Discrimination of Maize Haploid Seeds from Hybrid Seeds Using Vis Spectroscopy and Support Vector Machine Method

LIU Jin[1], GUO Ting-ting[1], LI Hao-chuan[2], JIA Shi-qiang[3],
YAN Yan-lu[3], AN Dong[3], ZHANG Yao[1], CHEN Shao-jiang[1]*

1. National Maize Improvement Center of China, China Agricultural University, Beijing 100193, China
2. Agronomy College, Henan Agricultural University, Zhengzhou 450002, China
3. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

Abstract Doubled haploid (DH) lines are routinely applied in the hybrid maize breeding programs of many institutes and companies for their advantages of complete homozygosity and short breeding cycle length. A key issue in this approach is an efficient screening system to identify haploid kernels from the hybrid kernels crossed with the inducer. At present, haploid kernel selection is carried out manually using the"red-crown" kernel trait (the haploid kernel has a non-pigmented embryo and pigmented endosperm) controlled by the *R*1-*nj* gene. Manual selection is time-consuming and unreliable. Furthermore, the color of the kernel embryo is concealed by the pericarp. Here, we establish a novel approach for identifying maize haploid kernels based on visible (Vis) spectroscopy and support vector machine (SVM) pattern recognition technology. The diffuse transmittance spectra of individual kernels (141 haploid kernels and 141 hybrid kernels from 9 genotypes) were collected using a portable UV-Vis spectrometer and integrating sphere. The raw spectral data were preprocessed using smoothing and vector normalization methods. The desired feature wavelengths were selected based on the results of the Kolmogorov-Smirnov test. The wavelengths with p values above 0.05 were eliminated because the distributions of absorbance data in these wavelengths show no significant difference between haploid and hybrid kernels. Principal component analysis was then performed to reduce the number of variables. The SVM model was evaluated by 9-fold cross-validation. In each round, samples of one genotype were used as the testing set, while those of other genotypes were used as the training set. The mean rate of correct discrimination was 92.06%. This result demonstrates the feasibility of using Vis spectroscopy to identify haploid maize kernels. The method would help develop a rapid and accurate automated screening-system for haploid kernels.

Keywords Vis spectroscopy; Maize; Haploid kernel discrimination; Support vector machine

中图分类号: O657.3　　**文献标识码:** A　　**DOI:** 10.3964/j.issn.1000-0593(2015)11-3268-07

Introduction

During the last several years, the *in vivo* induction of maternal haploids has become an indispensable method in maize (*Zea mays* L.) research and breeding. The development of homozygous lines is a key step in maize hybrid breeding. Using the conventional method, the breeding of inbred lines usually takes 6～10 generations, which requires substantial time, manpower, and material resources[1, 2]. With

Received: 2014-08-09; **accepted**: 2014-12-16

Foundation item: the National Projects (China Agriculture Research System, CARS-02-09, 2011AA10A103)

Biography: LIU Jin, (1980—), Doctor of National Maize Improvement Center of China, China Agricultural University
GUO Ting-ting, female, (1981—), Post-Doctor of National Maize Improvement Center of China, China Agricultural University
LIU Jin and GUO Ting-ting: joint first authors　　* Corresponding author　　e-mail: chen368@126.com

haploid induction technology, the source germplasm (as female parent) is crossed with a specific inbred line called the haploid inducer (as pollinator parent), which leads to a proportion of haploid seeds from maternal ears. The chromosome set of the selected haploid seed is then doubled, to eventually produce completely homozygous doubled haploid (DH) inbred lines within 1 year. Maize haploid technology can not only acceleratethe breeding process, but also helps to simplify logistics,and thus engineering operations[2,3]. In recent years, hybrid maize industries throughout the world have successfully implemented the large-scale application of DH technology. It is worth mentioning that most seed companies and research institutes in China are also gradually beginning to adopt this approach. Due to the induction rates of modern maize inducers (only about 8% on average)[4-6], as well as the increasing application of DH technology in maize breeding, certain challenges remain to be overcome. For example, plant breeders must determine how to accurately and rapidly screen the haploid kernels from a large number of seeds obtained after induction. At present, the most efficient means of haploid selection employs the *R*1-*nj* dominant anthocyanin color marker system carried by the inducer lines, which leads to a purple embryo and endosperm in the hybrid seed, in contrast to the uncolored embryo and purple endosperm in haploid seeds after pollination with the pollen of inducers[3, 7, 8]. However, the expression of anthocyanin markers is greatly affected by the environment and the genetic background of the parents, which creates variability in the extent and intensity of pigmentation. In addition, the purple embryos are concealed by changes in pericarp thickness across various maize seed genotypes. For these reasons, the discrimination of haploid kernels using the naked eye would be considerably more labor intensive, with more false positives. The oil content of seeds is also recognized as another rapid and accurate approach for maize haploid identification[9,10]. However, the oil-content marker method requires use of a unique high-oil inducer. Moreover, to our knowledge, it is almost impossible to discriminate the haploid seeds when the female donor parents are germplasm resources with high oil content. There is thus an urgent need to develop a more efficient and rapid nondestructive identification approach for commercial application in order to save resources and to improve the efficiency of DH production.

Modern spectroscopy is a rapid, effective, and non-destructive analytical technique widely adopted in disparate fields such as agriculture and food[12, 13]. Notably, the visible (Vis) absorption spectra (400～780 nm) mainly reflects the composition and structure of the chromophores absorbing visible light[14, 15]. Vis absorption spectra serve as the physical basis of haploid kernel identification, based on differences in the embryo pigments of haploid and hybrid kernel. Numerous researchers have attempted to separate haploid seeds through the use of image recognition or computer vision software[16]. Unfortunately, anthocyanin under the pericarp cannot be characterized through the use of photography. In contrast, visible diffuse transmittance spectroscopy can be used to identify embryo pigments that cannot be seen by the human eye. Researchers have even pursued soft independent modeling of class analogy (SIMCA) modeling together with near-infrared (NIR) spectroscopy to differentiate haploid and hybrid maize kernels[17]. However, separate models will be required for haploid and hybrid kernels with different genetic backgrounds, and existing models cannot be used to distinguish kernels from the backgrounds that are not involved in the model training. This modeling method can hardly be used in the actual process of haploid identification. During the application of haploid induction, a given inducer can induce several different donors. A model that cannot identify other backgrounds will need to be re-modeled each time.

The main objective of this work was to establish a method with which to distinguish haploid from hybrid maize seeds using a single kernel and the visible spectrum. We further sought to enhance the accuracy of this identification process by collecting information on anthocyanin hidden beneath the pericarp. In this paper, we also evaluate whether the model can identify haploids from different genetic backgrounds. The findings presented here should ultimately make it possible to develop a rapid and accurate automated system with which to screen for haploid kernels. With further research, the findings presented here should help to improve screening efficiency, reduce costs and facilitate the rapid development of DH breeding so as to ultimately replace the existing manual method.

1 Materials and methods

1.1 Seed source and spectra collection

For this study, 141 haploid and 141 hybrid seeds from nine genetic backgrounds (No. D101—D109) were used. The No. D107 seeds were obtained from a hybrid crossed by inducer line CAUHOI[18]; D108 seeds were crossed by inducer line UH400[2]; the remaining seeds were produced by inbred and DH lines crossed with inducer line UH400. After spectrum acquisition, all tested kernels were planted in the field in order to verify false positives based on growth potential[19].

Diffuse transmittance spectra are measured using a QE65000 spectrometer(Ocean Optics, Inc., Dunedin, Florida, USA) with the integrating sphere accessory. The hole of the light source is covered with a piece of aluminum foil, and then a corn seed is placed through the hole in the aluminum foil, with the embryo facing the hole in the light source. The light transmitted through the kernel is collected into the integrating sphere and returned to the detector. The 400～780 nm range of the original spectra (497 sampling points) and an integration time of 60 ms were used for all measurements. Five spectra were averaged to develop the final spectra for each sample.

1.2 Data preprocessing and characteristic wavelength analysis

The use of a reasonable preprocessing method is very important in order to build an accurate and consistent model. The single-kernel spectra are affected by seed shape and size, as well as pigment content and distribution range, which results in a broader range of absorbance among various kernels at the same wavelength. For this reason, using the noise-smoothing method (smoothing window width of 17 data points), each of the spectral data points was used for vector normalization, in order to remove the vertical translation caused by uneven seed morphology.

The statistical differences in absorbance values between the two classes of samples at a given wavelength were identified by using the Kolmogorov-Smirnov test(KS test)[20]. The difference in cumulative distribution function between two groups of samples is the statistics K value for the KS test. The computation expression for the K value is as follows:

$$K = \max \mid F_1(x) - F_2(x) \mid \tag{1}$$

where $F_1(x)$ and $F_2(x)$ are the cumulative distribution functions of absorbance for two sets of samples, respectively, at the same wavelength. K is thus the maximum value of the difference between $F_1(x)$ and $F_2(x)$. If p is<0.05, then the absorbance distributions of two types of samples are considered to differ significantly at this particular wavelength. The pre-processed data were then subjected to principal components analysis (PCA), in order to reserve the minimum number of principal components, remove variable collinearity, and reduce the dimensionality of the data.

1.3 Development and validation of the model

The model for distinguishing two classes of samples was established by support vector machine (SVM)[21]. Therefore, in this study, we used the SVM method to establish a haploid and hybrid identification model based on the use of individual maize kernels. The main criterion used to evaluate the model performance was how well the model could identify haploids that had not been used for modeling.

All samples were divided into nine groups based on genetic background. During model estimation, each background sample was used as a test data set, with the remaining background samples serving as the training data set. Each SVM model was established using the same parameters, including adopting aradial basis function (RBF), γ value$=1$, and c value $= 150$. After the number of correctly discriminated kernels was recorded, uhe accuracy was computed as follows

$$\text{Accuracy} = \frac{AN + YN}{T} \times 100\% \tag{2}$$

where AN, YN and T represent the numbers of correctly discriminated haploids, correctly discriminated hybrids, and total samples in the test data set, respectively. All of the calculation programs were performed using the Matlab R2010b (Mathworks, Natick, MA, USA) platform. The SVM source program was created using the OSU SVMs Toolbox (version 3.0).

2 Results and Discussion

2.1 The spectral data analysis

The original spectra and the averaged spectra for haploid and hybrid kernels are shown in Fig. 1 and Fig. 2 below. The original spectra have 497 wavelengths; absorbance is distributed from 0.6 to 2.9. The maximum absorption peak of haploid and hybrid kernels is approximately 520 nm, and so the difference between two types of spectral absorbance is mainly reflected in the range of 510～780 nm. The point of maximum disparity occurred at 600 nm, at which point the spectral region absorption of haploid kernels was significantly lower than that of hybrid kernels. The spectral region should be the range of visible light absorption by the embryo pigment. Although studies have reported that anthocyanins absorb maximum visible spectrain the range of 500～550 nm[22], the anthocyanins in various materials exhibit a range of colors due to differences in the levels of various pigments or other components. The complete maize kernel spectra presented here may therefore differ from the absorption spectra of anthocyanins in other aqueous materials.

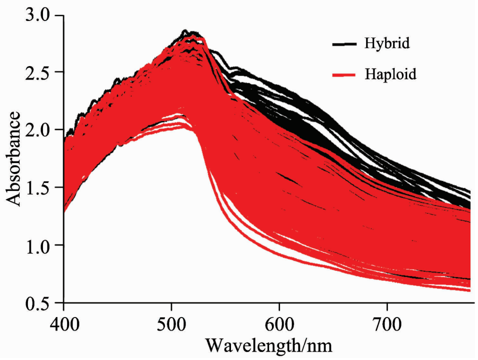

Fig. 1 The raw spectra of individual haploid kernels and hybrid kernels

For the spectra preprocessed through smoothing and vector normalization (Fig. 3), we used the KS test to identify any significant difference between the two types of sample absorbance distribution at each wavelength. With the exception of wavelengths ranging from 518～529 nm (14 wavelengths), all p-values were lower than 0.05. This finding indicates that the absorbance distribution of two samples will differ significantly at all wavelengths of the visible spectra; almost all wavelengths contain the information for two distinct samples. The data for wavelengths that yielded p-values＜0.05 were used to build the discrimination model.

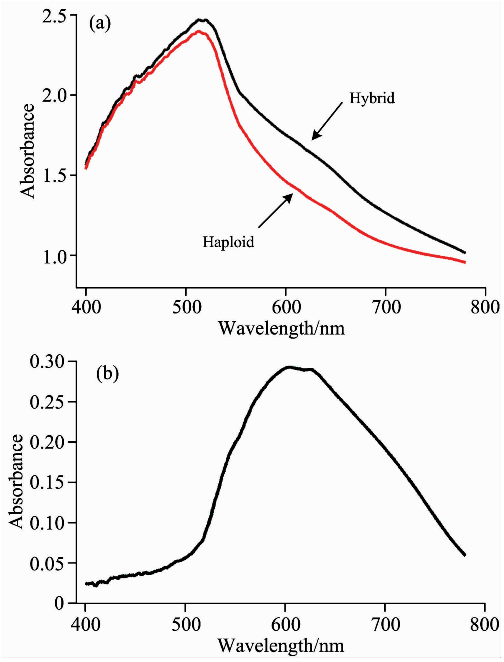

Fig. 2　The mean spectra of individual maize haploid and hybrid kernels（a）, and the differences between hybrid and haploid mean spectra（b）

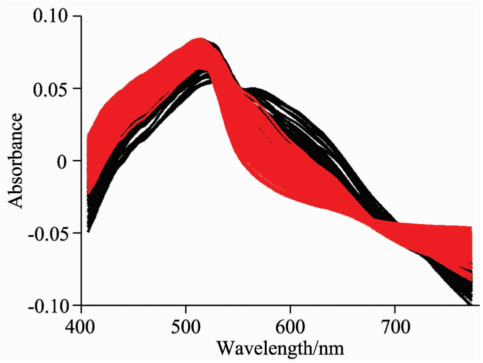

Fig. 3　The spectra were preprocessed through smoothing and vector normalization

The premise for building a cross-background identification model based on the visible spectra is the capacity to distinguishthe spectral samples of two types of kernels from the same background. The spectral differences between haploid and hybrid kernels from the same background should derivefrom differences in kernel phenotypic (including color, shape, and composition) that stem from ploidy differences in the embryonic genetic material. If two types of samples from the same background are difficult to distinguish, the difference between haploid and hybrid kernels cannot be extracted in the visible spectrum, let alone contribute to the construction of a model that can identify samples from different backgrounds.

2.2　The model establishment and evaluation

We first analyzed the principal component of samples from each background, to determine whether two types of spectral samples from the same background could be separated in the principal component space. Figure 4 is a two-dimensional distribution diagram for two types of samples, in the first and second main components. As shown, the principal component space of each background features a hyperplane (e.g. straight line in two-dimensional space) that divides the entire sample space into two sub-spaces, so that the two types of samplesfall into independent sub-spaces. This result shows that the visible spectra can separate two types of kernels fromthe same background. However, if the genetic background of the material is different, the expression of pigment genes will differ, and kernel traits (such as color and shape) from different backgrounds will affect the visible spectra. A cross-background model can be built only if there are common differences between haploid and hybrid kernels of different backgrounds. PCA was performed after mixing samples from nine genetic backgrounds. Figure 5 constituted by the first two principal components shows two distinct types of samples, which demonstrates that differences captured by the visible spectra are applicable to a variety of background material. Furthermore, spectral differences derived from the embryo ploidy should be greater in magnitude than the spectral differences caused by unrelated genetic backgrounds. Otherwise, the sample will cluster in accordance with its background. There is a certain degree of aliasing between two types of samples, which requires the use of an appropriate pattern recognition method to build a more accurate discrimination model. In order to avoid over-fitting, the model must also carry outcross-validation.

The haploid identification model based on the SVM method was tested using cross-validation, and the results are shown in Table 1. For this study, 141 haploid and141 hybrid kernels, respectively, from D101—D109 were used. For cross-validation, the number of haploid kernels with correct recognition was 133; the number of hybrid kernels with correct recognition was 127. The average Accuracy for nine cycles of cross-validationwas 92.06%. Eight tests had Accuracy over 85%, which is more than six times Accuracy at 90%. Accuracy was the lowest when using the D106 test set, as 50% of hybrid kernels were mislabeled as haploid kernels. Characteristic information of the genotype was mixed in with themodeling spectral information, which suggests that the optimum classification hyperplane modeled by other samples does not apply to D106.

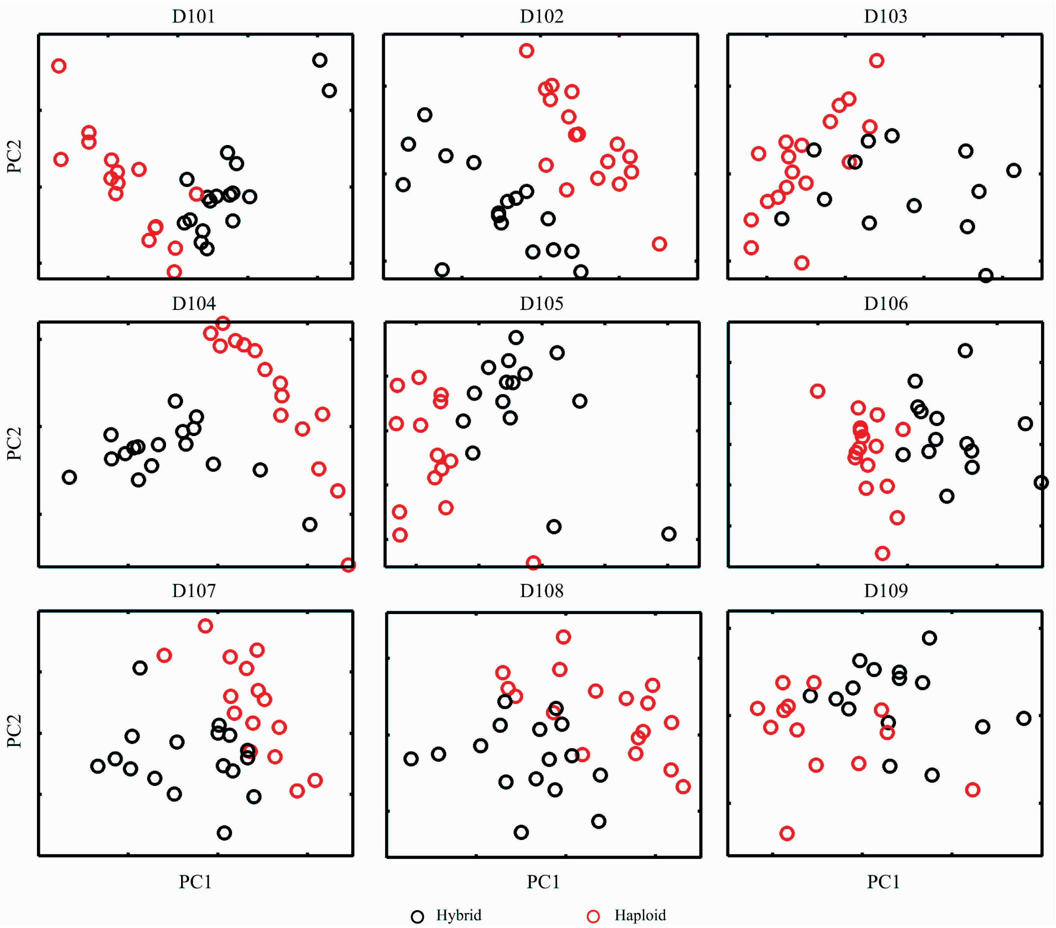

Fig. 4 PCA plots of spectra for separating individual haploid and hybrid kernels from D101—D109

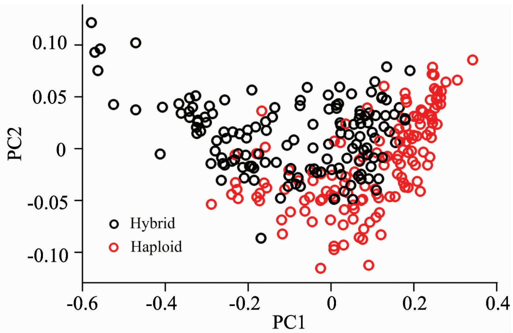

Fig. 5 The PCA plot for separating haploid and hybrid spectra from all genotypes

Current haploid selection mainly depends on seed embryo pigment. The main reasons for misrecognition are as follows: (1) the pigment is difficult to observe with the naked eye, as it is concealed deep within the embryo; (2) pigment gene expression is weak because of the influence of the environment, which results in lower pigment levels that cannot easily be detected by the human eye; and (3) the dominant pigment inhibitors carried by the female parent can suppress expression.

Table 1 Evaluation results of the SVM model

Name	Number of kernels		Number of correctly discriminated kernels		Accuracy /%
	Hybrid	Haploid	Hybrid	Haploid	
D101	17	16	17	16	100.00
D102	17	17	17	15	94.12
D103	13	18	12	18	96.77
D104	17	15	14	14	87.50
D105	15	14	15	14	100.00
D106	14	16	7	16	76.67
D107	17	15	16	15	96.88
D108	16	17	15	15	90.91
D109	15	13	14	10	85.71
Total	141	141	127	133	Avg. 92.06

This research attempts to solve the first two problems through the use of a novel, rapid, and nondestructive methodology. Compared with image recognition[16] and NIR spectroscopy[17], Vis spectroscopy was considered as most suitable for this endeavor. Automated imaging cannot collect anthocyanin information hidden under the seed coat, nor can it improve the accuracy rate of haploid screening. NIR spectra-can provide information on seed structure and contents, including starch, oil, and protein, which are far more abundant than embryo pigment. In addition, the contents of seeds from various genetic backgrounds are so different that extracting pigmentation differences and cross-background modeling can be quite challenging. Jones et al. (2012) built a model of individual background that examines haploids using NIR spectroscopy. Cross-background model building requires the proper chemometric methodology in order to eliminate the interference created by background variance. In contrast, pigment characteristics as captured by the Vis spectraare relatively simple; the absorbance variance of two types of samples can be observed in the original spectra presented in Fig. 1 and Fig. 2. Variance between two types of samples can be shown directly using PCA. With SVM model, the average Accuracy for cross-validation can reach as high as 90%. This study explored the viability of using the visible transmittance spectra to identify small samples of haploid maize kernels. Future research should attempt to increase sample quantity and range, thereby improving the spectral feature selection method.

3 Conclusion

Here, we have presentedan identification method to screen single haploid corn kernels using Vis diffuse transmittance spectra. In this study, up to 282 hybrid and haploid kernels from nine genetic backgrounds were used as experimental material. The Vis diffuse transmittance spectra of single corn kernels were recorded using a portable fiber Uv/Vis spectrometer. The original spectral data were preprocessed by smoothing and vector normalization. The feature wavelengths were selected by KS test. The discrimination models were built based on SVM method with the data through PCA. In each time of validation, one of genotypes seeds were used as the test set, and the remaining genotypes samples were reserved as the modeling set for cross-validation. The average Accuracy for modeling cross-validation reached as high as 92.06%. This result demonstrates the feasibility of using Vis spectroscopy to identify haploid maizekernels. An automatic rapid-screening system for haploid maize seeds based on this method will probably be established to meet large-scale DH breeding demands.

Acknowledgements

We also greatly thank Pro. Albrecht E. Melchingerat University of Hohenheim for contributing to inducer line UH400.

References

[1] Chase S, Nanda D. Economic Botany, 1969, 23(2): 165.
[2] Prigge V, Xu X, Li L, et al. Genetics, 2012, 190(2): 781.
[3] Geiger H, Gordillo G. Maydica, 2009, 54(4): 485.
[4] Röber F, Gordillo G, Geiger H. Maydica, 2005, 50(3/4): 275.
[5] Prigge V, Sanchez C, Dhillon B S, et al. Crop Science, 2011, 51(4): 1498.
[6] Zhang Z, Qiu F, Liu Y, et al. Plant Cell Reports, 2008, 27(12): 1851.
[7] Eder J, Chalyk S. Theoretical and Applied Genetics, 2002, 104(4): 703.
[8] Choe E, Carbonero C H, Mulvaney K, et al. Plant Breeding, 2012, 131(3): 399.
[9] CHEN Shao-jiang, SONG Tong-ming. Acta Agronomica Sinica, 2003, 29(4): 587.
[10] LIU Jin, GUO Ting-ting, YANG Pei-qiang, et al. Transactions of the Chinese Society of Agricultural Engineering, 2012, (s2): 233.
[11] Melchinger A E, Schipprack W, Würschum T, et al. Scientific Reports, 2013, 3: 2129.
[12] GUO Ting-ting, XU Li, LIU Jin, et al. Spectroscopy and Spectral Analysis, 2013, 33(6): 1501.
[13] Christenson B S, Schapaugh W T, An N, et al. Crop Science, 2014, 54(4): 1585.
[14] Christie R M. Colour Chemistry. Great Britain: The Royal Society of Chemistry, 2001.
[15] Jacquemin D, Perpete E A, Scuseria G E, et al. Journal of Chemical Theory and Computation, 2007, 4(1): 123.
[16] ZHANG Jun-xiong, WU Zhan-yuan, SONG Peng, et al. Transactions of the Chinese Society of Agricultural Engineering, 2013, 29(4): 199.
[17] Jones R W, Reinot T, Frei U K, et al. Applied Spectroscopy, 2012, 66(4): 447.
[18] Li L, Xu X, Jin W, et al. Planta, 2009, 230(2): 367.
[19] Auger D L, Ream T S, Birchler J A. Theoretical and Applied Genetics, 2004, 108(6): 1017.
[20] Smirnov N. The Annals of Mathematical Statistics, 1948, 19(2): 279.
[21] Cortes C, Vapnik V. Machine Learning, 1995, 20(3): 273.
[22] Harborne J B. The Biochemical Journal, 1958, 70(1): 22.

基于可见光光谱高效鉴别玉米单倍体籽粒

刘　金[1]，郭婷婷[1]，李浩川[2]，贾仕强[3]，严衍禄[3]，安　冬[3]，张　垚[1]，陈绍江[1*]

1. 中国农业大学农学与生物技术学院，国家玉米改良中心，北京　100193
2. 河南农业大学农学院，河南 郑州　450002
3. 中国农业大学信息与电气工程学院，北京　100083

摘　要　单倍体技术已发展成为玉米遗传研究及现代玉米育种的重要技术之一，单倍体籽粒的鉴别筛选是其中的重要环节。目前单倍体籽粒主要是依赖于籽粒的 $R1\text{-}nj$ 遗传标记通过人工肉眼观察颜色的有或无进行鉴别，费时费工。而且部分材料由于标记颜色很难从籽粒外部观察到，导致人工筛选准确率较低。基于可见光光谱分析建立玉米单倍体籽粒鉴别方法，探索利用可见光光谱鉴别玉米单倍体籽粒的可行性。同时，由于每季用于诱导单倍体的育种材料不尽相同，模型须能够鉴别未参加建模的材料的单倍体。本研究以 9 个遗传背景的单倍体和杂交籽粒共 284 粒作为试验材料，利用便携式紫外-可见光光纤光谱仪采集单个玉米籽粒的可见光漫透射光谱。光谱数据经平滑、矢量归一化预处理和主成分分析，基于支持向量机方法建立单倍体和杂交籽粒判别模型。每次选择 1 个背景的样本作为测试集，其余背景的样本作为建模集对模型进行交叉验证。模型交叉验证平均正确判别率达到 92.06%。其中 8 次测试正确判别率在 85%以上。结果表明利用可见光光谱分析建立玉米单倍体籽粒鉴别方法，并使模型可鉴别未参与建模材料的单倍体具有可行性。并且基于该方法有望建立玉米单倍体籽粒的自动化快速筛选系统，提高玉米单倍体育种效率。

关键词　可见光光谱；玉米；单倍体鉴别；模式识别

（收稿日期：2014-08-09，修订日期：2014-12-16）

刘　金，郭婷婷：并列第一作者

*通讯联系人