RESEARCH ARTICLE

Classification of Peruvian highland maize races using plant traits

R. Ortiz · José Crossa · Jorge Franco · Ricardo Sevilla · Juan Burgueño

Abstract The maize of Latin America, with its enormous diversity, has played an important role in the development of modern maize cultivars of the American continent. Peruvian highland maize shows a high degree of variation stemming from its history of cultivation by Andean farmers. Multivariate statistical methods for classifying accessions have become powerful tools for classifying genetic resources conservation and the formation of core subsets. This study has two objectives: (1) to use a numerical classification strategy for classifying eight Peruvian highland races of maize based on six vegetative traits evaluated in two years and (2) to compare this classification with the existing racial classification. The numerical classification maintained the main structure of the eight races, but reclassified parts of the races into new groups (Gi). The new groups are more separated and well defined with a decreasing accession within group × environ-ment interaction. Most of the accessions from G1 are from Cusco Gigante, all of the accessions from G3 (except one) are from Confite Morocho, and all of the accessions from G7 are from Chullpi. Group G2 has four accessions from Huayleño and four accessions from Paro, whereas G4 has four accessions from Huayleño and five accessions from San Geronimo. Group G5 has accessions from four races, and G6 and G8 formed small groups with two and one accession each, respectively. These groups can be used for forming core subsets for the purpose of germplasm enhancement and assembling gene pools for further breeding.

Keywords Ear height · Leaf length · Leaf number · Leaf number above the ear · Leaf width · Modified Location Model · Plant height

R. Ortiz (✉) · J. Crossa · J. Burgueño
International Maize and Wheat Improvement Center
(CIMMYT), Apdo. Postal 6-641, Mexico, DF, Mexico
e-mail: r.ortiz@cgiar.org

J. Franco
Facultad de Agronomia, Universidad de la Republica
Oriental del Uruguay, Avd. Garzon 780, CP 12900
Montevideo, Uruguay

R. Sevilla
Universidad Nacional Agraria, Av. La Universidad s/n, La
Molina, Lima, Peru

Introduction

The maize of Latin American with its enormous diversity has played an important role in the development of modern maize cultivars of the American continent. To understand the history of the different indigenous maize cultivars and the preservation of the variability within the genus is of fundamental importance in maize improvement. The maize races of Latin American have been

extensively studied and classified mainly based on specific traits related to grain characteristics such as grain shape, color, and texture. Maize accessions (landraces and cultivars) are classified into races to facilitate the use of their genetic variation (Anderson and Cutler 1942). A race consists of accessions that share like traits that allow their grouping together based on their genes, which are kept within the race through random mating, that control such characteristics. Maize races are defined mostly by characteristics such as vegetative traits, tassels, ears, and kernels, in which group accessions show similar phenotypes.

The germplasm complexes of Mexico were initially classified by Wellhausen et al. (1952), who suggested that the race Tuxpeño was introgressed into the Corn Belt maize through its Southern Dent parentage. In general it is agreed that racial classification should be done using traits that reflect, as much as possible, genetic differences between entries rather than traits highly influenced by the environment or genotype × environmental interaction. Taba et al. (1998) have classified Caribbean maize accessions with the purpose of forming non-overlapping groups and to find patterns of phenotypic diversity within clusters so that representative core subsets of accessions can be selected.

Landraces from Uruguay were classified using a statistical classifications method and a visual assessment of ear and grain characters (Gutierrez et al. 2003). Two criteria for comparing these two classification method were used: (1) to determine the continuous and categorical variables that better discriminated the groups or races, and (2) to determine the distance between the group or races. Results showed that the races were almost entirely maintained by the numerical classification, except one race that was redistributed into two groups. The authors justified these results based on the geographical origin of the accessions that were separated and concluded that the statistical classification obtained in this study gave a more detailed insight into the relationships among Uruguayan maize landraces as compared to visual racial classification.

Peruvian highland maize shows a high degree of variation stemming from its history of cultivation by Andean farmers (Grobman et al. 1961). Abu-Alrub et al. (2004) studied the variation in kernel and ear descriptors of 12 Peruvian highland maize races to identify the best morphological descriptors to characterize and classify such germplasm. The authors found that kernel traits are the best descriptors. Some races were well classified, but others overlap due to possible interactions and gene flow among races.

Plant genetic resources are the most fundamental component of all agricultural systems, thus preservation, evaluation, and enhancement of germplasm is an essential activity. With advances in computing technology, numerical taxonomy and multivariate statistical methods for classifying individuals have become powerful tools for classifying genetic resources conservation and the formation of core and mini core subsets. There are several classification methods that can be used in genetic resources conservation, in the formation of core subsets, and in defining possible heterotic groups. In all these methods the main objective is to select accessions that best represent the entire collection with the minimum loss of genetic diversity. Therefore, the best numerical classification strategy is the one that produces the most compact and well-separated groups (i.e., minimum variability within each group and maximum variability among groups). Lawrence and Krzanowski (1996) used the Location Model (LM), originally proposed by Olkin and Tate (1961), for classifying n individuals when p continuous traits and q categorical traits are measured in one environment. The LM combines the levels of all q categorical traits in one multinomial variable, W, with m levels ($w = 1,2,...,m$). Franco et al. (1998) modified the LM and proposed the Modified Location Model (MLM), by assuming that the m levels of the W variable and the p-multinormal variables, for each subpopulation, are independent. Franco et al. (1998) proposed using the MLM in a two-stage clustering strategy, where in the first stage, initial groups are defined by a hierarchical method known as the Ward method. The second stage consists of improving these groups by means of the MLM. This classification strategy has the following advantages: (1) the process responds to the optimization of two related objective functions, in the first stage the sum of squares within groups and in the second stage the likelihood function of the observations; (2) it is linked to a method for

defining the optimum number of groups; (3) it allows calculation of a measure of the group's precision (quality) because it assigns to each observation a probability of membership to the group; and (4) it uses all the available information, that is, the continuous variables as well as the categorical variables. The use of all the information produces better classification results than the use of only part of the data. The distances between groups are maximized and better differentiated, and more compact groups are obtained (Franco et al. 1998).

Franco et al. (1999) extended the Ward-MLM strategy to the case of clustering three-way data of accession × environment × trait. Thus, the vector of each observation for r environments is $1 \times (rp + 1)$. The three-way Ward-MLM clustering strategy considers the same trait measured in different environments as different variables (environment-trait combinations). Since the Ward-MLM strategy clusters individuals with consistent response across all variables in all the environments, it is reasonable to expect that the three-way cluster strategy should form groups with negligible group × environment interaction for all or some of the traits included in the study. Franco et al. (2003) studied the use of the three-way Ward-MLM clustering strategy in three different data sets for clustering cultivars into groups with low GEI and for studying how the relationship between a pair of traits changes across environments. The three-way Ward-MLM strategy produced groups of cultivars with low interaction with the environment. The increment of the correlation coefficient values between-groups with respect to the total correlation coefficients indicated that the groups formed by the three-way Ward-MLM strategy comprised subsets of individuals that had similar trait responses across environments. A detailed summary of the Ward-MLM statistical methodology for classifying genetic resources have been recently presented by Crossa and Franco (2004).

The main objectives of this study were: (1) to complement the racial classification of eight Peruvian highland maize races obtained through visual assessment with a numerical classification based on a sequential Ward-MLM statistical classification strategy using six vegetative traits measured in two consecutives years, and (2) to compare this classification with the existing racial classification in term of distance between groups and between races.

Materials and methods

Plant material

Eight Peruvian highland maize landraces were evaluated using two consecutive years at one site (Carhuaz, Ancash, northern Peruvian highlands, within the inter-Andean valley of the Callejón de Huaylas) (Table 1). The eight races are Confite Morocho (R1), Chullpi (R2), Uchuquilla (R3), Cusco Gigante (R4), Huayleño (R5), Paro (R6), San Gerónimo-Huancavelicano (R7), and Shajatu (R8). A detail description of these races can be found in Grobman et al. (1961) and Manrique Chavez (1997). Confite Morocho (R1) is a primitive race. Huayleño (R5), Chullpi (R2), Paro (R6), Shajatu (R8), Uchuquilla (R3) are anciently derived races. San Gerónimo-Huancavelicano (R7), Cuzco Gigante (R4) are lately derived races. However, some accessions of San Gerononimo can be considered an imperfectly defined with short plant, relatively early, with few leaves, 9 to 11, short and narrow, ear position very low. Populations derived from San Gerónimo are often found variable ear types, which show a similarity to Cuzco, to Huancavelicano, or to San Gerónimo-Huancavelicano. A brief description of these eight races (after Grobman et al. 1961) follows.

Confite Morocho (R1)

Is a very short race with an average number of leaves 11.9, length of the ear leaf of 53.4 cm; width 6.3 cm; very narrow, leaf area small; ear located at the middle of the stalk at a height of 56 cm. This race shows very slender white cobs with 12 mm diameter on average and the highest (2.3) cob/rachis index among all Peruvian races. Its small, usually pointed or beaked upward kernels (although some accessions show round shape) with striations but without denting arise from spreading spikelets. The flinty yellow endosperm belongs to the pop-type, and the aleurone is colorless.

Table 1 Peruvian highland maize accessions distributed according to race and geographical origin—Peruvian departments are listed from northwest (left) to southeast (right)

Peruvian maize race	La Libertad	Ancash (2400–3000)	Huancavelica (3000–3600)	Ayacucho (2400–3000)	Apurimac (2400–3000)	Cusco (2400–3000)	Puno (>3000)
Confite Morocho				Composite Aya 4, Aya 20, Aya 38	Apuc 54		
Cusco Gigante		Composite, Anc 617, Anc 622			Apuc 111	Cuz 195, Cuz 201, Cuz 218, Blanco Urumbamba, Blanco Urubamba Est. Fen.	
Chullpi			Composite, Hvca 118, Hvca 146, Chullpoide			Cuz 175, Cuz 207	
Huayleño		Composite, Anc 181, Anc 382, Anc 395, Anc 398, Anc 399, Anc 430, Comp. Terciopelo, Terciopelo					
Paro					Composite, Apuc 96, Apuc 97, Apuc 113		
San Gerónimo			Composite, Hvca 28, Hvca 52, Hvca 119, Hvca 134, San Gerónimo				
Shajatu	Lib 87	Composite, Composite Precoz, Anc 120, Anc 152, Anc 601					
Uchuquilla		Anc 544					Composite, Pun 15, Pun 16, Pun 18

Altitude ranges (m, approx.) are given in brackets below the name of each department

Chullpi (R2)

Tall, attaining greater heights than other Sierra races, medium long and narrow leaves, intermediate in numbers, leaf area low, ear located at a short distance above center of the plant, on a medium to wide stalk. The extremely long (16.5 mm), relatively wide, intermediately thick, sugary (*su* gene), beaked kernels with a depression in the cap and without striations appear to be irregularly arranged in 18 rows (if re-arranged), and show a yellow endosperm (in typical Chullpi accessions) and colorless aleurone. Its white, brown, red, variegated red or mosaic red cobs have on average 29.2 mm diameter, and this race shows an intermediate (1.6) cob/rachis index. Paro is a Sierra race very closely related to Chullpi, and not certain

whether Paro and Chullpi developed independently from a common ancestor or whether one preceded the other. The fact that Chullpi, Huayleño and Paro, postulated as descending from a common popcorn ancestor Confite Chavinense, are all consumed as kcancha, the closest form of utilization to popping, is suggestive of their racial relationship, and possibly of the parallel development of these races under selection for a common form of utilization.

Uchuquilla (R3)

Short, average height 1.33 m, low number of short and narrow leaves. Its medium length, width and thickness, non-imbricated, flinty (with soft starchy center) kernels show slight denting with medium striation, and yellow to orange endosperm (in the external cell layers). The white, purple or red cobs have on average 18.6 mm, and a medium (1.64) cob/rachis index. Uchuquilla originated from Kcarapampa.

Cuzco Gigante (R4)

Medium tall, average number of leaves 10.5, of medium width and length. Cuzco Gigante is a hybrid race selected from the original Cuzco complex . This race has the largest kernels of any known maize race. The long, wide and very thick kernels show a slight to medium depression, and slight striation, and possess white floury endosperm, and usually a colorless aleurone. The white, cherry, variegated red, red mosaic or brown cob has a 22.4 mm diameter on average, and a medium (1.68) cob/rachis index. It shares with related races Huancavelicano, and Uchuquilla common morphological ear features.

Huayleño (R5)

Very short, average number of leaves: 10.7, intermediate in length and width, leaf area low; 5.3. leaves above the ear; ear located at an average of 87 cm above ground. Huayleño accessions possess short, wide, ovoid-conical ear with irregular arrangement of white, floury kernels that lack any depression or striation, white soft endosperm, and mostly colorless aleurone. The cob shows a variable color with a 21 mm-diameter on average, and a high (1.84) cob/rachis index. Huayleño is a direct floury maize derivative of Confite Chavinense.

Paro (R6)

Short, average height 1.24 cm, low average number of leaves 9.3, 4.5 of which are above the ear node, leaf area intermediate to low, leaf length and width intermediate. The very long and small width kernels show intermediate thickness, terminating in a pronounced beak and exhibited a high degree of imbrication. The kernels, which also show a soft, floury endosperm, are very loosely attached to a very variable color cob, and therefore shell-off easily. The cob average diameter is 29.4 mm and this race shows a low (0.31) cob/rachis index. Hybridization of Chullpi with Confite Puntiagudo could very well have been the starting point for the origin of Paro. Alternative hypothesis that Chullpi and Paro are sister races, branching off from some common ancestor as the result of differential selection.

San Gerónimo-Huancavelicano (R7)

Plants short, low number of short and narrow leaves, leaf area low, ear position low on fifth node at an average height of 50 cm above ground, stalk slender. This race shows long, wide and thick, non-imbricated kernels with medium denting and slight striations, whose endosperms are white, floury and fairly soft, and an often colorless aleurone. The white, brown, red or variegated red cobs have 22.7 mm diameter on average, and a low (1.3) cob/rachis index. San Gerónimo-Huancavelicano is apparently an incipient new synthetic race which arose through hybridization of Huancavelicano with other sympatrically distributed races, particularly Paro and to lesser extent Chullpi.

Shajatu (R8)

Medium short, average height 1.52 m, low number of leaves 9.5, with 5.1 of them above ear node; length and width of leaves medium, leaf area medium, ear borne on eight node, at an average height of 0.8 m from the ground. The accessions belonging to this race show a non-imbricated, medium size (in all dimensions), floury and very soft kernels without striations having a white endosperm and purple aleurone. The cobs can be white, light brown and red with an average diameter

of 25.6 mm and medium (1.72) cob/rachis index. Shajatu is evidently very closely related to Ancashino, from which it probably originated by selection for the purple aleurone character. (Huayleño and Ancashino are postulated as sister races arising at the same time)

Different numbers of accessions were used for each race, making a total of 50 accessions (Table 1) arranged in a randomized complete block design with four replicates each. Data were recorded for 10 plants per replicate. The plant traits recorded were plant height (cm) (PH), ear height (cm) (EH), leaf number (LN), leaf number above ear (LE), leaf length (cm) (LL), and leaf width (cm) (LW). The combination of the six traits measured (PH, EH, LN, LNE, LL, and LW) in two consecutive years provides the 12 variables used for clustering the accessions of the different races which will be abbreviated as PH1, EH1, LN1, LE1, LL1, LW1, PH2, EH2, LN2, LNE2, LL2, and LW2.

Two-stage Ward- MLM classification method

The two-stage Ward-MLM strategy was proposed by Franco et al. (1998), where the initial groups are generated using the Ward (1963) minimum variance within-groups hierarchical method. Then the MLM, using the Ward's groups as the starting (initial) point, is applied with the objective of improving the classification of the observations to those groups. The initial groups formed by the Ward technique are based on the principle of minimum variance within group, forming spherical groups. Then the mixture distribution method (MLM) acts upon the previous clusters, particularly regarding shape, direction, and volume of the clouds of points that make up the groups in the p dimensional space. The maximization of the likelihood function begins at a point that has been reached using the geometric technique and it will then reach a peak (which could be local) near the starting point that contains the characteristics of the Ward technique. Since the geometric technique uses all available information, it will be useful if the mixture distribution model procedure also uses all the available information.

Three-way Ward- MLM classification method

Consider a random $n \times rp$ matrix of n observations where, for each observation, p traits are measured in each of the r environments. This forms a matrix with n rows (observations) and rp columns. These rp environment-trait combinations will be named 'variables.' Using the two-stage Ward-MLM approach, it is possible to avoid the ''independence across sites'' assumption and therefore to estimate variance of the traits within sites and their covariances across sites, provided that n (if homogeneity for the variance-covariance matrix is assumed for each subpopulation), or n_i (if heterogeneity for the variance-covariance matrix is assumed for each subpopulation) is greater than $rp + 1$ (Mardia et al. 1979). The continuous and discrete traits are combined into an $rp + 1$ vector that contains the rp values of the continuous variables plus the values of the multinomial variable $W = s$, which combines the information from the categorical traits in all the environments.

Parameter estimation by maximum likelihood, the probability of membership for each observation in each subpopulation to be used in the expectation maximization (EM) algorithm (Dempster et al. 1977), and other theoretical details of the three-way Ward-MLM are shown in Franco et al. (1999).

Determining the optimum number of clusters

The number of groups is defined using the pseudo-F and pseudo-t^2 criteria (SAS, 1999) combined with the likelihood profile associated with the likelihood-ratio test (Mardia et al. 1979). The pseudo-F and pseudo-t^2 criteria (SAS, 1999) founded to be the best two criteria out of 30 by Milligan and Cooper (1985), and the likelihood profile, associated with the likelihood ratio test (Mardia et al. 1979) were used to determine the optimal number of groups. The likelihood profile is used as a graphical display for observing the changes to the log-likelihood function in relation to the number of groups. The optimal number of clusters occurs when the log-likelihood function shows its highest increase.

The likelihood ratio test requires the support of a statistical model, and can only, therefore, be applied in that case. The idea is to test the null hypothesis, H_0, which states that the number of components in a mixture of distributions (groups) is g, versus the alternative hypothesis, H_a, that the number being $g'(g' > g)$. Wolfe (1971), based on Monte Carlo simulations, recommended the following criterion $\chi^2 = -(\frac{2}{n})(n - 1 - p - \frac{1}{2}g)\log\lambda \sim \chi^2_f$ where $\lambda = L_g/L_{g'}$ is the likelihood ratio of both hypotheses and the degrees of freedom $f = 2p(g'-g)$. Binder (1978) showed

that the asymptotic distribution is not χ^2. Everitt (1981), quoted by McLachlan and Basford (1988), suggests that the test may be used if the relation n/p (number of observations to number of variables) is greater than 5 and $n > 50$. In any case, the likelihood ratio or the growth of the likelihood is a useful guide in defining the number of groups. We have used the graph of the likelihood profile for different values of g and observed the maximum growth point as a criterion for defining the number of groups.

Canonical analysis of the groups formed by the Ward-MLM strategy

The canonical analysis finds canonical variables that maximize the ratio of the variance among groups to the within groups, allowing to better differentiation of the groups (or clusters) of individuals. Thus, results of the three-way Ward-MLM classification method can be represented in a graphical display of the canonical analysis, where the accessions comprising the groups are plotted in a two (or three) dimensional diagram of the first two (or three) canonical variables. The biological interpretation is provided by the correlation between the original variables and the canonical variables. Since the first canonical variable is the one that best separates the accessions, the original variables most correlated with the first canonical variables are those that best discriminate the accessions.

Results and discussion

Three-way Ward- MLM classification method

The pseudo-F and pseudo-t^2 criteria (SAS 1999) found that the optimum number of groups were four or eight. The likelihood profile, associated with the likelihood ratio test, showed that the two highest increase of the likelihood function were at the four and eight group level (Fig. 1) with increases of 56.2231 and 50.5962, respectively (Table 2). Because the difference in the value of the likelihood function between four and eight groups was negligible, and, from a biological perspective the formation of eight groups is more logical than the four groups, the eight group level was selected and the groups are named

G1–G8. The results show the need for a better biological understanding for classifying crop diversity before its utilization, even though if classification results from successive approximations as per in this research. In this regard, if race will remain as the unit for conservation, evaluation and utilization in a region, the classification system should be precise and should consider previous knowledge, e.g. the Peruvian maize races as defined by Grobman et al. (1961).

The sequential Ward-MLM classification strategy reveals that some races are well defined by vegetative traits. This is the case for accessions from races Confite Morocho (R1), Uchuquilla, (R3) and Paro (R6), which are all in groups G3, G5, G2, respectively (Table 3). Other races such as Cuzco Gigante (R4) and San Geronimo (R7) were split into two groups and races Chullpi (R2), Huayleño (R5), and Shajatu (R8) were split into three groups. Grobman et al. 1961 used morphology (especially ear traits), agro-ecology, and cultural use for grouping maize accessions into races, e.g. the race Huayleño was always in the same cluster with others showing a brown pericarp, which may explain why this race appears in the groups following numerical classification. Race Chullpi (R2) that comprised accessions that did not show much variability was not well defined by these six vegetative traits. Abu-Alrub et al. (2004) already reported that the race Chullpi was not well defined by tassel traits.

The distribution of the races and the groups along the first two canonical variables (which explained 85% of the total variability) is depicted in Fig. 2. The first canonical variable maximizes the ratio {variability among groups} versus {variability between groups} and thus it maximizes the discrimination of the accessions along the first canonical axis. The original variables most associated with the first canonical variable (CAN1) were PH, EH, and LN during the second year (PN2, EH2, LN2) with correlations of 0.537, 0.538, and 0.573, respectively (Table 4). Race Cusco Gigante (R4) split into G1 and G5, whereas San Geronimo (R7) split into G3 and G4. In the canonical graph of Fig. 2 G1 and G5 are non overlapping neighbours (along the first canonical axis), which is expected because they are from race R4 and G3 and G4 are neighbours that belong to race R7, respectively. Race Huayleño (R5) split into three neighbouring groups G2, G4, and G6 (Fig. 2) and

Fig. 1 Profile of the log-likelihood function against number of groups. The two arrows mark the two highest increases of the log-likelihood function (at the 4 group and at the 8 group levels)
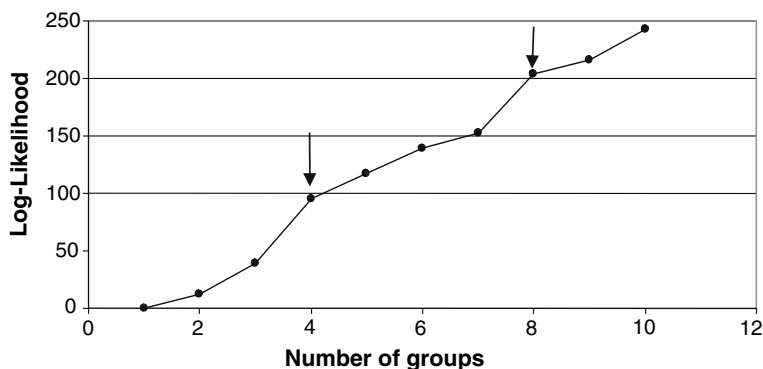

Table 2 Number of groups formed by the Ward-MLM method, logarithm of the likelihood function (log-likelihood) and their increments

Number of groups	Log-likelihood	Increment
1	−1158.7665	0
2	−1146.6813	12.0853
3	−1120.3171	26.3641
4	−1064.094	56.2231[a]
5	−1042.1467	21.9474
6	−1019.9974	22.1492
7	−1005.9321	14.0653
8	−955.3358	50.5963
9	−942.8695	12.4664
10	−915.7815	27.0880

[a] The highest increment

Shajatu (R8) split into groups G1 and G5 (like R4). Only one accession from race Shajatu (R8) appears as an outlier forming group G8. All the accessions from R3 (Uchuquilla), three accessions from R4 (Cusco Gigante), two accessions from Chullpi (R2), and most of the accessions from R8 (Shajatu) formed group G5. Group G7 with three accessions from Chullpi (R2) formed a very different group with the lowest values of all six vegetative traits in the second year of evaluation (Table 4). The canonical representation of the races and groups in Fig. 2 approximates the relationship between accessions in the numerical classification with respect to the racial classification.

Abu-Alrub et al. (2004) found that Cusco Gigante (R4) and Uchuquilla (R3) were placed at the extreme of the first principal component axis for tassel traits due to their high branch number and long branching space. Ludeña (1974) reported that chromosome knobs suggest that the racial complex Cusco derived from races such as Uchuquilla and Huancavelicano, which may explain above groupings. Our results showing all accessions from Confite Morocho (R1) falling into group G3, were similar to those of Also Abu-Alrub et al. (2004), who found that all accessions from this race were together when performing the principal component analysis using tassel traits. This result was expected since this is a primary race

Table 3 Distribution of the accessions of the eight Peruvian races (R1, R2, R3, R4, R5, R6, R7, and R8) into the four groups (G1, G2, G3, and G4) formed by the three-way Ward-MLM

Group	Group							
	R1[a] (5)	R2 (6)	R3 (5)	R4 (9)	R5 (9)	R6 (4)	R7 (6)	R8 (6)
G1 (22)	–	2	5	9	–	–	–	6
G2 (19)	–	1	–	–	9	4	5	–
G3 (6)	5	–	–	–	–	–	1	–
G4 (3)	–	3	–	–	–	–	–	–

Number of accessions in groups and races are in parenthesis

[a] R1 = Confite Morocho, R2 = Chullpi, R3 = Uchuquilla, R4 = Cusco Gigante, R5 = Huayleño, R6 = Paro, R7 = San Gerónimo, and R8 = Shajatu

Fig. 2 Plot of the first two canonical variables CAN1 and CAN2 and the distribution of the eight groups (G1, G2, G3, G4, G5, G6, G7, and G8) and the eight races R1 = Confite Morocho, R2 = Chullpi, R3 = Uchuquilla, R4 = Cusco Gigante, R5 = Huayleño, R6 = Paro, R7 = San Gerónimo, and R8 = Shajatu. In parenthesis is the percent of total variability account for by the first and second canonical variables
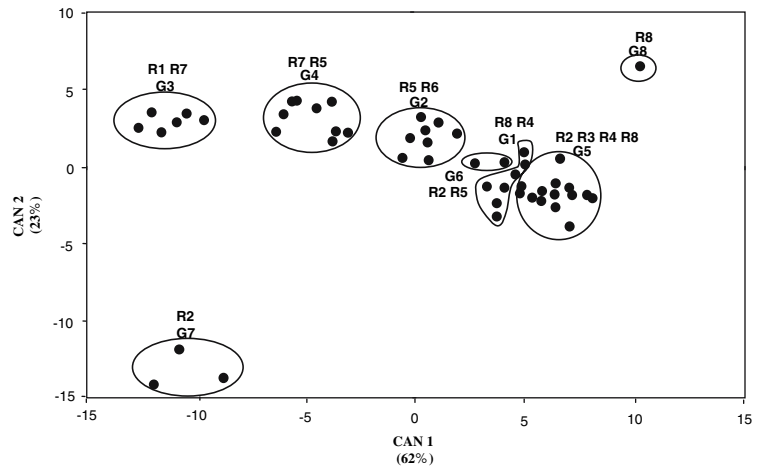
that tends to be distinct from others, and perhaps close to some derivatives but distant enough to stand almost alone. The kernel traits, that was the best descriptor for Abu-Alrub et al. (2004), also grouped together accessions within each race for Confite Morocho, Uchuquilla, and Paro (similar to the results of this study) as well as for San Gerónimo (two neighbouring groupings in this study) and Shajatu (two non neighbouring groups with one outlier accession). Chullpi was the only race that gave

results different than those reported by Abu-Alrub et al. (2004), because it separated into three groups G5, G6, and G7 in which G7 is not a neighbour of any other.

Ortiz and Sevilla (1997) recommended kernel and ear traits for classifying Peruvian highland germplasm due to their heritability, repeatability and low coefficients of variation. Results from this study show, however, that it is feasible to classify such germplasm using vegetative traits, if appropriate

Table 4 Average of 12 variables for each of the eight groups

Variable	Group								CAN1
	G1	G2	G3	G4	G5	G6	G7	G8	
PH1[a]	211.8	180.2	128.8	159.3	213.9	191.0	192.0	193.0	0.493
EH1	128.0	102.2	55.3	76.2	134.5	117.5	112.3	116.0	0.397
LN1	12.3	10.7	8.2	9.4	12.4	11.6	11.7	12.0	0.398
LE1	5.8	5.3	4.8	5.2	5.8	5.2	5.6	5.9	0.175
LL1	77.7	71.3	63.0	70.5	75.3	64.0	73.3	75.0	0.181
LW1	10.2	9.3	7.1	8.2	10.1	9.8	9.7	11.0	0.378
PH2	217.5	205.8	148.1	182.2	230.5	213.5	94.6	200.0	0.537
EH2	127.8	119.2	69.1	98.3	148.9	125.0	42.0	117.0	0.583
LN2	12.5	10.8	8.5	9.8	12.8	11.0	7.4	11.6	0.573
LE2	6.7	5.9	5.3	6.0	6.2	5.9	4.0	6.2	0.250
LL2	84.5	79.7	67.6	77.6	78.5	71.5	42.0	84.0	0.298
LW2	9.4	8.9	7.1	8.6	9.2	8.8	4.6	14.5	0.419

Correlation between the variables and the first canonical variable (CAN1)

[a] PH = Plant height (cm); EH = ear height (cm); LN = Leaf number (number);

LE = leaf number above the ear; LL = Leaf length (cm); LW = Leaf weight (cm);

1 = year 1: 2 = year 2

statistical classification methods are used with traits showing enough heriTable genetic variation; e.g. those included for the analysis of this experiment .

The characterization of the eight groups based on the six traits showed that G5 had high values for traits PH, EH, and LN in both years (Table 4). These traits best discriminated the accessions and had the highest positive correlations with the first canonical variable (which is the axis that most discriminates the accessions). Group G1 had the highest values for LE and LL in both years. On the other hand, accessions from groups G3 had the lowest values for all the variables in year 1 and year 2, respectively, whereas accessions from Group G7 show an intermediate response for all variables in year 1 but had the lowest performance of all variables in year 2.

Racial classification versus numerical classification—Mahalanobis distance

One objective of this study was to compare the racial classification of the eight Peruvian highland maize races based on visual assessment with the numerical classification using the three-way Ward-MLM sequential clustering strategy based on the distance between the groups as compared to the distances between the races. Since all the traits are continuous, the distance measure to be used is the Mahalanobis (1930) distance between groups (or between races) $D^2 = (Y_i - Y_j)' \hat{\sum} (Y_i - Y_j)$, where Y_i and Y_j are the vectors of means of the two groups (or races), assuming a common variance-covariance matrix Σ.

The results show that 21 out of 28 pair-wise D^2 were higher between groups than between races (Table 5). The average D^2 between groups was 264.50, whereas the average D^2 between races was

78.26. These results indicate that the groups formed by the numerical classification are much more separated, compact, and well defined in terms of these variables than the visual assessment used to define the races. Gutierrez et al. (2003) compared a racial classification obtained through visual assessment of maize accessions collected in Uruguay with the numerical classification obtained using the Ward-MLM strategy. Among others, one criterion used to determine the best classification strategy was the Mahalanobis distance between groups for continuous variables. The authors found that the average Mahalanobis distance between the groups formed by the Ward-MLM numerical classification was 49% greater than for the visual racial classification. This result shows that the Ward-MLM groups were better separated than the preliminary races and that these groups can be appropriately used for forming core subsets of these Peruvian highland races of maize.

Racial classification versus numerical classification—Accession within group × Environment interaction

Since the Ward-MLM strategy clusters individuals with consistent responses across all variables in all the environments, it is expected that the three-way cluster strategy should form groups of accessions with small accession within group × environment interaction for all the traits included in the study (Franco et al. 2003). The three-way Ward-MLM strategy should tend to form groups of accessions with high genotypic correlations across environments for most of the traits. Franco et al. (2003) employed the three-way Ward-MLM clustering strategy in three different data sets and found that it formed groups of

Table 5 Mahalonobis distance between groups below diagonal and between races above diagonal

–	132.23	151.51	208.61	48.84	118.16	29.48	265.43
41.17	–	20.03	38.25	44.42	37.88	74.30	73.11
283.77	150.06	–	32.77	62.97	26.00	81.63	56.98
111.03	33.61	54.11	–	78.32	31.02	100.25	22.18
32.66	53.59	336.29	152.63	–	52.28	16.90	144.50
63.66	35.96	267.00	114.12	42.75	–	52.22	44.45
413.08	364.88	276.88	310.85	435.45	408.16	–	146.53
362.87	333.77	722.99	454.74	331.76	256.80	961.62	–

cultivars with negligible cultivar within group × environment interaction for all continuous and discrete traits. In this study, variance component estimation show that the variance component of the accessions within group was smaller than the accession within race for traits PH, EH, and LN (Table 6) which indicate that for these traits the accessions within groups are more compact and cohesive than the accessions within races. The variance component of accessions within group × environment interaction was smaller than the accessions within race × environment interaction for all the traits indicating the Ward-MLM method did form groups with accessions that interact less with the environments than the accessions comprising the races.

Conclusions

In general, the numerical classification using the three-way ward-MLM sequential strategy maintained the main structure of the more differentiated races, but reclassified parts of the races into new groups so that more separated and well defined groups were formed. All of the accessions from G1 (except one) are from Cusco Gigante (R4); all of the accessions from G3 (except one) are from race Confite Morocho (R1); and all of the accessions from G7 are from Chullpi (R2). Group G2 has four accessions from Huayleño (R5) and four accessions from Paro (R6), whereas G4 has four accessions from Huayleño (R5) and five accessions from San Geronimo(R7). Group G5 has accessions from four races (R2, R3, R4, and R8), and G6 and G8 formed small groups with two

and one accession each, respectively. These groups can be used to forme core subsets for the purpose of germplasm enhancement and assembling gene pools for further breeding. The three-way Ward-MLM classification produced groups with very distinct characteristics in terms of the vegetative traits measured and with accessions that minimize the interaction with environment.

The above results allow addressing some of the hypothesis about the ancestry of Peruvian highland maize races. For example, Chullpi, Huayleño, and Paro are postulated to have descended from a common ancestor (Confite Chavinese). The results from numerical classification suggest that this postulate is confirmed for Huayleño and Paro but not for Chullpi. Similarly, Cusco Gigante (deriving from the Cusco Complex) and Uchuquilla (whose ancestor was Kcarapampa) share common morphological features, which confirms previous reports about both races sharing common tassel and ear traits (Abu-Alrub et al. 2004), and support the view of the racial Cusco Complex could derive from Uchuquilla (Ludeña 1974).

This study produced an adequate classification of the maize races of Peru, and the results of the three-way Ward-MLM classification method can be considered a further refinement and a complement of the racial classification based on visual assessment. The groupings obtained based on vegetative traits can be also used to form core subsets or mini core subsets of these races. The use of vegetative as well as reproductive traits should further help to classify the Peruvian highland maize races.

Table 6 Variance components (×10) of accessions within race ($\sigma^2_{A(R)}$), accession within race × environment interaction ($\sigma^2_{A(R)E}$), accessions within group ($\sigma^2_{A(G)}$), and accession within group × environment interaction ($\sigma^2_{A(G)E}$) for plant height (PH), ear height (EH), leaf number (LN), leaf number above the ear (LE), leaf length (LL), and leaf width (LW) measured in eight highland Peruvian maize races

Variance component	PH	EH	LN	LE	LL	LW
Group						
$\sigma^2_{A(G)}$	234	240	0.15	0.30	31	23
$\sigma^2_{A(G)E}$	523	491	2	0.46	62	0.31
Race						
$\sigma^2_{A(R)}$	1468	1757	3	0.12	14	5
$\sigma^2_{A(R)E}$	2768	1561	4	1	296	2

References

Abu-Alrub I, Christiansen J, Madsen S, Sevilla R, Ortiz R (2004) Assessing tassel, kernel and ear variation in Peruvian highland maize. Plant Genet Resour Newsl 137:34–41

Anderson E, Cutler HC (1942) Races of *Zae Mays*, *L*. Their recognizition and classification. Ann Mo Bot Gard 29:69–88

Binder DA (1978) Bayesian cluster analysis. Biometrika 65:31–38

Crossa J, Franco J (2004) Statistical methods for classifying genotypes. Euphytica 137:19–37

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc Ser B 39:1–38

Everitt BS (1981) A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. Multivar Behav Res 16:171–180

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, Villaseñor J, Castillo A, Taba S, Eberhart SA (1999) A two stages, three-way method for classifying genetic resources in multiple environments. Crop Sci 39:259–267

Franco J, Crossa J, Taba S, Shands H (2003) A multivariate method for classifying cultivars and studying group × environment × trait interaction. Crop Sci 43:1249–1258

Grobman A, Salhuana W, Sevilla R, Mangelsdorf E (1961) Races of maize in Peru; their origin, evolution and classification. Nat Acad Sci – Nat Res Council Pub. 915, Washington DC

Gutierrez L, Franco J, Crossa J, Abadie T (2003) Comparing a preliminary classification with a numerical classification of the maize landraces of Uruguay. Crop Sci 43:718–727

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

Ludeña L (1974) Posible origen del maíz en base a presencia de nudos cromosómicos. Inf Maiz (UNALM) 3:6–7

Mahalanobis PC (1930) On tests and measures of group divergences. J Proc Asiatic Soc Bengal 26:541–588

Manrique Chavez PA (1997) El maíz en Perú. CONCYTEC, Lima, Perú

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, London

McLachlan GJ, Basford KE (1988) Mixture models, inference and applications to clustering. Marcel Dekker, New York

Milligan GW, Cooper M (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50–2:159–179

Olkin I, Tate RF (1961) Multivariate correlation models with mixed discrete and continuous variables. Ann Math Stat 32:448–465

Ortiz R, Sevilla R (1997) Quantitative descriptors for classification and characterization of highland Peruvian maize. Plant Genet Resour Newsl 110:49–52

SAS Institute Inc (1999) SAS OnlineDoc®, Version 8. SAS Institute Inc., Cary, North Carolina

Taba S, Díaz J, Franco J, Crossa J (1998) Evaluation of Caribbean maize accessions to develop a core subset. Crop Sci 38:1378–1386

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

Wolfe JH (1971) A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Naval Personnel Training Res. Lab. Tech. Bull, STB 72–2

Wellhausen EJ, Roberts LM, Hernandez E, Mangelsdorf PC (1952) Races of maize in Mexico. The Bussey Institution, Harvard Univ. Press, Cambridge, Massachusetts