Use of NMR for Predicting Protein Concentration in Soybean Seeds Based on Oil Measurements

A.D. Weir, J. Omielan, E.A. Lee, and I. Rajcan*

Department of Plant Agriculture, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

ABSTRACT: Increasing the level of protein in soybean seeds has been a major target for soybean [*Glycine max* (L.) Merr.] breeders. The objective of this study was to examine the potential of predicting soybean seed protein based on oil values as determined by NMR. Seed protein and oil concentrations were determined in an F_2 population generated from the cross between a *G. max* (NK S08-80) and a *G. soja* (PI 458536) cultivar. The protein concentration in the population ranged from 40.4 to 52.6%. Protein–oil regression analysis was used to generate an equation for predicting seed protein concentration based on oil readings. The regression equation Protein = 62.3 – 1.3[Oil] (R^2 = 0.46) was developed, with a corresponding correlation of –0.69 between the traits. With this equation, the mean protein concentration of the selected 25% of the population (a simulated breeding pressure) was greater than the mean of the unselected population (46.1%, SE = 0.13) by about 1.9%. Individual F_2 plants that exceeded the mean protein value of the population constituted 86.4% of the selected samples. Selection based on oil concentration, however, failed to include 27.1% of the plants that were among the top 25% for protein concentration. Selection of high-protein plants based on NMR oil measurement was reasonably effective in the test population and might offer a new and rapid method of selecting high-protein individuals in soybean populations derived from the wild soybean progenitor, *G. soja*. If further tested on other populations and samples, it might be used as an analytical alternative for an indirect measurement of protein concentration based on NMR measurements of the oil.

Paper no. J10884 in *JAOCS 82*, 87–91 (February 2005).

KEY WORDS: *Glycine max*, *Glycine soja*, NIR, NMR, oil concentration, protein concentration, selection, soybean breeding, trait prediction.

Determination of soybean seed oil and protein concentrations is important for commercial trade and when evaluating germplasm in a plant breeding program. However, direct measurement of oil and protein using conventional chemical methods is constrained by cost, time to conduct analyses, and destruction of seed to complete the analysis. Oil is typically extracted over a period of 6 to 8 h from a sample of ground seed immersed in a petroleum ether solvent using a Soxhlet apparatus (AOCS Method Ac 3-44) (1). Protein concentration can be determined by the Kjeldahl procedure (AOCS Method Ac 4-91), which uses corrosive reagents and toxic catalysts (1). Alternatively, combustion nitrogen analysis (AOCS Method Ba 4e-93) may be used, which uses a thermal conductivity cell to quantify elemental nitrogen liberated from a sample that is burned at high temperature (1). These techniques provide accurate measurements but are relatively expensive, time-consuming, and undesirable when a limited amount of seed is available.

Indirect measurements of oil and protein have been used for quality control and evaluation purposes. Seed density and specific gravity were used to select for high-protein and high-oil soybeans, but these methods were not highly correlated with wet chemistry measurements (2). Li and Burton (3) found that single-plant selection for seed density was ineffective in improving protein concentration and proposed an alternative method based on selection of self-pollinated half-sib families.

Near-infrared reflectance spectroscopy (NIRS) has been widely used to measure the oil, protein, and moisture concentrations in whole soybean seeds simultaneously (4–8). Difficulties with these methods are that they require a large seed sample size, or that the seed must first be destroyed through grinding before analysis. Additionally, NIRS readings are influenced by seed coat color and may not provide accurate measurements when analyzing material segregated for seed coat color. This is the case with interspecific soybean crosses where various shades of brown, green, yellow, and black seed coats are encountered. In early generations, such as the F_2, large numbers of plants must be phenotyped, and the amount of available seed is limited by the production capacity of single plants.

NMR spectrometry is a rapid, nondestructive method that can be used to accurately determine the oil and moisture concentrations of whole seeds (2,9,10). The NMR technique measures the total hydrogen in a sample and is capable of distinguishing between hydrogen atoms in the oil fraction from those bound in water, carbohydrate, and protein (11). Sample analysis can be conducted in a few seconds and the results are highly reproducible (10). Large sample sizes are not required, as accurate readings can be obtained from individual seeds (9). NMR has been used to measure the oil concentrations of many crops, such as maize (9), soybeans (12), flax, and canola (13). Oil readings generated by NMR were reported to be highly correlated with values obtained by extraction with petroleum ether (r = 0.993) (2) and with gravimetric oil analysis (r = 0.995 to 0.999) (9). NMR has become a recognized method (AOCS Method Cd 16b-93) of analyzing the seed oil and moisture

*To whom correspondence should be addressed at Department of Plant Agriculture, Crop Science Bldg., University of Guelph, Guelph, Ontario, N1G 2W1, Canada. E-mail: irajcan@uoguelph.ca concentrations of rapeseed, sunflower seed, flax, and soybeans (1). The ability to confidently predict seed protein concentration based on NMR oil readings would be useful in soybean breeding programs. NMR can be used to analyze seeds quickly from large populations using small sample sizes. Seed is not destroyed during analysis and can be saved for future planting or other analyses.

Strong negative correlations between seed oil and protein concentrations have commonly been reported in the literature with correlation coefficients ranging from $r = -0.61$ to -0.98. (e.g., Refs. 6,14,15). This inverse relationship has been used previously for indirect trait selection. As an example, high-protein (44.0%) and low-protein (40.7%) lines were obtained from BC_1- and BC_2-derived *G. max* populations by selection based on NMR oil concentration readings (14). In another study that used an F_2 population, it was reported that selection based on NMR oil readings was adequate for obtaining low-protein (40.8%) and high-protein (45.7%) plants (16). The ability to select effectively for high protein based on NMR oil measurements would hasten the analysis of large population sizes and conserve seed that would otherwise be destroyed by using conventional chemical methods.

Previous reports that used NMR oil measurements to select for high-protein plants were based solely on the analysis of *G. max* germplasm (14,16). It remains to be determined whether selection for greater seed protein can be conducted in populations containing *G. soja* germplasm. The range of protein and oil in *G. soja* and populations derived from crossing *G. soja* with *G. max* frequently falls outside of the range found in *G. max*. This presents a challenge for both the breeders and seed analysts in that equipment and method calibrations that were developed for *G. max*-based ranges may not be effective for the traits' ranges present in *G. soja*. The objective of this study was to make an initial attempt to determine whether NMR oil readings could be used to select for high-protein soybeans in a soybean population containing wild *G. soja* germplasm. If successful, the method may be tested in other populations and samples for validation purposes.

MATERIALS AND METHODS

Plant material. Hybridization between a *G. max* cultivar, NK S08-80, with 41% protein, and a *G. soja* plant accession, PI 458536, with 51.4% protein, was used to generate an F_2 soybean population used in the study. Individual F_2 plants were grown in the field at the Woodstock Research Station in Woodstock, Ontario, Canada, during the 2000 season. The soil at this maturity group location was a Guelph loam. All seeds were treated prior to planting with the fungicide Vitaflow 280™ (Gustafson LLC, Plano, TX) at a rate of 2.6 mL/kg using a Micro Batch Coater (Vector Corporation, Marion, IA*). Glycine soja* parental seed was immersed in liquid nitrogen for 30 s before seed treatment to break the hard seed coat and improve percent germination in the field.

Two seeds were planted every 50 cm within rows at a depth of 3 to 5 cm. Rows were 4.5 m in length and spaced 53 cm apart. OAC Bayfield, a Maturity Group 0 cultivar, was included as a single-plant border to eliminate border effects on F_2 plants. Soil Implant+®, a *Bradyrhizobium japonicum* inoculant (Nitragin, Milwaukee, WI), was incorporated with the seed at the time of planting.

Plants were harvested and threshed individually upon reaching physiological maturity, and those that yielded sufficient seed for both oil and protein analyses ($n = 235$) were included in this study.

Oil and protein measurement. Seed protein was measured using the combustion nitrogen analysis method (1) on an FP-428 Automatic Nitrogen Analyzer manufactured by LECO Corp. (St. Joseph, MI). Seed samples were first ground to a fine powder using a Moulinex (Paris, France) coffee bean grinder and dried overnight at 90°C in a forced-air dryer. Total percent nitrogen was measured on a 250-mg sample and converted to percent crude protein on a dry-seed basis by multiplying the nitrogen value by 6.25 (1). Standard samples that were used as controls included peach leaves (National Institute of Standards and Technology, Gaithersburg, MD), alfalfa (LECO Corp.), and ethylenediaminetetraacetic acid (ISO-MASS Scientific, Calgary, Canada).

Oil concentration was measured using pulsed NMR spectrometry (1) on a Bruker Minispec Mq10 NMR Analyzer (Bruker, Karlsruhe, Germany). A calibration curve covering the range 8.3 to 15.3% oil was developed using a set of six soybean seed samples with known oil concentrations as determined by extraction in petroleum ether (1). This range of oil content is lower than that of most commercial soybean cultivars but is representative of a *G. soja* × *G. max* population (17,18), which may be used by soybean breeders. The accuracy of the calibration was validated using an additional set of five soybean seed samples maintained separate from those used in the calibration. These validation samples were selected for oil concentrations that covered the calibration range. Oil was measured on 5-g whole-seed samples that were dried overnight at 90°C in a forced-air dryer prior to analysis. The mean oil concentration for each sample was calculated for two replications over time.

Data analysis. Data were analyzed using the PROC UNIVARIATE, PROC CORR, PROC REG, and PROC GLM procedures in SAS software version 6.12 (19) with a Type I error rate (α) of 0.05. The Shapiro-Wilk statistic (W) was calculated for protein and oil measurements to determine whether the data followed a normal distribution. Protein–oil regression analysis was conducted, and a prediction equation was generated to estimate the seed protein concentration based on NMR oil measurements. The effectiveness of indirect selection for high protein based on low oil concentration was evaluated in a simulated selection study using the measured oil and protein values for the F_2 population. Because most food-grade soybean breeders are concerned with increasing the protein content rather than oil, plants belonging to the lowest 25% of the population for oil were selected as the high-protein individuals. This is considered a relatively mild simulated selection pressure for a breeding program aimed at maintaining a certain level of genetic diversity, without culling out too many individuals. It is

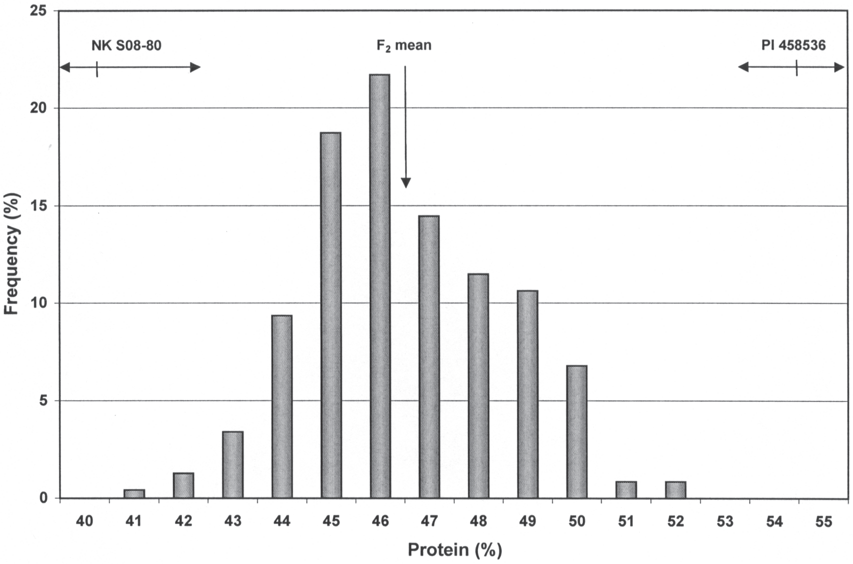

FIG. 1. Frequency distribution for seed protein in plants used for a protein–oil regression. *Glycine soja*-derived F_2 plants (n = 235) were from the cross NK S08-80 × PI 458536. Protein is expressed as percent crude protein of the total seed on a dry-seed basis. Arrows represent range of parental values. Observed mid-parent protein value was 47.7%.

important to note that the selected individuals from the population were not the same ones as those used for the development of the calibration and equation. The mean protein concentration of the selections was compared with that of the unselected population using a Welch's test, which does not assume equal variances between group means. The selected plants were compared with the entire F_2 population to determine the number of low-protein plants chosen and high-protein plants excluded using the indirect selection method based on oil concentration.

RESULTS AND DISCUSSION

A *G. soja*-derived F_2 soybean population was generated that segregated for both protein and oil concentration. Protein concentration was normally distributed (W = 0.98, P = 0.133) and ranged from 40.4 to 51.5% with a mean value of 46.1% (SE = 0.13) (Fig. 1). Oil concentration was also normally distributed (W = 0.97, P = 0.063) and ranged from 10.4 to 15.9% with a mean value of 12.8% (SE = 0.07) (Fig. 2). No transgressive

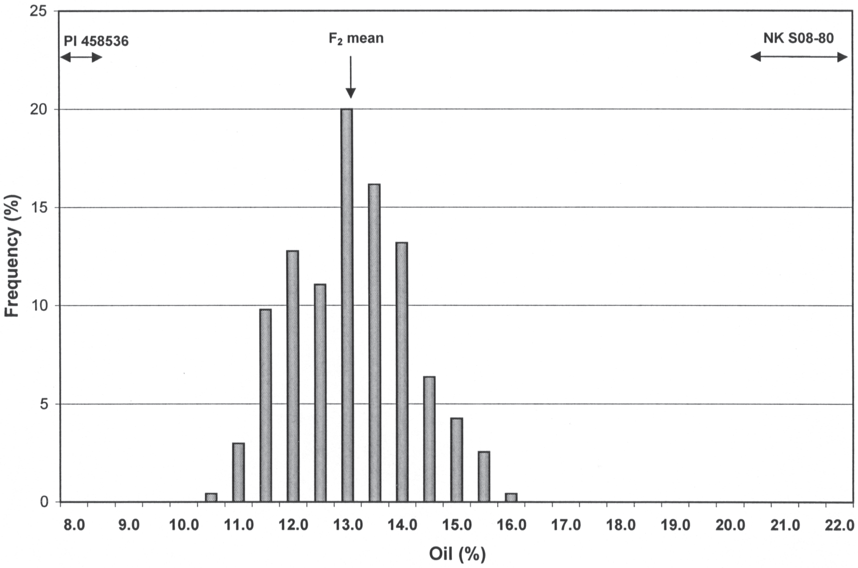

FIG. 2. Frequency distribution for seed oil concentration in plants used for a protein–oil regression. *Glycine soja*-derived F_2 plants (n = 235) were from the cross NK S08-80 × PI 458536. Protein is expressed as percent crude protein of the total seed on a dry-seed basis. Arrows represent range of parental values. Observed mid-parent oil value was 14.8%.

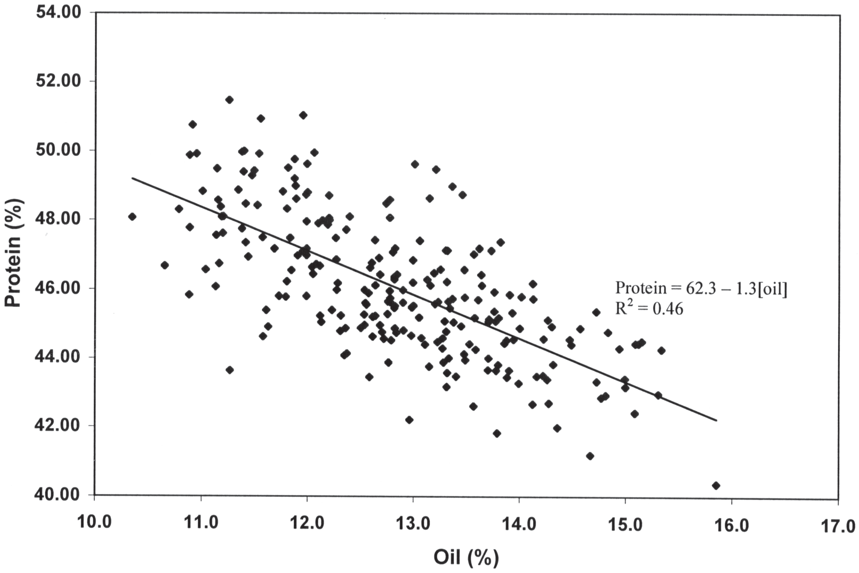

FIG. 3. Linear protein–oil regression for a *Glycine soja*-derived F_2 soybean population (n = 235) from the cross NK S08-80 × PI 458536. The linear regression equation predicted Protein = 62.3 (SE = 1.12) – 1.3[Oil] (SE = 0.09) (R^2 = 0.46). Protein and oil are expressed as percent of total seed on a dry-seed basis.

segregants were observed that exceeded either parent for protein or oil concentration. Previous studies using *G. soja* germplasm reported similar ranges and segregation patterns for seed protein and oil (17,18). The prediction equation was based on a significant range of values for both protein and oil that may reflect the specific range of the experimental population, rather than all soybean breeding populations. The latter issue would have to be further tested before any broader extrapolations could be made.

Linear regression analysis using protein and oil data was used to develop a prediction equation. The prediction equation for protein based on oil concentration was Protein = 62.3 (SE = 1.12) –1.3 (SE = 0.09) [Oil] (R^2 = 0.46, P = 0.0001) (Fig. 3). The R^2 value indicated that 46% of the observed variation in seed protein could be explained by oil concentration. It is possible that the rest of the variation depends on other compounds in the seed, such as carbohydrates, and their competition for photo-assimilates during the seed fill period. Chung *et al.* (20) reported that the genetically based oil/protein ratio of 1.6 was smaller than the 2.0 calorically based oil/protein ratio, likely because the remaining 0.4 units of carbon and/or energy was used for other dry matter in seed. This is in agreement with our findings.

The negative slope of the regression (with a correlation of –0.69) supports previous findings of an inverse relationship between protein and oil concentration (6,13,14). For each increase in oil concentration there was a general decrease in protein concentration on the magnitude of 1.26 times the observed increase in oil concentration. A negative correlation between protein and oil (r = –0.69) was found for the F_2 population. This negative association may be the result of a competition for limited resources by different genetic systems in the plant (21). The deviation of points from the regression line indicated that the value of the regression coefficient did not hold true for all F_2 plants (Fig. 3). Exceptions were observed where the values of both traits were relatively increased or reduced among individual F_2 plants, which is consistent with a correlation between traits that is different from unity.

A simulated selection for high-protein soybeans was conducted based on NMR oil measurements. Plants from the lowest 25% of the population (formed by cross NK S08-80 × PI 458536) for oil concentration were selected to obtain high-protein plants. The 59 F_2 plants selected had low oil concentrations that ranged from 10.3 to 12.0%. The mean oil concentration of the selected individuals was 11.5% (SE = 0.05). The protein concentration of the selected plants ranged from 43.6 to 51.5% (expressed as percent crude protein on a dry-seed basis) and had a mean value of 48.0% (n = 59, SE = 0.22). This was an improvement of 1.9% compared with the mean protein of the unselected population (46.1%, n = 235, SE = 0.13). A Welch's test comparing the protein means indicated that they were significantly different (F = 56.19, P < 0.0001). Individual F_2 plants that exceeded the mean protein value of the population constituted 86.4% of the selected samples. Additionally, selection based on oil concentration failed to include 27.1% of the plants that were among the top 25% for protein concentration in the *G. soja*-derived population.

Previous indirect selection for protein using NMR oil measurements was successful in obtaining BC_1- and BC_2-derived high-protein (44.0%) lines (13) and individual high-protein F_2 plants (45.7%) (15). These studies were conducted on populations generated with *G. max* plants. In contrast, the current study used exotic germplasm from *G. soja* as a source of high-protein concentration in a *G. soja*-derived F_2 population. NMR oil readings were used in the selection of plants that had a mean protein concentration of 48.0%. This study demonstrated that NMR, as an indirect selection tool, could be used effectively to select for high protein concentration in a *G. soja*-derived F_2 population to achieve protein levels higher than those previously reported. Although it is possible that some high-protein lines could be missed in selecting for high protein based on the NMR measurement of oil concentration, it may be a risk that is considered worth taking by soybean researchers wishing to incorporate the high-protein trait (alleles) from *G. soja*. This finding is particularly relevant because the amount of seed generated in such populations is often more limited than that derived from *G. max*.

ACKNOWLEDGMENTS

This work was supported with funding provided by the Ontario Soybean Growers and the Ontario Ministry of Agriculture and Food. It was conducted as partial fulfillment of the degree requirements for an M.S. thesis of A.D.W. at the University of Guelph. Julia Zilka is gratefully acknowledged for performing crosses between *G. soja* and *G. max* genotypes in the growth room. Excellent technical help in the field was provided by Wade Montminy, Cal Klager, and Yasenia Salazar.

REFERENCES

1. American Oil Chemists' Society, *Official Methods and Recommended Practices of the AOCS*, 5th edn., edited by D. Firestone, AOCS Press, Champaign, 1997.
2. Fehr, W.R., F.I. Collins, and C.R. Weber, Evaluation of Methods for Protein and Oil Determination in Soybean Seed, *Crop Sci. 8*:47–49 (1968).
3. Li, H., and J.W. Burton, Selecting Increased Seed Density to Increase Indirectly Soybean Seed Protein Concentration, *Ibid. 42*:393–398 (2002).
4. Panford, J.A., P.C. Williams, and J.M. deMan, Analysis of Oilseeds for Protein, Oil, Fiber and Moisture by Near-Infrared Reflectance Spectroscopy, *J. Am. Oil Chem. Soc. 65*:1627–1634 (1988).
5. Pazdernik, D.L., L.L. Hardman, J.H. Orf, and F. Clotaire. Comparison of Field Methods for Selection of Protein and Oil Content in Soybean, *Can. J. Plant Sci. 76*:721–725 (1996).
6. Sebolt, A.M., R.C. Shoemaker, and B.W. Diers Analysis of a Quantitative Trait Locus Allele from Wild Soybean That Increases Seed Protein Concentration in Soybean, *Crop Sci. 40*:1438–1444 (2000).
7. Csanádi, G., J. Vollmann, G. Stift, and T. Lelley, Seed Quality QTLs Identified in a Molecular Map of Early Maturing Soybean, *Theor. Appl. Genet. 103*:912–919 (2001).
8. Wilcox, J.R., Increasing Seed Protein in Soybean with Eight Cycles of Recurrent Selection, *Crop Sci. 38*:1536–1540 (1998).
9. Alexander, D.E., L.S. Silvela, F.I. Collins, and R.C. Rodgers, Analysis of Oil Content of Maize by Wide-Line NMR, *J. Am. Oil Chem. Soc. 44*:555–558 (1967).
10. Rubel, G., Simultaneous Determination of Oil and Water Contents in Different Oilseeds by Pulsed Nuclear Magnetic Resonance, *Ibid. 71*:1057–1062 (1994).
11. Conway, T.F., and F.R. Earle, Nuclear Magnetic Resonance for Determining Oil Content of Seeds, *Ibid. 40*:265–268 (1963).
12. Honeycutt, R.J., J.W. Burton, R.G. Palmer, and R.C. Shoemaker, Association of Major Seed Components with a Shriveled-Seed Trait in Soybean, *Crop Sci. 29*:804–809 (1989).
13. Reaney, M.J.T., N.J. Tyler, and K. Brown, Practical Nuclear Magnetic Resonance Analysis of Liquid Oil in Oilseeds: I. Factors Affecting Peak Width, *J. Am. Oil Chem. Soc. 76*:859–862 (1999).
14. Hartwig, E.E., and K. Hinson, Association Between Chemical Composition of Seed and Seed Yield of Soybeans, *Crop Sci. 12*:829–830 (1972).
15. Wilcox, J.R., and R.M. Shibles, Interrelationships Among Seed Quality Attributes in Soybean, *Ibid. 41*:11–14 (2001).
16. Hartwig, E.E., and T.C. Kilen, Yield and Composition of Soybean Seed from Parents with Different Protein, Similar Yield, *Ibid. 31*:290–292 (1991).
17. Thorne, J.C., and W.R. Fehr, Incorporation of High-Protein, Exotic Germplasm into Soybean Populations by 2- and 3-Way Crosses, *Ibid. 10*:652–655 (1970).
18. Erickson, L.R., H.D. Voldeng, and W.D. Beversdorf, Early Generation Selection for Protein in *Glycine max* × *G. soja* Crosses, *Can. J. Plant Sci. 61*:901–908 (1981).
19. SAS Institute, *SAS/STAT User's Guide*, version 6, 4th edn., SAS Institute, Inc., Cary, NC, 1989, Vol. 1.
20. Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, P.B. Cregan, R.C. Shoemaker ,and J.E. Specht, The Seed Protein, Oil, and Yield QTL on Soybean Linkage Group I, *Crop Sci. 43*:1053–1067 (2003).
21. Hanson, W.D., R.C. Leffel, and R.W. Howell, Genetic Analysis of Energy Production in the Soybean, *Ibid. 1*:121–126 (1961).

[Received June 11, 2004; accepted January 27, 2005]