_{disc}[b]	30	5	32												

[a] D^2: average Mahalanobis distance (using only the continuous attributes) between groups formed based only on nine RFLP marker-fragment, based on only three continuous variables, and based on nine RFLP marker-fragments and three continuous variables

[b] D_{disc}: average distance between groups (using only the discrete variables) formed based only on nine RFLP marker-fragments, based on only three continuous variables, and based on nine RFLP marker-fragments and three continuous variables

are formed based on both variables – discrete and continuous – the separation of the groups is fairly good (D^2=45). On the other hand, the average distance between groups using the discrete variables showed that when the groups are formed based on only the discrete variables, they are well differentiated (D_{disc}=32). When the groups are formed based on only the continuos variables, the separation is poor (D_{disc}=5); but when the groups are formed based on both variables – continuous and discrete – the groups are well separated (D_{disc}=30) and as good as when only the discrete variables are considered. This result indicates that the classification using both continuous and discrete variables produces well-defined and separated clusters.

It is interesting that two of the five RFLP markers that were selected, marker of fragments 5–6 and marker of fragments 7–8, mapped in genomic regions identified as being involved in the expression of a good synchrony mechanism between pollen shedding and silking emergence under drought conditions (Ribaut et al. 1996). This floral mechanism, known as the anthesis-silking interval (ASI), is one of the most important indicators of drought tolerance in maize. In addition, it has been observed in other maize crosses that the marker of fragment 1 also mapped on the top of the short arm of chromosome 7 which is also a key genomic region involved in the maize drought response (unpublished data).

For diallel 2 (MD2), the Ward (D), Ward-GM (C) and Ward-MLM (M) strategies identified five groups (Table 4). Based on Ward-MLM, lines in group 1, and to some extent lines in group 2, are low yielding and early maturing as compared to lines in group 4, which had the highest mean yield, late maturity, and were taller. The patterns on the relevant marker-fragment showed that for some fragments, lines in group 4 were scored for the presence of most of the 15 fragments, whereas lines from groups 1 and 2 tended to be scored for their absence. For example, fragments 11, 13 and 14 are present in the three lines of group 1 and absent in the four lines of group 4. By contrast, fragment 4 is absent in lines of group 1 and present in lines of group 4.

Similar to diallel 1, in diallel the average distance between groups based only on the continuous variables (D^2) and based only on the discrete variables (marker-fragment) (D_{disc}) balanced out the influence of the discrete and continuuos variables and gave both types of attributes the chance to contribute to the formation of the groups. Similar to previous results, the use of the Ward-MLM (M) produced compact and well-separated groups with respect to all continuous and categorical variables compared with classifications obtained based only on categorical (D) or continuous (C) variables.

In summary, the results of this study indicate the advantage of classifying genotypes simultaneously, and including categorical and continuous variables, in order to obtain a good formation of clusters. The groups are well-characterized based on both sets of variables.

Table 4 Values of days to anthesis (ANTH), days to silking (SILK), plant height (PLHT), ear height (EARHT) and grain weight (GRAIN), and 15 relevant RFLP marker-fragment (1–15), for a group of 15 maize lines of diallel 2 (MD2) obtained using the Ward-MLM strategy on the mixture of variables (M), using the Ward-GM strategy for the five continuous variables (C), and using the Ward method for the 15 relevant RFLP marker-fragments (D)

Entry	M	C	D	ANTH	SILK	PLHT	EARHT	GRAIN	Relevant fragment														
				– days –		– cm –		– t/ha –	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
CML202	1	1	1	87.6	91.3	97.0	36.3	0.726	0	1	1	0	0	1	0	1	0	0	1	0	1	1	1
CML216	1	3	1	87.4	94.4	119.3	56.3	0.453	1	1	1	0	0	1	0	1	0	0	1	0	1	1	0
CML311	1	1	5	83.8	88.0	101.6	41.9	0.701	0	0	1	0	1	0	1	0	1	0	1	0	1	1	1
CML206	2	2	2	89.9	96.2	110.0	38.4	0.779	0	0	1	0	0	0	0	1	0	1	0	1	0	0	1
CML321	2	1	3	84.9	87.0	113.6	41.5	1.328	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0
CML78	2	1	2	83.3	84.7	103.6	31.6	0.843	1	1	0	1	0	0	0	0	0	1	0	0	1	0	1
P501	2	2	5	89.4	91.5	113.6	42.5	1.525	0	1	0	1	1	0	0	1	0	1	0	1	1	0	1
P502	2	1	2	86.1	87.8	97.3	40.1	1.549	0	1	0	0	0	0	0	1	0	1	0	1	1	0	1
CML247	3	2	3	92.9	96.7	100.3	35.7	0.776	0	0	0	1	0	0	1	0	1	0	0	1	1	0	0
CML254	4	4	4	90.7	91.3	120.3	50.4	1.875	1	0	1	1	1	0	1	0	1	0	0	1	0	0	1
CML258	4	4	4	90.8	94.3	117.3	50.1	1.139	1	0	1	1	1	1	1	0	0	1	0	1	0	0	0
SPLC7	4	4	4	86.8	87.7	126.6	47.9	1.381	1	0	1	1	1	1	1	0	1	0	0	1	0	0	0
CML339	4	5	4	91.3	92.3	119.3	37.2	1.395	1	0	1	1	1	1	1	0	0	1	0	1	0	0	0
CML341	5	4	5	87.9	88.0	115.0	53.7	1.349	0	0	0	0	1	0	0	0	1	0	1	0	1	0	1
CML343	5	5	5	91.5	92.3	106.6	36.6	1.203	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1
D^2 [a]	45	61	3																				
D_{disc} [b]	30	5	32																				

[a] D^2: average Mahalanobis distance (using only the continuous attributes) between groups formed based only on 15 RFLP marker-fragments, based on only three continuous variables, and based on 15 RFLP marker-fragments and three continuous variables

[b] D_{disc}: average distance between groups (using only the discrete variables) formed based only on 15 RFLP marker-fragments, based on only three continuous variables, and based on 15 RFLP marker-fragment and three continuous variables

Conclusions

The retrospective method for selecting the most important fragments (and markers) for the classification of individuals into homogeneous clusters found few relevant fragments (and markers) that “best” discriminated the individuals. It is noteworthy that, although a substantial reduction in the required number of fragments and markers was generally achieved by the proposed strategy, there may be cases in which, due to the large genetic variability of the germplasm, a large proportion of the total fragments may be required to achieve an appropriate classification of the germplasm.

Furthermore, the results of this study suggest that a large number of RFLP fragments are not informative for the purpose of discriminating genotypes. Thus, when no information on the ability of the fragments (and markers) to discriminate genotypes is available, a larger number of these markers must be used. Although the relevant fragments identified in each experiment are germplasm-dependent, and therefore the markers associated with those fragments must be defined for each new analysis, the identification of the number of significant fragments by the proposed retrospective analysis can be used to predict the number of markers that should be considered in future studies, when working with the same type of germplasm and the same range of diversity.

This study shows that when simultaneously using genetic markers and phenotypic variables to classify genotypes, it is possible to obtain a relevant minimum subset of marker-fragments that can be used in conjunction with available morpho-agronomic data to better classify genotypes, compared to using only the continuous or only the discrete variables.

Acknowledgements The authors appreciate the contribution of Isabel Almanza and Karen Aitken in producing some of the data sets analyzed in this study.

References

Basford K, McLachlan GJ (1985) The mixture method of clustering applied to three-way data. J Classif 2:109–125

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred for metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Franco JE, Crossa J, Díaz J, Taba T, Villaseñor J, Eberhart SA (1997) A sequential clustering strategy for classifying gene bank accessions. Crop Sci 37:1656–1662

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

Franco J, Crossa J, Villasenor J, Castillo A, Taba S, Eberhart SA (1999). A two-stage, three-way method for classifying genetics resources in multiple environments. Crop Science 39:259–267

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–874

Hartl DL, Clark AG (1997) Principles of population genetics (3rd edn). Sinauer Associates, Sunderland, Massachusetts

Kaufman L, Rousseeuw PJ (1990) Finding groups in data. John Wiley and Sons, New York

Lawrence CJ, Krzanowski WJ (1996) Mixture separation for mixed-mode data. Stat Comput 6:85–92

McCullagh P, Nelder JA (1983) Generalized linear models. Chapman and Hall, London

Milligan GW, Soon SC, Sokol LM (1983) The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol PAMI-5, No.1, pp 40–47

Mojena R (1977) Hierarchical grouping methods and stopping rules: an evaluation. Comput J 20:359–363

Ribaut J-M, Hoisington DA, Deutsch JA, Jiang C, Gonzalez-de-Leon D (1996). Identification of quantitative trait loci under drought conditions in tropical maize. I. Flowering parameters and the anthesis-silking interval. Theor Appl Genet 92:905–914

Rohlf FJ (1997) NTSYSpc, numerical taxonomy and multivariate analysis system, version 2.02i. Applied Biostatistics Inc, New York

SAS Institute Inc (1993) SAS technical report p-243, SAS/STAT software: the GENMOD procedure, release 6.09, SAS Institute Incorporated, Cary, North Carolina

Schneider S, Roessli D, Excoffier L (2000) Arlequin ver. 2.000: a software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Taba S, Díaz J, Franco J, Crossa J, Eberhart SA (1999) A core subset of LAMP, from the Latin American Maize Project. CD-rom. CIMMYT, Mexico DF, Mexico

Ward J (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Wishart D (1987) CLUSTAN user manual, 3rd edn. Program Library Unit, University of Edinburgh, Scotland

Wolfe JH (1970) Pattern clustering by multivariate mixture analysis. Multivariate Behav Res 5:329-350