Spanish Journal of Agricultural Research
16 (2), e0203, 10 pages (2018)
eISSN: 2171-9292
https://doi.org/10.5424/sjar/2018162-11805
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, O.A, M.P. (INIA)

RESEARCH ARTICLE

OPEN ACCESS

Screening of transgenic maize using near infrared spectroscopy and chemometric techniques

Xuping Feng[1,2], Haijun Yin[3], Chu Zhang[1], Cheng Peng[4] and Yong He[1]

[1]Zhejiang University, College of Biosystems Engineering and Food Science. Hangzhou 310058, China. [2]China Jiliang University, College of Life Sciences. Hangzhou 310018, China. [3]Jiangsu Mingtian Seeds Science and Technology Co., LTD., Nanjing 210014, China. [4]Institute of Quality and Standard for Agro-products, Zhejiang Academy of Agricultural Sciences; Hangzhou 310021, China.

Abstract

The applicability of near infrared (NIR) spectroscopy combined with chemometrics was examined to develop fast, low-cost and non-destructive spectroscopic methods for classification of transgenic maize plants. The transgenic maize plants containing both cry1Ab/cry2Aj-G10evo proteins and their non-transgenic parent were measured in the NIR diffuse reflectance mode with the spectral range of 700–1900 nm. Three variable selection algorithms, including weighted regression coefficients, principal component analysis -loadings and second derivatives were used to extract sensitive wavelengths that contributed the most discrimination information for these genotypes. Five classification methods, including K-nearest neighbor, Soft Independent Modeling of Class Analogy, Naive Bayes Classifier, Extreme Learning Machine (ELM) and Radial Basis Function Neural Network were used to build discrimination models based on the preprocessed full spectra and sensitive wavelengths. The results demonstrated that ELM had the best performance of all methods, even though the model's recognition ability decreased as the variables in the training of neural networks were reduced by using only the sensitive wavelengths. The ELM model calculated on the calibration set showed classification rates of 100% based on the full spectrum and 90.83% based on sensitive wavelengths. The NIR spectroscopy combined with chemometrics offers a powerful tool for evaluating large number of samples from maize hybrid performance trials and breeding programs.

Additional keywords: facile screening method; *Zea mays*; transgenic maize selection; discrimination model.

Abbreviations used: ANN (Artificial Neural Network); *Bw* (Weighted Regression Coefficient); ELM (Extreme Learning Machine); KNN (K-nearest Neighbor); KS (Kennard-Stone); NBC (Naive Bayes Classifier); NIR (Near Infrared); PC (Principal Component); PCA (Principal Component Analysis); PCR (Polymerase Chain Reaction); PLS-DA (Partial Least Squares Discrimination Analysis); RBFNN (Radial Basis Function Neural Network); SIMCA (Soft Independent Modeling of Class Analogy); SVM (Support Vector Machine).

Authors' contributions: Conceived and designed the experiments: XF and HY. Performed the experiments: XF, CZ and CP. Analyzed the data: XF. Wrote the paper: XF and YH.

Citation: Feng, X.; Yin, H.; Zhang, C.; Peng, C.; He, Y. (2018). Screening of transgenic maize using near infrared spectroscopy and chemometric techniques. Spanish Journal of Agricultural Research, Volume 16, Issue 2, e0203. https://doi.org/10.5424/sjar/2018162-11805

Supplementary material (Fig. S1) accompanies the paper on SJAR's website.

Received: 31 May 2017. **Accepted:** 19 Jun 2018.

Funding: 863 National High-Tech Research and Development Plan (Project No: 2013AA10030401); State Key Laboratory Breeding Base for Zhejiang Sustainable Pest and Disease Control (No. 2010DS700124-KF1712).

Competing interests: The authors declare no competing financial interests.

Correspondence should be addressed to Yong He: yhe@zju.edu.cn

Introduction

Plant breeding uses molecular biology to produce new crop varieties or lines with desirable properties by using techniques to select and introduce genetic modifications and desirable traits into plants (Liu *et al.*, 2015; Yadav *et al.*, 2015; Yang *et al.*, 2017). One major technique of plant breeding is selection, the process of selectively propagating plants with desirable traits and eliminating those with less desirable traits (Schart *et al.*, 2016). This requires plant breeders to screen large populations of crops for individuals that possess the characteristics of interest. Currently, there are various molecular methodologies for plant breeding, such as polymerase chain reaction (PCR) (Taverniers *et al.*, 2004), enzyme linked immunosorbent assays (Kamle *et al.*, 2011) and microarrays (Xu *et al.*, 2005). However, these DNA- and protein-based methods for identification of transgenic plants are time consuming and costly when studying large numbers of samples, and thus unsuitable for on-line application. Therefore a method for the selection of transgenic samples after

transformation that does not require any wet chemistry, particularly the procedure of leaf DNA extraction, would be advantageous where many sample analyses are required.

Near infrared (NIR) spectroscopy is an alternative to traditional chemistry procedures for qualitative and quantitative analysis of biological materials (Wu *et al.*, 2014). The NIR region of 700–2500 nm can gather information related to different hydrogen bonds (C–H, N–H and O–H), which are the primary structures of organic molecules. In contrast to biochemical assays, NIR spectroscopy does not require technical expertise or complex techniques, and the spectrophotometer can be installed anywhere with no requirement of reagents or complicated protocols (García-Molina *et al.*, 2016).

The NIR spectroscopy has been widely used for decades for qualitative and quantitative analysis in agriculture and food research, and has been used for determining the moisture content of peanut kernels (Jin *et al.*, 2015), rice wine composition (Yu *et al.*, 2015), vine water potential (De Bei *et al.*, 2011) and, more recently, to estimate carotenoids in tomato products (Saad *et al.*, 2017) and berry shrivel (Beghi *et al.*, 2015). The application of NIR spectroscopic technology in the genetic field and especially in transgenic foods is now feasible (Alishahi *et al.*, 2010). García-Molina *et al.* (2016) applied NIR spectroscopy to discriminate transgenic wheat lines with low gliadin content from non-transgenic lines. Guo *et al.* (2014) identified clear differences between transgenic and non-transgenic tomatoes using VIS-NIR together with discriminant partial least squares regression with excellent classification accuracy of up to 100%. The basis of this technology for application in transgenic field is that it can identify phenotypic changes caused by genotypic changes that ultimately bring about changes on organic molecular bonds (Alishahi *et al.*, 2010). However, due to the overlapping bands in the NIR region, the spectral analysis is not straightforward and requires chemometric methods to extract important information and classify the mass data set from transgenic and non-transgenic samples (Murayama *et al.*, 2000). Chemometric approaches applied to spectra, using principal component analysis (PCA) and partial least squares discrimination analysis (PLS-DA) as well as support vector machines (SVM), have proved effective in distinguishing transgenic plants and food from non-transgenic samples (Liu *et al.*, 2014; García-Molina *et al.*, 2016; Feng *et al.*, 2017).

Thus, the objectives of this study were to (1) evaluate the possibility and accuracy of using NIR spectra to discriminate transgenic maize plants for breeding screening purposes, (2) identify sensitive wavelengths that attribute differences between transgenic and non-transgenic maize plants and (3) evaluate the performance of five discriminate models and establish an optimal model for classification.

Material and methods

Leaf samples

Seeds of transgenic maize (*Zea mays* L.) (containing both *cry1Ab*/*cry2Aj-G10evo* genes) and its parental line were provided by the Institute of Insect Sciences, Zhejiang University, China. The transgenic maize line contained both herbicide and insect tolerance traits created by *Agrobacterium tumefaciens* mediated transformation. The seeds were sown in plastic buckets in a 1:1:1 mix of soil:calcined clay:torpedo sand. The plants were grown in a greenhouse for 2 months. The youngest fully expanded leaf on a shoot and the second or third leaf formed were selected for NIR scanning. PCR was used to check the integrity of copies of the genes introduced during the breeding phase and the expression of the inserted exogenous gene.

NIR scanning and pretreatment

Maize leaf samples were scanned using a field portable NIR spectroradiometer NIRez (Isuzuoptics, Taiwan, China) with spectra range of 900–1700 nm. Reflectance spectra were collected every 10 nm within 900–1700 nm. Each sample was analyzed in three duplicates to reduce measurement errors. Maize leaf samples were placed directly in the diffuse reflection accessory. A total of 326 maize leaves were sampled, comprising 163 transgenic and 163 non-transgenic samples, with at least one leaf collected from each plant. Using the Kennard-Stone (KS) algorithm (Saptoro *et al.*, 2012), the whole dataset was divided into two groups: calibration and prediction sets. The KS algorithm calculates the Euclidean distance of every two NIR spectra and chooses two spectra with farthest distance as the first pairs, then calculates the Euclidean distances of the rest samples with the first pairs, which made the samples in both sets were representatively of the population and could avoid overfitting to some extent. Therefore, based on the KS method, 120 transgenic and 120 non-transgenic samples were chosen for calibration set. The remaining 43 transgenic and 43 non-transgenic samples were selected to form the prediction set. Samples were classified according to the genetic background using a classification model, which were preferably close to the values used to codify the class.

Unscrambler x10.1 (CAMO PROCESS AS, Oslo, Norway) and MATLAB version R2010b (The MathWorks, Natick, MA, USA) were used to process the data. In addition, origin Pro 7.0SR0 (Origin Lab Corporation, Northampton, MA, USA) software was used to design graphs. Model performances were evaluated by the classification accuracy of the calibration and prediction sets.

Chemometrics and data analysis

The first step involving classification was carried out using an exploratory analysis with PCA (Bryant & Yarnold, 1995). The PCA developed on the whole NIR spectral data was used to visualize the possible clusters and trends in the PCA score plot. In the second step, five classification methods including K-nearest neighbor (KNN) (Gil-Pita & Yao, 2009), Soft Independent Modeling of Class Analogy (SIMCA) (Waddell *et al.*, 2014), Naive Bayes Classifier (NBC) (Islam *et al.*, 2007), Extreme Learning Machine (ELM) (Huang *et al.*, 2012) and Radial Basis Function Neural Network (RBFNN) (Kosic, 2015) were applied on the original raw spectral data (90 bands) to identify the transgenic samples. Variable (wavelength) selection in multivariate analysis is an important step because the removal of highly correlated variables produces better predictions and a simpler process. Here, three variable selection algorithms [weighted regression coefficient (*Bw*), PCA-loadings and second derivative (2nd derivative)] were used to extract sensitive wavelengths that contributed the most discrimination information to these genotypes. In the final stage of this study, the actual roles of the extracted sensitive wavelengths were evaluated by establishing discrimination models based on the sensitive wavelengths. Classification methods were carried out using only a few wavelengths selected in the previous step as input, and the results were compared with the classification obtained by using the whole spectra. Figure 1 illustrates the main steps for the whole procedure.

PCA

PCA was used to reduce the dimensions of the original spectra into a low dimensional subspace, and an alternative set of coordinates called principal components (PCs) was projected (Rinnan *et al.*, 2009). The number of PCs is less than or equal to the number of original variables, and the first few PCs contain most

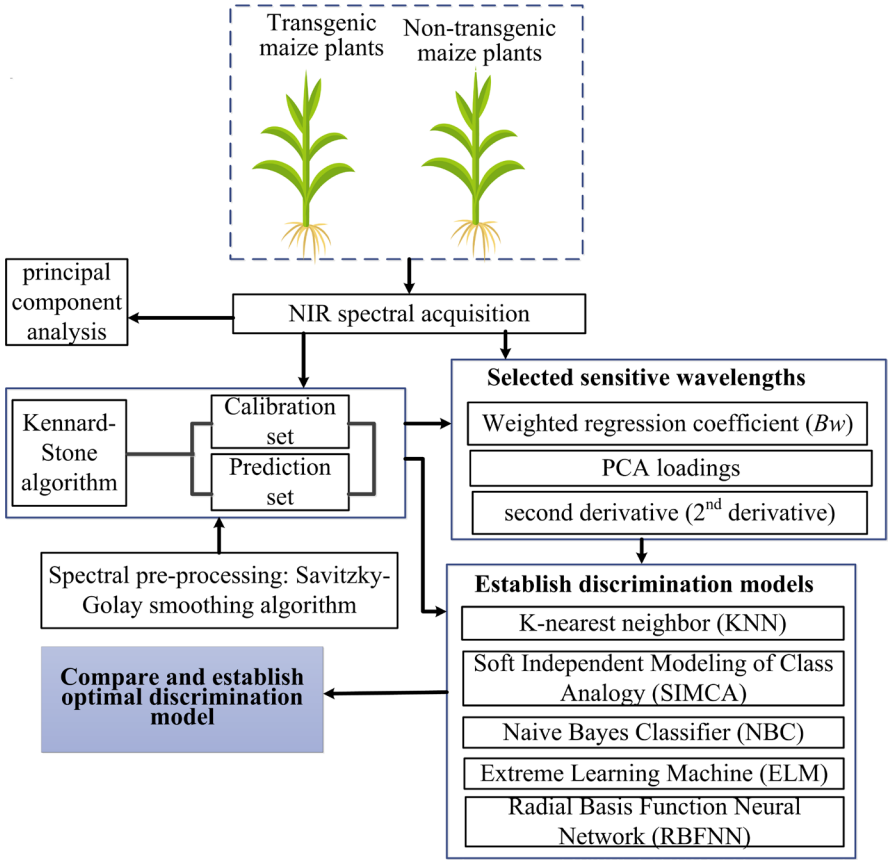

Figure 1. Flowchart of NIR spectral data analysis for discrimination of transgenic maize plants.

of the spectral information. For visual discrimination, we projected each of the spectra in the newly formed coordinate space of selected PCs (score plot), and the scores of the most significant PCs corresponding to each NIR spectra were used. PCA is described in detail by Rinnan *et al.* (2009).

Important wavelength selection

Variable selection is quite efficient in spectra analysis for handling collinearity problems and extracting the most important information. Many approaches are available for selecting sensitive wavelengths; and identifying prominent peaks and/or valleys with *Bw*, 2nd derivative and PCA-loading are among the most commonly used (Barbin *et al.*, 2012; Rodríguez-Pulido *et al.*, 2013; Zhang *et al.*, 2015). In the present study, important wavelengths were selected from the *Bw* plot in the PLS regression model (Zhang *et al.*, 2015). The 2nd derivative by Savitzky–Golay method was used to identify key wavelengths related to variations in classification (Barbin *et al.*, 2012). Loadings resulting from PCA of the raw spectral data represent the regression coefficient, and indicate the most dominant wavelength (Rodríguez-Pulido *et al.*, 2013). Simplified classification models were then developed using the selected wavelengths from the above three methods, and the results were compared with the classification accuracy obtained with the whole spectral data.

Discriminate models

To accurately identify transgenic plants from the parental line, pattern recognition approaches, including KNN, SIMCA, NBC, ELM and RBFNN, were used to establish discriminate models. These mentioned methods are the most commonly used in classification models. The details of related theory for these methods is found in the literature (Islam *et al.*, 2007; Gil-Pita & Yao, 2009; Huang *et al.*, 2012; Waddell *et al.*, 2014; Kosic, 2015). Other applied discriminate models such as PLS-DA and SVM have been used by other researchers for discrimination of transgenic maize kernels and transgenic rice seeds (Liu *et al.*, 2014; Feng *et al.*, 2017).

The KNN method is used to classify objects based on the closest training examples in the feature space. By comparing the distance between unknown samples (testing set) and samples in the training set, samples are classified based on proximity to training set samples (Gil-Pita & Yao, 2009). For each row (spectra data) in the target dataset (the set to be classified), the K closest members (*i.e.* the KNNs) of the training dataset are located. A Euclidean distance measure is used to calculate how close each member of the training set is to the target row that is being examined.

SIMCA is performed to describe each group separately based on their similarities in a principal component space (Waddell *et al.*, 2014). Objects are considered to belong to the class if their Euclidean distance from the constructed PC space is not significantly larger than the Euclidean distance of the class objects from their PC space.

NBCs are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features (Islam *et al.*, 2007). NBC is calculated based on the simplifying assumption that the attribute values are conditionally independent of a given target value.

ELMs are feedforward neural networks for classification or regression with a single layer of hidden nodes, where the weights connecting inputs to hidden nodes are randomly assigned and never updated (Huang *et al.*, 2012). ELM has one input layer and one linear hidden layer, and the optimal weights between the input and hidden layers are randomly chosen by minimal norm least square method.

RBFNN can separate a set of objects having different class memberships, which presents certain advantages including better approximation capabilities and shorter computational time (Kosic, 2015). In RBFNN, a radial basis function is used as the activation function for each node in the hidden layer, and nonlinear transformation from the input space to the hidden unit space applying a linear combination of the radial basis function is used in the network.

Results and discussion

Spectroscopic analysis

The spectral data were collected over the range of 900–1700 nm. Only spectra of 947.07–1666.49 nm were used for analysis as the head and the end of the spectra showed obvious noise caused by the instrument and the environment (Fig. S1 [suppl]). To eliminate the noise of the spectral data and improve the predictive ability for samples, raw spectra went through noise suppression by Savitzky–Golay smoothing algorithm with a window size of 7 and polynomial of order 2 (Pan *et al.*, 2010). The trend of spectra between transgenic and non-transgenic plants was very similar, with similar peak and valley positions (Fig. 2A & B). Slight differences were found between the mean spectral reflectance value of transgenic and non-transgenic maize (Fig. 2C). As most of the spectral information overlapped, it was difficult to discriminate the transgenic maize plants directly by their characteristic spectral feature. Therefore, chemometric

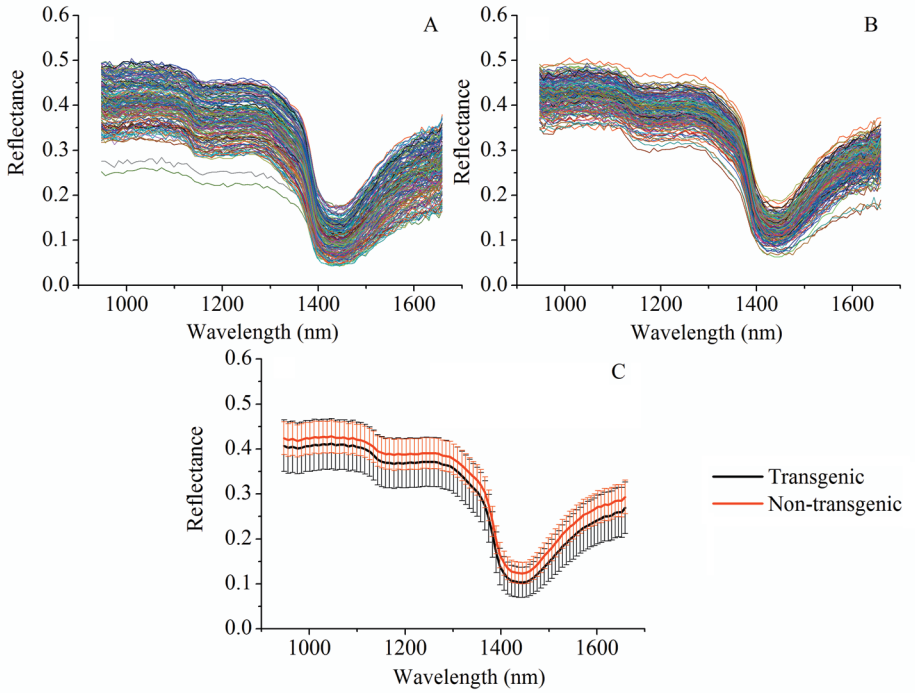

Figure 2. Profiles of original spectra (A: transgenic maize, B: non-transgenic maize) and mean spectra of transgenic and non-transgenic maize plants (C). The shaded areas represent the standard deviation in each wavelength.

methods were introduced to build a qualitative model for classification.

A PC model for exploratory purposes was first created to examine the qualitative difference of transgenic and non-transgenic maize leaves in PC space (Fig. 3). No distinct clustering was shown by scatter plots of PC1 *vs*. PC2 and PC3 *vs*. PC4 of transgenic and non-transgenic maize plants after PCA analysis (Fig. 3A & B). Transgenic and non-transgenic maize were clustered together in the projection of PC5 with PC6 and could not be effectively separated (Fig. 3C). The discrimination based on PCA was not effective in classing transgenic samples. It is worth mentioning that the overexpression of *cry1Ab*/*cry2Aj-G10evo* gene by transgenic editing technology improves glyphosate and insect resistance and ultimately changes organic molecular bonds, but there is no other phenotypic difference between transgenic and non-transgenic maize (Feng *et al.*, 2017). As the PCA program failed to class transgenic maize from its parental line, other discriminant models were utilized for improved separation.

Classification performance based on entire spectral bands

Five discriminate models (KNN, SIMCA, NBC, ELM and RBFNN) were established on the full NIR spectra to evaluate the classification performance and their comparisons are listed in Table 1. The prediction accuracy for each model was analyzed by the accuracy (in percentage) for the calibration and prediction sets. The accuracy of the classification was expressed as the fraction of correctly predicted samples to the total samples. Sensitivity of accuracy showed significant differences among the discriminate models calculated on entire spectral bands.

The best performance was for ELM, with classification accuracy of calibration and prediction sets exceeding 95%. The RBFNN model was less accurate than the ELM model, but was still acceptable. RBFNN and ELM are typical artificial neural networks (ANNs) (Lian *et al.*, 2014) and can learn nonlinear functions from the NIR spectral data. In the calibration set, the respective accuracies were both 100% for the two ANNs. The SIMCA, KNN and NBC models of the two sample sets were not satisfactory, with classification accuracies of the calibration set less than 80%. The discrimination performance by NBC was the lowest with accuracy of approx. 55% – many problems encountered by modern analytical chemists are nonlinear, and approaches such as NBC do not apply well. It is noteworthy that previous studies attempts to discriminate transgenic plants have also shown that linear classification methods were less satisfactory compared to those of SVM (Liu *et al.*, 2014).

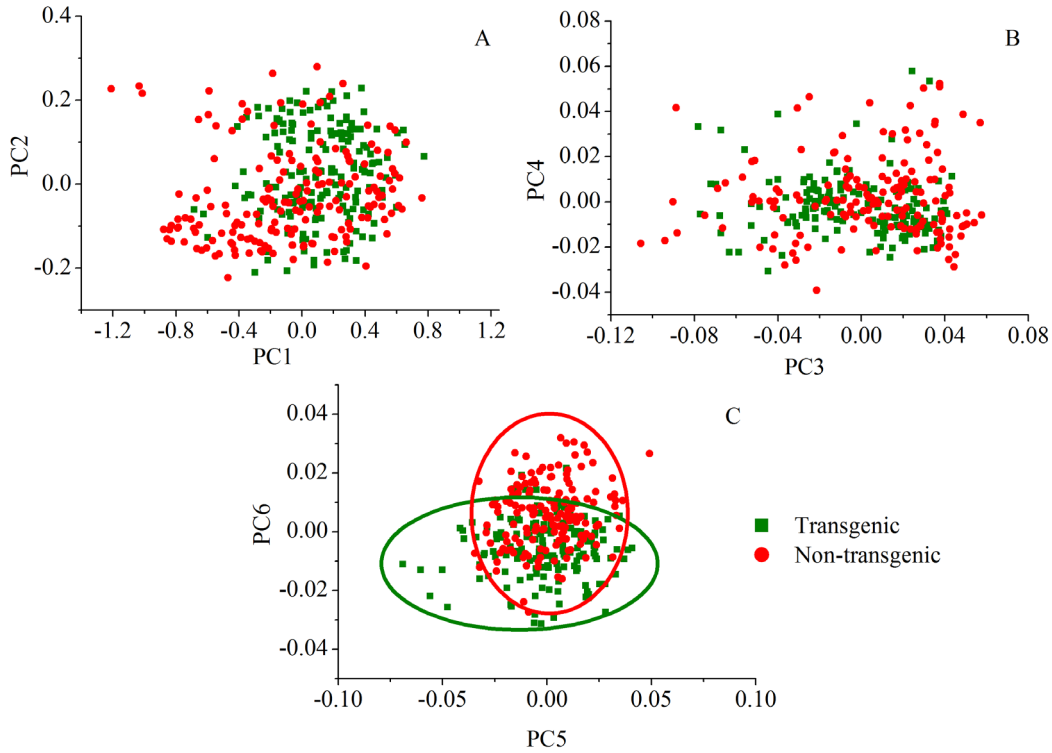

Figure 3. Score scatter plots of (A) PC1 *vs.* PC2, (B) PC3 *vs*. PC4, and (C) PC5 *vs*. PC6 for transgenic and non-transgenic maize plants.

We used a chemometrics approach because the discriminate models were used to highlight the chemical differences between transgenic and non-transgenic maize plants. NIR spectroscopy can be used to identify transgenic samples as this technology can capture the phenotypic changes in chemical bonding of organic molecules that are altered as a result of genetic changes (Alishahi *et al.*, 2010). Feng *et al.* (2017) developed a successful model to discriminate transgenic maize kernels based on the NIR hyperspectral imaging with the spectral range of 874.41–1733.91 nm. They demonstrated that SVM and PLS-DA models established on the full range of NIR spectra had good classification performance. The hyperspectral imaging they used had the advantage of acquiring spectral and spatial information, which allowed the identification of transgenic maize kernels on the prediction maps. Compared to hyperspectral imaging, our simple instrument acquires small point-source information from the sample and does not contain spatial information which is also important for discrimination. However, the NIR system that we used was portable and could be used from a USB flash drive without need of any installation, which is very helpful for the transgenic crop selection purposes of crop breeding laboratories. García-Molina *et al.* (2016) used spectral sensing in the region of 400–2500 nm to discriminate transgenic wheat grain with excellent accuracy. Moreover, NIR combined with chemometrics has proved effective in identification of transgenic soybean oils (Luna *et al.*, 2013), rice mutant seeds (Liu *et al.*, 2014) and transgenic tomato (Xie *et al.*, 2007). That is to say, a discriminant analysis model based on NIR spectra obtained enough information to discriminate the transgenic from parental samples because of their differences in chemical components.

Table 1. Discriminant analysis results of transgenic and non-transgenic maize leaves based on entire spectral bands.

Discriminate models[1]	Par[2]	Calibration set %	Prediction set %
SIMCA	15,15	64.17	75.00
KNN	3	76.25	83.33
NBC		54.17	55.95
ELM	126	100	95.20
RBFNN	10	100	92.86

[1]SIMCA: Soft Independent Modeling of Class Analogy; KNN: K-nearest neighbor; NBC: Naive Bayes Classifier; ELM: Extreme Learning Machine; RBFNN: Basis Function Neural Network. [2]Par shows the parameters of the discrimination models, number of PCs for SIMCA, number of selected nearest neighbors for KNN, optimum number of hidden nodes for ELM and spread values for RBFNN.

This suggests that application of NIR spectroscopy with chemometrics could successfully identify transgenic crops, and it has advantages of being fast, time-saving and low cost compared with molecular methods.

Sensitive wavelength selection and classification analysis based on feature wavelengths

The neighboring NIR wavelengths are always collinear, therefore effective wavelength methods are applied to determine the contributions of individual wavelengths for identification (Feng *et al.*, 2017). Certain wavelengths with obvious peaks and valleys were selected as sensitive wavelengths. Figure 4 shows the effective wavelengths that were selected by 2nd derivative, PCA-loadings and *Bw* with the preprocessing method. The number of sensitive wavelengths was reduced to seven for PCA-loading, ten for 2nd derivative and eight for *Bw*. The loading line plot for these selection methods showed similar prominent positive peaks at 1125.6, 1167.55, 1413.97, 1444.34 and 1520.78 nm. The band at around 1125 nm belongs to the second overtone of the C–H stretch (Kumaravelu *et al.*, 2017). The peak near 1167 nm is caused by the C–H stretching 2nd overtone of CH_3 and -CH_2- groups, and that at 1413 nm by the C–H stretching and C–H deformation vibration of CH_3 and -CH_2- groups, respectively (Schaefer *et al.*, 2013). The peak near 1444 nm is consistent with the N–H stretch (Boyd *et al.*, 2006). Furthermore, a peak near 1520 nm is assigned to N–H stretch vibration (Minami & Iwahashi, 2011). These wavelengths are believed to correspond to NIR spectral bands relevant to maize property changes caused by the transgenic event.

Normally, the full spectra can contain hundreds of variables. According to Dai *et al.* (2015), sensitive wavelengths might be equally or more efficient than full spectra in multivariate analysis. The reduced number of wavelengths was sufficient to characterize most classification tasks. Judicious selection of wavelengths decreases sensitivity to non-linearity and discarding the uninformative wavelengths can expedite data processing and improve model accuracy and robustness. In the final stage of this study, the actual roles of the sensitive wavelengths selected by the above-mentioned three methods were evaluated. The newly proposed combined discriminate models were compared: PCA-loadings–SIMCA, 2nd derivative–SIMCA, *Bw*–SIMCA, PCA-loadings–KNN, 2nd derivative–KNN, *Bw*–KNN, PCA-loadings–NBC, 2nd derivative–NBC, *Bw*–NBC, PCA-loadings–ELM, 2nd derivative–ELM, *Bw*–ELM, PCA-loadings–RBFNN, 2nd derivative–RBFNN, and *Bw*–RBFNN (Table 2). The identification of sensitive wavelength algorithms can improve the model performance, but some algorithms can reduce recognition ability of the model. The strongest discriminant model was developed by *Bw*–ELM with a classification rate of 90.83% for the calibration set and 86.90% for the prediction set. A correct classification rate of 95% was obtained in the calculation set based on the ELM discriminant model, which indicated that these selected emission peaks had reliable discrimination

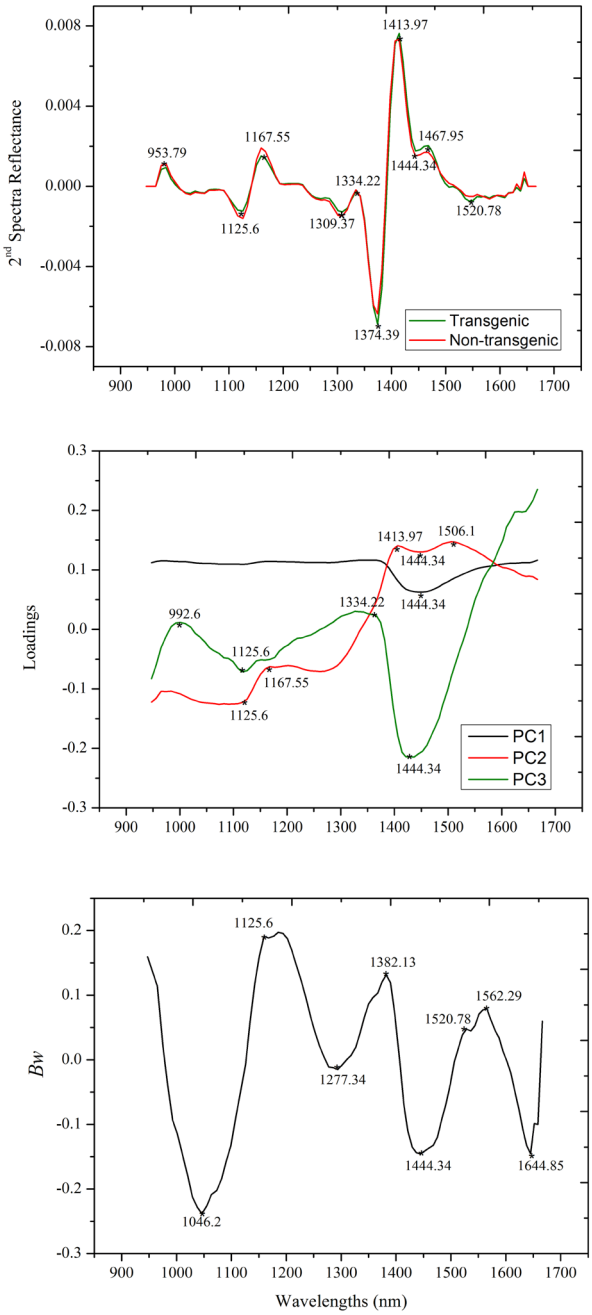

Figure 4. Distribution of sensitive wavelengths of transgenic and non-transgenic maize leaves selected by 2nd derivative, PCA-loadings and *Bw*.

Table 2. Results of discriminate models using important wavelengths

Discriminate models[1]	PCA-loadings			2nd derivative			*Bw*		
	Par[2]	Calibration set %	Prediction set %	Par[2]	Calibration set %	Prediction set %	Par[2]	Calibration set %	Prediction set %
SIMCA	3,3	64.58	83.00	5,5	63.75	73.81	4,4	64.17	72.62
KNN	3	74.17	84.52	3	78.33	88.10	3	75.83	86.90
NBC		65.00	69.05		66.00	71.43		65.00	71.43
ELM	101	90.83	84.52	144	95.00	83.33	102	90.83	86.90
RBFNN	9	96.00	76.19.5	4	91.25	82.14	4	88.75	80.95

[1,2] Model parameters and abbreviations as in Table 1.

power for distinguishing transgenic maize plants. The RBFNN model established on sensitive wavelengths had poorer classification accuracy compared to all wavelengths. The NBC and SIMCA models had poor classification performance, showing the correct classification rate in the range of 64.17%–83.00%, although they had better recognition capability when using the sensitive wavelengths for the calibration and prediction sets. The recognition ability of the KNN model established on sensitive wavelengths selected by 2nd derivative was higher than that for all wavelengths with classification rates of 78.33% and 88.10% for the calibration and prediction sets, respectively. These results showed that the most appropriate classification technique for the classification task was the ELM model, which tended to produce more robust results, although a good performance of prediction set was also obtained with RBFNN.

In summary, using NIR spectroscopy allowed us to monitor phenotypic changes in maize plants as a consequence of genetic changes. Seven classification methods were tested to determine which provided the best results. First, they were used on the entire spectral bands acquired by the system and then using only the most important selected wavelengths. Thus, in addition to obtaining the best combination of methods to select features and classify genotypes, the performance of the selected wavelengths was evaluated. The results showed an excellent classification by the neural network models ELM and RBFNN. An ELM model using the spectral and features peaks after appropriate data pretreatment had valuable and robust calibration and prediction abilities with a classification accuracy exceeding 90% on the calibration set. The use of NIR combined with chemometrics for screening transgenic maize in plant breeding programs is a very attractive platform and has potential for wide use in rapid and on-site screening because it is non-invasive, cost-effective and does not require pretreatment.

References

Alishahi A, Farahmand H, Prieto N, Cozzolino D, 2010. Identification of transgenic foods using NIR spectroscopy: a review. Spectrochim Acta Part A 75 (1): 1-7. https://doi.org/10.1016/j.saa.2009.10.001

Barbin D, Elmasry G, Sun DW, Allen P, 2012. Near-infrared hyperspectral imaging for grading and classification of pork. Meat Sci 90 (1): 259-268. https://doi.org/10.1016/j.meatsci.2011.07.011

Beghi R, Giovenzana V, Marai S, Guidetti R, 2015. Rapid monitoring of grape withering using visible near-infrared spectroscopy. J Sci Food Agr 95 (15): 3144-3149. https://doi.org/10.1002/jsfa.7053

Boyd DS, Entwistle JA, Flowers AG, Armitage RP, Goldsmith PC, 2006. Remote sensing the radionuclide contaminated Belarusian landscape: a potential for imaging spectrometry? Int J Remote Sens 27 (10): 1865-1874. https://doi.org/10.1080/01431160500328355

Bryant FB, Yarnold PR, 1995. Principal-components analysis and exploratory and confirmatory factor analysis. In: Reading and understanding multivariate statistics; Grimm LG & Yarnold PR (Eds.), pp: 99-136. Am Psychol Assoc, Washington DC.

Dai Q, Cheng JH, Sun DW, Zeng XA, 2015. Advances in feature selection methods for hyperspectral image processing in food industry applications: A review. Crit Rev Food Sci 55 (10): 1368-1382. https://doi.org/10.1080/10408398.2013.871692

De Bei R, Cozzolino D, Sullivan W, Cynkar W, Fuentes S, Dambergs R, Tyerman S, 2011. Non-destructive measurement of grapevine water potential using near infrared spectroscopy. Aust J Grape Wine R 17 (1): 62-71. https://doi.org/10.1111/j.1755-0238.2010.00117.x

Feng X, Zhao Y, Zhang C, Cheng P, He Y, 2017. Discrimination of transgenic maize kernel using NIR hyperspectral imaging and multivariate data analysis. Sensors 17 (8): 1894. https://doi.org/10.3390/s17081894

García-Molina MD, García-Olmo J, Barro F, 2016. Effective identification of low-gliadin wheat lines by near infrared

spectroscopy (NIRS): implications for the development and analysis of foodstuffs suitable for celiac patients. Plos One 11 (3): e0152292. https://doi.org/10.1371/journal.pone.0152292

Gil-Pita R, Yao X, 2009. Evolving edited k-nearest neighbor classifiers. Int J Neural Syst 18 (6): 459-467. https://doi.org/10.1142/S0129065708001725

Guo H, Pan T, Chen J, Wang J, Cao G, 2014. Vis−NIR wavelength selection for non-destructive siscriminant analysis of breed screening of transgenic sugarcane. Anal Methods-UK 6 (21): 8810-8816. https://doi.org/10.1039/C4AY01833H

Huang GB, Zhou H, Ding X, Zhang R, 2012. Extreme learning machine for regression and multiclass classification. IEEE T Syst Man CY B 42 (2): 513-529. https://doi.org/10.1109/TSMCB.2011.2168604

Islam M J, Wu QMJ, Ahmadi M, Sid-Ahmed MA, 2007. Investigating the performance of naive- bayes classifiers and k- nearest neighbor classifiers. J Converg Inform Technol 5 (2): 133-137. https://doi.org/10.1109/ICCIT.2007.148

Jin H, Li L, Cheng J, 2015. Rapid and non-destructive determination of moisture content of peanut kernels using hyperspectral imaging technique. Food Anal Method 8 (10): 1-9. https://doi.org/10.1007/s12161-015-0147-1

Kamle S, Ojha A, Kumar A. 2011. Development of an enzyme linked immunosorbant assay for the detection of Cry2Ab Protein in transgenic plants. Gm Crops 2 (2): 118-125. https://doi.org/10.4161/gmcr.2.2.16191

Kosic D, 2015. Fast clustered radial basis function network as an adaptive predictive controller. Neural Netw 63: 79-86. https://doi.org/10.1016/j.neunet.2014.11.008

Kumaravelu C, Ravi A, Gopal A, Joshi J, 2017. Estimation of oil content of single cotton seed using NIR spectrometer by area under curve method. Trends in Industrial Measurement and Automation (TIMA), IEEE Conf, pp: 1-4. https://doi.org/10.1109/TIMA.2017.8064798

Lian C, Zeng Z, Yao W, Tang H, 2014. Performance of combined artificial neural networks for forecasting landslide displacement. IEEE Conf, pp: 418-423. https://doi.org/10.1109/IJCNN.2014.6889497

Liu C, Liu W, Lu X, Chen W, Yang J, Zheng L, 2014. Nondestructive determination of transgenic Bacillus thuringiensis rice seeds (Oryza sativa L.) using multispectral imaging and chemometric methods. Food Chem 153 (12): 87-93. https://doi.org/10.1016/j.foodchem.2013.11.166

Liu Y, Qin L, Han L, Xiang Y, Zhao D, 2015. Overexpression of maize SDD1 (ZmSDD1) improves drought resistance in Zea mays L. by reducing stomatal density. Plant Cell Tiss Org 122 (1): 147-159. https://doi.org/10.1007/s11240-015-0757-8

Luna AS, Silva APD, Pinho JSA, Ferré J & Boqué R, 2013. Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy. Spectrochim Acta A 100: 115-119. https://doi.org/10.1016/j.saa.2012.02.085

Minami H, Iwahashi M, 2011. Molecular self-assembling of N-methylacetamide in solvents. Int J Specty 2011: 640121. https://doi.org/10.1155/2011/640121

Murayama K, Czarnikmatusewicz B, Wu Y, Tsenkova R, Ozaki Y, 2000. Comparison between Conventional Spectral Analysis Methods, Chemometrics, and Two-Dimensional Correlation Spectroscopy in the Analysis of Near-Infrared Spectra of Protein. Appl Spectrosc 54 (7): 978-985. https://doi.org/10.1366/0003702001950715

Pan T, Xie J, Chen J, Chen H, 2010. Joint optimization of savitzky-golay smoothing modes and PLS factors was applied to near infrared spectral analysis of serum cholesterol. IEEE Conf, pp: 1-4. https://doi.org/10.1109/ICBBE.2010.5514789

Rinnan Å, Berg FVD, Engelsen SB, 2009. Review of the most common pre-processing techniques for near-infrared spectra. Trac-Trend Anal Chem 28 (10): 1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

Rodríguez-Pulido FJ, Barbin DF, Sun DW, Gordillo B, González-Miret ML, Heredia FJ, 2013. Grape seed characterization by NIR hyperspectral imaging. Postharvest Biol Tec 76: 74-82. https://doi.org/10.1016/j.postharvbio.2012.09.007

Saad AG, Pék Z, Szuvandzsiev P, Gehad DH, Helyes L, Saad AG, Pék Z, Szuvandzsiev P, Gehad DH, Helyes L, 2017. Determination of carotenoids in tomato products using Vis/NIR spectroscopy. J Microbiol Biotechn Food Sci 7 (1): 27-31. https://doi.org/10.15414/jmbfs.2017.7.1.27-31

Schaefer C, Lecomte C, Clicq D, Merschaert A, Norrant E, Fotiadu F, 2013. On-line near infrared spectroscopy as a Process Analytical Technology (PAT) tool to control an industrial seeded API crystallization. J Pharmaceut Biomed Anal 83 (5): 194-201. https://doi.org/10.1016/j.jpba.2013.05.015

Schart JG, Wiel CCMVD, Lotz LAP, Smulders MJM, 2016. Opportunities for products of new plant breeding techniques. Trends Plant Sci 21(5): 438-449. https://doi.org/10.1016/j.tplants.2015.11.006

Saptoro A, Tadé MO, Vuthaluru H. 2012. A modified kennard-stone algorithm for optimal division of data for developing artificial neural network models: chemical product and process modeling. Chem Prod Process Model 7 (1): 1-14. https://doi.org/10.1515/1934-2659.1645

Taverniers I, Bockstaele EV, Loose MD, 2004. Cloned plasmid DNA fragments as calibrators for controlling GMOs: different real-time duplex quantitative PCR methods. Anal Bioanal Chem 378 (5): 1198-1207. https://doi.org/10.1007/s00216-003-2372-5

Waddell EE, Williams MR, Sigman ME, 2014. Progress toward the determination of correct classification rates in fire debris analysis II: utilizing soft independent modeling

of class analogy (SIMCA). J Forensic Sci 59 (4): 927-935. https://doi.org/10.1111/1556-4029.12417

Wu Z, Ouyang G, Shi X, Ma Q, Wan G, Qiao Y, 2014. Absorption and quantitative characteristics of C-H bond and O-H bond of NIR. Opt Spectrosc 117 (5): 703-709. https://doi.org/10.1134/S0030400X1411023X

Yang X, Lei L, Jiang X, Wei W, Cai X, Su J, Feng W, Lu BR, 2017. Genetically engineered rice endogenous 5-enolpyruvoylshikimate-3-phosphate synthase (epsps) transgene alters phenology and fitness of crop-wild hybrid offspring. Sci Rep-UK 7 (1): 6834. https://doi.org/10.1038/s41598-017-07089-9

Xie L, Ying Y, Ying T, Yu H, Fu X, 2007. Discrimination of transgenic tomatoes based on visible/near-infrared spectra. Anal Chim Acta 584 (2): 379-384. https://doi.org/10.1016/j.aca.2006.11.071

Xu X, Li Y, Zhao H, Wen SY, Wang SQ, Huang J, Luo Y B, 2005. Rapid and reliable detection and identification of GM events using multiplex PCR coupled with oligonucleotide microarray. J Agr Food Chem 53 (10): 3789-3794. https://doi.org/10.1021/jf048368t

Yadav UP, Ayre BG, Bush DR, 2015. Transgenic approaches to altering carbon and nitrogen partitioning in whole plants: assessing the potential to improve crop yields and nutritional quality. Front Plant Sci 6: 275. https://doi.org/10.3389/fpls.2015.00275

Yu HY, Niu X Y, Lin H J, Ying Y B, Li BB, Pan XX, 2015. A feasibility study on on-line determination of rice wine composition by Vis-NIR spectroscopy and least-squares support vector machines. Food Chem 113 (1): 291-296. https://doi.org/10.1016/j.foodchem.2008.06.083

Zhang C, Liu F, Kong W, He Y, 2015. Application of visible and near-infrared hyperspectral imaging to determine soluble protein content in oilseed rape leaves. Sensors 15 (7): 16576-16588. https://doi.org/10.3390/s150716576