Grouping and Identification of Argentine Maize Races by Principal Component Analysis of Zein Reversed-Phase HPLC Data

J. L. Robutti, F. S. Borras, M. E. Ferrer, and J. A. Bietz

ABSTRACT

Racial classification of maize has important taxonomic and phylogenetic implications. It may also serve to organize germ plasm inventory, thus helping breeders choose their stocks. Maize racial classification is usually based on phenotypic descriptors, which may not always accurately express genetic characteristics. In contrast, synthesis and expression of zeins is directly associated with genotype. This study was conducted to determine whether racial grouping and identification of maize can be done by applying principal component analysis to zein reversed-phase high-performance liquid chromatography (RP-HPLC) data. Zeins from samples of 97 landraces (primitive varieties) of the Argentine races Cristalino Colorado, Dentado Blanco, Avatí Moroti, Capia, and Pisingallo, stored at the Pergamino Active Maize Germplasm Bank, were analyzed by RP-HPLC. Data from the -21- to 53-min chromatogram region (total zeins [Z]), the -21- to 30 min region (zeins 2 [Z2]), or the -38- to 52-min region (zeins 1 [Z1]) were subjected to multivariate analysis based on principal component to group samples by race and to assign unknown samples to predetermined racial groups. Clearly differentiated racial groups were revealed, closely matching groups based on phenotype. Unknown samples could be assigned, with a low percentage of misidentification, to predetermined groups based on Mahalanobis distances. The shortest distances of unknown samples were almost always the distances to their respective groups. Approaches other than multivariate analysis were used to group and assign samples to defined races but they were not as effective. Results indicate the potential of this method as a complementary tool to perform racial grouping and identify maize materials with high genetic variability.

Maize genotypes are extraordinarily diverse. Classification is necessary to make order from the bewildering multiplicity of landraces (primitive varieties). Botanical, genetic, cytological, and agronomic studies have revealed relationships among landraces and have grouped them into more or less well-defined races. Because relationships are implicit in any natural system of classification, an attempt was made to determine the origin and relationships of recognized races. Besides its taxonomic and phylogenetic implications, racial classification can also serve to organize inventories of materials available to plant breeders, thus enhancing accuracy in their choice of stocks (Wellhausen et al 1952).

Racial grouping of maize is usually based primarily on ear and kernel phenotypic descriptors. Some of these descriptors may, however, be affected by environment, and thus may not accurately express genetic characteristics. Methods based on biochemical molecular markers present valid alternatives and are often preferable to cereal genetic identification and classification procedures based on phenotype. Such markers may be genes or gene products that directly reflect the genome or they may be proteins, especially invariant storage proteins, that are expressed.

In maize, the synthesis and expression of storage proteins is directly associated with genotype (Feix and Quayle 1993). Thus, zein characterization by RP-HPLC can reliably identify inbred and single hybrid maize genotypes (Paulis and Bietz 1994). RP-HPLC has also been used to group and reveal pedigrees of maize inbreds and hybrids (Smith 1988). Protein compositions so revealed are characteristic of genotype and depend little on environment (Smith and Smith 1986). This conclusion is supported by unpublished results from our laboratory at Pergamino, which showed that RP-HPLC zein patterns of two breeding lines grown during two consecutive years at two locations were little affected by environment during development. RP-HPLC zein patterns are affected neither by germi-
MATERIALS AND METHODS

Materials

Samples from the Pergamino Active Maize Germplasm Bank representing different native maize landraces were analyzed. Samples were originally collected in fields from different regions of Argentina. Eight to 30 ears (female plants) were collected per sample. Because materials in this study are open-pollinating populations, the male parents are unknown and may be considered infinite. Collected samples, packed in polyethylene or paper sealed bags, had been stored at 5-10°C and 40-50% rh for one to 20 years. After removal from storage, samples were kept in sealed bags in airtight containers at room temperature until analyzed.

Samples had previously been assigned to race (Solari and Gómez 1997) according to ear and kernel phenotypic descriptors such as texture type, kernel and cob color, ear and kernel shape, width, and length, and number of rows on ears (Wellhausen et al 1952, Goodman and Pateriani 1969, Ferrer 1982). The five races in this study were selected to represent the principal maize texture types. CC (flint) has medium-sized round kernels of orange color and horny texture. Its ears are cylindrical in shape with a medium number of rows. The cob is wide and white, and pericarp and aleurone are colorless. DB (dent) has cylindrical-conical ears that are long and wide with a high number of rows. Kernels are large and prism-shaped. The crown is indented and the relatively abundant floury endosperm is located at the center. Pericarp and aleurone are colorless. The cob is wide and reddish or pink in color. AM (subtropical floury) has long cylindrical ears. Kernels are wedge-shaped and flat. The endosperm is completely floury, the aleurone is yellow, and the pericarp is colorless. The cob is wide and reddish or pink in color. Ca (highland floury) has ears that are medium in length with a medium number of rows. Kernels are long, medium in width, with white floury endosperm indented at the crown. Pericarp and aleurone can be colorless or purple. Pi (corn) has cylindrical ears pointed at the apex of variable size, frequently fasciated, and of narrow diameter. Kernels are small with sharp tips. Endosperm is almost completely horny and white. Though pericarp is predominantly white, some samples are yellow, purple, red, or other color. Values for these parameters have been published for each race (Solari and Gómez 1997). Phenotypic descriptor values of samples in this study closely matched values typical of their respective races. Samples for which some phenotypic descriptors deviated substantially from values typical for their races were not included because such samples might indicate introgression from other races.

Sample Analysis by RP-HPLC

For zein characterization, sample size was 50 kernels for each landrace. Previous studies (unpublished data) had shown 30 kernels to be the minimum number representing variability of landraces in the Pergamino Active Maize Germplasm Bank. That is, chromatograms of more than 30 combined kernels were identical to results for 30-kernel samples. In an analogous situation, Kubiczek et al (1993) found, when using RP-HPLC for cultivar identification in rye (Secale cereale, an open-pollinating cereal), that 50-kernel samples adequately expressed the full range of genetic heterogeneity present. Similarly, Smith and Smith (1988) using RP-HPLC, showed that 25 kernels sampled from an F2 maize generation had the full genetic heterogeneity of its inbred progenitors. Bietz (1985a,b) also reported that 50-kernel samples of field-grown (F2) maize generations (exhibiting major kernel-to-kernel variation in protein composition) fully represented compositions of the two inbreds from which a hybrid (F1) was produced, and that the inbreds represented by the F2 generation could be identified from its RP-HPLC pattern.

Zeins from each material were characterized by RP-HPLC as previously described (Eyherabide et al 1996). A Hewlett-Packard 1050 HPLC system with a quaternary pump, autosampler, and UV detector was used. The column used (at 60°C) was a Vydac C18 (4.6 x 250 mm, 5-μm particles, 300 Å pores), preceded by a 22- x 3.5-mm precolumn. Samples (5 μL) were eluted at 1 mL/min with a linear 50-min 28-60.5% B gradient (solvent A, water + 0.1% [v/v] trifluoroacetic acid; solvent B, acetonitrile + 0.1% [v/v] trifluoroacetic acid), followed by 10-min isocratic elution at 60.5% B. Detection was by UV absorbance at 210 nm. Quantitative analysis was performed using HP ChemStation 3.0 software. Duplicate extracts were analyzed. Retention times and peak areas were reproducible within limits shown in earlier experiments. These earlier experiments, run on six replicates from a maize endosperm sample, indicated a maximum difference of 0.24 min for

Fig. 1. Characteristic reversed-phase HPLC chromatograms of zein-2 proteins from five maize races (Ferrer et al 1994).

Fig. 2. Grouping of maize races based on principal component analysis of total zein reversed-phase HPLC data. Ca, Capia; CC, Cristalino Colorado; DB, Dentado Blanco; AM, Avat/Morot; and Pi, Pisingallo.
retention time and a coefficient of variation in all cases <2% for absolute peak areas. Peaks with relative areas >2% were taken for this analysis. Using GRAMS32 software (Galactic Industries, Salem NH) for all steps, chromatograms were baseline-corrected with the auto-level method, aligned with respect to the peak of maximum absorbance at 26.163 min, normalized with respect to total zein content, and averaged from duplicates.

Experimental Design

To group races and assign samples to a given race, the PLS/IQ add-on application was used on mean-centered chromatograms.

Experiment A. For race grouping, 55 samples of CC, DB, AM, Ca, and Pi, as classified by phenotypic criteria, were used. Principal component analysis (PCA) was performed on complete RP-HPLC chromatograms of total zeins to cluster samples into racial groups using 10 principal components. Scores of the two first principal components were plotted against each other in search of graphical racial groupings. Resulting groups were then compared with groupings according to phenotype for the same samples.

Another approach used in an attempt to group these samples was by plotting values of characteristic peaks against one another. Thus, the relative area of peak 2 (Fig. 1) with maximum absorbance at t = 26.1 min was plotted against peak 3 (t = 27.2 min).

Experiment B. To assign unknown samples to predefined groups, 42 CC and Ca samples (different from those in Experiment A) were used. These races were selected because the most samples of these races remained after Experiment A, leaving an adequate number of samples for calibration and validation sets. Randomly selected samples from 12 landraces of CC formed the CC calibration group. Likewise, samples from 12 landraces of Ca formed the Ca calibration group. The 18 remaining samples (nine CC and nine Ca) formed the unknown set. All samples were truly representative of their respective race as judged by phenotypic criteria.

The program was asked to assign unknowns to CC or Ca using the Prediction option. Predictions were made from Mahalanobis distances calculated by the MD/PCA/R method (Mahalanobis Distance by Principal Component Analysis with Residuals) (Galactic Industries) on five principal components extracted from the total chromatograms from Z2 (=21–30 min) or zein-1 (Z1) (=38–52 min) regions. Because PCA reduces data into a smaller set of representative numbers (scores), the problem of overdiscrimination can be avoided while still using entire spectral regions. The Mahalanobis matrix is then done on principal component scores rather than actual spectral responses. An approach to improving sensitivity is to combine PCA with spectral residuals and use them all for Mahalanobis group-matrix calculations. These residuals are included in the Mahalanobis group-matrix by appending a column vector of extra scores that contains the mean-centered sum squared spectral residuals for each spectrum. Therefore, the relationship between PCA scores and spectral residuals is considered when a sample is predicted against the Mahalanobis matrix. The Mahalanobis distance is a very useful way of determining the similarity of values measured for unknown and known samples. Because the Mahalanobis distance is measured in terms of standard deviations from the mean of the training samples, the reported matching values give a statistical measure of how well unknown samples match training samples.

Other approaches were used to assign unknown samples to a given race. One was looking for peaks that would seem unique to a given race, in this case peak d (Fig. 1) for Pi. Another was the ratio of relative areas of peaks 2 and 3.

RESULTS AND DISCUSSION

Experiment A

By applying the first and second principal components from PCA of zein RP-HPLC data, samples of landraces of the five maize races analyzed clustered into five well-separated racial groups (Fig. 2). The first two principal components explain 77% of the total spectral variance. This grouping matches exactly the one based on phenotypic characters. Plotting of peak 2 versus peak 3 was not as effective (Fig. 3). The upper left region of Fig. 3 shows a cluster of mainly Ca samples, while lower right samples correspond mostly to CC. The intermediate region, however, shows considerable overlap and interracial (among races) discrimination is not as clear-cut as in Fig. 2.

![Fig. 3. Scatter plot of % area of reversed-phase HPLC peak 2 vs. peak 3 for 55 samples of five maize races: Capia (○); Avatl Morotl (*); Pisingallo (□); Dentado Blanco (△); Colorado (◆).](image)

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>Mahalanobis Distances* Between Unknown Samples and Calibration Groups†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC Ca Cal CC Ca CC Ca Cal Ca Cal CC Cal Ca Cal CC Cal Ca Cal CC Cal Ca</td>
</tr>
<tr>
<td>CC</td>
<td>42 2 0 1 4 12 52 1 8 157 41 2 4 19</td>
</tr>
<tr>
<td>32</td>
<td>2 1 1 10 41 0 11 267 32 1 1 23</td>
</tr>
<tr>
<td>45</td>
<td>18 204* 1 15 37 3 2 244 42 14 19* 31</td>
</tr>
<tr>
<td>43</td>
<td>2 1 1 11 26 3 3 139 35 2 10 21</td>
</tr>
<tr>
<td>41</td>
<td>4 1 20 37 3 1 371 39 4 1 36</td>
</tr>
</tbody>
</table>

* Rounded to nearest integer.
† Cal = calibration, CC = Cristalino Colorado race; Ca = Capia race; ZT = total zeins; Z1 = zein-1; and Z2, zein-2.
‡ Samples with unusually large Mahalanobis distance values (%).
Experiment B

Table I lists Mahalanobis distances between unknown materials and calibration groups CC and Ca, based on Z1, Z2, and ZT RP-HPLC analyses. When using Z1, all Ca samples were correctly classified as their distances to their calibration group are all shorter than those of the CC samples. This was not the case with the CC samples: two had distances (20 and 9) to their group that are equal to or larger than those shown by Ca samples to the CC group. These two samples can be considered misidentified. The percentage of misidentification was 22%. Analyses based on Z2 correctly classified all samples. In each case, the shortest distances of unknown samples within each racial group was to their respective calibration group. For ZT analyses, one CC sample had a distance (19) to CC that was slightly larger than the shortest distance (18) of one Ca sample to the CC group. In this case, the percentage of misidentification was 11%. It may be concluded that Z2 proteins identify the two races better than either Z1 or ZT. This should not be surprising because some Z2 proteins are highly associated with endosperm texture (Paulis et al. 1992, Moro et al. 1995, Mestres and Matencio 1996, Robutti et al. 1997). The two races in this experiment differ widely in endosperm texture (CC is predominantly flinty, whereas Ca is almost completely floury). These results demonstrate that methods used in this study efficiently identify maize races. Actual unknown samples would be identified as belonging to the racial group to which they had the shortest distance.

Of the unknown samples analyzed, a few (marked with § in Table I) had atypically large Mahalanobis distances to their calibration groups. One CC sample, when classified by ZT and Z1, gave distances of 19 and 20, respectively, to its calibration group (Table I). This sample was also the only unknown CC sample with highly atypical RP-HPLC peaks between 36 and 39 min (Fig. 4B), as compared with a typical CC sample (Fig. 4A). Similarly, another CC sample classified by Z2 had a Mahalanobis distance of 35 (Table I). When analyzed by RP-HPLC (Fig. 4C), this sample displayed a peak at 23.9 min that was absent in all other unknown CC samples. It is likely that these two samples did not truly represent their race. This may indicate introgression from other races not detectable during selection of samples based on phenotypic ear and kernel traits. Thus, PCA/MDR analysis of zein RP-HPLC data can detect racial genetic impurities not apparent on visual inspection of ears and kernels.

The other approaches used to assign unknowns to a certain race were not as effective. Peak d in Fig. 1 would appear unique to Pi, but this chromatogram is an average of several samples. In fact, some individual Pi chromatograms have this peak, while others do not (Fig. 5). This intraracial (within races) variability would make it difficult to assign a sample to a given race based solely on visual inspection of one chromatogram. On the other hand, the ratio of peak 2 to peak 3 may effectively differentiate CC and Ca, but this ratio is also similar for CC, DB, and Pi (Fig. 1). Such calculations could wrongly identify an unknown sample as DB or Pi, when it actually belongs to CC.

CONCLUSIONS

Our results show the potential of chemometric analysis of RP-HPLC zein data using multivariate statistical software such as GRAMS32. This method is an important tool for racial identification and grouping of highly variable genetic materials such as those in the Pergamino Active Maize Germplasm Bank. It can both complement and extend results from more traditional methods based on phenotypic traits.

The present study appears to be the first to apply such methodology to highly variable genetic materials. In view of the promising results obtained, further studies will include materials from more races and with different degrees of introgression among races.

LITERATURE CITED


Supplied by the U.S. Department of Agriculture, National Center for Agricultural Utilization Research, Peoria, IL.