Selection of Haploid Maize Kernels from Hybrid Kernels for Plant Breeding Using Near-Infrared Spectroscopy and SIMCA Analysis

ROGER W. JONES,* TONU REINOT, URSULA K. FREI, YICHIA TSENG, THOMAS LÜBBERSTEDT, and JOHN F. MCCLELLAND

Ames Laboratory–USDOE, Iowa State University, Ames, Iowa 50011 (R.W.J., T.R., J.F.M.); and Iowa State University, Department of Agronomy, Ames, Iowa 50011 (U.K.F., Y.T., T.L.)

Samples of haploid and hybrid seed from three different maize donor genotypes after maternal haploid induction were used to test the capability of automated near-infrared transmission spectroscopy to individually differentiate haploid from hybrid seeds. Using a two-step chemometric analysis in which the seeds were first classified according to genotype and then the haploid or hybrid status was determined proved to be the most successful approach. This approach allowed 11 of 13 haploid and 25 of 25 hybrid kernels to be correctly identified from a mixture that included seeds of all the genotypes.

Index Headings: **Near-infrared spectroscopy; NIR spectroscopy; Corn; Maize; Haploid selection; Partial least squares; PLS; Single kernel analysis; Soft independent modeling of class analogy; SIMCA.**

INTRODUCTION

Determination of phenotypical traits is a necessary step in the selective breeding process used to create improved plant varieties. Each generation of seed produced in the selective breeding process needs to be examined so that only the seed having the most promising properties is chosen for propagation. The difficulty is in finding an analytical technique that can examine the seed rapidly but accurately enough to allow a meaningful selection of the superior seed within the short time period between the harvesting of one generation and the planting of the next. The most selective technique would be one that examines each seed individually but that requires a technique that can determine the properties of interest very quickly without destroying the seed.

Diploid maize is typically cultivated as F1 hybrids, generated by crossing two highly homozygous inbred lines from different genetic pools. Recently doubled haploid lines have become an important breeding tool in maize breeding. Traditional homozygous line development in maize takes at least five to six generations of selfing heterozygous materials, whereas maternal induction of haploids and subsequent doubling and selfing can achieve this goal within two generations.

Single ears from two synthetic populations (P1 and P2) pollinated with the haploid inducing genotype were selected for haploid and hybrid seed, based on the expression of the *R1-nj* marker allele. For the experiments here, seed from three different donor genotypes (P1-102, P1-103, and P2-20) were chosen, as they had at least 20 haploid seeds per cob. The study included a total of 364 kernels, consisting of 81 P1-102 kernels (22 haploid, 59 hybrid), 83 P1-103 kernels (23 haploid, 60 hybrid), and 100 P2-20 kernels (33 haploid, 67 hybrid).

The identity of haploid seed is determined by the expression of the dominant marker gene *R1-nj*, which leads to a colored embryo in the hybrid seed versus an uncolored embryo in haploid seed. Visual inspection was used to determine haploid and hybrid status, even though this manual selection of seeds is time consuming and especially error prone in genetic backgrounds with colored seed. In our most recent planting of selected haploids of P1 and P2, selection errors were made in 15% of the P1 plants and 0.6% of the P2 plants. The donor population P1 has a high percentage of colored cobs, which makes visual selection difficult. For the present study, P1 kernels were selected only from uncolored cobs to avoid the high incidence of selection errors. Currently, the process of haploid selection in maize is based on one-by-one, visual (human) inspection of kernels, and plant breeding programs often require the inspection of hundreds of thousands of kernels, so an automated, spectroscopic procedure would be a leap forward. In addition, it would greatly reduce costs and potentially enable screening more kernels in the short time window between harvest in September and winter nursery planting at the end of October.

Near-infrared (NIR) spectroscopy is routinely used for determining the composition of bulk grain.[1] There has been increasing interest in applying NIR spectroscopy to the composition analysis of individual seeds. NIR spectroscopy has been used to determine moisture level,[2] total oil,[3–7] total protein,[4,5] and total starch[4,5] in individual, intact maize kernels. It has also been used to measure specific fatty acids in individual seeds of maize and other grains[7,8] and to categorize individual maize kernels according to the presence of toxins.[9,10] Determining composition-related functional characteristics, such as potential ethanol yield,[11] has also been demonstrated. These studies have involved both transmission[2–4,9,10] and reflection[4–7,9–11] spectroscopy. The reported reflection spectra extend deeper into the NIR range than the transmission spectra, usually reaching 1700 nm.[4,6,7,9,10] Reflection, however, interrogates only the surface region of the kernel and only the region struck by the NIR beam, while diffuse transmission examines most of the bulk of the kernel. Some of the reflection-based studies have partly compensated for this by tumbling or dropping the kernels during analysis so as to view a larger portion of the kernel.[5,6,11] This paper examines using NIR transmission spectroscopy to select haploid from hybrid maize kernels after maternal haploid induction. Because it is known that the haploid nature of a kernel is reflected in its embryo, we have chosen to use transmission spectroscopy, with its ability to examine the bulk of the kernel. In

Received 27 July 2011; accepted 6 January 2012.

* Author to whom correspondence should be sent. E-mail: jonesrw@ameslab.gov.

DOI: 10.1366/11-06426

0003-7028/12/6604-0447$2.00/0

addition, we have been able to extend the transmission analysis to 1700 nm, removing the spectral-range advantage of reflection spectroscopy.

EXPERIMENTAL

Seeds were equilibrated to 8% moisture and vacuum cleaned prior to spectroscopic measurements with an automated, in house-built, seed-screening system. Seeds were fed from a hopper with a vibratory feeder to a rotating drum which picked up single seeds by vacuum and rotated them sequentially to a position where a NIR light beam was transmitted through the seed in order to sample the whole seed volume. The transmitted light was collected into a fiber-optic cable and transmitted to a Carl–Zeiss NIR spectrometer (Model MCS-611 NIR). Five spectra were averaged to produce the final spectrum for each kernel. Between each of the five spectra, the kernel was shaken to reposition it so that successive spectra were acquired from different vantage points. The shaking was not strong enough to flip a kernel over. The time to acquire each spectrum was some multiple of 50 ms that depended on the transparency of the kernel, which was evaluated by a 1 ms illumination prior to the spectrum acquisition. Transparency of the kernels varied by a factor of 15. Spectrum acquisition time was lengthened as kernel transmission decreased so that the resulting spectra were all within a factor of two of the same strength. The time to acquire all five spectra for a kernel typically took 1.0 to 1.3 min. The spectra were converted to absorption prior to modeling and smoothed with a binomial function (5-point half-width). The full 943–1707 cm^{-1} range of the spectra was used in the modeling.

Soft independent modeling of class analogy (SIMCA)[12–14] was used to define both the genotypes and the haploid and hybrid classes according to their spectra. The SIMCA module in Pirouette (Pirouette Version 3.2; Infometrix, Bothell, WA) was used for this purpose. For purposes of the SIMCA modeling, six-sevenths of the seeds were used as the training sets for building the models, and the remaining one-seventh of the samples, which were chosen randomly, were reserved as unknowns for testing the models. Various pretreatments to the spectra (i.e., mean centering, variance scaling, autoscaling, multiplicative scatter correction, standard normal variate, and second derivatives) were tested for every analysis discussed below. The particular pretreatments listed for each analysis are those that gave the best results. In most cases, the number of factors determined by Pirouette as optimum was the number used for the analysis. Only when defining the haploid class for the P1-102 and P2-20 genotypes was the number of factors increased above the software-selected number (by two for P1-102 and by one for P2-20). These increases were made based on the test of whether all training-set spectra were correctly classified. Classification of the unknowns was not used as a guide in determining the number of factors to be used.

RESULTS AND DISCUSSION

Figure 1 shows typical spectra of whole kernels. There is very little difference between the spectra of the haploid and hybrid kernels or among the spectra of the three different genotypes. SIMCA or some other chemometric method is required to classify the kernels according to their spectra.

It was found that SIMCA could not produce a completely successful haploid/hybrid classification for all three genotypes simultaneously. It was necessary to separate the genotypes before making the haploid/hybrid classification. The model for this used variance scaling and multiplicative scatter correction (MSC) pretreatment.[15] The model used 20, 15, and 15 factors to define the P2-20, P1-102, and P1-103 classes, respectively. The class–distance plot for the training set is shown in Fig. 2. The P2-20 genotype is well separated from both P1 genotypes. By contrast, the separation between the two P1 genotypes is much smaller (although visually enlarged somewhat in the plot by the use of log scales), showing that they are much more similar to one another. This is in accordance with the fact that the two P1 genotypes originate from the same synthetic population and are genetically closely related. Nevertheless, the separation into genotypes is fully successful. The model places each spectrum into the class for which it has the smallest class

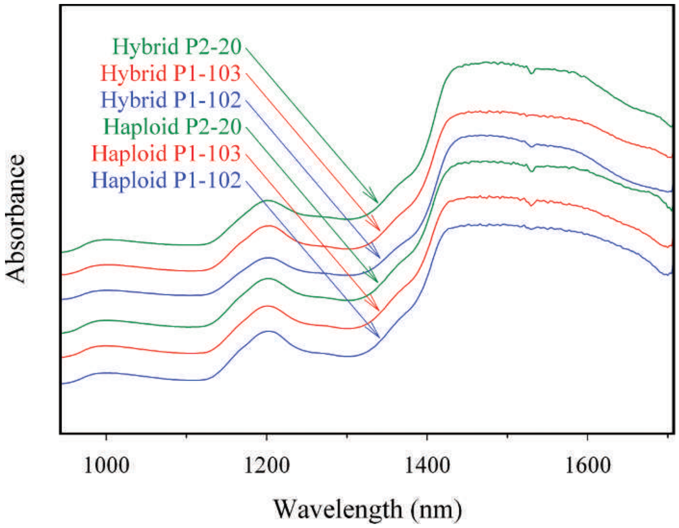

FIG. 1. Representative spectra of individual, intact corn kernels. The spectra are offset vertically for clarity but are otherwise on the same vertical scale.

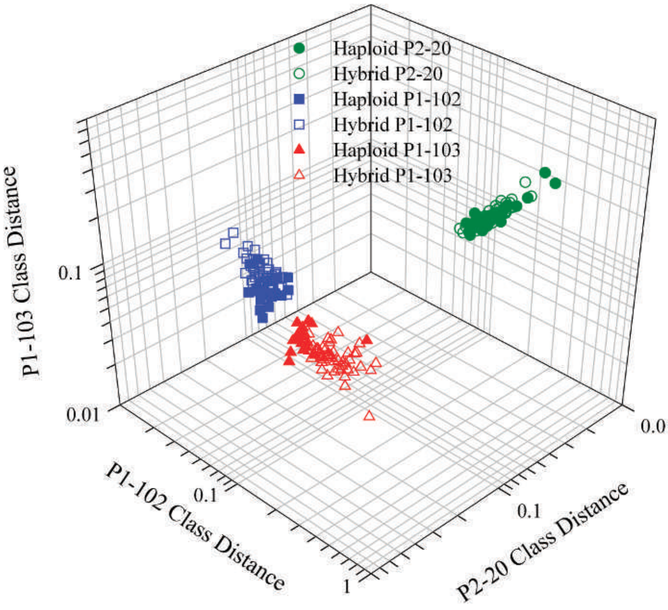

FIG. 2. Class–distance plot for separating the training-set spectra into their respective genotypes. The SIMCA model grouped all spectra in the correct genotype.

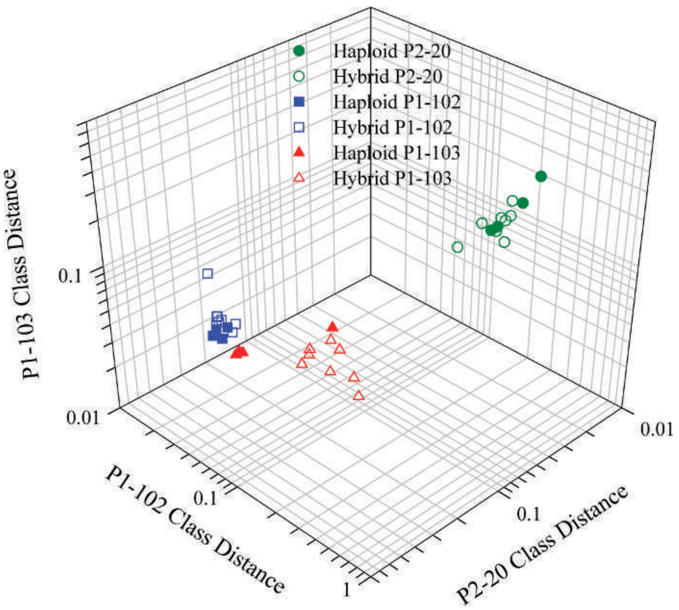

FIG. 3. Class–distance plot for separating the spectra of the unknowns into their respective genotypes. The SIMCA model grouped all spectra with the correct genotype.

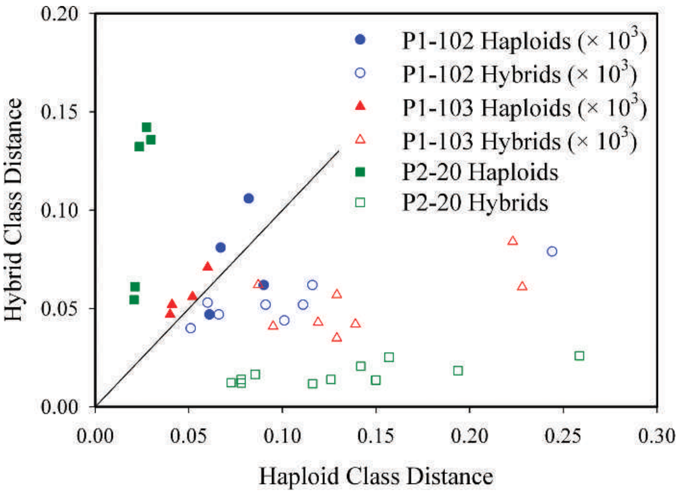

FIG. 5. Separation of the unknown spectra according to haploid versus hybrid status by the three genotype-specific SIMCA models. Two P1-102 haploids are misclassified (solid blue circles below the diagonal line separating classes).

distance, and that classification is fully correct for the training set. Figure 3 shows the corresponding plot for the set of unknowns. Again the separation between P2-20 and the others is quite large compared to the separation between P1-102 and P1-103, but all of the spectra fall into their correct class.

Once the spectra were classified into their separate genotypes, separate models were generated for each genotype that partitioned the kernels into haploid and hybrid classes. The P2-20 genotype proved relatively simple to separate into haploid and hybrid classes. The best model for P2-20 successfully separated all kernels correctly for both the training and unknowns sets. The model used variance scaling pretreatment and required 13 and 16 factors to define the haploid and hybrid classes, respectively. Figure 4 shows the class–distance plot for the P2-20 genotype training-set members, and Fig. 5 shows the class–distance plot of the P2-20 unknowns. The diagonal line in each figure (corresponding to $x = y$) is the border between the haploid and hybrid classes, and all P2-20 spectra fall on the correct sides of those lines.

The model for the P1-103 genotype also successfully separated all training-set samples and unknowns correctly into haploid and hybrid classes. Figures 5 and 6 show the class–distance plots for the unknowns and training set, respectively. The model used mean centering, MSC, and 17-point Savitzky–Golay second-derivative pretreatment on the spectra and required 10 and 15 factors for the haploid and hybrid classes, respectively.

Modeling for the P1-102 genotype was not as successful. The best model correctly separated all training-set members, but two of the four haploid members of the unknowns set fell within the hybrid class range, as shown in Figs. 5 and 7. A larger training set might improve the classification quality. The model used mean centering and 15-point Savitzky–Golay

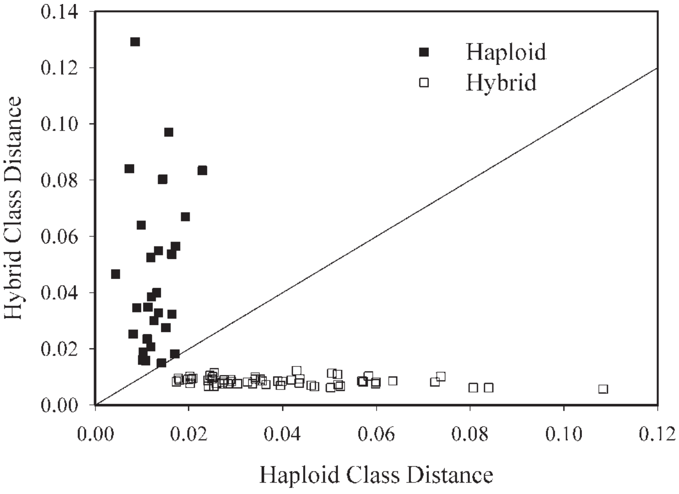

FIG. 4. Class–distance plot for separating haploid and hybrid training-set spectra for the P2-20 genotype. All samples are on the correct side of the diagonal line separating the two classes.

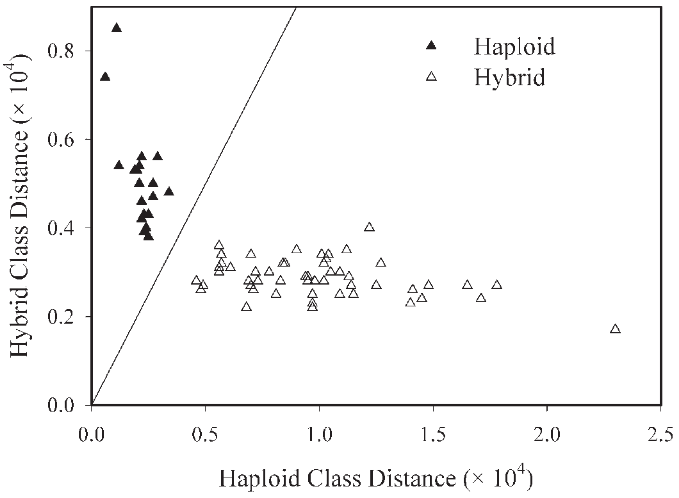

FIG. 6. Class–distance plot for separating haploid and hybrid training-set spectra of the P1-103 genotype. All samples are on the correct side of the diagonal line separating classes.

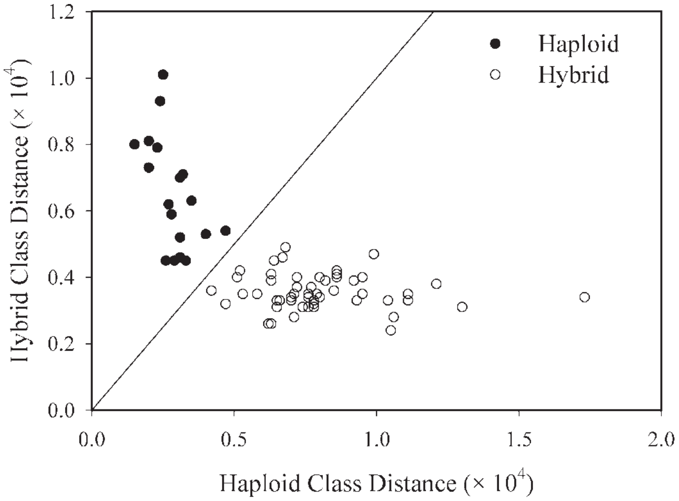

FIG. 7. Class–distance plot for separating haploid and hybrid training-set spectra of the P1-102 genotype. All samples are on the correct side of the diagonal line separating classes.

second-derivative pretreatment, with 10 and 12 factors for the haploid and hybrid classes, respectively.

CONCLUSION

Despite the problem with the P1-102 haploids, SIMCA modeling of NIR spectra appears to be a successful method of differentiating haploid and hybrid maize kernels. All of the kernels could be assigned to their proper genotype, and once that was done, all training-set kernels could be assigned to the correct haploid/hybrid class, as were all 25 hybrid and 11 of the 13 haploid kernels used in the unknowns sets for the various genotype-specific models. Future work is planned to test this screening method on a wider range of genotypes, to determine how small the number of calibration seeds used can be, and to increase the instrumental throughput.

ACKNOWLEDGMENTS

This research was supported by an Iowa State University Plant Sciences Institute Trans-disciplinary Innovative Research Grant. Work at the Ames Laboratory was supported by the Department of Energy-Basic Energy Sciences under Contract No. DE-AC02-07CH11358.

1. *Near-Infrared Transmittance Handbook (NIRT)* (Federal Grain Inspection Service, U.S. Department of Agriculture, Washington, D.C., 2006).
2. E. E. Finney, Jr., and K. H. Norris, Trans. ASAE **21,** 581 (1978).
3. B. A. Orman and R. A. Schumann, Jr., J. Am. Oil Chem. Soc. **69,** 1036 (1992).
4. T. M. Baye, T. C. Pearson, and A. M. Settles, J. Cereal Sci. **43,** 236 (2006).
5. G. Spielbauer, P. Armstrong, J. W. Baier, W. B. Allen, K. Richardson, B. Shen, and A. M. Settles, Cereal Chem. **86,** 556 (2009).
6. J. Janni, B. A. Weinstock, L. Hagen, and S. Wright, Appl. Spectrosc. **62,** 423 (2008).
7. B. A. Weinstock, J. Janni, L. Hagen, and S. Wright, Appl. Spectrosc. **60,** 9 (2006).
8. B. L. Tillman, D. W. Gorbet, and G. Person, Crop Sci. **46,** 2121 (2006).
9. T. C. Pearson, D. T. Wicklow, E. B. Maghirang, F. Xie, and F. E. Dowell, Trans. ASAE **44,** 1247 (2001).
10. F. E. Dowell, T. C. Pearson, E. B. Maghirang, F. Xie, and D. T. Wicklow, Cereal Chem. **79,** 222 (2002).
11. D. Haefele, D. Sevenich, D. Jones, J. Janni, and S. Wright, Int. Sugar J. **109,** 154 (2007).
12. S. Wold and M. Sjöström, "SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy," in *Chemometrics: Theory and Application*, B. Kowalski, Ed. ACS Symposium Series **Vol. 52** (American Chemical Society, Washington, D.C., 1977), Chap 12**,** pp. 243–282.
13. P. J. Gemperline, L. D. Webber, and F. O. Cox, Anal. Chem. **61,** 138 (1989).
14. D. E. Rubio-Diaz, T. De Nardo, A. Santos, S. de Jesus, D. Francis, and L. E. Rodriguez-Saona, Food Chem. **120,** 282 (2010).
15. P. Geladi, D. MacDougall, and H. Martens, Appl. Spectrosc. **39,** 491 (1985).