JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Article

pubs.acs.org/JAFC

Principal Variance Component Analysis of Crop Composition Data: A Case Study on Herbicide-Tolerant Cotton

Jay M. Harrison,[†] Delia Howard,[‡] Marianne Malven,[‡] Steven C. Halls,[‡] Angela H. Culler,[‡] George G. Harrigan,*,[‡] and Russell D. Wolfinger[§]

[†]Maritz Holdings Inc., 1395 North Highway Drive, Fenton, Missouri 63099, United States

[‡]Monsanto Company, 800 North Lindbergh Boulevard, St. Louis, Missouri 63167, United States

[§]SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, United States

Supporting Information

ABSTRACT: Compositional studies on genetically modified (GM) and non-GM crops have consistently demonstrated that their respective levels of key nutrients and antinutrients are remarkably similar and that other factors such as germplasm and environment contribute more to compositional variability than transgenic breeding. We propose that graphical and statistical approaches that can provide meaningful evaluations of the relative impact of different factors to compositional variability may offer advantages over traditional frequentist testing. A case study on the novel application of principal variance component analysis (PVCA) in a compositional assessment of herbicide-tolerant GM cotton is presented. Results of the traditional analysis of variance approach confirmed the compositional equivalence of the GM and non-GM cotton. The multivariate approach of PVCA provided further information on the impact of location and germplasm on compositional variability relative to GM.

KEYWORDS: *cotton (Gossypium spp.), genetically modified, genetically engineered, herbicide-tolerant, composition, principal variance component analysis, PVCA*

INTRODUCTION

Genetically modified (GM) crops are subjected to rigorous premarket regulatory assessments that include numerous laboratory and field studies.[1] Compositional analyses represent one component of these assessments and typically include the measurement of levels of key nutrients such as protein, oil, fiber, amino acids (AAs), minerals, fatty acids (FAs), and vitamins, as well as crop-specific metabolites such as gossypol and cyclopropenoid fatty acids (CPFAs) in cotton or isoflavones in soybean. Results to date from these studies have demonstrated consistently that the effect of transgenic breeding, if any, is notably less than the impact of environment or germplasm on crop composition variation in conventional crops.[2] Indeed, levels of all crop components are influenced markedly by environment,[2−5] and this observation provides context to the evaluation of new GM crops when compared to conventional counterparts.

The customary approach employed for comparative compositional assessments is to apply analysis of variance (ANOVA) to each analyte (nutrient or antinutrient of interest). The means and variances of new crop varieties are compared to the means and variances of compositional analytes from a parental or near-isogenic comparator with contrasts within the ANOVA model. Values from other varieties (either in-study or derived from the literature) may be used to represent natural variability in crop composition. In the univariate ANOVA approach, however, a statistically significant difference between the test and the control varieties is usually of small magnitude and rarely indicates a meaningful difference, especially when compared to other sources of variation, such as geographical location. New statistical methods that simplify or provide more meaningful interpretations have been developed and proposed for the evaluation of the composition of GM crops in a regulatory setting. Harrison et al.[6−8] provided an introduction to the use of Bayesian statistical methodology to compare GM crops to their controls without testing for significance. Harrison and Harrigan[3] used boxplots and other graphical summaries to demonstrate the extent of variation among and within varieties.

Other approaches for describing the relative magnitudes of natural variation among varieties and variation due to effects of environments and their interactions with the varieties can also be considered for providing useful contextual information for judging the effect of genetic modification of crop plants. For example, Li et al.[9] developed the method of principal variance component analysis (PVCA) in the context of quantifying batch effects on the variation of responses in microarray gene expression studies. The success of this approach implies suitability for more extensive applications. PVCA combines the application of two popular data analysis procedures, principal components analysis (PCA) and variance components analysis, and has three main goals: (i) summarize large data sets with a smaller set of relevant variables, (ii) describe the percentages of variance in the original data that are explained by the new variables, and then (iii) describe the relative amounts of variation in those variables that can be explained by factors in the experimental design and other covariates. This approach, when applied to a compositional assessment of new GM crops,

Received: February 11, 2013
Revised: April 30, 2013
Accepted: May 6, 2013
Published: May 6, 2013

 © 2013 American Chemical Society

would provide an easily interpretable summary of the relative contribution of GM to compositional variability.

In the following discussion, we applied PVCA to a compositional study on a new herbicide-tolerant cotton developed by Monsanto Company. This cotton, MON 88701, is tolerant to dicamba and glufosinate herbicides. MON 88701 contains a demethylase gene from *Stenotrophomonas maltophilia* that expresses a dicamba mono-oxygenase protein to confer tolerance to dicamba and a bialaphos resistance (*bar*) gene from *Streptomyces hygroscopicus* that expresses the phosphinothricin *N*-acetyltransferase protein to confer tolerance to glufosinate.

As part of a comparative assessment, compositional analyses were conducted on seed from MON 88701 and a parental control, as well as a total of nine commercially available varieties harvested from multiple replicated fields in the United States during the 2010 growing season. Seed components measured included proximates and fibers, amino acids, fatty acids, vitamin E, minerals, gossypol, and CPFAs. Statistical analysis included ANOVA as well as the novel application of PVCA.

■ MATERIALS AND METHODS

Cotton Samples for Compositional Analyses. Acid-delinted seed samples were collected from MON 88701 grown in a 2010 U.S. production. The field production included cotton variety Coker 130 as a parental conventional control as well as commercially available varieties, termed references, to provide further information on compositional variability inherent to cotton. A list of the references, as well as field locations, is presented in Supplementary Table 1A in the Supporting Information. Eight sites were planted with a randomized complete block design, with four replicates per site containing MON 88701 in two separate plots (treated and not treated with herbicide), the conventional control, and four of the commercial reference varieties. A total of 37 control and reference replicates (Supplementary Table 1B in the Supporting Information) were excluded from the analyses due to adventitious presence of the MON 88701 event.

Compositional Analyses. Components assessed included proximates (ash, calories, carbohydrates by calculation, protein, and total fat), acid detergent fiber (ADF), neutral detergent fiber (NDF), crude fiber (CF), total dietary fiber (TDF), AAs (18 components), FAs (C8−C22), minerals (calcium, copper, iron, magnesium, manganese, phosphorus, potassium, sodium, and zinc), vitamin E, and antinutrients including gossypol and CPFAs (dihydrosterculic, malvalic, and sterculic). Data were converted to dry weight expressions based on moisture measurements on the corresponding samples. Compositional analyses were conducted at Covance Laboratories Inc., in Madison, Wisconsin. Brief descriptions of the methods utilized for the analyses are described below. Measurements were based on single sample analysis.

Proximates and Calories. Protein levels were determined by the Kjeldahl method.[10] Protein and other nitrogenous compounds in the samples were reduced to ammonia by digestion of the samples with sulfuric acid containing a catalyst mixture. The acid digest was made alkaline, and the ammonia was distilled and titrated with a standard acid. The percent nitrogen was determined and converted to percent protein by multiplication with 6.25. The fat content of the grain was determined using the Soxhlet extraction method.[11] Seed samples were weighed into cellulose thimbles containing sand or sodium sulfate and dried to remove excess moisture. Pentane was dripped through the samples to remove the fat. Extracts were then evaporated, dried, and weighed. The ash content was determined by combustion at 550 °C and gravimetric quantitation of the nonvolatile matter remaining.[12] The moisture content was determined by the loss of weight after drying to a constant weight in a 100 °C vacuum oven.[13] Carbohydrates by calculation used the fresh weight-derived data and the following equation:

$$\%\,\text{carbohydrates} = 100\% - (\%\,\text{protein} + \%\,\text{fat} + \%\,\text{moisture} + \%\,\text{ash})$$

Calories were calculated using the Atwater factors with the fresh weight-derived data and the following equation:

$$\text{calories (kcal/100 g)} = (4 \times \%\,\text{protein}) + (9 \times \%\,\text{fat}) + (4 \times \%\,\text{carbohydrates})$$

Fiber Analysis. ADF was determined by washing the tissue with an acidic boiling detergent solution to dissolve the protein, carbohydrate, and ash.[14] An acetone wash was used to remove the fats and pigments. The remaining lignocellulose fraction was determined gravimetrically. NDF was determined by treating the tissue with a neutral boiling detergent solution to dissolve the protein, carbohydrate, and ash. Fats and pigments were removed using an acetone wash. The remaining hemicellulose, cellulose, and lignin fractions were measured gravimetrically.[14,15] CF was quantitated as the loss on ignition of dried residue remaining after digestion of the sample with 1.25% sulfuric acid and 1.25% sodium hydroxide solutions under specific conditions.[16]

AA Composition. The samples were hydrolyzed in 6 N HCl for approximately 24 h at approximately 106−110 °C. Phenol was added to the 6 N HCl to prevent halogenation of tyrosine. Cystine and cysteine were converted to *S*-2-carboxyethylthiocysteine by the addition of dithiodipropionic acid.[17] Tryptophan was hydrolyzed from proteins by heating at approximately 110 °C in 4.2 N NaOH for 20 h. The samples were analyzed by high-performance liquid chromatography (HPLC) after preinjection derivatization.[18] The primary AAs were derivatized with *o*-phthalaldehyde (OPA), and the secondary AAs were derivatized with fluorenylmethyl chloroformate (FMOC).

FAs. The lipid was extracted and saponified with 0.5 N sodium hydroxide in methanol. The saponification mixture was methylated with 14% boron trifluoride in methanol. The resulting methyl esters were extracted with heptane containing an internal standard. The methyl esters of the FAs were analyzed by gas chromatography using external standards for quantitation.[19]

Minerals. The sample was dried, precharred, and ashed overnight in a muffle set to maintain 500 °C. The ashed sample was reashed with nitric acid, treated with hydrochloric acid, taken to dryness, and put into a solution of 5% hydrochloric acid. The amount of each element was determined at appropriate wavelengths by comparing the emission of the unknown sample, measured on the inductively coupled plasma spectrometer, with the emission of the standard solutions.[20]

CPFAs. The total lipid fraction was extracted from the sample using chloroform and methanol. A portion of the lipid fraction was then saponified with a mild alkaline hydrolysis. The free FAs were extracted with ethyl ether and hexane. The free FAs were then converted to their phenacyl derivatives with 2-bromoacetophenone. The derivatives were quantitated on a HPLC system equipped with an ultraviolet detector. The amounts of malvalic, sterculic, and dihydrosterculic acids were determined by comparison to external calibration curves of similarly derivatized reference standards.[21]

Free and Total Gossypol. For free gossypol, the sample was extracted with aqueous acetone. The solution was then filtered, and the free gossypol was reacted with aniline. For total gossypol analysis, the sample was extracted using a complexing reagent containing acetic acid, 3-amino-1-propanol, and dimethylformamide. The solution was then filtered, and the total gossypol was reacted with aniline.[22] For both analyses, the dianilinogossypol was quantitated spectrophotometrically using a standard curve.

Vitamin E. The sample was saponified to break down any fat and release vitamin E. The saponified mixture was extracted with ethyl ether and then quantitated by HPLC using a silica column.[23]

Statistical Analysis of Composition Data. *Univariate Analysis.* Studentized PRESS residuals tests were applied to each data set to

identify outliers. Sodium values from two Coker 130 replicates from the ARTI site were removed (PRESS > 6). To complete a statistical analysis for a component in this study, at least 50% of the values for an analyte had to be greater than the assay limit of quantitation (LOQ). Thirteen FAs (caprylic, capric, lauric, myristoleic, pentadecanoic, pentadecenoic, heptadecanoic, heptadecenoic, γ-linolenic, eicosenoic, eicosadienoic, eicosatrienoic, and arachidonic acids) did not meet this criterion and were excluded from statistical analysis. These FAs are known to be present in only low amounts in conventional cottonseed, if present at all. This study confirmed that this observation extended to MON 88701. If less than 50% of the observations for a component were below the assay LOQ, individual analyses that were below the assay LOQ were assigned a value equal to one-half the assay LOQ. This affected only sodium (LOQ = 100 ppm fwt), and applied were the following assignments: Coker 130, two replicates at KSLA and one replicate each at LACH and NMLC; MON 88701 (not treated with herbicide), one replicate at all sites except NCBD.

Compositional analytes of MON 88701 and Coker 130 were analyzed using a mixed model ANOVA. The analysis used the model

$$Y_{ijk} = U + T_i + L_j + B(L)_{jk} + LT_{ij} + e_{ijk}$$

where Y_{ijk} = unique individual observation, U = overall mean, T_i = variety effect, L_j = random site effect, $B(L)_{jk}$ = random block within site effect, LT_{ij} = random site by variety interaction effect, and e_{ijk} = residual error. For each analyte, mean comparison tests were conducted. SAS software was used to generate all summary statistics (PROC MEANS and PROC MIXED) and to perform all analyses.

PVCA. An overview of the key steps in the PVCA procedure is presented prior to details of its implementation in this study. The first step of the PVCA procedure is to normalize the responses, if necessary. For example, a variance-stabilizing transformation, such as a logarithm, may be applied to individual analytes that have skewed distributions due to a few extreme values.

The next step is to standardize each of the responses by subtracting the mean of all observations for analyte, then dividing that difference by the sample standard deviation, or $Z = (X - \overline{X})/S_x$. This standardization removes the influence of the units of measurement on the results and provides easier interpretation of the relative contributions of the analytes in subsequent steps.

The third step is to check for linear dependencies in the data. This can be done by computing the correlation matrix of the standardized variables and then computing the rank of that matrix. If the rank is less than the number of variables, then some of the variables must be dropped from the analyses. They supply redundant information, and they could adversely affect subsequent matrix calculations or interpretations of the results.

The fourth step is to apply principal component analysis to the correlation matrix. Principal components are linear combinations of the standardized variables of the following form:

$$P^1 = a^{11}Z^1 + a^{12}Z^2 + ... + a^{1k}Z^k$$
$$P^2 = a^{21}Z^1 + a^{22}Z^2 + ... + a^{2k}Z^k$$
$$...$$
$$P^k = a^{k1}Z^1 + a^{k2}Z^2 + ... + a^{kk}Z^k$$

where k is the number of analytes. The new variables $\{P^1, P^2, ..., P^k\}$ are obtained from an eigenanalysis of the correlation matrix. With eigenanalysis, matrix operations are used to express the $k \times k$ correlation matrix as a sum of products of eigenvectors and eigenvalues. The new variables $\{P^1, P^2, ..., P^k\}$ are constructed in a way such that: (a) The coefficients $a^{i1}, a^{i2}, ..., a^{ik}$, which correspond to the i-th eigenvector of the correlation matrix, are all between −1 and 1, and the sum of their squared values is 1. (b) The variance of P^i is λ^i, the i-th largest eigenvalue of the correlation matrix. (c) The sum of the variances is equal to the number of variables; that is, $\mathrm{Var}(P^1) + \mathrm{Var}(P^2) + ... + \mathrm{Var}(P^k) = \lambda^1 + \lambda^2 + ... + \lambda^k = k$. Thus, the percentage of the total variability in the data that can be explained by principal component i is $(100\%)(\lambda^i/k)$, and the average amount of variance explained by one of the principal components is 1. (d) The new variables $\{P^1, P^2, ..., P^k\}$ are uncorrelated with each other and thus represent different dimensions or underlying facets contained in the data.

The large set of k analytes can be effectively replaced with a smaller subset of the first few principal components, since correlations among subsets of the variables $\{Z^1, Z^2, ..., Z^k\}$ affect the derivation of the new variables $\{P^1, P^2, ..., P^k\}$. For example, in a compositional study, we might expect subsets of AAs to be correlated with each other in response to nitrogen availability. In turn, one of the principal components may include a weighted sum of those AA values. That principal component would then be correlated with each of those AAs and would thus serve as an effective surrogate variable to replace those AAs in subsequent analyses without much loss of information. Li et al.[9] recommended the retention of enough principal components to explain 60−90% of the variation in the data, with a maximum of 10 retained components. For our application, we chose to retain enough principal components to explain 80% of the variation, with a maximum of 10 components.

Li et al.[9] continued with their development of the original PVCA procedure with the principal components, but we added an intermediate step to assist with the interpretation of the results. In principal component analysis, an analyte Z^i may be correlated with two or more of the principal components, so its role may be difficult to interpret. Factor analysis is a method of deriving new linear combinations of variables from the principal components via rotation, a mathematical operation involving matrix multiplication. The rotated principal components can be generated in a number of different ways to satisfy various mathematical criteria. For our procedure, we chose the varimax rotation method.[24] Varimax factors have the following form, which can be expressed in terms of either the standardized variables or the principal components:

$$F^1 = b^{11}Z^1 + b^{12}Z^2 + ... + b^{1k}Z^k = c^{11}P^1 + c^{12}P^2 + ... + c^{1k}P^k$$
$$F^2 = b^{21}Z^1 + b^{22}Z^2 + ... + b^{2k}Z^k = c^{21}P^1 + c^{22}P^2 + ... + c^{2k}P^k$$
$$...$$
$$F^j = b^{j1}Z^1 + b^{j2}Z^2 + ... + b^{jk}Z^k = c^{j1}P^1 + c^{j2}P^2 + ... + c^{jk}P^k$$

where j ($\leq k$) is the number of components to be retained. The coefficients are calculated in a way such that F^1 and each of the other factor scores have a mean of 0 and a standard deviation of 1. The relative importance of the variables in the calculation of each factor is indicated by the magnitudes of the coefficients $\{b^{11}, ..., b^{jk}\}$. Influential analytes for each factor are indicated by coefficients with large absolute values.

As with principal components, the new variables $\{F^1, F^2, ..., F^j\}$are also uncorrelated with each other and thus provide insight into different dimensions or facets of the data that may be influenced by common experimental effects. Collectively, $\{F^1, F^2, ..., F^j\}$ explain the same amount of variance as the first j principal components. The principal components maximize the amount of variance explained by each successive component; however, the varimax rotation maximizes the variance of the squared coefficients. This quantity is maximized when the squared coefficients are close to zero or one. Consequently, the varimax rotation tends to produce new variables $\{F^1, F^2, ..., F^j\}$ in which each analyte Z^i is strongly correlated with at most one of the factors. Thus, the role of the analyte Z^i with respect to the variation in the data is more easily distinguished. The proportion of variance explained by F^i is the sum of the squared coefficients $\{b^{i1}, b^{i2}, ..., b^{ik}\}$ divided by the number of analytes k.

In our presentation, the factors $\{F^1, F^2, ..., F^j\}$ are arranged in descending order by the amounts of variation explained. However, in general, the factor F^i does not explain the same amount of variation as the principal component P^i. Also, unlike principal components, factor F^i could change if the number of principal components specified for retention is changed. For this reason, the principal component analysis is applied, and the number of components to retain is determined from that analysis, before the varimax rotation is applied.

After applying a varimax rotation to the principal components, the next step is to apply variance components analysis to each of the

derived factor variables $\{F^1, F^2, ..., F^j\}$. An ANOVA model is applied, and all of the sources of variation of interest to the researcher are modeled with random effects. Fixed effects are used only when a source of variation is not of interest and is to be considered as a nuisance effect. Li et al.[9] recommend the restricted maximum likelihood (REML) approach, since REML is the most efficient and accurate method to estimate variance components, especially when the experimental design is unbalanced, as is the case in our study.

In our application of PVCA, the variation among cotton varieties is of interest. However, we wish to distinguish the amounts of variation due to differences among GM, control, and reference varieties and among the varieties within each of these three genotype groups. For this purpose, a nested model was used, with genotype group effects having three levels [GM (both MON 88701 T and NT), control, or reference] and variety effects nested within the genotype group. Next, for each factor score F^i, the sum of the variance components from the ANOVA model was calculated, and the fraction of variance for that factor that could be attributed to that source of variation was calculated by dividing the corresponding variance component by the total variance. The final step of our procedure is to combine the percentages of variance among the analytes explained by the varimax factor rotation and the percentages of variance explained by the sources of variation in the experiment by simple multiplication.

One advantage of PVCA is that it can be performed with existing routines in statistical software packages. The JMP Genomics package has a routine for performing PVCA with unrotated principal components, or the procedure can be implemented in the standard JMP software with the Principal Components platform (which includes varimax rotation capabilities) and the Fit Model platform. Using SAS Version 9 software, the following procedures were applied for the analyses that are presented here: PROC STDIZE for data standardization, PROC PRINCOMP for PCA, PROC FACTOR for varimax rotation, and PROC MIXED for obtaining variance components in ANOVA. Additionally, a script in R for performing PVCA without the varimax rotation is available.[25]

No transformations were applied to the data before standardization. Two outlying values above the mean for sodium were identified, but these were not transformed for this example. The sum of the percentages of the total FAs in seed was 100% in each sample. To eliminate this linear dependency, the variable representing behenic acid in seed was dropped from the analyses, since it was small in abundance and showed no significant differences between GM and control in the traditional univariate analysis.

RESULTS AND DISCUSSION

Univariate Analysis. The compositional evaluation of MON 88701 [herbicide-treated (T) and not treated (NT)] was conducted by analyzing seed data across all eight individual sites. Results from this combined-site analysis are presented in Tables 1−5. Overall, the data confirmed that transgene insertion had negligible impact on cotton composition and that MON 88701 (T and NT) is compositionally equivalent to a parental conventional control. The combined-site analysis highlighted extensive overlap in the range of values for respective components in the seed of MON 88701 (T and NT) and the parental conventional control. This overlap reflected similarities in geographical variability as well as generally small magnitudes of difference observed between MON 88701 (T and NT) and the parental conventional control components at the individual site level. The novel application of PVCA was therefore used to quantify the relative contribution of the experimental factors in this study and to evaluate its use as a complement or alternative method to traditional univariate methods.

PVCA on Compositional Data. For the following discussion, the principal components and factor scores will be said to be strongly correlated with individual compositional analytes if the absolute value of the linear correlation between the derived variable and the analyte exceeds 0.707, with the interpretation that at least 50% of the variation in the analyte can be explained by its linear relationship with the new variable (see Supplementary Table 2 and Figure 1 in the Supporting Information). Absolute correlations between 0.5 and 0.707 can be described as moderately correlated, with the amount of variation in the analyte that can be explained by the factor ranging between 25 and 50%.

The principal component analysis showed that seven principal components were needed to meet the threshold of 80% (as discussed in the Materials and Methods) of the total variance explained, so the first seven components were retained for subsequent examination (see Supplementary Figure 1 in the Supporting Information for PCA plot). Each of the first seven eigenvalues was greater than 1, indicating that each of these components explained more than the average amount of variance that could be explained by all 51 principal components.

The varimax rotation was applied to the first seven principal components to form a set of seven rotated factors. The advantage of the varimax rotation can be illustrated with an example involving stearic acid, which was not strongly correlated by any single principal component (P^6, $r = -0.43$; P^7, $r = -0.43$; all other correlations, $|r| \leq 0.30$). Thus, the role of stearic acid is difficult to discern from the principal components. However, stearic acid had a strong correlation with only one varimax factor (F^6, $r = 0.72$; all other correlations, $|r| \leq 0.19$), so the correspondence of stearic acid with factor F^6 is easier to discern. In our example, each analyte had a strong correlation with at most one factor.

Factor F^1: Protein and AAs. The first factor, F^1, obtained by varimax rotation of the seven principal components was strongly correlated with total protein and with 17 AAs. $F^1 = 0.05\ Z(\text{alanine}) + 0.05\ Z(\text{arginine}) + ... + 0.05\ Z(\text{valine})$, a linear combination of the Z scores of all 51 analytes. Coefficients for all factors are listed in Supplementary Table 2 in the Supporting Information with the information presented in heatmap and dendrogram format in Figure 1. F^1 was also strongly negatively correlated with total carbohydrates ($r = -0.79$). F^1 explained a total of 38% of the variation in the compositional data. Obviously, the protein and AAs should be correlated with each other, providing some validation of the PVCA application. The negative correlation of carbohydrates with protein is also expected.

Next, ANOVA with only random effects was used to estimate the variance in F^1 from the following sources: genotype group [GM (treated, T, or not treated, NT), control, or conventional], variety nested within genotype group, site, genotype group by site interaction, interaction of site by variety nested within genotype group, replicates nested within site, and residual error. This analysis showed that the largest source of variation in F^1 was variation among the sites, which explained 50% of the variance (Table 6). The next largest contributor to variation in F^1 was the variation among the three genotype groups, which explained 26% of the variance. The third largest contributor to variation was residual error, which explained 16% of the variance in F^1. Sources of variance that explain smaller relative amounts of variation than the residual error can be regarded as negligible and will not be explained further.

From these results, it can be concluded that the predominant source of variation in cottonseed composition in this study was due to the variation of proteinogenic AAs, which was mainly

dx.doi.org/10.1021/jf400606t | J. Agric. Food Chem. 2013, 61, 6412−6422

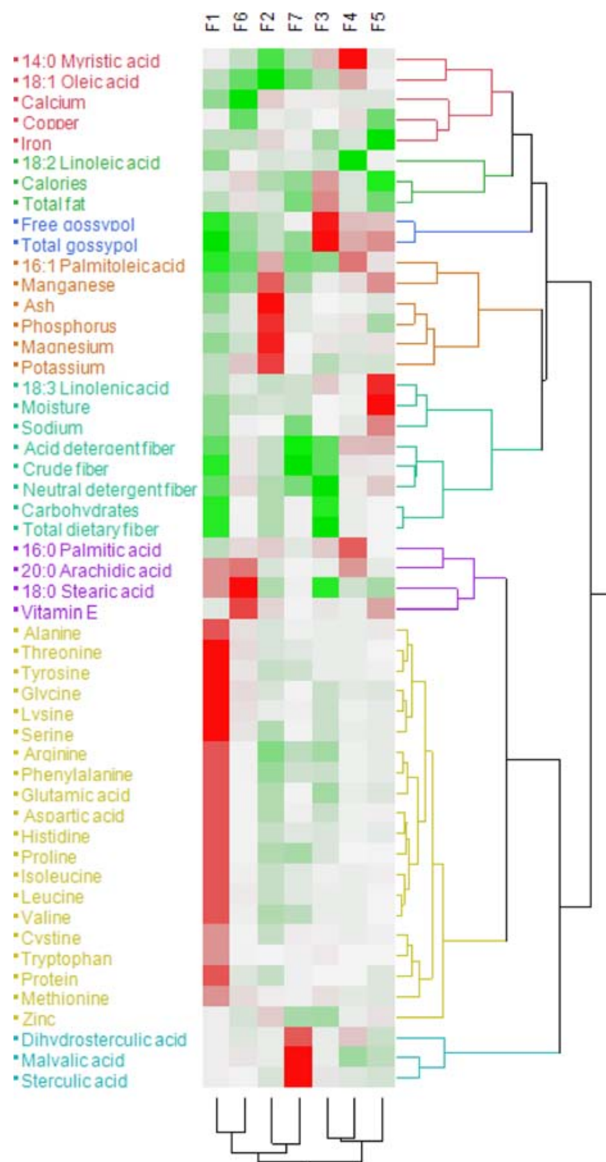

Figure 1. Heatmap and dendrogram of varimax coefficient scores. Dendrograms are schematic representations of multivariate distances, with joined segments indicating similarity.

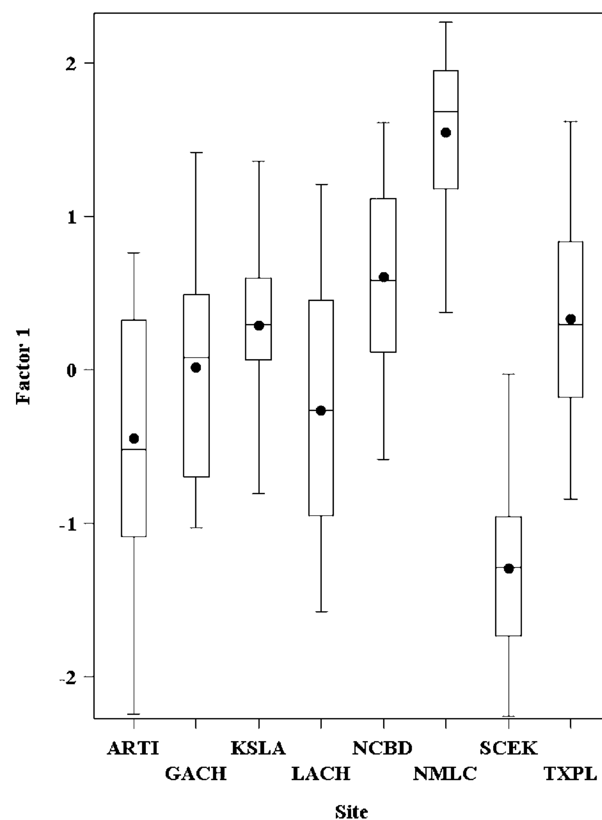

Figure 2. Variation of the first varimax factor F^1 among sites. Filled circles indicate means, and horizontal lines on a boxplot, from top to bottom, represent the maximum, third quartile, median, first quartile, and minimum.

associated with variation among sites. Figure 2 contains boxplots to display the variation in values of F^1 among the sites. Figure 3 shows the variation among varieties in F^1 and shows that, despite the variation among variety groups, MON 88701 (T and NT) was generally similar in protein composition to the parental control. These results are consistent with the lack or any significant difference between MON 88701 (T and NT) and control protein composition (Table 1).

The AA composition of MON 88701 (T and NT) and the parental control were also similar as observed in the original ANOVA. No significant differences for 15 of the 18 AAs were observed. Statistically significant differences were observed only for the following AAs: arginine (T and NT), methionine (T only), and proline (T only). The small magnitudes of difference between the MON 88701 and the control for these AAs (<3% relative, with respect to the control) reflect the lack of meaningful differences in protein values. The data provided here demonstrated that MON 88701 is not a major contributor to protein and AA variability in cottonseed and confirmed the compositional equivalence of MON 88701 to the parental control in levels of these analytes.

Protein expression is a quantitative trait, and levels are influenced by both genotype and environment.[26−28] The results obtained through PVCA support the conclusion of a lack of meaningful impact of transgenic breeding and are consistent with studies[29] that show that protein levels in cottonseed are affected by geography.

Factor F^2: Minerals. The factor F^2 explained an additional 9% of the variation in the composition data (Supplementary Table 3 in the Supporting Information). F^2 was strongly correlated with magnesium (r = 0.89), phosphorus (r = 0.83), ash (r = 0.91), and potassium (r = 0.82). Again, the major source of variation was due to site, which explained 69% of the variation in F^2. Residual error was the next largest contributor to variance (14%).

The ash and mineral composition of MON 88701 (T and NT) and the parental control were very similar. Mean values for ash in MON 88701 (T and NT) and the control were 4.31, 4.30, and 4.09% dwt, respectively (Table 2); a difference at the 5% significance level was observed. Differences between MON 88701 (both T and NT) and the control were observed for calcium, magnesium, manganese, potassium, and zinc and between MON 88701 (NT) and the control for copper. All differences were also characterized by small magnitudes (all <10% relative with respect to the control). Values were also characterized by extensive overlap, an observation most likely attributed to the established impact of location on mineral composition as highlighted by PVCA. Ash and mineral levels in

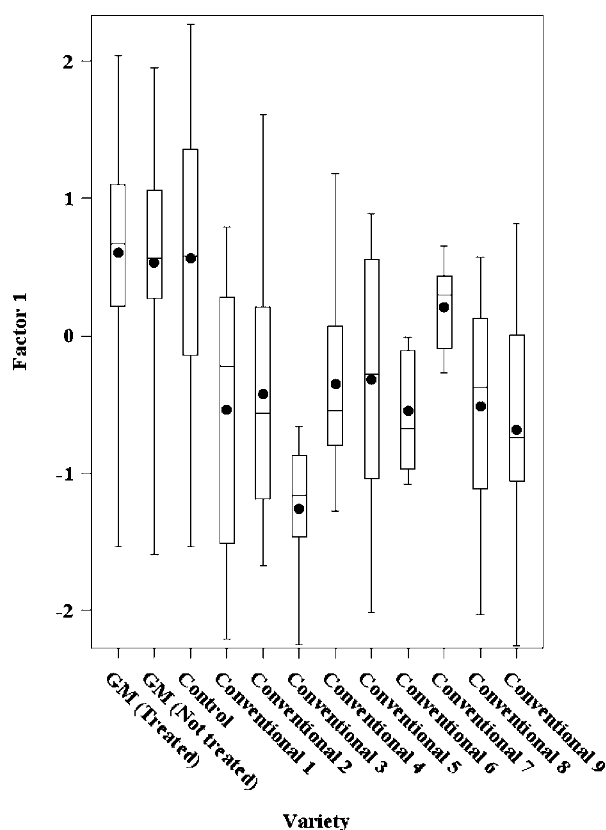

Figure 3. Variation of the first varimax factor F^1 among varieties.

cottonseed are known to vary extensively,[27] and the results from the univariate and PVCA approaches are consistent with that.

Factor F^3: Gossypol. The factor F^3 explained an additional 8% of the variation (Supplementary Table 4 in the Supporting Information). F^3 was strongly correlated with total gossypol (r = 0.85) and free gossypol (r = 0.84). The major source of variation in F^3 was due to differences between variety within genotype group, which explained 44% of the variation. Variation among sites (28%) and among genotype groups (17%) was also larger than the residual error variation (8%). Figure 4 shows that the variation among the nine conventional varieties was large as compared to the variation among the GM and control varieties.

Differences between MON 88701 (T and NT) and the control at the 5% significance level were observed for total gossypol. Mean values for total gossypol in MON 88701 (T and NT) and the control were 1.04, 1.03, and 0.97% dwt, respectively. Mature cottonseed demonstrates considerable variation in levels of gossypol. Environmental factors such as year effect, rainfall, and stress have been shown to impact gossypol levels.[30−32] Given the known variability of gossypol due to other factors, it can be concluded that MON 88701 does not contribute to variation in gossypol levels (Table 3).

Factors F^4 and F^5: FAs and Moisture. The factor F^4 explained an additional 7% of the variation (Supplementary Table 5 in the Supporting Information). F^4 was strongly positively correlated with myristic acid (r = 0.83) and palmitic acid (r = 0.71) and strongly negatively correlated with linoleic acid (r = −0.93). The largest source of variation was due to varieties within genotype groups (63%), followed by sites (29%) and residual error (4%). The factor F^5 was also influenced by FA composition and explained 7% of the variation (Supplementary Table 6 in the Supporting Information). F^5 was strongly positively correlated with linolenic acid (r = 0.77) and moisture (r = 0.83). Variation among sites (76%) was the primary source of variation,

Table 1. Summary of Protein and Amino Values for MON 88701

component[a]	mean[b] (range[c])			reference range[c]	main factor[d]	effect[e]
	MON 88701 (T)	MON 88701 (NT)	Coker 130			
protein	27.91 (22.71−31.47)	27.71 (22.49−31.29)	27.79 (23.53−31.27)	20.58−29.28	F^1	site
alanine	1.06 (0.91−1.14)	1.05 (0.88−1.15)	1.05 (0.88−1.17)	0.83−1.22	F^1	site
arginine	3.03[f] (2.33−3.60)	3.03[f] (2.31−3.62)	3.15 (2.41−3.77)	2.30−3.55	F^1	site
aspartic acid	2.39 (1.94−2.64)	2.39 (1.95−2.69)	2.40 (1.92−2.74)	1.79−2.72	F^1	site
cystine	0.41 (0.32−0.47)	0.40 (0.31−0.46)	0.40 (0.31−0.46)	0.29−0.47	F^1	site
glutamic acid	4.76 (3.80−5.38)	4.71 (3.79−5.57)	4.84 (3.66−5.70)	3.39−5.45	F^1	site
glycine	1.10 (0.93−1.19)	1.09 (0.92−1.19)	1.09 (0.91−1.20)	0.85−1.23	F^1	site
histidine	0.74 (0.58−0.85)	0.74 (0.58−0.84)	0.75 (0.61−0.84)	0.57−0.84	F^1	site
isoleucine	0.91 (0.75−1.01)	0.91 (0.76−1.01)	0.92 (0.77−1.03)	0.72−1.03	F^1	site
leucine	1.53 (1.29−1.70)	1.53 (1.28−1.68)	1.54 (1.28−1.69)	1.20−1.72	F^1	site
lysine	1.24 (1.05−1.38)	1.24 (1.03−1.37)	1.23 (1.06−1.39)	0.99−1.44	F^1	site
methionine	0.40[f] (0.35−0.46)	0.39 (0.33−0.44)	0.38 (0.32−0.46)	0.29−0.49	F^1	site
phenylalanine	1.43 (1.14−1.66)	1.43 (1.13−1.63)	1.45 (1.15−1.66)	1.10−1.63	F^1	site
proline	1.00[f] (0.82−1.21)	1.02 (0.78−1.16)	1.03 (0.81−1.25)	0.79−1.17	F^1	site
serine	1.43 (1.14−1.66)	1.08 (0.93−1.28)	1.09 (0.86−1.24)	0.81−1.24	F^1	site
threonine	0.87 (0.74−0.94)	0.87 (0.73−0.95)	0.86 (0.73−0.95)	0.67−0.96	F^1	site
tryptophan	0.41 (0.33−0.52)	0.41 (0.34−0.50)	0.42 (0.37−0.52)	0.31−0.46	F^1	site
tyrosine	0.81 (0.67−0.92)	0.81 (0.68−0.88)	0.81 (0.67−0.91)	0.63−0.91	F^1	site
valine	1.21 (1.00−1.40)	1.21 (0.98−1.38)	1.23 (1.00−1.40)	0.97−1.36	F^1	site

[a]Expressed as % dwt. [b]Least-square mean. [c]Minimum to maximum of individual replicate values across all sites. [d]Factor to which an analyte shows the highest coefficient. [e]Greatest source of variance for a given factor. This table shows that the greatest source of variance for these analytes was site, even when a significant mean difference (α = 0.05) was observed between MON 88701 (T and NT) and the parental control, Coker 130. [f]Mean different from control at α = 0.05.

Table 2. Summary of Ash and Mineral Values for MON 88701

component[a]	mean[b] (range[c])			reference range[c]	main factor[d]	effect[e]
	MON 88701 (T)	MON 88701 (NT)	Coker 130			
ash	4.31[f] (3.77−4.74)	4.30 (3.76−4.88)	4.09 (3.34−5.00)	3.18−4.68	F^2	site
calcium	0.15[f] (0.10−0.22)	0.15f (0.10−0.21)	0.13 (0.08−0.19)	0.08−0.18	F^6	site
copper	8.90 (5.22−11.91)	8.94[f] (5.02−12.15)	8.92 (5.40−11.92)	4.46−11.62	F^1	site
iron	67.21 (41.96−83.17)	72.43 (41.73−109.70)	71.16 (45.03−95.10)	39.49−114.34	F^5	site
magnesium	0.40[f] (0.35−0.44)	0.40[f] (0.35−0.45)	0.38 (0.33−0.44)	0.31−0.46	F^2	site
manganese	12.81[f] (10.18−14.81)	13.63[f] (10.59−17.47)	11.73 (8.61−14.11)	9.07−17.14	F^2	site
phosphorus	0.72 (0.56−0.84)	0.71 (0.58−0.87)	0.72 (0.54−0.87)	0.48−0.87	F^2	site
potassium	1.12[f] (0.98−1.24)	1.13[f] (0.99−1.32)	1.07 (0.79−1.27)	0.90−1.26	F^1	site
sodium	0.03 (0.02−0.12)	0.026 (0.0053−0.082)	0.03 (0.01−0.10)	0.01−0.08	F^5	site
zinc	37.58[f] (27.31−46.74)	37.81[f] (27.60−46.04)	40.14 (28.22−52.95)	25.07−48.49	F^1	site

[a]Expressed as % dwt except for copper, iron, manganese, and zinc, which are expressed as mg/kg dwt. [b]Least-square mean. [c]Minimum to maximum of individual replicate values across all sites. [d]Factor to which an analyte shows the highest coefficient. [e]Greatest source of variance for a given factor. This table shows that the greatest source of variance for these analytes was site location, even when a significant mean difference (α = 0.05) was observed between MON 88701 (T and NT) and the parental control, Coker 130. [f]fMean different from control at α = 0.05.

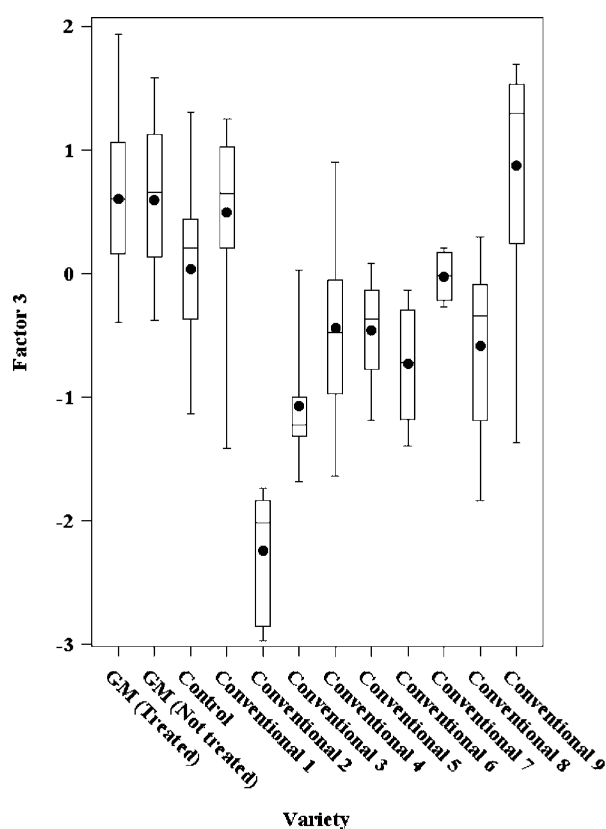

Figure 4. Variation of factor F^3 among varieties.

followed by variation within genotype groups (14%) and residual error (7%).

The FA composition of MON 88701 (T and NT) and the control was very similar (Table 4). Differences between MON 88701 and the control were observed at the 5% significance level for myristic acid (T and NT) and linoleic acid (T). These statistically significant differences reflected small differences in mean values for these FAs in MON 88701 (T and NT) and the control; 0.02% total FA for myristic acid and 0.39% total FA for linoleic acid.

The major FAs in terms of abundance in cottonseed are palmitic acid, stearic acid, oleic acid, linoleic acid, palmitoleic acid, and myristic acid. In efforts to change chemical properties of cottonseed oil, these FAs are targets of conventional breeding efforts to modify their relative levels.[33] In an ANOVA among cottonseed cultivars, it was observed that most of phenotypic variance of FA was due to genotype and environmental effects.[33,34] Environmental differences such as crop year influence FA profiles,[35] while variety differences such as glanded and glandless also have been shown to impact FA profiles in cottonseed.[36] This extensive literature on FA composition, as well as the results from PVCA, confirms that transgenic breeding in cotton is not a major source of variability in these components.

Factor F^6: Stearic Acid and Vitamin E. The factor F^6 explained 6% of the variation. F^6 was strongly positively correlated with stearic acid (r = 0.72) and vitamin E (r = 0.71). The largest contributor to variation in F^6 was sites (67%), followed by varieties within genotype groups (14%) and residual error (7%) (Supplementary Table 7 in the Supporting Information).

No difference between MON 88701 (T and NT) and the control was observed for stearic acid, consistent with the observation noted above that the overall FA composition between the comparators was very similar. It is intriguing, however, that stearic acid segregates in the factor analysis from the other FAs as well as showing a correlation with vitamin E. These findings suggest that PVCA could be a useful exploratory tool in evaluating relationships between different seed components that may be of value to breeding programs.

A difference between MON 88701 (T and NT) and the control was observed at the 5% significance level for vitamin E. This difference reflected mean values of 140.14, 139.01, and 131.33 mg/kg dwt; MON 88701 (T and NT) and the control, respectively. This small magnitude of difference can be evaluated in the context of variability in individual replicate values for the control: 91.78 to 162.98 mg/kg dwt. The large variation in values can be attributed to the influence of sites and is consistent with literature observations.

Vitamin E levels are known to be affected by environment and genotype. For example, Smith and Creelman[37] showed a significant difference in vitamin E levels across 18 varieties grown at a single location in the United States over a period of 2 years. Smith and Creelman[37] also point out that the environmental factors associated with variability in vitamin E levels include year and temperature. The results obtained

Table 3. Summary of Antinutrient Values for MON 88701

component[a]	mean[b] (range[c])			reference range[c]	main factor[d]	effect[e]
	MON 88701 (T)	MON 88701 (NT)	Coker 130			
free gossypol	0.94[f] (0.80−1.18)	0.93 (0.76−1.10)	0.89 (0.68−1.20)	0.50−1.41	F^3	variety (genotype)
total gossypol	1.04[f] (0.84−1.24)	1.03[f] (0.84−1.52)	0.97 (0.74−1.10)	0.56−1.61	F^3	variety (genotype)
malvalic acid	0.39 (0.20−0.55)	0.39 (0.24−0.50)	0.37 (0.26−0.49)	0.11−0.59	F^7	residual
sterculic acid	0.22 (0.13−0.29)	0.22 (0.12−0.27)	0.21 (0.17−0.27)	0.06−0.34	F^7	residual
dihydrosterculic acid	0.15[f] (0.11−0.19)	0.15[f] (0.12−0.19)	0.14 (0.11−0.17)	0.04−0.23	F^7	residual

[a]Gossypols expressed as % dwt, acids expressed as total FAs. [b]Least-square mean. [c]Minimum to maximum of individual replicate values across all sites. [d]Factor to which an analyte shows the highest coefficient. [e]Greatest source of variance for a given factor. This table shows that the greatest source of variance for these analytes was substance within a genotype (GM or non-GM group) and/or residual error, even when a significant mean difference (α = 0.05) was observed between MON 88701 (T and NT) and the parental control, Coker 130. [f]Mean difference from control at α = 0.05.

Table 4. Summary of Fat, FA, and Vitamin E Values for MON 88701

component[a]	mean[b] (range[c])			reference range[c]	main factor[d]	effect[e]
	MON 88701 (T)	MON 88701 (NT)	Coker 130			
total fat	23.14[f] (19.79−26.78)	23.51[f] (20.99−25.54)	22.31 (20.71−25.20)	16.58−25.25	F^c	variety (genotype)
myristic	0.77[f] (0.66−0.95)	0.77[f] (0.66−0.95)	0.79 (0.71−0.98)	0.45−1.04	F^4	variety (genotype)
palmitic	23.95 (22.34−25.28)	23.93 (22.30−25.45)	23.80 (22.69−25.05)	19.11−26.73	F^4	variety (genotype)
palmitoleic	0.50 (0.44−0.54)	0.50 (0.45−0.55)	0.50 (0.45−0.54)	0.44−0.67	F^4	variety (genotype)
stearic	2.54 (2.29−2.85)	2.52 (2.16−2.93)	2.47 (2.15−2.76)	1.98−2.97	F^6	site
oleic	15.10 (14.15−16.45)	15.05 (14.05−16.29)	14.96 (14.06−16.44)	13.71−18.39	F^2	site
linoleic	55.77[f] (54.24−58.22)	55.84 (54.22−58.48)	56.15 (54.04−57.93)	49.78−59.61	F^4	variety (genotype)
linolenic	0.18 (0.14−0.34)	0.18 (0.11−0.38)	0.17 (0.12−0.21)	0.72−1.03	F^5	site
arachidic	0.29 (0.23−0.32)	0.29 (0.23−0.32)	0.28 (0.23−0.32)	1.79−2.72	F^4	variety (genotype)
behenic	0.15 (0.12−0.19)	0.15 (0.12−0.17)	0.15 (0.13−0.21)	0.29−0.47		
vitamin E	140.14[f] (86.23−179.34)	139.01[f] (87.22−184.47)	131.33 (91.78−162.98)	84.07−162.76	F^6	site

[a]Expressed as % dwt for total fat, % total FA for all FAs, and mg/kg dwt for vitamin E. [b]Least-square mean. [c]Minimum to maximum of individual replicate values across all sites. [d]Factor to which an analyte shows the highest coefficient. [e]Greatest source of variance for a given factor. This table shows that the greatest source of variance for these analytes was site location and/or variety within a genotype (GM or non-GM group), even when a significant mean difference (α = 0.05) was observed between MON 88701 (T and NT) and the parental control, Coker 130. [f]Mean different from control at α = 0.05.

Table 5. Summary of Fiber and Proximate Values for MON 88701

component[a]	mean[b] (range[c])			reference range[c]	main factor[d]	effect[e]
	MON 88701 (T)	MON 88701 (NT)	Coker 130			
ADF	25.27[f] (23.26−27.74)	25.53[f] (23.30−30.43)	26.58 (22.08−29.58)	23.42−31.62		
NDF	30.73[f] (25.13−34.42)	31.43[f] (28.05−37.27)	32.59 (28.87−35.89)	29.27−40.63	F^1	site
TDF	39.44 (36.91−42.13)	39.75[f] (36.22−43.22)	41.13 (39.05−44.37))	37.29−48.60	F^1	site
CF	18.17 (15.97−21.66)	17.78[f] (14.54−20.73)	18.54 (16.06−21.70)	16.92−23.32	F^1	site
calories	498.5[f] (482.46−517.46)	500.37[f] (487.62−511.92)	495.24 (487.70−512.65)	466.09−509.91	F^5	site
carbohydrates by calculation	44.64[f] (41.40−48.89)	44.47[f] (41.07−48.81)	45.83 (42.14−50.30)	43.28−54.90	F^1	site

[a]Expressed as % dwt except calories, which are expressed as kcal/100g dwt. [b]Least-square mean. [c]Minimum to maximum of individual replicate values across all sites. [d]Factor to which an analyte shows the highest coefficient. [e]Greatest source of variance for a given factor. This table shows that the greatest source of variance for these analytes was site location, even when a significant mean difference (α = 0.05) was observed between MON 88701 (T and NT) and the parental control, Coker 130. [f]Mean different from control at α = 0.05.

through PVCA support the conclusion of a lack of meaningful impact of transgenic breeding.

Factor F^7: Cyclopropenoid Fatty Acids. The factor F^7 was strongly correlated with the three cyclopropenoid fatty acids (CFPA): dihydrosterculic (r = 0.76), sterculic (r = 0.94), and malvalic (r = 0.92). The largest contributor to variation in F^7 was residual error (87%) (Supplementary Table 8 in the Supporting Information). Differences between MON 88701 (T and NT) and the control at the 5% significance level were observed for dihydrosterculic acid.

CPFAs in cotton include malvalic acid, sterculic acid, and dihydrosterculic acid. Geographic location and variety have been shown to contribute to natural variability in CPFA levels in conventional cotton.[36,38]

Concluding Remarks. As previously described, variation among sites was the primary contributor to variation in four of the seven factors, while variation among the varieties within a genotype was the largest component of variance for two other factors. In cases where the variation among the variety groups was larger than residual error variation, boxplots of the values showed that the two GM varieties were, in fact, similar to their near-isogenic control variety, while other conventional varieties varied widely. Variation among replicates and variation due to interaction between varieties and sites was negligible for all

Table 6. Summary of Factor 1 Sources of Variance

source of variance	fraction of variance due to source (%)	PVC[a] contribution (%)
site	50.43	19.26
genotype_group	25.84	9.87
residual	15.88	6.06
variety(genotype_group)	4.40	1.68
rep(site)	1.82	0.69
site*genotype_group	1.63	0.62
site*variety(genotype)	0.00	0.00
	100	38.19

[a]PVC = principal variance component.

seven factors. Figure 5 shows the cumulative contributions of each variance component over all factors, while Figure 6 shows the same information when grouped by factor.

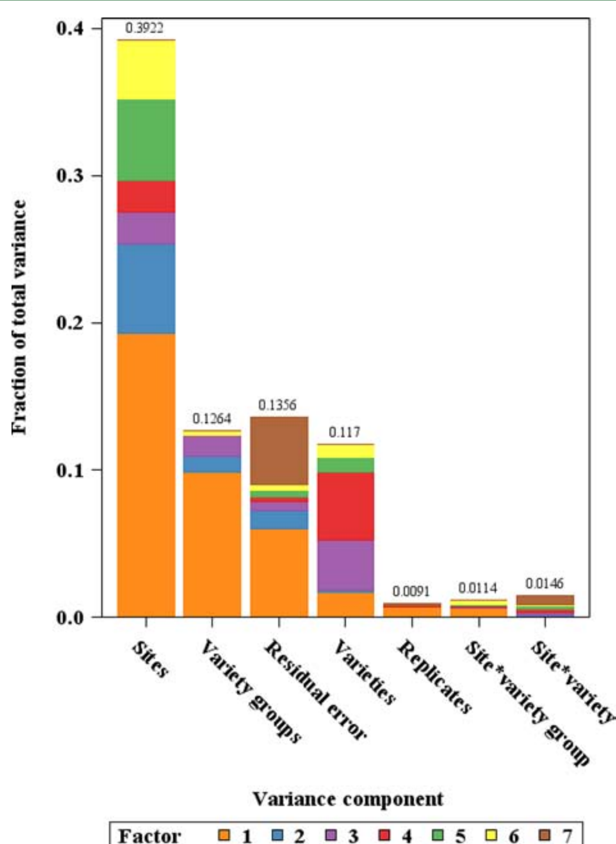

Figure 5. Proportions of variance explained by variance components for each factor. Numbers in the plot represent the cumulative proportions of variation in the data that are attributed to each variance component, totaled over factors.

Our implementation of PVCA offers the following advantages for interpreting composition data.

1. The use of multivariate analyses allows more unified interpretations of the effects of genetic modification, environment, and natural genotypic variation on biochemical systems than we can obtain from the customary univariate ANOVA for individual analytes. For example, PVCA showed that several minerals were correlated with each other and could be considered as measurements of one phenomenon that was most strongly associated with environmental variation.

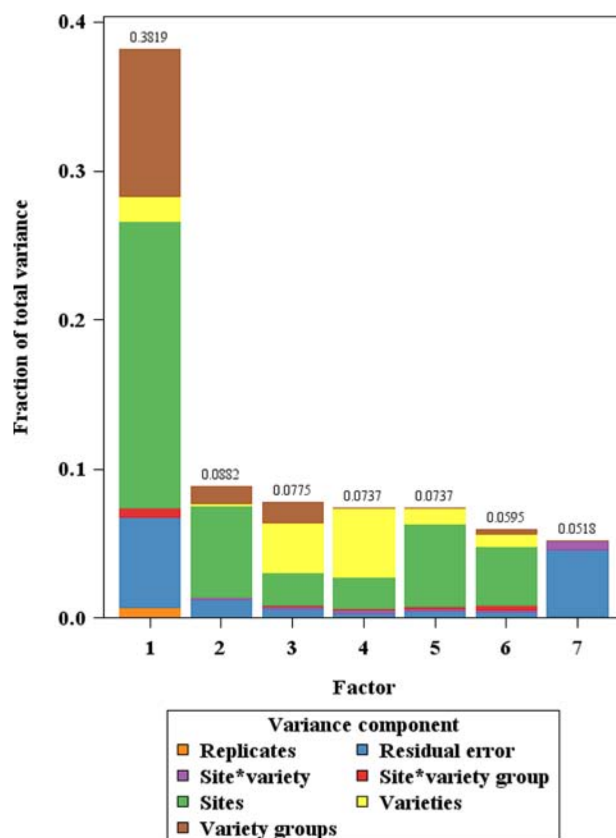

Figure 6. Proportions of variance explained by each factor, with relative contributions from each variance component. Numbers in the plot represent the cumulative proportions of variance in the data that are attributed to each factor.

2. The use of variance components allows the judgments of differences between GM and conventional plants to be made within the context of natural genotypic and environmental variation.

3. PVCA offers an effective bridge between the traditional ANOVA approach that is used with designed field trials in agronomy and the multivariate approach that is favored in chemistry and genetics.

4. The principal components and varimax factor rotations could highlight relationships among the analytes to prompt further scientific investigation.

5. Extraneous or nuisance variables that were not a part of the design of the experiment but are anticipated to affect the levels of the analytes can be modeled with fixed effects to remove their influence on the estimation of the variance components.

6. Even though the intermediate calculations involve some complicated mathematical concepts, the final results are phrased in term of percentages and are easy to convey to a broad audience. Misinterpretations of statistical significance are less likely to occur with this simplified presentation of findings.

7. PVCA can be applied with no additional modification for situations in which the intended effect of transgene insertion is a change in composition, such as an oil quality trait.

8. PVCA is relatively easy to perform using standard software packages such as JMP, SAS, and R.

If an appropriate standardized framework can be established, the application of PVCA to composition assessments for regulatory purposes could be considered especially in studies

where the comparative assessment applies to GM and non-GM comparators grown at multiple locations. In cases where modification of seed composition is the intended effect to achieve a quality trait, PVCA could also be used to confirm that, for example, soy oil profiles are major sources of variation that can be linked directly to the GM.

We propose that PVCA also provides a useful method for analyzing data from metabolomic profiling studies. Such studies offer some additional challenges, such as the inclusion of many more analytes than samples, missing values, censored values when analytes fall below the limit of detection, and the inclusion of large numbers of analytes that represent unknown compounds. Despite these challenges, PVCA can be applied to such data with some additional steps. These approaches are under development.

■ ASSOCIATED CONTENT

Supporting Information

Tables of commercial reference varieties, number of missing replicates, coefficients for varimax factor scores, summaries of sources of variance for factors 2-7, and accompanying figures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: george.g.harrigan@monsanto.com.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Kalaitzandonakes, N.; Alston, J. M.; Bradford, K. J. Compliance costs for regulatory approval of new biotech crops. *Nat. Biotechnol.* **2007**, *25*, 509−511.

(2) Harrigan, G. G.; Lundry, D.; Drury, S.; Berman, K.; Riordan, S. G.; Nemeth, M. A.; Ridley, W. P.; Glenn, K. C. Natural variation in crop composition and the impact of transgenesis. *Nat. Biotechnol.* **2010**, *28*, 402−404.

(3) Harrigan, G. G.; Harrison, J. M. Assessing compositional variability through graphical analysis and Bayesian statistical approaches; Case studies on transgenic crops. *Biotechnol. Genet. Eng. Rev.* **2012**, *28*, 15−32.

(4) Zhou, J.; Berman, K. H.; Breeze, M. L.; Nemeth, M. A.; Oliveira, W. S.; Braga, D. V.; Berger, G. U.; Harrigan, G. G. Compositional variability in conventional and glyphosate-tolerant soybean varieties grown in different regions in Brazil. *J. Agric. Food Chem.* **2011**, *59*, 11652−11656.

(5) Zhou, J.; Harrigan, G. G.; Berman, K. H.; Webb, E. G.; Klusmeyer, T. H.; Nemeth, M. A. Stability of the compositional equivalence of grain from insect-protected corn and seed from herbicide-tolerant soybean over multiple seasons, locations and breeding germplasms. *J. Agric. Food. Chem.* **2011**, *59*, 8822−8828.

(6) Harrison, J. M.; Breeze, M. L.; Harrigan, G. G. Introduction to Bayesian statistical approaches to compositional analyses of transgenic crops 1. Model validation and setting the stage. *Regul. Toxicol. Pharmacol.* **2011**, *60*, 381−388.

(7) Harrison, J. M.; Breeze, M. L.; Berman, K. H.; Harrigan, G. G. Bayesian statistical approaches to compositional analyses of transgenic crops 2. Application and validation of informative prior distributions. *Regul. Toxicol. Pharmacol.* **2013**, *65*, 251−258.

(8) Harrison, J. M.; Culp, D.; Harrigan, G. G. In *Bayesian MCMC analyses for regulatory assessments of food composition*, Proceedings of the 24th Annual Conference on Applied Statistics in Agriculture, Manhattan, Kansas, 2012.

(9) Li, J.; Bushel, P. R.; Chu, T.-M.; Wolfinger, R. D. Principal variance components analysis: Estimating batch effects in microarray gene expression data. In *Batch Effect and Experimental Noise in Microarray Studies: Sources and Solutions*; Scherer, A., Ed.; John Wiley & Sons: West Sussex, United Kingdom, 2009.

(10) Association of Official Analytical Chemists (AOAC). Nitrogen (total) in fertilizers. Methods 955.04. In *Official Method of Analysis of AOAC International*, 17th ed.; The Association of Official Analytical Chemists International: Gaithersburg, MD, 2000; Chapter 2, pp 14−15.

(11) Association of Official Analytical Chemists (AOAC). Fat (crude) or ether extract in meat. Method 960-39. In *Official Method of Analysis of AOAC International*, 17th ed.; Horwitz, W., Ed.; The Association of Official Analytical Chemists International: Gaithersburg, MD, 2000; Chapter 39, p 2.

(12) Association of Official Analytical Chemists (AOAC). Ash of flour. Method 923.03. In *Official Method of Analysis of AOAC International*, 17th ed.; Horwitz, W., Ed.; The Association of Official Analytical Chemists International: Gaithersburg, MD, 2000; Chapter 32, p 2.

(13) Association of Official Analytical Chemists (AOAC). Moisture in cheese. Method 926.08. In *Official Method of Analysis of AOAC International*, 17th ed.; Horwitz, W., Ed.; The Association of Official Analytical Chemists International: Gaithersburg, MD, 2000; p 70.

(14) United States Department of Agriculture (USDA). Forage fiber analysis. In *Agricultural Handbook*, Number 379; United States Department of Agriculture: Washington, DC, 1970.

(15) American Association of Cereal Chemists (AACC). Insoluble dietary fiber. Method 32.20. In *American Association of Cereal Chemists*, 9th ed.; AACC: Minneapolis, MN, 1998.

(16) Association of Official Analytical Chemists (AOAC). Fiber (Crude) in Animal Feed and Pet Food. Method 962.09. In *Official Method of Analysis of AOAC International*, 18th ed.; Horwitz, W., Ed.; The Association of Official Analytical Chemists International: Gaithersburg, MD, 2012; p 70.

(17) Barkholt, V.; Jensen, A. L. Amino acid analysis: Determination of cysteine plus half-cystine in proteins after hydrochloric acid hydrolysis with a disulfide compound as additive. *Anal. Biochem.* **1989**, *177*, 318−322.

(18) Schuster, R. Determination of amino acids in biological, pharmaceutical, plant and food samples by automated precolumn derivatization and high-performance liquid chromatography. *J. Chromatogr.* **1988**, *431*, 271−284.

(19) American Oil Chemists Society (AOCS). Fatty acid composition by gas chromatography. Method Ce 1-62. In *Official Methods and Recommended Practices of the American Oil Chemists Society*, 5th ed.; Firestone, D., Ed.; AOCS: Champaign, IL, 1997.

(20) Association of Official Analytical Chemists (AOAC). Multielement Methods 984.27 and 985.01. In *Official Method of Analysis of AOAC International*, 18th ed.; Horwitz, W., Ed.; The Association of Official Analytical Chemists International: Gaithersburg, MD, 2011.

(21) Randall, W. High performance liquid chromatography analysis of cyclopropene fatty acids. *Biochem. Arch.* **1986**, *2*, 63−71.

(22) *Official Methods and Recommended Practices of the American Oil Chemists' Society, Methods Ba 7-58 and Ba 8-78*; American Oil Chemists' Society: Champaign, IL.

(23) Speek, A. J.; Schijver, J.; Schreurs, W. H. P. Vitamin E composition of some seed oils as determined by high-performance liquid chromatography with fluorometric detection. *J. Food Sci.* **1985**, *50*, 121−124.

(24) Gorsuch, R. L. *Factor Analysis*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1983.

(25) National Institute of Environmental Health Sciences. Principal variance component analysis; 2011. Available at http://www.niehs.nih.gov/research/resources/software/biostatistics/pvca/index.cfm.

(26) Leffler, H. R.; Elmore, C. D.; Hesketh, J. D. Seasonal and fertility-related changes in cottonseed protein quantity and quality. *Crop Sci.* **1977**, *17*, 953−956.

(27) OECD. *Consensus Document on Compositional Considerations for New Varieties of Cotton (Gossypium hirsutum and Gossypium barbadense): Key Food and Feed Nutrients and Anti-Nutrients*; Organisation for Economic Co-operation and Development: Paris, France, 2009.
(28) Song, X.-L.; Zhang, T.-Z. Identification of quantitative trait loci controlling seed physical and nutrient traits in cotton. *Seed Sci. Res.* **2007**, *17*, 243−251.
(29) Turner, J. H.; Ramey, H. H., Jr.; Worley, S., Jr. Influence of environment on seed quality of four cotton cultivars. *Crop Sci.* **1976**, *16*, 407−409.
(30) Bertrand, J. A.; Sudduth, T. Q.; Condon, A.; Jenkins, T. C.; Calhoun, M. C. Nutrient content of whole cottonseed. *J. Dairy Sci.* **2005**, *88*, 1470−1477.
(31) Sunilkumar, G.; Campbell, L. M.; Puckhaber, L.; Stipanovic, R. D.; Rathore, K. S. Engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. *Proc. Natl. Acad. Sci.* **2006**, *103*, 18054−18059.
(32) Stipanovic, R. D.; Puckhaber, L. S.; Bell, A. A.; Percival, A. E.; Jacobs, J. Occurrence of (+) and (−) gossypol in wild species of cotton and in *Gossypium hirsutum* var. *marie-galante* (Watt) Hutchinson. *J. Agric. Food Chem.* **2005**, *53*, 6266−6271.
(33) Dowd, M. K.; Boykin, D. L.; Meredith, W. R., Jr.; Campbell, B. T.; Bourland, F. M.; Gannaway, J. R.; Glass, K. M.; Zhang, J. Breeding and genetics: Fatty acid profiles of cottonseed genotypes from the national cotton variety trials. *J. Cotton Sci.* **2000**, *14*, 64−73.
(34) Pettigrew, W. T.; Dowd, M. K. Varying planting dates or irrigation regimes alters cottonseed composition. *Crop Sci.* **2011**, *51*, 2155−2164.
(35) Cherry, J. P. Cottonseed oil. *J. Am. Oil Chem. Soc.* **1983**, *60*, 312A−319A.
(36) Lawhon, J. T.; Cater, C. M.; Mattil, K. F. Evaluation of the food use potential of sixteen varieties of cottonseed. *J. Am. Oil Chem. Soc.* **1977**, *54*, 75−80.
(37) Smith, C. W.; Creelman, R. A. Vitamin E concentration in upland cotton seeds. *Crop Sci.* **2001**, *41*, 577−579.
(38) Bailey, A. V.; Harris, J. A.; Skau, E. L. Cycloproprenoid fatty acid content and fatty acid composition of crude oils from twenty-five varieties of Cottonseed. *J. Am. Oil Chem. Soc.* **1966**, *43*, 107−110.