Non-destructive genotypes classification and oil content prediction using near-infrared spectroscopy and chemometric tools in soybean breeding program

Daniel Carvalho Leite[1], Aretha Arcenio Pimentel Corrêa[1], Luis Carlos Cunha Júnior[2], Kássio Michell Gomes de Lima[3], Camilo de Lelis Medeiros de Morais[4], Viviane Formice Vianna[1], Gustavo Henrique de Almeida Teixeira[3], Antonio Orlando Di Mauro[1], Sandra

[1]Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias (FCAV), Campus de Jaboticabal. Via de acesso Prof. Paulo Donato Castellane s/n, Jaboticabal _ SP, Brazil. CEP: 14.870-900.

[2]Universidade Federal de Goiás (UFG), Escola de Agronom ia (EA), Goânia _ GO, Rodovia Goiânia/Nova Veneza Km 0 Campos Samambaia, Goiania _ GO, Brazil. CEP: 74001-970.

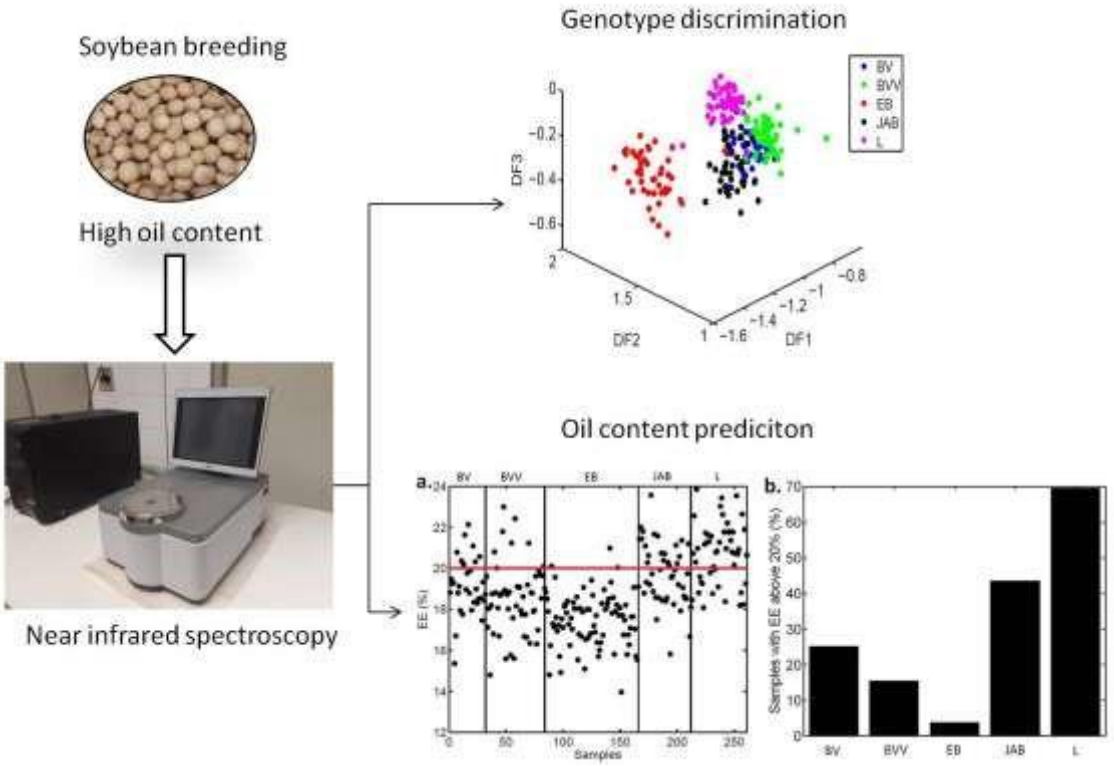
[3]Universidade Federal do Rio Grande do Norte (UFRN), Instituto de Química, Química Biológica e Quimiometria, Avenida Senador Salgado Filho, n° 30 00, Bairro de Lagoa Nova, CEP: 59.078-970, Natal, Rio Grande do Norte, Brazil.

[4]School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston, Lancashire, PR1 2HE, United Kingdom.

*Corresponding author: shu.trevisoli@unesp.br

Graphical abstract

Helena Unêda-Trevisoli[1,*]



Highlights

- NIRS can be used in soybean breeding programs to discriminate superior genotypes;
- NIRS can be used in breeding programs topredict the oil content of intact grains.
- GA-LDA resulted in a high discrimination accuracy (88.89% - prediction set);
- PLSR oil prediction models presented low RMSEP (0.96%) and adequate R^2 (0.66);

Abstract

In soybean (*Glycine max* L.) breeding programs, segregation is normally observed, and it is not possible to have replicates of individuals because each genotype is a unique copy. Therefore, near-infrared spectroscopy (NIRS) was used as a nondestructive tool to classify soybeans by genotypes and to predict oil content. A total of 260 soybean genotypes were divided into five classes, which were composed of 32, 52, 82, 46, and 49 samples of the BV, BVV, EB, JAB, and L class, respectively. NIR spectra were obtained using oven-dried samples (80 g) in a reflectance mode. A successive projection algorithm and genetic algorithm with linear discriminant analysis discriminated genotypes of the low (L class) from the high (EB class) for oil content (88.89% accuracy). The partial least square regression models for oil content were considered good (root mean square error of prediction of 0.96%). Therefore, NIRS can be used as a non-destructive tool in soybean breeding programs, but further investigation is necessary to improve the robustness of the models. It is important to note that to use the models, it is necessary to collect NIR spectra from dry soybean samples.

Keywords: *Glycine max* L., principal component analysis (PCA), PCA with linear discriminant analysis (PCA-LDA), successive projection algorithm (SPA) with LDA (SPA-LDA), and genetic algorithm (GA) with LDA (GA-LDA).

1. Introduction

According to FAOSTAT (2019), in 2017 the world soybean [*Glycine max* L. (Merrill)] production was 353 million tons (MT). In 2012, Brazilian soybean production surpassed that of the United States (Palmer & Hymowitz, 2016); however, in 2017, production in the United States reached 119 MT, followed by Brazil (114 MT), Argentina (55 MT), and

China (13 MT). Soybeans are considered one of the main cultivated oilseeds worldwide (Conab, 2019) and account for 67% of the protein meal in the world (Palmer & Hymowitz, 2016). Similarly, soybeans are an important source of oil, protein for both humans and animals, and other products, such as biodiesel in Brazil (Woyann et al., 2019).

In general, soybean breeding is conducted to create variability for desired traits, identification of superior genotypes, and production of commercial seeds (Miladinović et al., 2011). Plant breeding programs generally select the best genotypes based on the most important agronomic and commercial traits. Thus, for the soybean crop, one of the most important traits is the oil content in the seed, allied with agronomic traits (Bezerra et al., 2017).

The soybean reproduces by self-fertilization, being considered a perfect autogamous species (Silva et al., 2017a). Therefore, the system of conducting a conventional breeding program consists of artificial hybridizations, to obtain variability and subsequent self-fertilization for the selection of genotypes with superior traits. In the initial process of obtaining segregated generations with the existence of autogamy, from the first generation (F_1) onwards, the plants self-fertilize again and the segregated generations produce unique genotypes, as long as segregation lasts. The fixed genotypes are obtained predominantly in the F_6 generation. Thus, during the genotype segregation process, there is no possibility of having replicates for the individuals because each genotype produced is a unique copy **(**Silva et al., 2017a). Therefore, the existence of nondestructive methods for the evaluation and selection of agronomic traits, such as nearinfrared spectroscopy (NIRS), is a process of extreme importance for the early stages of a breeding program. Additionally, the non-destruction of seeds allows these populations to be advanced to high levels by inbreeding, allowing the selection of superior strains for the future release of commercial cultivars

carrying agronomic traits of interest, with high oil content among them. This process is the main focus of the present study (Silva et al., 2017b).

Because our soybean breeding program is producing advanced lines of different genealogies from two-way, four-way, and eight-way crosses, with a high degree of endogamy, non-destructive evaluation of seeds would be an important tool for the improvement of the genetic gain. Thus, the main objective of this study was to use nearinfrared spectroscopy (NIRS) and chemometric tools as a non-destructive method to classify intact grains by genotypes and to predict oil content in soybean breeding program.

2. Material and methods

2.1. Plant material

A total of 260 soybean (*Glycine max* (L.) Merrill) genotypes from 2012/2013 and 2013/2014 seasons were produced in the soybean breeding program of UNESP – FCAV (21°15′17″S, 48°19′20″W, 595 m above sea level). The 260 genotypes consisted of advanced lines and some commercial checks, which were placed into five different groups according to the research line from which they were originated. The genotypes from the class L were obtained by crossing parents predominantly with characteristics of precocity and high grain yield. The genotypes from the class JAB were obtained from a study that had a wide genetic basis for soybeans; thus, two-way, four-way and eight-way crosses were synthesized from commercial parents with high yield and traditional soybean germplasms. The genotypes from the classes BV and BVV were obtained from crosses between parents with high yield and parents with resistance to soybean rust (*Phakopsora pachyrhizi*). Finally, the genotypes from the class EB were obtained from crossings between parents with high yield and parents with resistance source to root-knot nematodes (*Meloidogyne incognita*) and soybean cyst nematode (*Heterodera glycines*), Table A.1.

From each genotype, a sample of 80 g of intact soybean grains was obtained for the NIR spectra collection and reference analysis.

2.2. NIR spectra collection

Before the NIR spectra collection, the samples were oven-dried at 105°C for 24 h to obtain grains with similar moisture content. After temperature stabilization at ~ 25°C, the intact soybean samples were poured into a glass vial, which was set on a rotary accessory. The NIR spectra were obtained using a Bruker spectrometer (model Tango, Ettlingen, Germany) using the reflectance mode on the wavenumber range of 12,000– 4,000 cm^{-1}, with 64 scans and a resolution of 16 cm^{-1}. A total of five spectra were collected from each genotype sample totaling 1,300 NIR spectra. The reflectance spectra were converted into a pseudo-absorbance scale Log(1/R), where R was the measured reflectance signal.

2.3. Reference analysis

2.3.1. Moisture content

The soybean grains were dried in an oven (FANEM Model 320-SE, São Paulo, Brazil) at 105°C for 24 h and the moisture content was determined according to the recommendations of the Brazilian Ministry of Agriculture and Husbandry (Brasil, 2009). The results are expressed in percentages (%) in Table 1.

2.3.2. Oil content

After NIR spectra collection the dried soybean grain samples were ground in a knife mill (Metalúrgica Roma, model MR 340, São Paulo, Brazil) with a 1 mm screen. The oil extraction was carried using a Soxhlet extractor (Tecnal, model TE 044-5/50, Piracicaba, Brazil) according to the reference method reported by A.O.A.C. (1997). The results are expressed in percentages (%) in Table 1.

2.4. Chemometric analysis

The NIR spectral data pre-processing, classification, and prediction models were performed using MATLAB® software R2012b (MathWorks, USA). The NIR spectra were pre-processed using the multiplicative scatter correlation as described by Geladi et al. (1985), which is intended to reduce light scattering influences. Similarly, the standard normal variate was applied to the spectra as stated by Barnes et al. (1989). To correct the spectra baseline, the first (1SG) and second (2SG) derivative of Savitzky-Golay (Savitzky and Golay, 1964) were used. All NIR spectra were mean-centered and 1,300 spectra were reduced to their average spectra, totaling 260 spectra. These spectra were divided into calibration (70%), validation (15%), and prediction (15%) sets by applying the classical Kennard-Stone algorithm (Kennard and Stone, 1969).

2.4.1. Classification model development

For the development of the classification models, principal component analysis (PCA) with linear discriminate analysis (PCA-LDA), partial least squares discriminant analysis (PLS-DA), successive projection algorithm (SPA) with LDA (SPA-LDA), and a genetic algorithm (GA) with LDA (GA-LDA) (Costa et al., 2016) were used. The optimum number of variables for SPA-LDA and GA-LDA was obtained using the average G risk of incorrect classification of LDA. The cost function (G) was obtained from the validation set, as follows:

$$G = \frac{1}{N_V}\sum_{n=1}^{N_V} g_n \qquad (1)$$

where N_V is the number of validation samples and g_n is defined as:

$$g_n = \frac{r^2(x_n, m_{I(n)})}{\min_{I(m)\neq I(n)} r^2(x_n, m_{I(m)})} \qquad (2)$$

where $I(n)$ is the index of the truth class for the nth validation object, x_n; $r^2(x_n, m_{I(n)})$ is the square Mahalanobis distance between the object x_n (class index) and

the average of a sample $m_{I(n)}$ in its truth class; $r^2(x_n, m_{I(m)})$ is the square Mahalanobis distance between the object x_n and the sample average $m_{I(m)}$ in its wrong class.

The GA routine was conducted for 40 generations with 80 chromosomes each. The crossing probabilities and mutations were adjusted to 60% and 10%, respectively. Therefore, the algorithm was repeated three times from random initial populations. The best solution (in terms of aptitude value) resulting from the three GA routines was used. The LDA scores, loadings, and discriminant function (DF) were obtained for the different genotypes.

2.4.2. Prediction model development

The NIR spectral datasets were correlated with oil content using the partial least squares (PLS) regression and cross-validation Venetian blinds (six latent variables - LVs). The spectra were only mean-centered before model construction. To evaluate the performance of the calibration models, the root mean square error of the calibration (RMSEC) and root mean square error of the prediction (RMSEP) were calculated, according to the following equation:

$$\text{RMSEC or RMSEP} = \sqrt{\frac{\sum_{i=1}^{np} (yi - yi')^2}{n-K-1}}$$

where yi represents the value predicted by the multivariate model, yi represents the reference value, and n corresponds to the number of samples.

The performance of the calibration models was also evaluated based on the determination coefficient R^2, both for the calibration and prediction set (Pasquini, 2003).

3. Results and discussion

3.1. NIR spectra

All models were built from the spectral data transformed into a pseudo-absorbance scale. The raw NIR spectra of all genotypes (n = 260) are shown in Figure 1a. Similarly, the mean NIR spectra of the genotype classes are shown in Figure 1b. The five genotype classes were composed of 32 samples of the BV class, 52 BVV class, 82 EB class, 46 JAB class, and 49 L class (Figure 1b).

As shown in Figure 1, the spectral differences between genotypes (Figure 1a) and genotype classes (Figure 1b) were minimal, only the BVV class presented a lower apparent spectral intensity and a slight shift on wavenumbers higher than 5,000 cm^{-1}. However, it was not possible to discriminate the different genotypes by only evaluating the NIR spectra.

The raw NIR spectra presented broad light scattering (Figure 1a), but it was possible to identify seven main peaks at 4024, 4288, 4789, 5192, 5712–5824, 6752, and 8320 cm^{-1} (Figure 1b). The absorption bands around 4789 cm^{-1} arose from R-OH, 5192 cm^{-1} from the OH combinations, 5824 cm^{-1} from CH first overtone, the 6752 cm^{-1} from the OH first overtone caused by the presence of H_2O, and the 8320 cm^{-1} from the CH second overtone (Firmani et al. 2019). Similar NIR spectra were reported by Bras et al. (2005) in soybean flour, but milling can provide better results by reducing the heterogeneity from intact soybean grains. Based on the NIR spectra features, to develop classification models for intact soybean grain classification based on oil content, the most important wavenumbers were 5712–5824 and 8320 cm^{-1} because they were related to CH bounds commonly present in fatty acids (Cozzolino et al. 2005; Firmani et al. 2019).

However, due to sophisticated softwares the entire NIR spectra could be tested.

3.2. Reference analysis: Moisture and oil content

The moisture and oil content obtained using the reference analysis are shown in Table 1. Because the soybean grains were dried out before NIR spectra collection, the moisture content was very low and ranged from 5.04 to 8.88% with an average value of 6.66% (Table 1). It is important to note that the average standard deviation values were very low, being 0.83%, and 1.88% for moisture, and oil content, respectively.

According to Palmer & Hymowitz (2016) soybeans are the most important source of edible vegetable oil and high-quality vegetable protein in the world. In general, soybeans contain 40% protein and 6.5% to 28.7% oil. Regarding oil content, an average value of 19% is commonly reported in soybeans, which represents 360-610 kg of oil per hectare (Huang et al., 2016). For biodiesel production, it is important to develop cultivars with higher oil content. Cavalcante et al. (2011) evaluated 19 soybean lines and five cultivars and observed an average oil content of 16.75% with the highest oil content of 21.59%. These values agreed with the range of 14.62 to 20.67% reported by Lundry et al. (2008), but Marro et al. (2020) studying different soybean cultivars in Argentina reported higher oil contents (20-27%). Therefore, the oil content of the developed soybean genotypes was in the range of what is commonly reported for this species, but the identification of superior genotypes is important for obtaining superior genotypes. It was possible to accomplish this using multivariate classification techniques.

3.4. Chemometrics: Oil prediction

The oil content of intact soybean grains was predicted using partial least squares regression (PLSR) with cross-validation Venetian blinds using 10 data splits ($RMSE_{CV}$ = 1.68%). Mean-centering was the only pre-processing applied to the spectral dataset because this produced the best $RMSE_{CV}$ value. The $RMSE_C$ and $RMSE_P$ were considered low, 1.42 and 0.96%, respectively. The correlation coefficients (R^2) were below 0.70,

with 0.51 and 0.66 for the calibration and prediction sets. The measured *versus* predicted oil content by PLS is shown in Figure 2. Similar results for soybean oil prediction were reported by Ferreira et al. (2013) and Xu et al. (2020).

3.3. Chemometrics: classification

To develop the classification models, two approaches were used. First, the raw NIR spectra were used and then the NIR spectra were pre-processed with different preprocessing techniques. The performance of these pre-processing techniques was evaluated by comparing the accuracy, sensitivity, and specificity with the validation set for each classification model tested. The best classification metrics were obtained with the raw NIR spectra; hence, all models were built without pre-processing the spectra.

3.3.1. PCA

The PCA was performed with mean-centered NIR spectra. Principal component 1 (PC1) accounted for 94.58% of the variance in the data and principal component 2 (PC2) accounted for 4.52%, which together represented 99.10% of the variance of the data. The scores plot for PC1 versus PC2 is shown in Figure 3c. The scores plot did not exhibit a very clear discrimination profile for the genotype classes, although the EB genotype appeared to be clustered mostly below zero for scores on PC2 and the L genotype was above this mark. On the other hand, on PC1, the scores for the JAB genotype were clustered primarily on the right side and the BVV genotype on the left side. The scores for the BV genotype were scattered inside the entire confidence ellipse at the 95% confidence level (Figure 3c). Therefore, other multivariate techniques needed to be tested to improve classification results.

3.3.2. PCA-LDA

PCA with linear discriminant analysis (PCA-LDA) was applied to the raw NIR spectra. To perform the discrimination using PCA-LDA it was used with five PCs because they

accounted for 98.29% of the data variance (Figure 3d). The final number of PCs was defined according to the distribution of the variance for each PC, such that the minimum number of PCs accounted for the maximum variance, which occurred before the variance reached a small and constant trend, was selected. By applying the PCA-LDA it was not possible to obtain satisfactory discrimination between genotype groups, and the classification rates for the calibration, validation, and prediction sets were 67.72, 76.47, and 86.11%, respectively (Figure 3d). It was not possible to visualize a clear separation among genotype groups, although good classification parameters, especially for the EB (F-score = 100%, area under the curve (AUC) = 1.00) and JAB (F-score = 94.6%, AUC = 0.948) classes, were observed in the prediction set (Table 2). The F-score depicts the overall classification performance for each class considering the unbalanced size. The ROC curves for each chemometric model tested are shown in the Supplementary Material. Using the same chemometric method, Carvalho et al. (2018) also could not differentiate macadamia cultivars produced by plant breeding. Therefore, more sophisticated chemometric techniques were tested.

3.3.3. PLS-DA

PLS-DA was applied to the mean-centered raw spectra using five latent variables selected by Venetian blinds (10 data splits) cross-validation (cumulative spectral variance of 99.65%). The PLS-DA performance with the prediction set is displayed in Table 2, whereas the regression coefficients and DF graph are shown in Figure 4. PLS-DA exhibited good predictive performance for BVV (F-score = 92.3%), EB (F-score = 100%), JAB (F-score = 88.5%), and L (F-score = 91.7%) classes with AUC values of 0.812, 1.00, 0.894, and 0.937, respectively. However, the BV class was poorly classified with an F-score of 68.1% and AUC of 0.719.

3.3.4. SPA-LDA

Successive projection algorithms with linear discriminant analysis (SPA-LDA) were applied to the NIR spectra and to develop the SPA-LDA models, four variables were selected according to the minimum of the cost function (Figure 5a). The selected variables were the wavenumbers 4008, 4536, 5256, and 9920 cm^{-1} (Figure 5a). By applying the SPA-LDA, it was possible to discriminate the genotype EB from the other genotypes and these genotypes were also grouped and a slight separation appeared (Figure 5b). The classification rates increased when the SPA-LDA was used and values of 69.62, 85.29, and 88.89% were obtained for the calibration, validation, and prediction sets, respectively. The model performance using the external prediction set is depicted in Table 2, in which the F-scores and AUC values ranged 87.6–100% and 0.857–1.00, respectively. To improve the classification of the soybean genotypes, the genetic algorithm with linear discriminant analysis (GA-LDA) was also tested.

3.3.5. GA-LDA

The GA-LDA was applied to the raw NIR spectra and 11 variables were selected (Figure 6a). The selected variables were the wavenumbers 4040, 4696, 5392, 5952, 6032, 7224, 7328, 7776, 7864, 10264, and 10632 cm^{-1} (Figure 6a). The GA-LDA was considered the best technique to develop discriminate models to classify the different soybean genotypes. The genotype EB was also segregated from the other genotypes, but by applying GA-LDA it was possible to improve the grouping and segregation of the other genotypes. Genotypes L and BVV formed two distinct classes, but JAB and BV intermingled and their discrimination was not possible (Figure 6b). The classification rates increased considerably when GA-LDA was used and it was possible to obtain values of 82.28, 97.01, and 88.89% for the calibration, validation, and prediction sets, respectively. The classification results for the five groups of genotypes in the prediction

set are shown in Table 2, where GA-LDA produced F-scores of 57.1–100% and AUC values of 0.700–1.00. The algorithm presented a poor classification for the BV class, similar to that of PCA-LDA, although it showed much better classification for the other four classes (BVV, EB, JAB, and L).

The classification results for the five groups of genotypes are shown in Table A.2. A possible explanation for the separation of the genotype EB from the other genotypes might be related to its lower oil content (Figure 6a). Additionally, some of their parents were considered sources of resistance to root-knot nematode (Sambaíba parent) and cyst nematode (Matrinxã and Kinoshita parents), which were not present in the genealogy of other groups of evaluated genotypes (L, JAB, BV, and BVV), according to Table 1S.

By examining the number of samples that had oil content above 20% (Figure 7), the genotype L stands out because almost 70% of its sample had an oil content above 20% (early genotypes, from FT-Cometa, IAC-Foscarin parents and Monsoy 7501, COODETEC 205, IAC 23 checks). On the other hand, the EB genotypes had the lowest percentage of samples (3.66%) having an oil content above 20% (Figure 7). The 20% oil content was chosen as the threshold value and superior genotypes should have an oil content above this limit.

4. Conclusions

The NIRS can be use to as a non-destructive method to classify intact grains by genotypes and to predict oil content in soybean breeding program. Classification chemometric techniques, especially SPA-LDA, can be applied to discriminate soybean genotypes with high accuracy.

The PLSR models for oil content prediction were considered good. Therefore, NIRS could be used to predict the oil content of intact soybean grains, but further

investigation must address other sources of variability to improve the robustness of the prediction models.

The use of oven-dried samples is not a problem for practical application because the seeds can be dried using other drying methods, such as in desiccators with silica gel. The important aspect is to only use dry soybean samples.

Conflict of interest

The authors would like to state that we do not have any conflict of interest.

5. Acknowledgements

The authors would like to thank the FAPESP for the financial support (Proc 2011/12958-9) and the Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) for the fellowship of the first author , and for the financial support (grant 88881.128982/2016-01) given to Camilo de Lelis Medeiros de Morais.

6. References

A.O.A.C., 1997. Official Methods of Analysis of the Association of Official Analytical Chemists, sixteenth ed. Patricia Cuniff, Arlington.

Firmani, P., Bucci, R., Marini, F., Biancolillo, A. 2019. Authentication of 'Avola almonds' by near infrared (NIR) spectroscopy and chemometrics. J. Food Comps. Anal. 82, 103235.

Bezerra, A.R.G., Sediyama, T., Silva, F.L., Borém, A., Silva, A.F., Silva, F.C.S., 2017. Agronomical aspects of the development of cultivars, in: Silva, F.L., Borém, A., Sediyama, T., Ludke, W.H. (Eds.), Soybean Breeding. Springer International Publishing, Chan, pp. 395-411.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777.

Brás, L.P., Bernardino, S.A., Lopes, J.A., Menezes, J.C., 2005. Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour. Chemometr. Intell. Lab. Syst. 75, 91– 99.

Brasil, 2009. Ministério da Agricultura, Pecuária e Abastecimento. Regras para Análises de Sementes / Ministério da Agricultura Pecuária e Abastecimento. Secretaria de Defesa Agropecuária, Mapa/ACS, Brasília.

Carvalho, L.C., Morais, C.L.M, Lima, K.M.G., Leite, G.W.P., Oliveira, G.S., Casagrande, I.P., Santos Neto, J.P. Teixeira, G.H.A., 2018. Using intact nuts and near infrared spectroscopy to classify macadamia cultivars. Food Anal. Methods. 11, 1857– 1866.

Cavalcante, A.K., Sousa, L.B., Hamawaki, O.T., 2011. Determination and evaluation of oil content in soybean seeds by nuclear magnetic resonance methods and Soxhlet. Bioscience J. 27, 8-15.

Conab. 2019. Acompanhamento da safra brasileira: Grão, V. 6 - safra 2018/19- N. 12 - Décimo segundo levantamento. https://www.conab.gov.br/info-agro/safras/graos. (accessed 07 October 2019).

Costa, R.C., Cunha Júnior, L.C., Morgenstern, T.B., Teixeira, G.H.A., Lima, K.M.G., 2016. Classification of jaboticaba fruits at three maturity stages using NIRS and LDA. Anal. Methods. 8, 2533-2538.

Cozzolino, D., Murray, I., Chree, A. Scaife, J.R., 2005. Multivariate determination of free fatty acids and moisture in fish oils by partial least-squares regression and near-infrared spectroscopy. LWT – Food Sci. Technol. 38, 821-828.

Faostat, 2019. FAO Statistics, Food and Agriculture Organization of the United Nations. http://faostat.fao.org/ (accessed 07 October 2019).

Ferreira, D.S., Pallone, J.A.L., Poppi, R.J., 2013. Fourier transform near-infrared spectroscopy (FT-NIRS) application to estimate Brazilian soybean [*Glycine max* (L.) Merril] composition. Food Res. Int. 51, 53-58.

Geladi, P., MacDougall, D., Martens, H., 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Appl. Spectrosc. 39, 491–500.

Xu, R., Hu, W., Zhou, Y., Zhang, X., Xu, S., Guo, Q., Qi, P., Chen, L., Yang, X., Zhang, F., Liu, L., Qiu, L., Wang, J. 2020. Use of near-infrared spectroscopy for the rapid evaluation of soybean [*Glycine max* (L.) Merri.] water soluble protein content. Spectrochim. Acta A. 224, 117400.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics. 11, 137–148.

Miladinović, J., Burton, J.W., Balešević Tubić, S., Miladinović, D., Djordjević, V., Djukić, V., 2011. Soybean breeding: comparison of the efficiency of different selection methods. Turk. J. Agric. For. 35, 469-480.

Marro, N., Cofré, N., Grilli, G., Alvarez, C., Labuckas, D., Maestri, D., Urcelay, C. 2020. Soybean yield, protein content and oil quality in response to interaction ofarbuscular mycorrhizal fungi and native microbial populations from mono-and rotation-cropped soils. Appl. Soil Ecol. 152, 103575.

Lundry, D.R., Ridley, W.P., Meyer, J.J., Riordan, S.G., Nemeth, M.A., Trujillo, W.A., Breeze, M.L., Sorbet, R., 2008. Composition of grain, forage, and processed fractions from second-generation glyphosate-tolerant soybean, MON 89788, is equivalent to that of conventional soybean (*Glycine max* L.). J. Agric. Food Chem. 56, 4611–4622.

Palmer, R.G., Hymowitz, T., 2016. Soybean: germplasm, breeding, and genetics, in: Wrigley, C., Corke, H., Seetharaman, K., Faubion, J. (Eds.), Encyclopedia of Food Grains, second ed., volume 4, Academic Press. pp.333-342.

Pasquini, C., 2003. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. 14, 198-219.

Savitsky, A., and Golay, M.J.E., 1964. Smoothing and differenciation by simplified least squares procedures. Anal. Chem. 36, 1627-1632.

Silva, F.L., Ludke, W.H., Del Conte, M.V., Bueno, T.V., Silva, A.S.L., 2017a. Methods for advancing segregating populations, in: Silva, F.L., Borém, A., Sediyama, T., Ludke, W.H. (Eds.), Soybean Breeding. Springer International Publishing, Chan, pp. 149-170.

Silva, A.F., Sediyama, T., Borém, A., Silva, F.L., Silva, F.C.S., Bezerra, A.R.G., 2017b. Registration and protection of cultivars, in: Silva, F.L., Borém, A., Sediyama, T., Ludke, W.H. (Eds.), Soybean Breeding. Springer International Publishing, Chan, pp. 427-440.

Huang, H., Long, S., Singh, V. 2016. Techno-economic analysis of biodiesel and ethanol co-production from lipid-producing sugarcane. Biofuels Bioprod. Bioref. 10, 299–315.

Woyann, L.G., Meira, D., Zdziarski, A.D., Matei, G., Milioli, A.S., Rosa, A.C., Madella, L.A., Benin, G. 2019. Multiple-trait selection of soybean for biodiesel production in Brazil. Ind. Crop. Prod. 140, 111721.

Figures

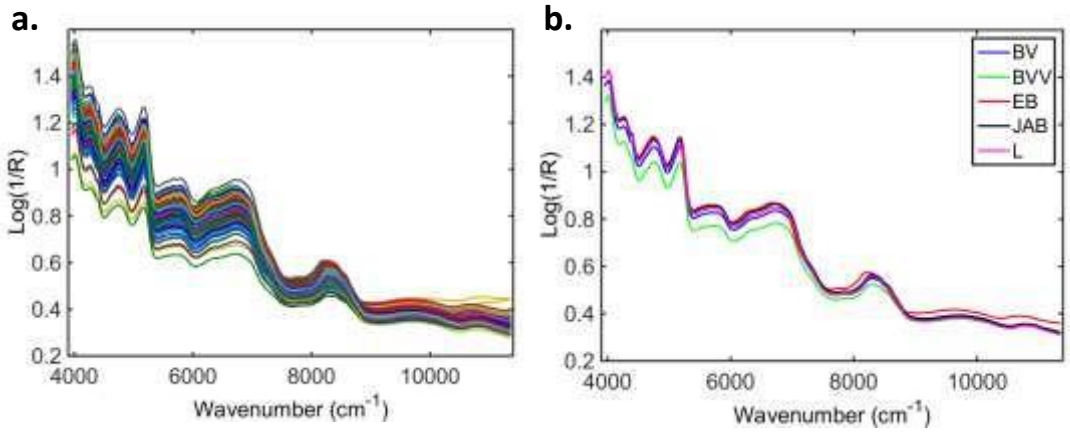


Figure 1. Near-infrared (NIR) spectra of intact soybean grains from 260 different genotypes (a), and the mean NIR spectra of the correspondent genotype groups (b). The spectra are plotted as pseudo-absorbance Log(1/R), where R is the reflectance.

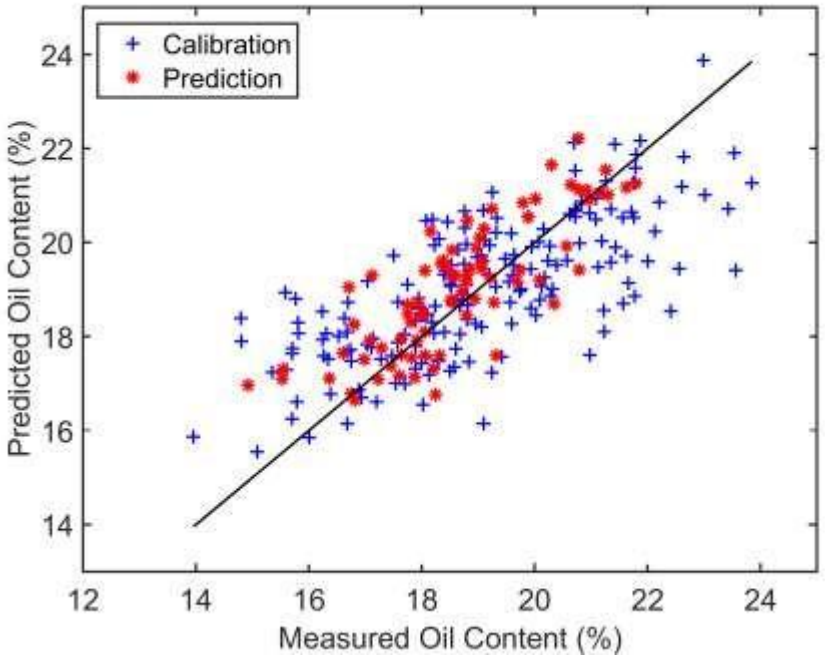


Figure 2. Measured and predicted oil content (%) in intact soybean genotypes.

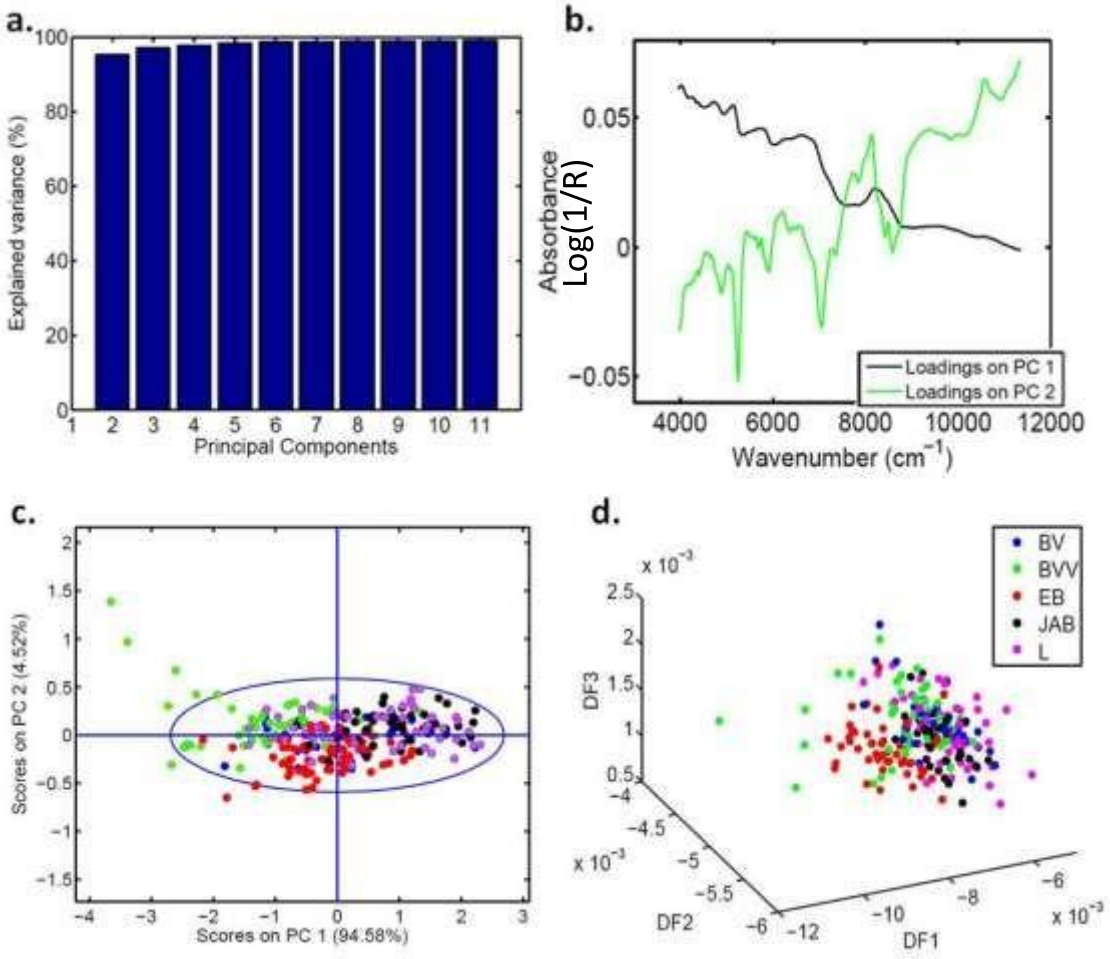


Figure 3. (a) explained variance for each principal component (PC) for the PCA. (b) loadings on the first and second PCs for the PCA. (c) Scores plot on PC1 versus PC2 for classes BV (●), BVV (●), EB (●), JAB (●) and L (●). (d) discrimination factor (DF) plot of PCA-LDA with raw NIR spectra of 260 soybean genotypes grouped in five classes.

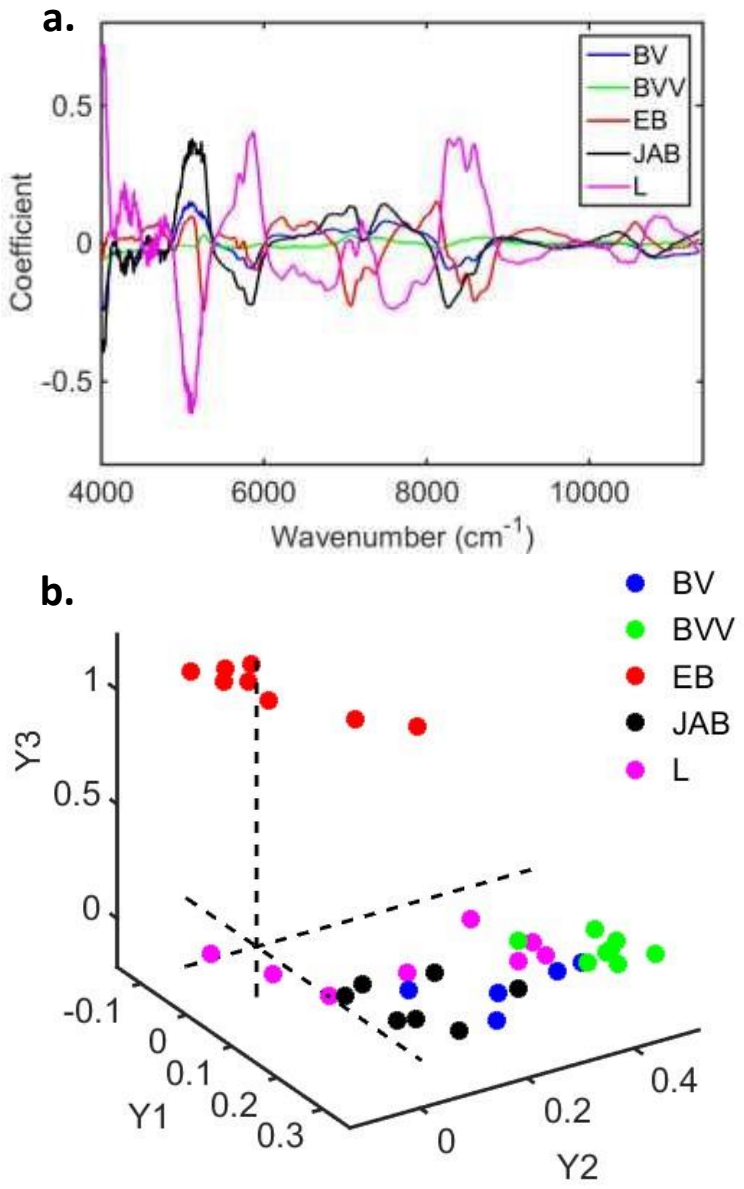


Figure 4. (a) PLS -DA regression coefficients for each genotype class. (b) PLS -DA predicted classes (Y) for the prediction set.

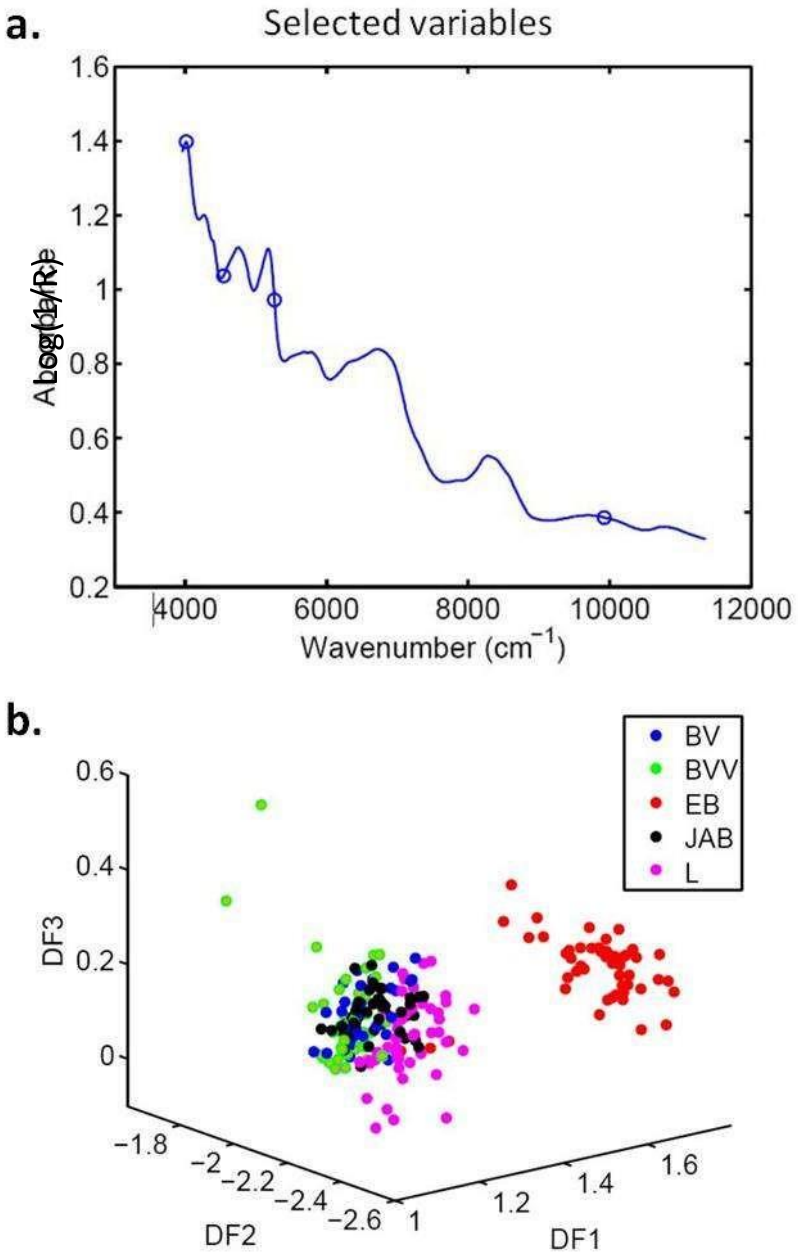


Figure 5. (a) Selected variables (wavenumbers _ cm^{-1}) using successive projection algorithm (SPA). (b) Discrimination factor (DF) plot of SPA-LDA with raw NIR spectra of 260 soybean genotypes grouped in five classes.

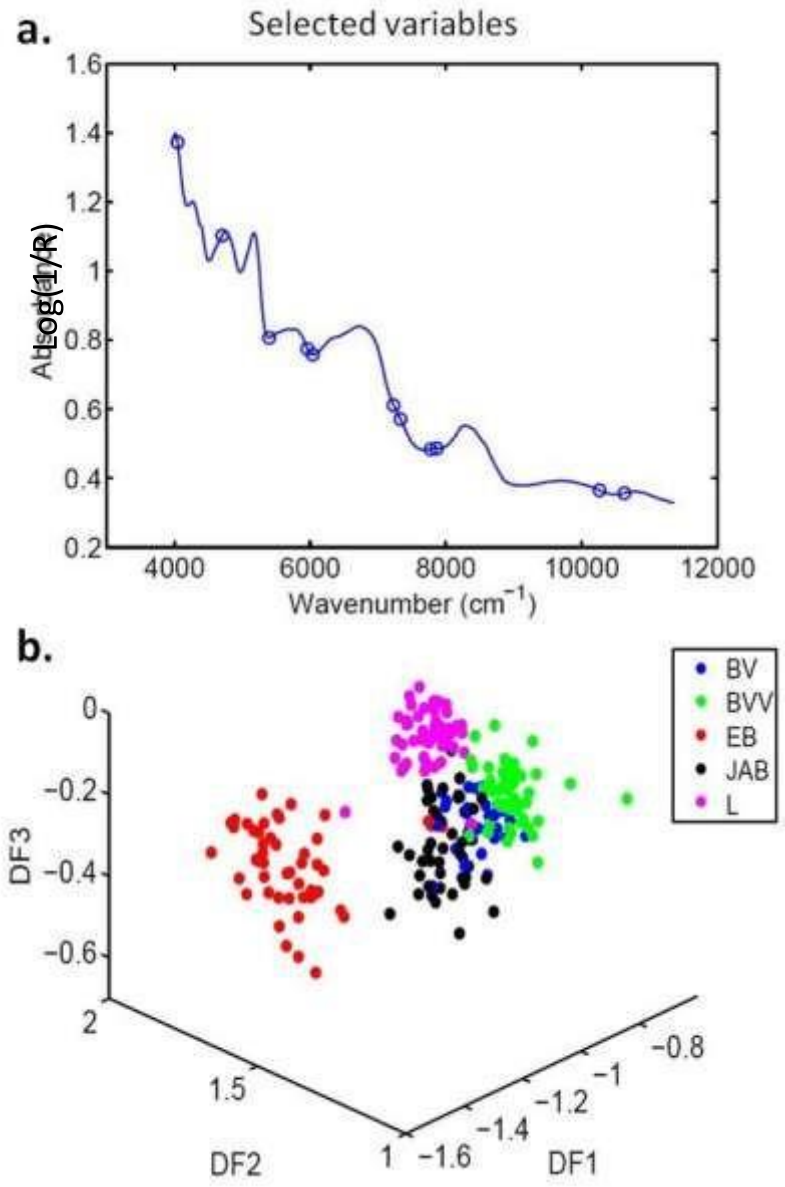


Figure 6. (a) Selected variables (wavenumbers _ cm^{-1}) using genetic algorithm (GA). (b) Discrimination factor (DF) plot of GA -LDA with raw NIR spectra of 260 soybean genotypes grouped in five classes.

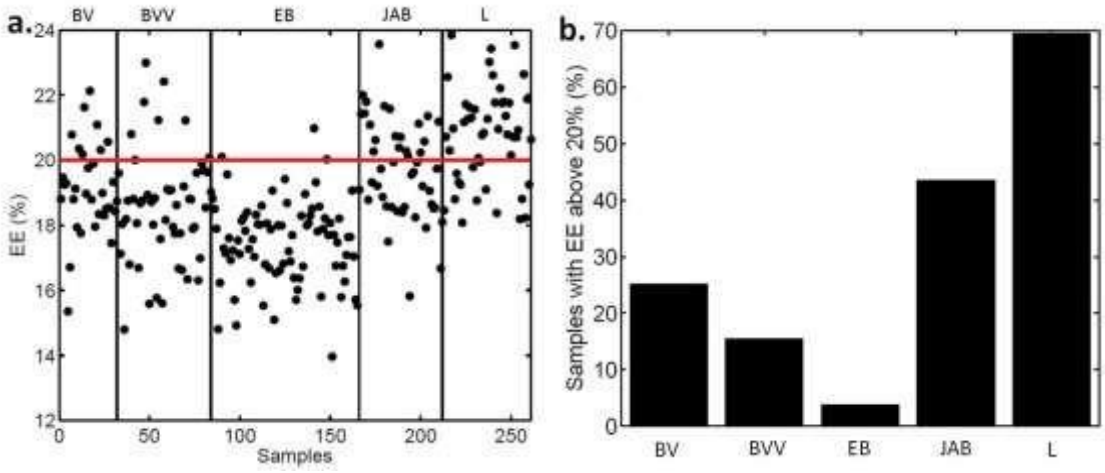


Figure 7. (a) Distribution of oil content between the five soybean genotypes classes, (b) percentage of samples within classes with oil content above 20%.

Tables:

Table 1. Descriptive statistics of intact soybean grains from the different genotypes: moisture, dry matter, and oil content.

Parameters	**Maximum**	**Minimum**	**Average**	**SD[a]**	**number**
Moisture (%)					
BV	7.06	6.75	6.73	0.18	31
BVV	7.10	5.86	6.40	0.27	50
JAB	7.83	6.65	6.65	2.08	45
L	8.88	6.27	7.25	0.84	48
EB	7.26	5.04	5.98	0.41	80
Dry matter (%)					
BV	93.57	93.25	93.27	0.18	31
BVV	94.14	92.90	9360	0.27	50
JAB	92.50	93.83	92.50	0.93	45
L	91.12	93.73	92.75	0.84	48
EB	92.74	94.96	94.02	0.41	80
Oil (%)					
BV	22.13	18.73	19.15	1.39	31
BVV	22.99	14.80	18.54	1.73	50
JAB	19.73	23.57	19.73	1.51	45
L	18.08	23.85	20.75	1.49	48
EB	13.96	20.98	17.43	1.27	80

[a]SD = standard deviation.

Table 2. Figures of merit for the classification models of intact soybean grains from the different genotypes. AUC stands for 'area under the curve' of the receiver operating characteristic curve (ROC).

Algorithm	Class	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-score (%)	AUC
PCA-LDA	BV	94.4	80.0	96.8	87.6	0.884
	BVV	94.4	75.0	100	85.7	0.875
	EB	100	100	100	100	1.00
	JAB	91.7	100	89.7	94.6	0.948
	L	91.7	75.0	96.4	84.4	0.857
PLS-DA	BV	75.8	100	51.6	68.1	0.719
	BVV	92.8	100	85.7	92.3	0.812
	EB	100	100	100	100	1.00
	JAB	89.6	100	79.3	88.5	0.894
	L	91.9	87.5	96.4	91.7	0.937
SPA-LDA	BV	94.4	80.0	96.8	87.6	0.884
	BVV	97.2	87.5	100	93.3	0.937
	EB	100	100	100	100	1.00
	JAB	94.6	100	93.1	96.4	0.966
	L	91.7	100	93.1	96.4	0.857
GA-LDA	BV	91.7	40.0	100	57.1	0.700
	BVV	91.7	87.5	92.9	90.1	0.902
	EB	100	100	100	100	1.00
	JAB	94.6	100	93.1	96.4	0.966
	L	100	100	100	100	1.00