Received: 27 September 2019 | Accepted: 20 July 2020

DOI: 10.1111/pbr.12857

ORIGINAL ARTICLE

Plant Breeding WILEY

Classification approaches for sorting maize (*Zea mays* subsp. *mays*) haploids using single-kernel near-infrared spectroscopy

Jeffery L. Gustin[1] | Ursula K. Frei[2] | John Baier[1] | Paul Armstrong[3] | Thomas Lübberstedt[2] | A. Mark Settles[1]

[1]Department of Horticultural Sciences, University of Florida, Gainesville, FL, USA

[2]Department of Agronomy, Iowa State University, Ames, IA, USA

[3]Grain Marketing Production Research Center, USDA-ARS, Manhattan, KS, USA

Correspondence
Jeffery L. Gustin, Department of Horticultural Sciences, University of Florida, Gainesville, FL, USA.
Email: jgustin@ufl.edu

Funding information
This work was supported by the National Science Foundation Plant Genome Research Program (NSF-PGRP) award 1444456, a United States Department of Agriculture cooperative agreement, the National Institute of Food and Agriculture Small Crops Research Initiative (NIFA-SCRI) award 2018-51181-28419 and the Vasil-Monsanto Endowment.

Communicated by: Wolfgang Friedt

Abstract

Doubled haploids (DHs) are an important breeding tool for creating maize inbred lines. One bottleneck in the DH process is the manual separation of haploids from among the much larger pool of hybrid siblings in a haploid induction cross. Here, we demonstrate the ability of single-kernel near-infrared reflectance spectroscopy (skNIR) to identify haploid kernels. The skNIR is a high-throughput device that acquires an NIR spectrum to predict individual kernel traits. We collected skNIR data from haploid and hybrid kernels in 15 haploid induction crosses and found significant differences in multiple traits such as percent oil, seed weight, or volume, within each cross. The two kernel classes were separated by their NIR profile using Partial Least Squares Linear Discriminant Analysis (PLS-LDA). A general classification model, in which all induction crosses were used in the discrimination model, and a specific model, in which only kernels within a specific induction cross, were compared. Specific models outperformed the general model and were able to enrich a haploid selection pool to above 50% haploids. Applications for the instrument are discussed.

KEYWORDS

doubled haploid, germplasm enhancement of maize, haploid classification, partial least squares regression, *R1-nj*, single-kernel near-infrared spectroscopy

1 | INTRODUCTION

Doubled haploids (DHs) have been a valuable tool in maize breeding for decades (Geiger & Gordillo, 2009; Liu et al., 2016; Röber, Gordillo, & Geiger, 2005; Smith et al., 2008). DHs can generate fully homozygous inbred lines in just two generations, while traditional breeding methods require six or more generations of self-pollinations to achieve nearly homozygous inbred lines. DHs can also reduce the breeding cycles needed to introgress new genetic loci into elite germplasm (Lübberstedt & Frei, 2012). The time and resource savings obtained from using DH technology allows breeders to evaluate substantially more inbred lines for hybrid performance than would be practical using traditional breeding practices.

In maize, in vivo haploids are produced by pollinating a female donor line contributing the haploid genome with a male haploid inducer line. A haploid kernel is produced in this cross when the female egg cell is stimulated to develop without proper fertilization by the male pollen grain (Li et al., 2017; Zhao, Xu, Xie, Chen, & Jin, 2013). Haploid kernels, therefore, possess a haploid embryo containing the maternal genome and a fertilized triploid endosperm. The frequency of haploid kernels on typical maize ears is less than 0.1% (Chase, 1949). In 1959, Coe identified the maize line, Stock 6, which produces haploids at a rate 3% of progeny (Coe, 1959). Modern haploid inducers have been selected for haploid induction frequencies up to 15% (Cai et al., 2007; Liu et al., 2016; Rotarenco, Dicu, State, & Fuia, 2010). Despite the 5-fold increase in induction frequency, haploids remain a small fraction of total kernels in an induction cross. Consequently, DH production efficiency is limited, in part, by the techniques used to identify haploid kernels among a larger pool of hybrid siblings.

Haploid kernels per se are not visually distinguishable from hybrid siblings. To discern the two classes, dominant colour markers are used to mark the paternal genome in hybrid embryos (Geiger & Gordillo, 2009; Liu et al., 2016). The *R1-nj* allele causes purple anthocyanin accumulation in both the embryo and the endosperm aleurone. When an *R1-nj* inducer is crossed to a colourless female donor, haploids kernels can be sorted from diploids visually based on the lack of anthocyanin in the embryo. Unfortunately, kernel pigmentation is suppressed in some genetic backgrounds. The *C1-I* allele is a dominant suppressor of anthocyanin accumulation in kernels, and *C1-I* is common in tropical and Flint germplasm (Chaikam et al., 2015; Rotarenco, Kirtoca, & Jocota, 2007). In seedlings, the dominant "red root" *Pl1* allele identifies haploids based on absence of root coloration (Chaikam, Martinez, Melchinger, Schipprack, & Boddupalli, 2016; Rotarenco et al., 2010). The *R1-nj* and *Pl1* visual markers require that each kernel or seedling be scored by trained personnel and, in the case of the red root marker, require transplanting haploid seedlings soon after scoring. Manually sorting tens of thousands of induction cross progeny is a bottleneck for larger-scale DH operations (Geiger & Gordillo, 2009).

Automated systems are an attractive alternative to manual sorting. Automated systems differentiate hybrid and haploid kernels based on physical or chemical differences between the kernel genotypes. Pilot studies on automated colour sorting based on expression of *R1-nj* in the embryo scutellum showed feasibility to identify more than 80% of haploids (Boote et al., 2016; De La Fuente, Carstensen, Edberg, & Lübberstedt, 2017; Song et al., 2018). However, the effectiveness of colour sorting on a broad range of induction crosses with variable expression of *R1-nj* as well as testing randomly oriented kernels is needed to determine accuracy in a practical context. There are also physical and compositional differences between haploid and hybrid kernels. Haploid kernels have reduced weight and relative oil content as compared to hybrid siblings (Rotarenco et al., 2007; Smelser et al., 2015). These differences allow alternative approaches for sorting haploids that do not rely on genetic colour markers (Chen & Song, 2003; Melchinger, Schipprack, Friedrich Utz, & Mirdita, 2014; Melchinger, Schipprack, Mi, & Mirdita, 2015; Melchinger, Schipprack, Würschum, Chen, & Technow, 2013; Smelser et al., 2015).

Near-infrared reflectance (NIR) spectroscopy is an established method for rapid, non-destructive determination of the chemical composition of grains (Osborne, 2006, Gustin & Settles, 2015). Near infrared light, consisting of wavelengths between 700 nm and 2,500 nm, is absorbed by water and organic chemical bonds within the grain and penetrates further into the sample than the more strongly absorbed mid and far infrared light (Lodder, 2002). The NIR absorbance profile of a sample can be converted to estimates of chemical or physical characteristics of the kernel such as protein content, oil content, grain weight, or density using chemometrics approaches.

Jones et al. (2012) demonstrated that single-kernel near-infrared transmittance spectroscopy could discriminate haploids from hybrids. Although haploid classification was accurate, the transmittance spectra were acquired over 1 min as the kernel was vibrated into multiple positions. Additional studies have verified that near-infrared transmittance can classify maize haploids with long integration times (Lin, Yu, Li, & Qin, 2017). The long data acquisition time required for accurate transmittance spectra is not well suited for high-throughput processing of single kernels (Baye, Pearson, & Settles, 2006).

Armstrong (2006) developed a rapid single-kernel NIR (skNIR) device that acquires an NIR spectrum with a 20 millisecond integration time from an individual kernel as it tumbles down an illuminated light tube (Tallada, Palacios-Rojas, & Armstrong, 2009). The skNIR device can predict single-kernel traits including oil, protein, density, weight and volume (Armstrong & Tallada, 2012; Gustin et al., 2013; Spielbauer et al., 2009). The skNIR device has also been shown to predict composition traits for other large seed crops such as soybean and common bean (Hacisalihoglu et al., 2016; Hacisalihoglu, Larbi, & Settles, 2010). The objective of this study was to test the accuracy of this rapid skNIR device for sorting haploid kernels.

2 | MATERIALS AND METHODS

2.1 | Kernel samples

Haploid and diploid kernels from 15 induction crosses derived from the Germplasm Enhancement of Maize (GEM) project were used to generate the data for the study (http://www.public.iastate.edu/~usda-gem/GEM_Project/GEM_Project.htm) (Brenner, Blanco, Gardner, & Lübberstedt, 2012). The female donor parents were backcross generation 3 (BC3) plants from introgressions of BGEM inbred lines into the expired PVP lines, PHB47 and PHZ51, as recurrent parents. BGEM inbred lines contain introgressions of various tropical accessions into PHB47 and PHZ51 (Brenner et al., 2012; Sanchez, Liu, Ibrahim, Blanco, & Lübberstedt, 2018; Smelser, Gardner, Blanco, Lübberstedt, & Frei, 2016; Vanous et al., 2018; Vanous et al., 2019). PHB47 is from the stiff stalk (SS) heterotic group, while PHZ51 is a non-stiff stalk (NSS). BHI201 was the haploid inducer. It has the *R1-nj* and *Pl1* colour marker genes and 12%–14% induction rate (Liu et al., 2016).

2.2 | Single-kernel NIR spectra and determination of kernel traits

From each of the 15 induction crosses, 48 diploid and 48 haploid kernels were visually identified and arrayed in 48-well microtitre plates. Kernel weights and skNIR spectra were collected from each kernel using the skNIR platform described by Armstrong & Tallada (2012) and Spielbauer et al. (2009). NIR reflectance values were recorded at 1 nm intervals between 907 and 1,689 nm and absorbance values were calculated as log(1/R). Each spectrum was centred to an arbitrary mean of 1. Two weights and two spectra were recorded from each kernel. Seven kernel composition and quality traits were determined for each kernel from the averaged NIR spectra using previously derived PLS regression coefficients (Gustin et al., 2013; Spielbauer et al., 2009). The traits were relative oil, protein and starch content, measured on a fresh weight basis, as well as total and material density, and total and material volume. Total density and volume include air space, whereas material density and volume do not. Single-kernel weight was measured with an in-line microbalance on the skNIR platform.

2.3 | Validating ploidy

After skNIR data collection, the kernels were planted and seedling leaf tissue was sampled for DNA extraction. Seedlings that could not be genotyped were not included in the analysis. The ploidy of each DNA sample was evaluated by amplifying two insertion/deletion (INDEL) markers that were polymorphic between the inducer line and the PHB47 and PHZ51 recurrent female donor lines. In most cases, haploid kernels contained only the recurrent donor line allele. In some induction crosses, the allele from the exotic non-recurrent parent of the female donor line was polymorphic with the recurrent parent. In such cases, a third marker was used to determine ploidy. In all, 1,354 kernels were confirmed as either haploid or diploid.

2.4 | General model

Partial least squares linear discriminant analysis (PLS-LDA) was used to construct models to separate haploid and hybrid kernels based on skNIR data. PLS models extract relevant latent variables from highly dimensional, highly covariate data such as spectral data and are standard for NIR spectral calibration models (Frank & Friedman, 1993; Wold, Sjostrom, & Eriksson, 2001). PLS-LDA models are divided into the PLS and LDA operations. PLS identifies latent factors that best explain variance between the haploid and hybrid classes. LDA calculates a discriminant vector from the PLS factors that best separates the classes. PLS-LDA was implemented in the R package "plsgenomics" (Boulesteix, Durif, Lambert-Lacroix, Peyre, & Strimmer, 2015). Prior to PLS-LDA model construction, the sample set was adjusted to reflect a haploid induction frequency of 11% by randomly subsampling the haploid class to fit a 1:8 ratio of haploids:hybrids. The models were calibrated on 14 of 15 induction crosses, while the 15th induction cross was held out for external validation. Leave-one-out validation was repeated 15 times so that all induction crosses were held out. Calibration and validation steps were repeated 100 times using random subsampling of haploid kernels to ensure that all haploid kernels were included in the external validation set.

Kernel ploidy (haploid vs. hybrid) was the dependent variable. Predictor variables were either mean-centred spectra or eight kernel composition and quality traits. No additional pretreatments beyond mean centring were applied to the spectra. Predictor variables were the average of two technical replicates of skNIR spectra for each kernel; whereas validation was tested with a single, randomly selected spectrum. This approach reflected a likely usage case where a sorting instrument could be calibrated using kernels with multiple skNIR technical replicates, while classification and sorting would rely on a single spectrum per kernel.

The PLS models were fit using the SIMPLS algorithm. PLS factors were chosen based on minimization of the standard error of prediction (SEP) for the cross-validation set. Class assignments for the validation samples were made using the [predict.lda] function with prior probabilities set to 0.11 and 0.89 for haploid and hybrid classes, respectively, to reflect expected proportions within the sample set. Accuracy was evaluated with the False Discovery Rate (FDR), which was the fraction of predicted haploids that were actually hybrid kernels, and the False Negative Rate (FNR), which was the fraction of actual haploid kernels predicted to be hybrid. Reported results are based on validation datasets.

2.5 | Induction cross-specific models

Independent PLS-LDA models were calibrated for each induction cross population by subsampling 50% of the hybrid and haploid kernels in a 1:1 ratio of haploid:hybrid genotypes. External validation kernels were randomly drawn from the remaining kernels in a 1:8 ratio of haploids:hybrids. Calibration and validation were repeated 100 times using random subsampling to ensure that most haploid kernels were evaluated in the validation set. PLS-LDA models were fit and evaluated using the same approach as general models.

2.6 | Statistical analysis

MANOVA and ANOVA analyses were conducted using "manova" and "aov" functions, respectively, in the R statistical package "stats" (Team, 2017). To address MANOVA assumptions, outlier data were removed, by identifying values within groups that were greater than three standard deviations from the group mean. On average, less than one outlier was removed per group. Group distribution normality was evaluated using the Shapiro–Wilk test in the package "stats", and homogeneity of variance/covariance matrices among groups was tested using Box's M-test in the package "biotools" (da Silva, Malafaia, & Menezes, 2017; Team, 2017). Twenty of 270 groups (15 induction crosses × 9 kernel composition traits × 2 ploidy classes) had a Shapiro-Wilk test *p*-value less than .01 indicating that they were not normally distributed and 3 of 15 induction populations had heterogeneous covariance matrices. These populations were BGEM-0014-S × inducer, BGEM-0112-S × inducer and BGEM-0071-S × inducer. While limited, these violations of MANOVA assumptions suggest that the MANOVA and ANOVA results for these inducer populations may be inaccurate.

3 | RESULTS

3.1 | NIR spectral variation between haploid and hybrid kernels

NIR-based discrimination of haploid from hybrid kernels requires chemical and/or structural differences between the two classes resulting in predictable variation within the NIR spectra. Figure 1 shows mean NIR spectra of haploid and hybrid kernels. Differences between these means are slight when all induction crosses are combined (Figure 1a). Within specific induction crosses, spectral profiles can differ more substantially. Figure 1b shows mean haploid and hybrid spectra from the BGEM-0055-S × inducer population. The ends of the spectral profiles are largely non-overlapping.

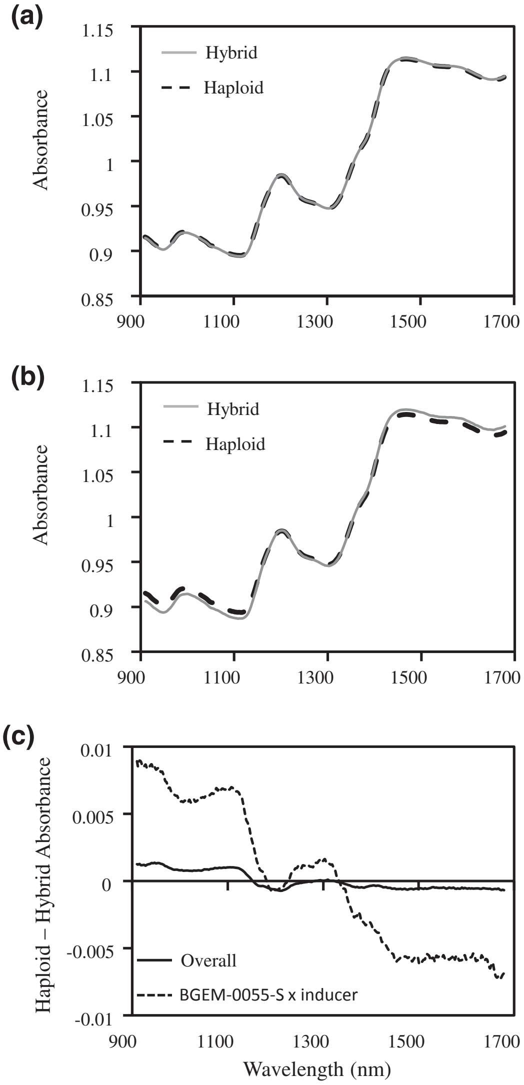

FIGURE 1 Differences between haploid and hybrid NIR spectra. (a) Overall mean absorbance spectra of haploid and hybrid kernels. (b) Mean absorbance spectra of haploid and hybrid kernels from BGEM-0055-S x inducer induction population. (c) Difference between the absorbance values of the mean haploid and hybrid spectra from all kernels (overall) and from BGEM-0055-S × inducer cross. Absorbance is reported as (log(1/reflectance)

A difference plot between haploid and hybrid wavelength values shows the distinctions more clearly (Figure 1c). The difference between the overall means ranges from approximately −0.001 to 0.001, while the BGEM-0055-S × inducer population ranges from approximately −0.008 to 0.008, an 8-fold increase in spectral separation. These data suggest that spectral features distinguishing haploid kernels from hybrid are not generally conserved among all induction crosses.

3.2 | Compositional and quality variation in GEM induction crosses

Kernel oil content and weight are known to be reduced in haploids (Melchinger et al., 2013; Rotarenco et al., 2007; Smelser et al., 2015). However, additional signatures may exist that could help facilitate NIR-based discrimination. The induction crosses were evaluated for eight kernel composition and quality traits and compared by ploidy and by the GEM recurrent parent (Table 1). The largest difference between haploid and hybrid kernels was relative oil content. Haploid oil content was reduced 0.88% on average, which represents a 27% relative reduction compared to hybrid kernels.

Weight and volume were also reduced in haploid kernels, but by a relatively smaller amount than kernel oil. Kernel weight was reduced by 18 mg, on average, in haploids. Unlike oil content, not all induction crosses produced lighter haploids. Only, 7 of the 15 crosses had significant differences in kernel weight between ploidy classes (Table 2). Eight of the 15 inductions crosses had a significant difference for total volume, which includes air spaces within the kernel, or material volume, which excludes air space. Haploids in one cross, BGEM-0110-N x inducer, had a larger predicted total kernel volume without a significant change in material volume, suggesting increased air space in the haploid kernels of this specific donor by inducer combination. Increased air space in the kernel has been shown to be associated with haploids in certain backgrounds due to malformed or missing embryos (Irani, Knapp, Lubberstedt, Frei, & Askari, 2016).

Relative protein content was increased in haploids in both donor backgrounds and was significant in five crosses. Starch and density varied between GEM backgrounds, but at least one of these traits was significant in 9 of the 15 crosses. Reduced material density combined with increased protein content are rough indicators of haploid kernels. While these general observations show that kernel oil content was the best marker for haploid status among the traits that were analysed, haploids in individual induction crosses had pronounced differences in traits other than oil (Table S1). The data show that the female donor genotype substantially influenced the size, chemical and quality differences between haploid and hybrid kernels, in agreement with previous observations (Melchinger et al., 2014; Rotarenco et al., 2007; Smelser et al., 2015).

To test if the cross-specific and haploid versus hybrid trait differences were significant, an interaction term between ploidy and induction cross was included in MANOVA and two-way ANOVA. All three terms were highly significant when all kernel traits were evaluated

TABLE 1 Kernel composition and quality trait measurements of haploid and hybrid kernels grouped by recurrent female parent/heterotic group and haploid inducer line

		PHZ51/NSS		PHB47/SS		
		Haploid	**Hybrid**	**Haploid**	**Hybrid**	**Inducer**
Weight (mg)	Mean	345	364	349	364	168
	SD	50.1	47.9	46.0	48.2	23.0
	Range[a]	293–370	314–379	298–377	325–401	113-214
Oil (%)	Mean	2.37	3.27	2.55	3.47	3.23
	SD	0.7	0.82	0.659	0.765	0.713
	Range	2.33–2.95	2.73–3.73	1.85–3.42	3.14–4.14	1.00-4.59
Protein (%)	Mean	10.9	10.7	13.3	13.0	16.2
	SD	1.38	1.42	1.32	1.55	1.18
	Range	9.99–11.2	10.3–11.2	11.7–14.3	11.2–14.5	13.5–20.3
Starch (%)	Mean	63	62.1	56.3	56.5	53.7
	SD	3.76	3.87	3.59	3.62	2.82
	Range	59.5–65.1	58.9–66.0	52.9–60.6	53.6–59.1	47.7-62.9
Density (t) (g/cm^3)	Mean	1.57	1.61	1.57	1.58	1.43
	SD	0.06	0.07	0.080	0.07	0.05
	Range	1.56–1.60	1.57–1.70	1.45–1.62	1.52–1.62	1.29-1.59
Density (m) (g/cm^3)	Mean	1.54	1.53	1.57	1.57	1.57
	SD	0.03	0.03	0.04	0.04	0.03
	Range	1.52–1.56	1.51–1.55	1.54–1.60	1.53–1.59	1.44-1.63
Volume (t) (cm^3)	Mean	262	264	245	257	161
	SD	35.0	37.1	29.7	31	28.7
	Range	236–292	239–297	211–255	234–278	73.8-218
Volume (m) (cm^3)	Mean	242	250	222	231	123
	SD	31.2	28.1	28.5	26.7	22.6
	Range	220–259	228–257	193–235	216–254	57.4-182

[a]Range shows minimum and maximum values of induction cross population means and the minimum and maximum kernel values for the inducer.

simultaneously in the MANOVA model (Table S2). Evaluation of each trait using two-way ANOVA found that the cross by ploidy interaction term was significant for all traits except for material volume and percent starch (Table S3). The ploidy term was significant ($p < .05$) for kernel oil, weight, protein, total density and total and material volume. These analyses illustrate that the female donor genotype significantly influences the effect of a haploid embryo on mature kernel traits. While the relative effect size was small for all traits except oil, the MANOVA suggests that other traits predicted by skNIR can be distinguishing features between hybrid and haploid kernels. The additional NIR signatures could be leveraged to improve discrimination accuracy beyond what can be achieved by kernel oil content per se.

3.3 | Haploid discrimination using general skNIR models

Two PLS-LDA models were evaluated to determine the accuracy of haploid and hybrid classification for all crosses combined. The first model used the 770 wavelengths of the NIR absorbance spectra as predictors, while the second model used eight kernel composition and quality traits predicted by skNIR. Predicted kernel composition traits yielded lower FDR with slightly higher FNR (Table 3). However, the error rates for both general models were too high to be effective methods for sorting haploid kernels. By comparison, visual discrimination using the *R1-nj* was far more accurate than either general models.

Table 4 shows the general model accuracy within each induction cross. The data represent 10 iterations of subsampling the induction crosses with a 1:8 haploid:hybrid ratio. The general skNIR model classified some induction crosses with higher accuracy than others. For example, BGEM-0071-S × inducer was the most accurately classified induction cross with FDR and FNR of 0.59 and 0.53 respectively. There were no induction crosses in which the PLS-LDA model rivalled visual accuracy. There was also no significant correlation between the visual FDR and skNIR FDR ($r = .45$, $p = .09$, $df = 13$), indicating that visual classification accuracy was not a predictor of PLS-LDA model accuracy.

TABLE 2 Kernel trait differences between mean hybrid and mean haploid values for kernels from each induction cross

Induction cross	Recurrent parent	Weight (mg)	Oil (%)	Protein (%)	Starch (%)	Total density (g/cm^3)	Material density (g/cm^3)	Total volume (cm^3)	Total volume (cm^3)
BGEM-0000-S × inducer	PHB47	8	0.9[a]	0.1	0.4	−0.01	−0.01	4	2
BGEM-0014-S × inducer		5	0.5[a]	0.4	−2.0[a]	0.04[a]	0.03[a]	−6	6
BGEM-0029-S × inducer		3	1.7[a]	−1.9[a]	0.6	0.01	−0.02[a]	18[a]	15[a]
BGEM-0112-S × inducer		−13	1.1[a]	−0.7	1.7[a]	0.03	0.00	6	6
BGEM-0055-S × inducer		48[a]	1.0[a]	1.3[a]	0.0	−0.02	0.01	28[a]	15
BGEM-0071-S × inducer		47[a]	0.8[a]	0.6	0.2	0.02	0.01	25[a]	21[a]
BGEM-0095-S × inducer		1	0.9[a]	0.0	−1.8[a]	0.01	−0.01	8	11[a]
BGEM-0168-S × inducer		30[a]	0.8[a]	−0.3	−0.2	−0.01	−0.01	21[a]	18[a]
BGEM-0172-S × inducer		30[a]	1.0[a]	−0.6[a]	0.8	0.02	−0.01	13[a]	16[a]
BGEM-0213-S × inducer		25	0.4[a]	−0.8[a]	2.0[a]	−0.04[a]	0.00	11	1
BGEM-0046-N × inducer	PHZ51	14	0.7[a]	−0.1	0.9	0.01	−0.01	6	2
BGEM-0110-N × inducer		24[a]	1.8[a]	−0.5	−6.5[a]	0.04[a]	0.01	−39[a]	10
BGEM-0181-N × inducer		38[a]	0.7[a]	0.4	0.2	0.02	0.00	21[a]	19[a]
BGEM-0207-N × inducer		11	0.9[a]	0.0	−1.3	0.01	−0.02[a]	3	6
BGEM-0231-N × inducer		5	0.4[a]	−0.8[a]	1.9[a]	0.00	−0.02[a]	14	0

[a] Student's *t* test *p* value less than the Benjamini–Hochberg critical value for False Discovery Rate of 0.05.

TABLE 3 General model haploid discrimination accuracies for single genotypic holdout external validation kernels

Models	Predictors	Factors	FDR[a]	FNR[b]
PLS-LDA	NIR	9	0.70	0.73
PLS-LDA	Kernel Composition	8	0.36	0.89
Visual	Kernel Colour	–	0.17	0.17

[a] False Discovery Rate.

[b] False Negative Rate.

3.4 | Haploid discrimination with induction cross-specific models

The cross-specific variation in NIR spectral differences between haploids and hybrid kernels suggests that higher classification accuracy could be achieved using specific classification models for each induction cross. To test this, induction cross-specific PLS-LDA models using either the spectra or the kernel composition traits as predictors were constructed. In addition, an oil-only approach was evaluated whereby 20% of kernels with the lowest predicted oil content were classified as haploid. The 20% oil threshold matches the fraction of kernels predicted to be haploid from induction cross-specific PLS-LDA models using spectra as the predictors. Comparisons between the oil-only approach and the other models can evaluate if adding traits in addition to oil improves the predictive power.

The PLS-LDA models using NIR predictors performed the best with average FDR and FNR of 0.54 and 0.16, respectively (Table 5). These results suggest NIR spectra have discriminatory signal that is not captured by the latent vectors used to predict kernel traits. The FNR using NIR predictors was substantially lower than the other two approaches and was even lower than the FNR for visual classification, suggesting a larger proportion of true haploids can be recovered. Nine of the fifteen induction crosses had an FDR of less than 0.50, meaning that over 50% of the kernels classified as haploid were true haploids (Table 4). Each of these nine crosses also had an FNR less than 0.23, meaning that over 75% of true haploids were classified correctly. The BGEM-0110-N x inducer cross had near perfect separation of haploid kernels (FDR and FNR = 0.03).

There was also variation in the accuracy of visual sorting with six crosses having an FDR > 0.1. Interestingly, the FDR for cross-specific skNIR models was positively correlated with the FDR for visual separation ($r = .65$). The underlying cause for this relationship is not readily apparent, but it suggests that some aspect of the visual score, possibly anthocyanin accumulation in the embryo, is encoded in the NIR signal.

3.5 | skNIR haploid classification uses information beyond oil content

Relative oil content was the most consistent trait that differentiated the two kernel classes (Table 2). However, there was no significant relationship between the oil content difference between haploid and hybrid kernels, hereafter, indicated by oil^{Δ}, and the FDR or FNR for the cross-specific models (Figure 2). The two induction crosses with the largest oil^{Δ}, BGEM-0029-S × inducer ($\text{oil}^{\Delta} = 1.86\%$) and BGEM-0110-N × inducer ($\text{oil}^{\Delta} = 1.88\%$), had divergent FDR values of 0.50

TABLE 4 External validation classification accuracy of each haploid induction cross using general and cross-specific models

Induction cross	Recurrent parent	General model		Cross-specific model		Visual	Oil$^{\Delta}$ (%)
		FDR[a]	FNR[b]	FDR	FNR	FDR[c]	
BGEM-0000-S x inducer	PHB47/SS	0.77	0.87	0.34	0.07	0.04	0.78
BGEM-0014-S x inducer		0.94	0.92	0.68	0.23	0.25	0.46
BGEM-0029-S x inducer		0.83	0.78	0.50	0.02	0.14	1.86
BGEM-0112-S x inducer		0.69	0.48	0.64	0.22	0	1.01
BGEM-0055-S x inducer		0.66	0.65	0.42	0.12	0.06	0.81
BGEM-0071-S x inducer		0.59	0.53	0.59	0.17	0.13	0.87
BGEM-0095-S x inducer		0.91	0.85	0.75	0.28	0.27	0.81
BGEM-0168-S x inducer		0.54	0.60	0.70	0.32	0.14	0.72
BGEM-0172-S x inducer		0.49	0.56	0.61	0.13	0.14	1.00
BGEM-0213-S x inducer		0.52	0.72	0.25	0.23	0	0.33
BGEM-0046-N x inducer	PHZ51/NSS	0.39	0.75	0.42	0.12	0	0.77
BGEM-0110-N x inducer		0.78	0.78	0.03	0.03	0	1.88
BGEM-0181-N x inducer		0.76	0.75	0.58	0.12	0	0.51
BGEM-0207-N x inducer		0.52	0.82	0.46	0.03	0	1.31
BGEM-0231-N x inducer		0.80	0.78	0.29	<0.01	0.04	0.48

[a] False Discovery Rate.

[b] False Negative Rate.

[c] Visual FDR was the fraction of hybrid kernels visually scored as haploids and confirmed by genotyping.

TABLE 5 Haploid discrimination accuracy using induction cross-specific models. Values from PLS-LDA models represent single-kernel holdout external validation. Lowest oil method uses the 20% lowest oil values per induction cross

Method	Predictors	Factors	FDR[a]	FNR[b]
PLS-LDA	NIR	9	0.53	0.16
PLS-LDA	Kernel Composition	8	0.57	0.49
Lowest Oil	Oil	—	0.67	0.39

[a] False Discovery Rate.

[b] False Negative Rate.

and 0.03, respectively (Table 4). There were also induction crosses with low oil$^{\Delta}$ and relatively low FDR. However, there was a trend for induction crosses with high oil$^{\Delta}$ values to have low FNR values, but low FNR values were not unique to high oil$^{\Delta}$ crosses. These data indicate that skNIR provides information beyond oil composition to classify haploids in many induction crosses.

4 | DISCUSSION

4.1 | Multiple NIR signals discriminate haploid kernels

Rotarenco et al. (2007) originally proposed that lack of xenia in the haploid embryo would alter kernel composition relative to hybrid kernels in an induction cross. Haploid embryos can reduce kernel oil content by 19% on average and embryo weight by 8%–14% when multiple donor genotypes are used (Rotarenco et al., 2007; Smelser et al., 2015). For the PHZ51 and PHB47 recurrent parent backgrounds in this study, haploid kernels reduced oil content by

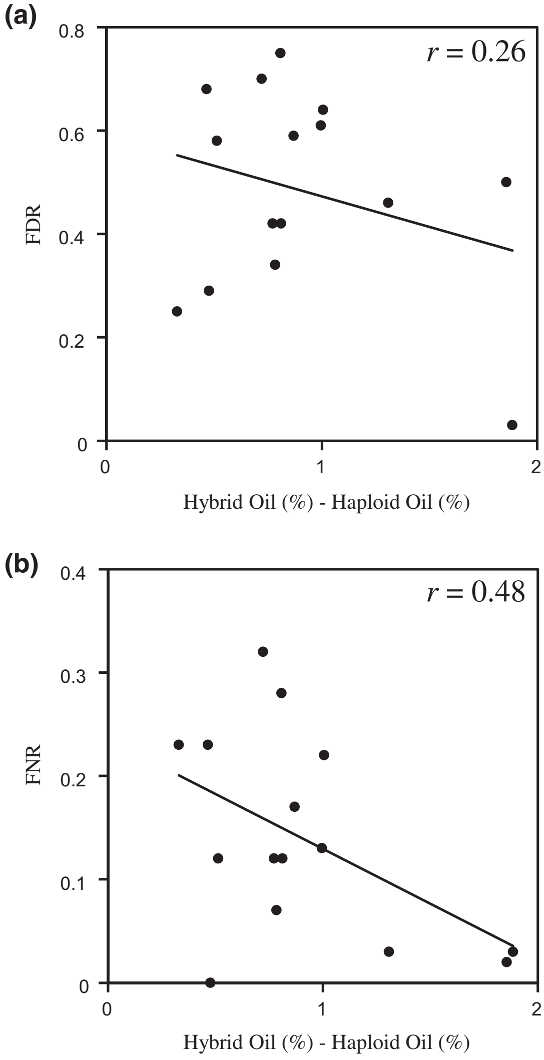

FIGURE 2 Scatterplot of the difference between hybrid and haploid oil content and the (a) FDR and (b) FNR for the induction cross-specific models. Linear regression trend line is shown

27% and 26% and reduced kernel weight by 5% and 4% respectively. Reduced kernel oil content in haploids can be explained by reduced embryo size relative to hybrid. The embryo contains a large proportion of total kernel oil. Reducing embryo size would reduce both the total and relative oil content, as long as the oil concentration in the embryo remains constant.

Other kernel traits were also significantly altered in haploids of many induction populations including kernel volume, density and relative protein content. The embryo has higher density than the endosperm (Gustin et al., 2013). This suggests that haploids could have a lower density than hybrids due to a smaller embryo size. Indeed, two induction crosses had haploids with significantly lower density. Overall, however, density traits did not show consistent trends in different induction crosses. In 4 of the 15 crosses, protein content was increased in the haploid kernels. It is not clear why haploid embryos would increase a kernels protein content. These induction cross-specific changes in kernel traits provide additional signals, beyond oil, for haploid kernel classification.

The lack of consistent trends beyond reduced oil content within the skNIR spectra precluded the development of a general PLS-LDA model to sort haploids accurately. This is consistent with other automated haploid sorting studies, which found donor, inducer and donor by inducer-specific effects rendered general classification models inaccurate (Jones et al., 2012; Melchinger et al., 2015). However, we found that constructing a PLS-LDA model specific to donor by inducer cross substantially improved classification accuracy. Eight of the 15 induction crosses achieved FDR < 0.50 and FNR < 0.23. For induction populations with a haploid sorting accuracy of FDR < 0.50, incorporating a skNIR instrument would reduce the number of kernels for manual inspection by 80% or more. In a practical usage case, a user would first calibrate the skNIR sorter using a small sample of the induction population by visually sorting with the *R1-nj* colour marker. The small sample would then be run on the skNIR and if the FDR of the calibration model was less than 0.50, the skNIR sorter could be used to enrich haploids from a large population of the specific cross prior to manual sorting.

The accuracy of haploid sorting using skNIR could be greatly improved by crossing donors with a high-oil haploid inducer, such as UHM600 or UHM601, which have 10%–12% kernel oil content (Melchinger et al., 2013). In an induction cross, the high oil genes contributed by the inducer increase hybrid kernel oil content, while haploid kernels develop maternal-level oil content (Rotarenco et al., 2007). When UHM600 or UHM601 is used as inducers, the oil$^{\Delta}$ averaged 1.78% with diverse donors (Melchinger et al., 2014). This increase in oil$^{\Delta}$ was enough to separate haploid and hybrid kernels into largely non-overlapping oil distributions. Here, the oil$^{\Delta}$ averaged 0.9% with two induction crosses having oil$^{\Delta}$ greater than 1.7% (Figure 2). Both high oil$^{\Delta}$ populations had an FNR of less than 0.05 and one had an FDR of 0.1 supporting the high oil$^{\Delta}$ effect in improving sorting accuracy. The increased oil$^{\Delta}$ provided by a high-oil haploid inducer combined with the additional signals embedded with the NIR spectra have the potential to reduce skNIR FDR and FNR close to visual accuracy.

Nuclear Magnetic Resonance (NMR) spectroscopy is an alternative to NIR for sorting haploids based on kernel composition. Single-kernel NMR spectroscopy provides higher accuracy measurements of oil content in grains than NIR and several studies have shown that haploid sorting using NMR spectroscopy can match or even improve upon visual sorting in high-oil induction crosses with diverse female donors (Melchinger et al., 2014; Melchinger et al., 2015; Melchinger et al., 2013). Several automated NMR single-kernel sorting platforms have been constructed that are capable of sorting haploids (Wang et al., 2016; Melchinger et al., 2018). Although these NMR sorters have a more accurate determination of seed oil content, these systems are slow as compared to skNIR. Current designs of NMR sorters require 4–6 s to process each kernel. The skNIR platform, on the other hand, was designed to process 10 kernels per second or 36,000 kernels per hour (Armstrong, 2006). Increased sorting speed makes the skNIR an attractive option for high-throughput sorting of maize haploids, particularly if combined with a high oil inducer that can drive larger oil$^{\Delta}$. We are constructing a single-kernel sorting device based on the light tube design of the platform used in this study. This device will be used to test the practical application of skNIR in the DH production process.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

J.G., A.M.S. and T.L. conceived and designed the experiments and J.G. and A.M.S. wrote the manuscript. T.L., U.F. and P.A. reviewed and edited the manuscript. T.L. and U.F. provided the haploid induction populations for the analysis. U.F, J.B., P.A. and J.G. preformed the experiments. J.G. preformed the data analysis and modelling. All authors have read and approved the manuscript.

ORCID

Jeffery L. Gustin https://orcid.org/0000-0002-5913-0200
Paul Armstrong https://orcid.org/0000-0002-4012-0010
Thomas Lübberstedt https://orcid.org/0000-0002-0526-0798
A. Mark Settles https://orcid.org/0000-0002-5846-0996

REFERENCES

Armstrong, P., & Tallada, J. (2012). Prediction of kernel density of corn using single-kernel near infrared spectroscopy. *Applied Engineering in Agriculture*, *28*(4), 569–574. https://doi.org/10.13031/2013.42071

Armstrong, P. R. (2006). Rapid single-kernel NIR measurment of grain and oil seed attributes. *Applied Engineering in Agriculture*, *22*(5), 767–772. https://doi.org/10.13031/2013.21991

Baye, T. M., Pearson, T. C., & Settles, A. M. (2006). Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. *Journal of Cereal Science*, *43*(2), 236–243. https://doi.org/10.1016/j.jcs.2005.11.003

Boote, B. W., Freppon, D. J., De La Fuente, G. N., Lübberstedt, T., Nikolau, B. J., & Smith, E. A. (2016). Haploid differentiation in maize kernels based on fluorescence imaging. *Plant Breeding*, *135*(4), 439–445. https://doi.org/10.1111/pbr.12382

Boulesteix, A.-L., Durif, G., Lambert-Lacroix, S., Peyre, J., & Strimmer, K. (2015).plsgenomics: PLS Analyses for Genomics (Version 1.3-1).

Brenner, E. A., Blanco, M., Gardner, C., & Lübberstedt, T. (2012). Genotypic and phenotypic characterization of isogenic doubled haploid exotic introgression lines in maize. *Molecular Breeding*, *30*(2), 1001–1016. https://doi.org/10.1007/s11032-011-9684-5

Cai, Z., Xu, G. L., Liu, X. H., Dong, Y. L., Dai, Y. X., & Li, S. H. (2007). The breeding of JAAS3-haploid inducer with high frequency parthenogenesis in maize. *Journal of Maize Science*, *151*, 1–4.

Chaikam, V., Martinez, L., Melchinger, A. E., Schipprack, W., & Boddupalli, P. M. (2016). Development and validation of red root marker-based haploid inducers in maize. *Crop Science*, *56*(4), 1678–1688. https://doi.org/10.2135/cropsci2015.10.0653

Chaikam, V., Nair, S. K., Babu, R., Martinez, L., Tejomurtula, J., & Boddupalli, P. M. (2015). Analysis of effectiveness of R1-nj anthocyanin marker for in vivo haploid identification in maize and molecular markers for predicting the inhibition of R1-nj expression. *Theoretical and Applied Genetics*, *128*(1), 159–171. https://doi.org/10.1007/s00122-014-2419-3

Chase, S. S. (1949). Monoploid frequencies in a commercial double cross hybrid maize, and in its component single cross hybrids and inbred lines. *Genetics*, *34*, 328–332.

Chen, S. J., & Song, T. M. (2003). Identification haploid with high oil xenia effect in maize. *Acta Agronomica Sinica*, *29*, 587–590.

Coe, E. H. (1959). A line of maize with high haploid frequency. *American Naturalist*, *93*(873), 381–382. https://doi.org/10.1086/282098

da Silva, A. R., Malafaia, G., & Menezes, I. P. (2017). Biotools: An R function to predict spatial gene diversity via an individual-based approach. *Genetics and Molecular Research*, *16*(2). https://doi.org/10.4238/gmr16029655

De La Fuente, G. N., Carstensen, J. M., Edberg, M. A., & Lübberstedt, T. (2017). Discrimination of haploid and diploid maize kernels via multispectral imaging. *Plant Breeding*, *136*(1), 50–60. https://doi.org/10.1111/pbr.12445

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 27.

Geiger, H. H., & Gordillo, G. A. (2009). Doubled haploids in hybrid maize breeding. *Maydica*, *54*, 485–499.

Gustin, J., Jackson, S., Williams, C., Patel, A., Armstrong, P., Peter, G., & Settles, A. (2013). Analysis of maize (Zea mays) kernel density and volume using microcomputed tomography and single-kernel near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry*, *61*(46), 10872–10880. https://doi.org/10.1021/jf403790v

Gustin, J. L., & Settles, A. M. (2015) Seed phenomics. In R. Fritsche-Neto, & A. Borém (Eds.), *Phenomics* (pp. 67–82). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-13677-6_5

Hacisalihoglu, G., Gustin, J., Louisma, J., Armstrong, P., Peter, G., Walker, A., & Settles, A. (2016). Enhanced single seed trait predictions in soybean (Glycine max) and robust calibration model transfer with near-infrared reflectance spectroscopy. *Journal of Agricultural and Food Chemistry*, *64*(5), 1079–1086. https://doi.org/10.1021/acs.jafc.5b05508

Hacisalihoglu, G., Larbi, B., & Settles, A. M. (2010). Near-infrared reflectance spectroscopy predicts protein, starch, and seed weight in intact seeds of common bean (Phaseolus vulgaris L.). *Journal of Agricultural and Food Chemistry*, *58*(2), 702–706. https://doi.org/10.1021/jf9019294

Irani, S., Knapp, A. D., Lubberstedt, T., Frei, U., & Askari, E. (2016). Increasing seed viability of maize haploid inducing lines by genetic and non-genetic approaches. *Crop Science*, *56*(4), 1940–1947. https://doi.org/10.2135/cropsci2015.11.0713

Jones, R. W., Reinot, T., Frei, U. K., Tseng, Y., Lübberstedt, T., & McClelland, J. F. (2012). Selection of haploid maize kernels from hybrid kernels for plant breeding using near-infrared spectroscopy and SIMCA analysis. *Applied Spectroscopy*, *66*(4), 447–450. https://doi.org/10.1366/11-06426

Li, X., Meng, D., Chen, S., Luo, H., Zhang, Q., Jin, W., & Yan, J. (2017). Single nucleus sequencing reveals spermatid chromosome fragmentation as a possible cause of maize haploid induction. *Nature Communications*, *8*(1), 991. https://doi.org/10.1038/s41467-017-00969-8

Lin, J., Yu, L., Li, W., & Qin, H. (2017). Method for identifying maize haploid seeds by applying diffuse transmission near-infrared spectroscopy. *Applied Spectroscopy*, *72*(4), 611–617. https://doi.org/10.1177/0003702817742790

Liu, Z., Wang, Y., Ren, J., Mei, M., Frei, U. K., Trampe, B., & Lübberstedt, T. (2016). Maize doubled haploids. *Plant Breeding Reviews*, *40*, 123–166. https://doi.org/10.1002/9781119279723.ch3

Lodder, R. A. (2002). Handbook of Near-Infrared Analysis, 2nd ed., Revised and Expanded. Practical Spectroscopy Series Volume 27 Edited by D. A. Burns (NIR Resources) and E. W. Ciurczak (Purdue Pharma LP). Dekker: New York. 2001. ISBN: 0-8247-0534-3. *Journal of the American Chemical Society*, *124*(19), 5603–5604. https://doi.org/10.1021/ja015320c

Lübberstedt, T., & Frei, U. K. (2012). Application of doubled haploids for target gene fixation in backcross programmes of maize. *Plant Breeding*, *131*(3),449–452.https://doi.org/10.1111/j.1439-0523.2011.01948.x

Melchinger, A. E., Böhm, J., Utz, H. F., Müller, J., Munder, S., & Mauch, F. J. (2018). High-throughput precision phenotyping of the oil content of single seeds of various oilseed crops. *Crop Science*, *58*(2), 670–678. https://doi.org/10.2135/cropsci2017.07.0429

Melchinger, A. E., Schipprack, W., Friedrich Utz, H., & Mirdita, V. (2014). In vivo haploid induction in maize: Identification of haploid seeds by their oil content. *Crop Science*, *54*(4), 1497–1504. https://doi.org/10.2135/cropsci2013.12.0851

Melchinger, A. E., Schipprack, W., Mi, X., & Mirdita, V. (2015). Oil content is superior to oil mass for identification of haploid seeds in maize produced with high-oil inducers. *Crop Science*, *55*(1), 188–195. https://doi.org/10.2135/cropsci2014.06.0432

Melchinger, A. E., Schipprack, W., Würschum, T., Chen, S., & Technow, F. (2013). Rapid and accurate identification of in vivo-induced haploid seeds based on oil content in maize. *Scientific Reports (Nature Publisher Group)*, *3*, 2129. https://doi.org/10.1038/srep02129

Osborne, B. G. (2006). Applications of near infrared spectroscopy in quality screening of early-generation material in cereal breeding programmes. *Journal of Near Infrared Spectroscopy*, *14*(2), 93–101. https://doi.org/10.1255/jnirs.595

Rotarenco, V., Dicu, G., State, D., & Fuia, S. (2010). New inducers of maternal haploids in maize. *Maize Genetics Cooperation Newsletter*, *84*, 21–22.

Rotarenco, V., Kirtoca, I., & Jocota, A. (2007). Using oil content to identify kernels with haploid embryos. *Maize Genetics Cooperation Newsletter*, *81*, 11.

Röber, F. K., Gordillo, G. A., & Geiger, H. H. (2005). In vivo haploid induction in maize - Performance of new inducers and significance of doubled haploid lines in hybrid breeding. *Maydica*, *50*(3), 275-283.

Sanchez, D. L., Liu, S., Ibrahim, R., Blanco, M., & Lübberstedt, T. (2018). Genome-wide association studies of doubled haploid exotic introgression lines for root system architecture traits in maize (Zea mays L.). *Plant Science*, *268*, 30–38. https://doi.org/10.1016/j.plantsci.2017.12.004

Smelser, A., Blanco, M., Lübberstedt, T., Schechert, A., Vanous, A., Gardner, C., & Tuberosa, M. (2015). Weighing in on a method to discriminate maize haploid from hybrid seed. *Plant Breeding*, *134*(3), 283–285. https://doi.org/10.1111/pbr.12260

Smelser, A., Gardner, C., Blanco, M., Lübberstedt, T., & Frei, U. (2016). Germplasm enhancement of maize: A look into haploid induction and chromosomal doubling of haploids from temperate-adapted tropical sources. *Plant Breeding*, *135*(5), 593–597. https://doi.org/10.1111/pbr.12397

Smith, J. S. C., Hussain, T., Jones, E. S., Graham, G., Podlich, D., Wall, S., & Williams, M. (2008). Use of doubled haploids in maize breeding: Implications for intellectual property protection and genetic diversity in hybrid crops. *Molecular Breeding*, *22*(1), 51–59. https://doi.org/10.1007/s11032-007-9155-1

Song, P., Zhang, H., Wang, C., Luo, B. I. N., & Xiong Zhang, J. U. N. (2018). Design and experiment of a sorting system for haploid maize kernel. *International Journal of Pattern Recognition and Artificial Intellegence*, *32*(3), 1855002. https://doi.org/10.1142/S0218001418550029

Spielbauer, G., Armstrong, P., Baier, J., Allen, W., Richardson, K., Shen, B., & Settles, A. (2009). High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. *Cereal Chemistry*, *86*(5), 556–564. https://doi.org/10.1094/CCHEM-86-5-0556

Tallada, J., Palacios-Rojas, N., & Armstrong, P. (2009). Prediction of maize seed attributes using a rapid single kernel near infrared instrument. *Journal of Cereal Science*, *50*(3), 381–387. https://doi.org/10.1016/j.jcs.2009.08.003

Team, R. C. (2017).R: A language and environment for statistical computing.

Vanous, A., Gardner, C., Blanco, M., Martin-Schwarze, A., Lipka, A. E., Flint-Garcia, S., … Lübberstedt, T. (2018). Association mapping of flowering and height traits in germplasm enhancement of maize doubled haploid (GEM-DH) lines. *The Plant Genome*, *11*(2), 170083. https://doi.org/10.3835/plantgenome2017.09.0083

Vanous, A., Gardner, C., Blanco, M., Martin-Schwarze, A., Wang, J., Li, X., … Lübberstedt, T. (2019). Stability analysis of kernel quality traits in exotic-derived doubled haploid maize lines. *The Plant Genome*, *12*(1), 1–14. https://doi.org/10.3835/plantgenome2017.12.0114

Wang, H., Liu, J., Xu, X., Huang, Q., Chen, S., Yang, P., … Song, Y. (2016). Fully-automated high-throughput NMR system for screening of haploid kernels of maize (Corn) by measurement of oil content. *PLOS ONE*, *11*(7), e0159444. https://doi.org/10.1371/journal.pone.0159444

Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*, 22.

Zhao, X., Xu, X., Xie, H., Chen, S., & Jin, W. (2013). Fertilization and uniparental chromosome elimination during crosses with maize haploid inducers. *Plant Physiology*, *163*(2), 721.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Gustin JL, Frei UK, Baier J, Armstrong P, Lübberstedt T, Settles AM. Classification approaches for sorting maize (*Zea mays* subsp. *mays*) haploids using single-kernel near-infrared spectroscopy. *Plant Breed*. 2020;00:1–10. https://doi.org/10.1111/pbr.12857