



- (51) **International Patent Classification:**
G01N 21/359 (2014.01) *G01N 33/00* (2006.01)
- (21) **International Application Number:**
PCT/US2017/022310
- (22) **International Filing Date:**
14 March 2017 (14.03.2017)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
62/308,549 15 March 2016 (15.03.2016) US
- (71) **Applicant: UNIVERSITY OF FLORIDA RESEARCH FOUNDATION, INCORPORATED** [US/US]; 223 Grinter Hall, Gainesville, FL 32611 (US).
- (72) **Inventors: SETTLES, Andrew M.**; c/o University of Florida, Bldg. 885, P.O. Box 110690, Gainesville, FL 32611 (US). **GUSTIN, Jeffrey Lynn II**; 2606 NW 19 Way, Gainesville, FL 32605 (US).
- (74) **Agents: FRANK, Louis, C.** et al.; Saliwanchik Lloyd & Eisenschenk, P.O. Box 142950, Gainesville, FL 32614-2950 (US).

- (81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) **Title:** HAPLOID SEED CLASSIFICATION USING SINGLE SEED NEAR-INFRARED SPECTROSCOPY

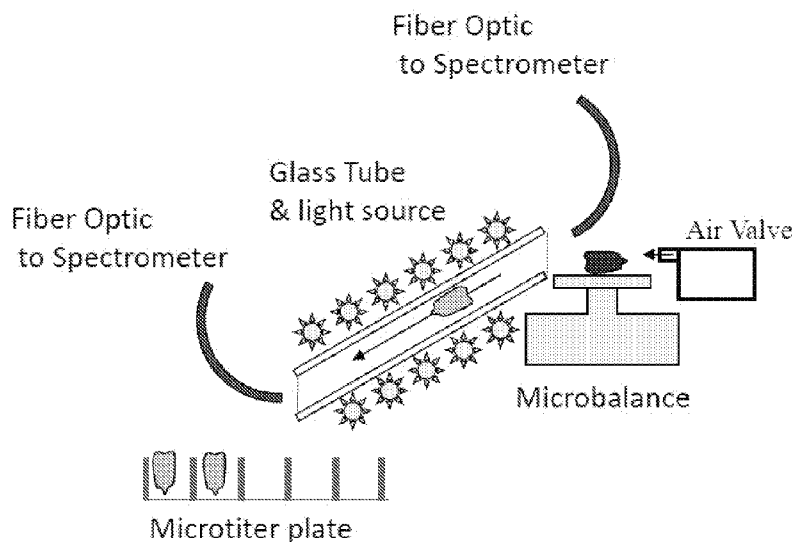


FIG. 1

(57) **Abstract:** Methods for sorting haploid maize kernels in haploid induction crosses are provided. A method of sorting haploid kernels can include capturing near-infrared (NIR) spectra and applying a general multivariate statistical model to the acquired NIR spectra to discriminate haploid kernels from diploid-hybrid kernels mixed therein. NIR spectra can be collected on a single kernel using a high-throughput apparatus such that the amount of time required to analyze individual kernels can be significantly reduced in comparison to existing NIR technology and in comparison to traditional manual sorting.

DESCRIPTION

HAPLOID SEED CLASSIFICATION USING SINGLE SEED NEAR-INFRARED SPECTROSCOPY

5

CROSS REFERENCE TO RELATED APPLICATION

This application claims the benefit of U.S. Provisional Application Serial No. 62/308,549, filed March 15, 2016, the disclosure of which is hereby incorporated by
10 reference in its entirety, including any figures, tables, and drawings.

GOVERNMENT SUPPORT

This invention was made with government support under grant No. 1444456 awarded by the National Science Foundation and grant No. 2011-67003-30215 awarded
15 by the National Institute of Food and Agriculture. The government has certain rights in the invention.

BACKGROUND

Doubled haploids (DHs) are used extensively in maize breeding programs.
20 Haploid seeds are generated using pollen from a genetic stock known to induce haploid embryo development. Haploid seedlings can then be treated with colchicine to induce genome doubling and a new inbred breeding stock. Steps involved in DH production can be inefficient. Sorting of haploid seeds is conventionally based on qualitative manual identification by R-nj color markers. For certain crosses possessing dominant repressors
25 of the R-nj color marker, such as tropical landraces and sweet corn, difficulty in manual haploid sorting can greatly restrict DH breeding in these germplasm groups.

BRIEF SUMMARY

Embodiments of the subject invention provide methods for identifying haploid
30 maize kernels in haploid induction crosses. A method of sorting haploid kernels can include capturing near-infrared (NIR) spectra and applying a general multivariate statistical model to the acquired NIR spectra to discriminate haploid kernels from diploid-

hybrid kernels mixed therein. In an embodiment, NIR spectra can be collected on a single kernel using a high-throughput apparatus such that the amount of time required to analyze individual kernels is significantly reduced in comparison to existing NIR-based sorting technology and traditional manual sorting. In a preferred embodiment, a Linear Discriminant Analysis (LDA) can be applied to the acquired spectroscopic data to derive a model for reliable classification of haploid seeds from diverse germplasm sources scanned by the NIR spectrometer.

Advantageously, methods provided herein enable quantitative and accurate sorting of single-kernel haploid seeds that are traditionally identified qualitatively using R-nj color markers, thereby increasing the efficiency of the sorting process and allowing a broader set of germplasm sources to be available for DH production.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a schematic illustration of a high-throughput single-kernel NIR analyzer according to an embodiment of the subject invention.

Figure 2 shows an image of a high-throughput single-kernel NIR analyzer according to an embodiment of the subject invention.

Figure 3 shows images of hybrid and haploid seeds from different pedigrees

Figure 4 shows a graph of the false discovery rate and false negative rate of sorting accuracy of kernel oil content alone (i.e., “Oil FDR” and “Oil FNR”) and the linear discriminate model (i.e., “LDA FDR” and “LDA FNR”) according to an embodiment of the subject invention.

DETAILED DESCRIPTION

Embodiments of the subject invention provide methods for identifying haploid maize kernels in haploid induction crosses. In some embodiments, methods provided herein can be used in assisting in the development of DH techniques in other cereal crops for which double haploids can be selected at the mature seed stage, with non-limiting examples including sorghum and millet. Advantageously, methods provided herein enable automated accurate sorting of haploid kernels that are traditionally identified qualitatively using R-nj color markers, thereby increasing the efficiency of the sorting process and allowing a broader set of germplasms to be identified for DH production.

In an embodiment, a method of sorting haploid kernels comprises determining kernel composition and quality values using a near-infrared (NIR) spectrometer and applying a general multivariate statistical model to the acquired NIR spectra to discriminate haploid kernels from diploid-hybrid kernels mixed therein. NIR spectra can
5 be collected on a single kernel using a high-throughput apparatus such that the amount of time required to analyze individual kernels is significantly reduced in comparison to existing NIR-based sorting technology and in comparison to traditional manual sorting.

In an exemplary embodiment, the NIR platform employs a conventional open-source design such as, for example, one obtained from the U.S. Department of
10 Agriculture-Agriculture Research Service in Manhattan, KS. As illustrated in Figure 1, a single-kernel NIR (skNIR) device can include an air valve, two 400-micro fiber optics connecting the device to an InGaAs array-based spectrometer, a 12 mm x 67 mm borosilicate glass tube, a light source, and a microtiter plate. Optionally, the NIR device can comprise a microbalance used to measure seed weight.

15 In certain embodiments, the method of acquiring NIR spectra of single kernels can include weighing each kernel on the microbalance, followed by blowing the kernel into the borosilicate glass tube (*e.g.*, with an airstream provided by the air valve). The instrument can then collect an NIR spectral profile within a predetermined range of wavelengths as the kernel falls through the glass tube, which can be illuminated by a light
20 source (*e.g.*, a halogen light source). Light reflected by the kernel can subsequently be captured and relayed to a diode array-based spectrometer (*e.g.*, by fiber optic cables located at the top and/or the bottom ends of the glass tube). The kernel can then be collected and indexed in the microtiter plate. In an embodiment, the NIR spectrum can be acquired for a range of wavelengths from about 907 nm to about 1689 nm (or any
25 subrange located therewithin) in, for example, 1 nm intervals.

In one embodiment, a solenoid sorter can be attached to the NIR device to automate haploid seed selection. The data collection time for each single kernel can be on the order of seconds such as, for example, 1 second, 2 seconds, 3 seconds, 4 seconds, 5
30 seconds, or 6 seconds. Data collection time can also be less than 1 second when seed weight is not collected using the microbalance, and such collection time can be, for example, 0.3 second, 0.5 second, 0.6 second, or 0.8 second per kernel. When an exemplary

automated embodiment of the skNIR device is employed, the spectra acquisition time can be as short as 0.25 second.

NIR spectra acquired using an apparatus discussed herein can reveal signals for major chemical components of the maize kernel such as, for example, protein, starch, and oil. In addition, kernel quality traits such as kernel volume, weight, density, and air space within the kernel can be estimated from the NIR spectra based on careful calibration results as previously reported (see, for example, Gustin *et al.*, Journal of Agricultural and Food Chemistry 61, 10872-10880, 2013). Kernels comprising haploid embryos can display changes in kernel composition consistent with reduced embryo size, manifested as reduced kernel weight and/or relative oil content. Furthermore, minor, but statistically significant, changes in other kernel traits including protein and starch content can also be observed in haploid seeds. These compositional changes in haploid seeds provide a strong basis for discrimination using infrared light.

The use of skNIR for measuring the chemical composition of seeds has been previously reported (see, for example, Armstrong, P.R., 3 Applied Engineering in Agriculture 22: 767-772, 2006; Spielbauer, G., Armstrong, P. *et al.*, Cereal Chemistry 86: 556-564, 2009); however, the spectral measurements were not used in combination with a multivariate statistical model to separate haploid seeds from their hybrid counterparts.

In some embodiments, the skNIR can distinguish haploid kernels from diploid kernels in haploid induction crosses using a general statistical regression model. In a preferred embodiment, a model created by Linear Discriminant Analysis (LDA) can be applied to the sorting process, resulting in reliable classification of haploid seeds from diverse germplasm sources scanned by the skNIR spectrometer.

LDA is a supervised classification method that, as opposed to an unsupervised method such as the principal component analysis (PCA), utilizes knowledge of haploid or diploid-hybrid class to calculate latent variables that emphasize spectral differences between species needing to be classified. In some embodiments, an LDA model requires a set of training spectra to which the haploid/diploid-hybrid class is known in order to build a calibration model that can then be used to predict the class of an unknown kernel based on its NIR spectra. In an embodiment, an LDA calibration model can be built using the spectral information from kernels of a randomly chosen portion of genotypes within a

larger set of genotypes. Data of the remaining genotypes can be used to validate the calibration model.

In an embodiment of the subject invention, LDA can be a two-step process. First, partial least squares (PLS) regression models use the kernels' NIR spectra to generate kernel values for multiple traits. Table 1 shows exemplary ANOVA statistics for the effect of haploid status on each trait that can be derived from the NIR spectra. The traits impacted by haploid status are relative content of oil, starch, and protein; and density and volume. Second, these five kernel traits are then regressed on haploid/hybrid class structure using LDA implemented by the R statistical function 'lda' in package 'MASS' to identify the discriminant function that can best separate the classes.

Table 1. An exemplary two-way ANOVA statistics (Type I) modeling kernel traits with induction cross (cross) and kernel ploidy (ploidy) as the main effects. "Coef" denotes the β coefficient of the ploidy term and indicates the haploid effect on a given trait.

Seed trait	Effect	df ^a	Coef	Std Error Coef	Mean Sq	F-stat	p value
Weight (mg)	cross	14	-	-	45538	25.0	<0.0001
	ploidy	1	-18.4	2.34	111633	61.2	<0.0001
Oil (%)	cross	14	-	-	9.71	20.3	<0.0001
	ploidy	1	-0.90	0.04	270	563	<0.0001
Protein (%)	cross	14	-	-	149	81.7	<0.0001
	ploidy	1	0.19	0.07	12.2	6.65	0.01
Starch (%)	cross	14	-	-	1104	102	<0.0001
	ploidy	1	0.21	0.18	15.0	1.40	0.2
Density (total) (g/cm ³)	cross	14	-	-	0.10	22.5	<0.0001
	ploidy	1	-0.015	0.003	0.07	17.0	<0.0001
Density (material) (g/cm ³)	cross	14	-	-	0.05	47.4	<0.0001
	ploidy	1	0.002	0.002	0.002	1.6	0.2
Volume (total) (g/cm ³)	cross	14	-	-	21650	24.8	<0.0001
	ploidy	1	-8.72	1.62	25212	28.9	<0.0001
Volume (material) (g/cm ³)	cross	14	-	-	18545	26.7	<0.0001
	ploidy	1	-9.60	1.45	30522	43.9	<0.0001

The LDA model discussed herein uses composition and seed size traits to complete the sorting instead of using anthocyanin colors obtained from, for example, R-nj color markers. Prior works (see, for example, Jones, R.W., Reinot, T. *et al.*, Applied Spectroscopy 66: 447-450, 2012; Melchinger, A.E., Winter, M. *et al.*, Crop Science 55:

1076-1086, 2015; Melchinger, A.E., Schipprack, W. *et al.*, Crop Science 54: 1497-1504, 2014) have demonstrated methods of haploid seed sorting based on the kernels' chemical composition using alternative devices including, for example, nuclear magnetic resonance (NMR) and near-infrared transmission (NIT) spectroscopy. The data collection time on a
5 single kernel afforded by these devices varies from seconds to minutes, which can be orders of magnitude slower than the methods provided herein.

A major advantage of the skNIR sorting provided herein is the speed at which the seeds can be sorted. Efficient sorting is integral to the DH breeding process. Induction crosses need to be rapidly sorted soon after harvest before being planted in the next
10 available nursery. The population size of haploid induction crosses is typically 15,000 to 30,000 kernels, and a small proportion thereof, typically 1-10%, will have haploid embryos. When a color marker such as the R-nj marker is used, the haploid seeds are selected visually with manual labor. Experienced personnel manually sorting haploids can process approximately 1000 seeds per hour, provided that the color markers are clearly
15 expressed in the kernels, requiring approximately 15 to 30 man-hours to process a single haploid induction population in related art methods.

Advantageously, spectra acquired using the skNIR device provided herein can be integrated over 20 milliseconds, and sorting individual kernels can take less than one second. The theoretical throughput of an automated embodiment of the skNIR device can
20 be nearly 10-fold higher than the speed of manual sorting. In the design of the skNIR device discussed herein, the rate-limiting step for spectral acquisition of a single seed is the time needed for the seed to pass through the light tube, which is approximately 0.25 second. At this rate, the skNIR device could conservatively process 10,000 seeds per hour, and several haploid induction populations could be sorted in a single day.
25 Advantageously, faster sample processing time can reduce labor costs and minimize the time required for DH production pipeline.

The skNIR platform and the general LDA model discussed herein, being built upon diverse genetic materials, can identify an optimal LDA vector to distinguish haploid from diploid kernels in any cross without prior knowledge of the genotypic class structure
30 thereof. Advantageously, the LDA model can be optimized to conduct efficient haploid selection and to allow a broader set of germplasms to be included in doubled haploid (DH)

production, potentially including sweet corn and tropical germplasm sources that currently cannot be bred efficiently with DH technology.

Furthermore, the accuracy of haploid selection based on chemical composition of the kernels may be more stable across a wide diversity of germplasm than visual sorting
5 based on a color marker. Manual sorting based on visual kernel markers is not perfectly accurate as the accuracy depends largely upon expression of the color markers in the kernel. Previous reporting has shown wide variations in visual sorting accuracy in induction crosses using diverse female donors (see Melchinger, A.E., Schipprack, W. *et al.*, Crop Science 54: 1497-1504, 2014). Improving haploid sorting accuracy would add
10 efficiency to DH production process by allowing a “fine tuning” of the resources used to generate DH inbred lines.

Another advantage of the sorting method based on spectra (*e.g.*, skNIR spectra) is the reduced or removed requirement of visible color markers for haploid sorting. Typically, breeding efforts to improve inducer line performance is hindered by requiring
15 the inducer lines to carry either dominant color markers. Single-kernel NIR sorting, however, is based on the kernel composition changes resulting from haploid embryo development and will thus allow sorting in genetic backgrounds that suppress color markers. Dominant color markers depend upon multiple, unlinked loci, creating excessive linkage drag when attempting to improve the agronomic and induction performance by
20 traditional plant breeding. Advantageously, embodiments of the subject invention can lead to development and testing of new inducer lines with greater and more consistent haploid induction rate, and ultimately enable DH maize breeding for a broader international market by allowing development of inducer lines that are adapted to a wider range of agricultural environments.

25 Additionally, embodiments provided herein can enable the potential development of a standalone skNIR device equipped with a novel mathematical model for augmenting or replacing manual screening services and providing a less expensive alternative to NMR kernel sorters.

The terms “seeds” and “kernels” are used interchangeably herein in the context of
30 maize species. The terms “diploid-hybrid”, “diploid”, and “hybrid” are used interchangeably herein.

The methods and processes described herein (e.g., mathematical models such as LDA models) can be embodied as code and/or data. The software code and data described herein can be stored on one or more machine-readable media (e.g., computer-readable media), which may include any device or medium that can store code and/or data for use
5 by a computer system. When a computer system and/or processor reads and executes the code and/or data stored on a computer-readable medium, the computer system and/or processor performs the methods and processes embodied as data structures and code stored within the computer-readable storage medium.

It should be appreciated by those skilled in the art that computer-readable media
10 include removable and non-removable structures/devices that can be used for storage of information, such as computer-readable instructions, data structures, program modules, and other data used by a computing system/environment. A computer-readable medium includes, but is not limited to, volatile memory such as random access memories (RAM, DRAM, SRAM); and non-volatile memory such as flash memory, various read-only-
15 memories (ROM, PROM, EPROM, EEPROM), magnetic and ferromagnetic/ferroelectric memories (MRAM, FeRAM), and magnetic and optical storage devices (hard drives, magnetic tape, CDs, DVDs); network devices; or other media now known or later developed that is capable of storing computer-readable information/data. Machine-readable media (e.g., computer-readable media) should not be construed or interpreted to
20 include any propagating signals. A computer-readable medium of embodiments of the subject invention can be, for example, a compact disc (CD), digital video disc (DVD), flash memory device, volatile memory, or a hard disk drive (HDD), such as an external HDD or the HDD of a computing device, though embodiments are not limited thereto. A computing device can be, for example, a laptop computer, desktop computer, server, cell
25 phone, or tablet, though embodiments are not limited thereto.

The subject invention includes, but is not limited to, the following exemplified embodiments.

Embodiment 1. A method of sorting haploid seeds, the method comprising:
providing a population of seeds to be sorted, each seed being either a haploid seed
30 or a diploid-hybrid seed;
providing an apparatus comprising a light source and a spectrometer capable of examining each seed individually;

illuminating each seed with light generated by the light source;
acquiring at least one spectrum for each seed;
processing the at least one spectrum of each seed using computer software with
user-determined parameters and storing the data in a data matrix;

5 calibrating a regression model based on a portion of the data matrix that excludes a
selected genotype of seeds, the genotype being considered a held-out genotype;

cross-validating the regression model by holding out each genotype present in the
population of seeds and repeating the classification step; and

classifying a randomly sampled seed using the validated model.

10 Embodiment 2. The method according to embodiment 1, wherein the
haploid seeds are maize kernels.

Embodiment 3. The method according to any of embodiments 1-2, wherein
the regression model employs a linear discriminate analysis (LDA).

15 Embodiment 4. The method according to any of embodiments 1-3, wherein
the seeds to be sorted are of at least one induction cross.

Embodiment 5. The method according to any of embodiments 1-4, wherein
the seeds to be sorted are of the same genotypic background.

Embodiment 6. The method according to any of embodiments 1-4, wherein
the seeds to be sorted are of different genotypic backgrounds.

20 Embodiment 7. The method according to any of embodiments 3-6, wherein
the accuracy of the classification step is improved by assigning a score to each seed
following the LDA analysis, and setting a fraction threshold for the population of seeds
based on their scores and the predicted induction frequency and/or the size of the
induction population,

25 Embodiment 8. The method according to any of embodiments 1-7, wherein
the apparatus further comprises an air valve, a transparent seed passage tube, a balance
capable of measuring the weight of each seed, and fiber optics accompanying the
spectrometer.

30 Embodiment 9. The method according to embodiment 8, wherein the
apparatus further comprises an automated seed sorter.

Embodiment 10. The method according to embodiment 9, wherein the
automated seed sorter is a solenoid sorter.

Embodiment 11. The method according to any of embodiments 1-10, wherein the light source emits near-infrared (NIR) light.

Embodiment 12. The method according to embodiment 11, wherein the light source emits light in a wavelength range of between 907 nm and 1689 nm (or any
5 subrange therewithin).

Embodiment 13. The method according to any of embodiments 1-12, capable of differentiating at least one of the following traits of the population of seeds being sorted: relative oil content, relative protein content, relative starch content, seed density, and seed volume.

10 Embodiment 14. The method according to any of embodiments 1-13, wherein the time required for sorting each seed in the population is less than 1 second.

Embodiment 15. The method according to any of embodiments 1-13, wherein the time is required for sorting each seed is less than 0.25 second.

15 Embodiment 16. The method according to any of embodiments 1-13, wherein the time is required for sorting each seed is 0.25 second.

Embodiment 17. The method according to any of embodiments 3-16, wherein the LDA is implemented by an R-based statistical function capable of discriminating the haploid seeds from the diploid hybrid seeds.

20 Embodiment 18. An apparatus for sorting haploid seeds, the apparatus comprising:

a light source;

a spectrometer capable of examining single seeds individually;

an air valve;

a seed passage tube;

25 fiber optics accompanying the spectrometer; and

a (non-transitory) machine-readable medium (e.g., (non-transitory) computer-readable medium) in operable communication with the spectrometer, the machine-readable having machine-executable instructions (e.g., computer-executable instructions) stored thereon for performing a method (e.g., when executed by a processor, the
30 instructions perform the method) comprising the following steps:

processing at least one spectrum of each seed, obtained from the spectrometer, with user-determined parameters;

storing the data of the at least one spectrum of each seed in a data matrix;
and

calibrating a regression model based on a portion of the data matrix that
excludes a selected genotype of seeds, the genotype being considered a held-out
5 genotype.

Embodiment 19. The apparatus according to embodiment 18, wherein the
method performed by the machine-executable instructions further comprises:

cross-validating the regression model by holding out each genotype present
in the population of seeds and repeating the classification step; and

10 classifying a randomly sampled seed using the validated model.

Embodiment 20 The apparatus according to any of embodiments 18-19,
wherein the seed passage tube is a transparent seed passage tube.

Embodiment 21 The apparatus according to any of embodiments 18-20,
further comprising a balance capable of measuring the weight of single seeds individually.

15 Embodiment 22 The apparatus according to any of embodiments 18-21,
further comprising an automated seed sorter.

Embodiment 23. The apparatus according to embodiment 22, wherein the
automated seed sorter is a solenoid sorter.

Embodiment 24 The apparatus according to any of embodiments 18-23,
20 wherein the light source is a near-infrared (NIR) light source that emits NIR light.

Embodiment 25. The method according to embodiment 24, wherein the light
source emits light in a wavelength range of between 907 nm and 1689 nm (or any
subrange therewithin).

Embodiment 26 The apparatus according to any of embodiments 18-25,
25 wherein the regression model employs a linear discriminate analysis (LDA).

Embodiment 27. The method according to embodiment 26, wherein the
accuracy of the classification step is improved by assigning a score to each seed following
the LDA analysis, and setting a fraction threshold for the population of seeds based on
their scores and the predicted induction frequency and/or the size of the induction
30 population.

Embodiment 28 The apparatus according to any of embodiments 26-27, wherein the LDA is implemented by an R-based statistical function capable of discriminating the haploid seeds from the diploid hybrid seeds.

Embodiment 29. The apparatus according to any of embodiments 18-28, further comprising a processor in operable communication with the (non-transitory) machine-readable medium (e.g., (non-transitory) computer-readable medium), the processor being capable of causing the machine-executable instructions (e.g., computer-executable instructions) stored on the medium to perform the method.

A greater understanding of the present invention and of its many advantages may be had from the following example, given by way of illustration. The following example is illustrative of some of the methods, applications, embodiments and variants of the present invention. It is, of course, not to be considered as limiting the invention. Numerous changes and modifications can be made with respect to the invention.

EXAMPLE 1

To develop a classification algorithm, a set of 1354 kernels obtained from the Iowa State University (ISU) Doubled Haploid Facility were scanned with a skNIR spectrometer. A total of 14 induction crosses were sampled. The female donor for the induction cross originated from the Germplasm Enhancement of Maize (GEM) project (see, for example, Brenner *et al.*, 2012 Mol Breeding 30: 1001). For each cross, 48 haploids and 48 diploid-hybrids were visually identified by the ISU staff using traditional R-nj color markers. Hybrids have anthocyanin accumulation in the embryo, while haploids lack anthocyanins in the embryo (Figure 3). The ploidy of the kernels was subsequently confirmed by genotyping seedling tissue after the kernels had been analyzed by the skNIR.

Two NIR spectra were collected from each kernel. Each spectrum was mean-centered, and the average profile was used for modeling. Seed trait predictions showed that haploid seeds have statistically significant differences in kernel weight, density, and volume, as well as oil and protein content based on ANOVA (see Table 1). These data show that haploid embryos impact multiple kernel traits simultaneously.

Prior to model selection, the ratio of haploid kernels to hybrid kernels was adjusted to reflect a typical induction-cross haploid frequency of 12% haploids and 88% hybrids. Six haploid kernels were randomly sampled from the starting pool of 48 from each

induction cross. The kernels were split into calibration and external validation datasets with the kernels from all but one genotype used to calibrate the model. The calibration model was used to classify the kernels of the held-out genotype as haploid or diploid. Leave-one-out (LOO) cross-validations were run 14 times with each genotype being held
5 out for external validation in turn. This process was repeated 100 times with random sampling of the haploid kernels at each iteration. This method produced approximately 75,000 independent classifications.

Sorting accuracy is defined in terms of a false negative rate (FNR), which is the percentage of haploids that were classified as hybrids and a false discovery rate (FDR),
10 which is the percentage of classified haploids that are actually hybrid. The composite error rate (CER) is the sum of FDR and FNR and is a useful metric for comparing accuracy between models and induction populations. The general LDA model's FDR and FNR was 0.40 and 0.78, respectively, with a CER of 1.18.

Further improvements were made in classification accuracy by applying a
15 thresholding method based on the LDA scores derived from the LDA model. For this method, the kernels with the lowest LDA score in each induction cross were classified as haploid. Figure 4 shows how the FDR and FNR change as an increasing number of low LDA score kernels are included into the fraction. As expected, increasing the fraction of lowest LDA score kernels reduces FNR as more haploids are included and increases FDR
20 due to greater inclusion of hybrid kernels. Setting a fraction threshold based on induction frequency and the size of the induction population would be a tradeoff between the FDR and FNR. For example, in the induction cross populations experimented herein, selecting 15% of kernels with the lowest LDA score would yield a pool of kernels that contained 43% haploids (FDR = 0.57) and recover 59% of all the haploids (FNR = 0.41) on average.

25 Haploid discrimination approaches that use NMR rely only on oil content to classify haploids; however, it has been found that other kernel characteristics can also contribute to haploid discrimination. To test whether oil content alone could classify haploid and hybrid kernels as well as the LDA scores, the same threshold approach was taken except that kernels with the lowest oil content were binned into the predicted
30 haploid group rather than kernels with the lowest LDA scores. Figure 4 shows that the FDR and FNR curves for oil alone are very similar to the FNR and FDR curves based on LDA scores. However, the LDA score method generates a modestly higher accuracy of

between the 5% and 35% fractions. This is consistent with previous data showing that oil is the main discriminating variable. However, adding the additional skNIR-based traits improved discrimination.

The thresholding strategy using LDA scores enriched haploids in all induction crosses examined. At a threshold of 20%, CER among the induction crosses ranged from 0.58 to 1.43 (Table 2). All but one of the induction crosses had an FNR of less than 0.5, which means that over 50% of the haploids in the population were correctly identified. Binning 20% of the lowest LDA score kernels also enriched all induction crosses to at least twice the induction rate of the original population (12%) and in 6 induction crosses, the 20% bin would contain over 40% haploids. These data demonstrate the utility of the skNIR method in enriching an induction crosses for haploid kernels.

Table 2. Accuracy of classification of each induction cross when classifying kernels with the bottom 20% of LDA scores as haploid.

<u>Induction cross</u>	<u>FDR</u>	<u>FNR</u>	<u>CER</u>
(GEM1161 x PHZ51) x inducer	0.61	0.29	0.90
(GEM1171 x PHZ51) x inducer	0.55	0.25	0.79
(GEM1183 x PHZ51) x inducer	0.63	0.32	0.95
(GEM1206 x PHZ51) x inducer	0.69	0.39	1.08
(GEM1230 x PHZ51) x inducer	0.68	0.42	1.11
(GEM1239 x PHB47) x inducer	0.55	0.17	0.72
(GEM1244 x PHB47) x inducer	0.80	0.63	1.43
(GEM1253 x PHB47) x inducer	0.53	0.05	0.58
(GEM1262 x PHB47) x inducer	0.52	0.20	0.71
(GEM1271 x PHB47) x inducer	0.55	0.18	0.73
(GEM1276 x PHB47) x inducer	0.56	0.20	0.76
(GEM1284 x PHB47) x inducer	0.65	0.34	0.99
(GEM1300 x PHB47) x inducer	0.62	0.30	0.91
(GEM1302 x PHB47) x inducer	0.51	0.16	0.67
(GEM1320 x PHB47) x inducer	0.70	0.46	1.16

It should be understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application.

- 5 All patents, patent applications, provisional applications, and publications referred to or cited herein (including those in the “References” section) are incorporated by reference in their entirety, including all figures and tables, to the extent they are not inconsistent with the explicit teachings of this specification.

REFERENCES

1. Armstrong, P. R. 2006. Rapid single-kernel NIR measurement of grain and oil-seed attributes. 3 Appl. Eng. Agric. 22:767-772.
2. Jones et al., 2012 Selection of haploid maize kernels from hybrid kernels for plant breeding using near-infrared spectroscopy and SIMCA analysis. Appl Spectrosc. 66(4):447-50.
3. Brenner et al., 2011 Genotypic and phenotypic characterization of isogenic doubled haploid exotic introgression lines in maize. Mol Breeding (2012) 30: 1001.
4. BoMill (<http://www.bomill.com/>) has a patented single seed sorter which uses Near-infrared transmission (NIT) spectroscopy to sort wheat and barley seed at a rapid rate. European Patent, EC B07C5/34; G01N2/35G.
5. Brimrose COA has a single seed NIR sorter product called Luminar 3076 that uses Acousto Optic Tunable Filter Near-Infrared (AOTF-NIR). http://www.brimrose.com/products/nir_mir_spectrometers/sort_by_spec_trometers/3076.html.
6. Melchinger et al., 2014 In vivo haploid induction in maize: Identification of haploid seeds by their oil content. Crop Science 54(4):1497-1504.
7. Melchinger et al., 2015 Controlling misclassification rates in identification of haploid seeds from induction crosses in maize with high-oil inducers. Crop Science 55(3): 1076-1086.
8. Hacisalihoglu G., Gustin J.L., Louisma J., Armstrong P., Peter G.F., Walker A.R., Settles A.M. (2016) Enhanced Single Seed Trait Predictions in Soybean (*Glycine max*) and Robust Calibration Model Transfer with Near-Infrared Reflectance Spectroscopy. *J. Agric. Food Chem.* 64:1079-1086.
9. Gustin *et al.*, Journal of Agricultural and Food Chemistry 61, 10872-10880, 2013.
10. Hacisalihoglu et al., Journal of Agricultural and Food Chemistry 2010 – Common Bean
11. Spielbauer et al., Journal of Cereal Chemistry 2006 – Maize
12. Iowa State University Doubled Haploid Facility:
<http://www.plantbreeding.iastate.edu/DHF/DHF.htm>.

13. Germplasm Enhancement of Maize (GEM) Project: http://www.public.iastate.edu/~usda-gem/GEM_Project/GEM_Project.htm.
14. Spielbauer, G., Armstrong, P. *et al.*, Cereal Chemistry 86: 556-564, 2009.
15. Mevik, B.; Wehrens, R.; Liland, K. H. pls: Partial least squares and principal component regression. R package version 2.3-0. <http://CRAN.R-project.org/package=pls>, 2011.
16. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

CLAIMS

What is claimed is:

1. A method of sorting haploid seeds, comprising:
providing a population of seeds to be sorted, each seed being either a haploid seed or a diploid-hybrid seed;
providing an apparatus comprising a light source and a spectrometer capable of examining each seed individually;
illuminating each seed with light generated by the light source;
acquiring at least one spectrum for each seed;
processing the at least one spectrum of each seed using computer software with user-determined parameters and storing the data in a data matrix;
calibrating a regression model based on a portion of the data matrix that excludes a selected genotype of seeds, the genotype being considered a held-out genotype;
cross-validating the regression model by holding out each genotype present in the population of seeds and repeating the classification step; and
classifying a randomly sampled seed using the validated model.
2. The method according to claim 1, wherein the haploid seeds are maize kernels.
3. The method according to any of claims 1-2, wherein the regression model employs a linear discriminate analysis (LDA).
4. The method according to any of claims 1-3, wherein the seeds to be sorted are of at least one induction cross.
5. The method according to any of claims 1-4, wherein the seeds to be sorted are of the same genotypic background.

6. The method according to any of claims 1-4, wherein the seeds to be sorted are of different genotypic backgrounds.

7. The method according to any of claims 3-6, wherein the accuracy of the classification step is improved by assigning a score to each seed following the LDA analysis, and setting a fraction threshold for the population of seeds based on: their scores and the predicted induction frequency; the size of the induction population; or both.

8. The method according to any of claims 1-7, wherein the apparatus further comprises an air valve, a transparent seed passage tube, a balance capable of measuring the weight of each seed, and fiber optics accompanying the spectrometer.

9. The method according to claim 8, wherein the apparatus further comprises an automated seed sorter.

10. The method according to claim 9, wherein the automated seed sorter is a solenoid sorter.

11. The method according to any of claims 1-10, wherein the light source emits near-infrared (NIR) light.

12. The method according to claim 11, wherein the light source emits light in a wavelength range of between 907 nm and 1689 nm.

13. The method according to any of claims 1-12, capable of differentiating at least one of the following traits of the population of seeds being sorted: relative oil content, relative protein content, relative starch content, seed density, and seed volume.

14. The method according to any of claims 1-13, wherein the time required for sorting each seed in the population is less than 1 second.

15. The method according to any of claims 1-13, wherein the time is required for sorting each seed is less than 0.25 second.

16. The method according to any of claims 1-13, wherein the time is required for sorting each seed is 0.25 second.

17. The method according to any of claims 3-16, wherein the LDA is implemented by an R-based statistical function capable of discriminating the haploid seeds from the diploid hybrid seeds.

18. An apparatus for sorting haploid seeds, the apparatus comprising:
a light source;
a spectrometer capable of examining single seeds individually;
an air valve;
a seed passage tube;
fiber optics accompanying the spectrometer; and
a machine-readable medium in operable communication with the spectrometer, the machine readable having machine-executable instructions stored thereon for performing a method comprising the following steps:

processing at least one spectrum of each seed, obtained from the spectrometer, with user-determined parameters;

storing the data of the at least one spectrum of each seed in a data matrix;
and

calibrating a regression model based on a portion of the data matrix that excludes a selected genotype of seeds, the genotype being considered a held-out genotype.

19. The apparatus according to claim 18, wherein the method performed by the machine-executable instructions further comprises:

cross-validating the regression model by holding out each genotype present in the population of seeds and repeating the classification step; and
classifying a randomly sampled seed using the validated model.

20 The apparatus according to any of claims 18-19, wherein the seed passage tube is a transparent seed passage tube.

21 The apparatus according to any of claims 18-20, further comprising a balance capable of measuring the weight of single seeds individually.

22 The apparatus according to any of claims 18-21, further comprising an automated seed sorter.

23. The apparatus according to claim 22, wherein the automated seed sorter is a solenoid sorter.

24 The apparatus according to any of claims 18-23, wherein the light source is a near-infrared (NIR) light source that emits NIR light.

25. The method according to claim 24, wherein the light source emits light in a wavelength range of between 907 nm and 1689 nm.

26 The apparatus according to any of claims 18-25, wherein the regression model employs a linear discriminate analysis (LDA).

27. The method according to claim 26, wherein the accuracy of the classification step is improved by assigning a score to each seed following the LDA analysis, and setting a fraction threshold for the population of seeds based on: their scores and the predicted induction frequency; the size of the induction population; or both.

28 The apparatus according to any of claims 26-27, wherein the LDA is implemented by an R-based statistical function capable of discriminating the haploid seeds from the diploid hybrid seeds.

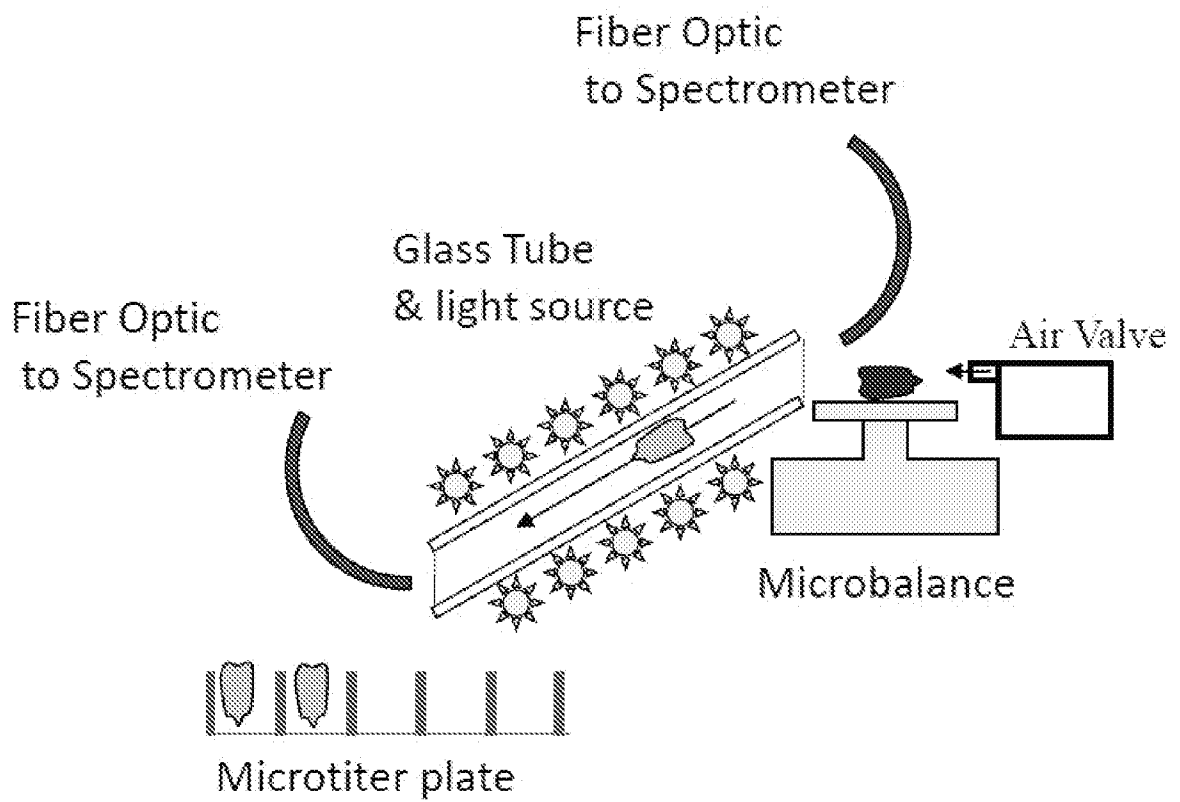


FIG. 1

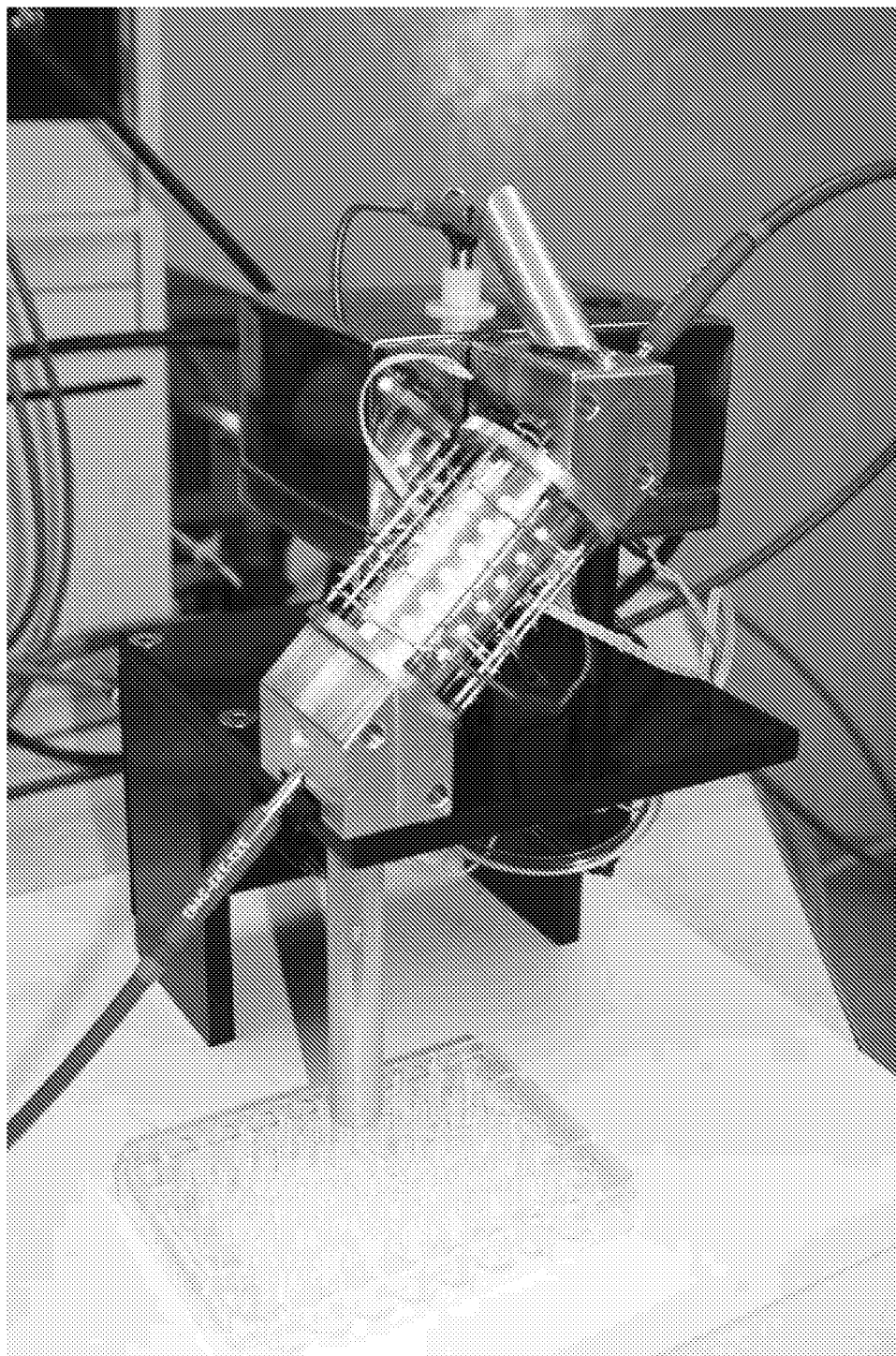


FIG. 2

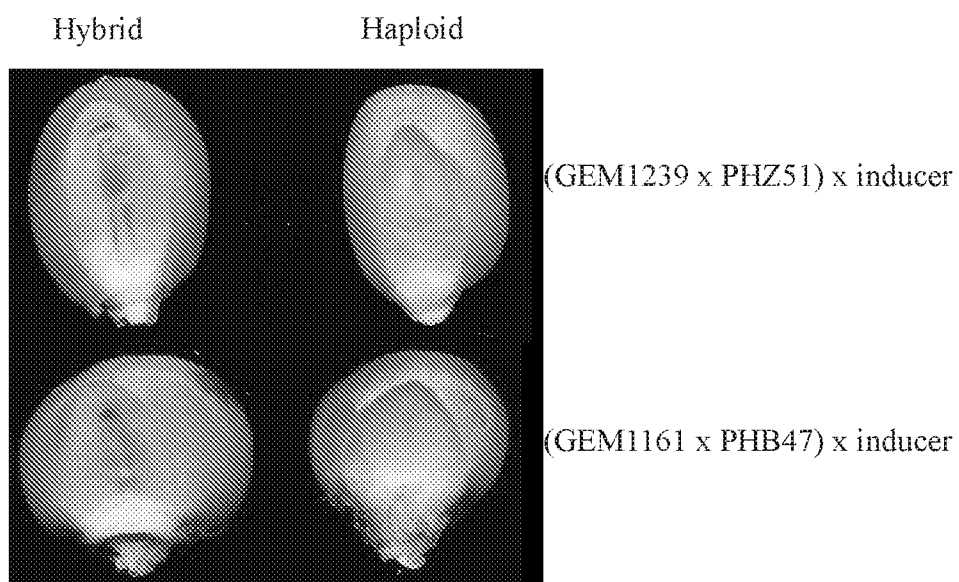


FIG. 3

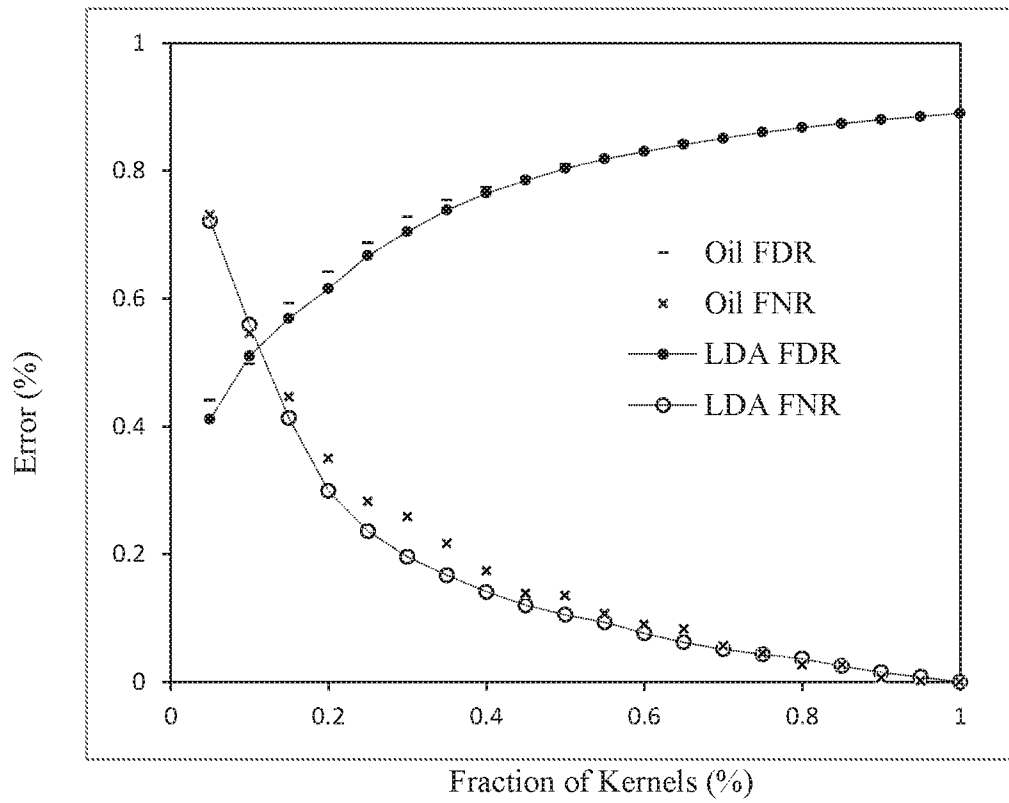


FIG. 4

A. CLASSIFICATION OF SUBJECT MATTER**G01N 21/359(2014.01)i, G01N 33/00(2006.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G01N 21/359; C12Q 1/02; B07C 5/34; B07C 5/36; G01N 21/35; A01C 1/00; B07C 5/02; A01H 5/10; G01N 33/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: sort, haploid seed, light source, spectrometer, regression model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 02-48687 A2 (MONSANTO TECHNOLOGY LLC) 20 June 2002 See page 4, lines 1-18, page 11, line 7 - page 26, line 29, page 31, line 19 - page 44, line 31, claims 41, 46, 58, 80, 111 and figures 1, 1a-1c, 2-3.	1-28
A	EP 2140749 A1 (AARHUS UNIVERSITET DET JORDBRUGSVIDENSKABELIGE FAKULTET et al.) 06 January 2010 See paragraphs [0088]-[0097] and figures 1-2.	1-28
A	US 2008-0035532 A1 (HUNTER et al.) 14 February 2008 See paragraphs [0097]-[0187] and figures 1, 2A-2C, 3A-3B, 4, 5A-5B, 6.	1-28
A	US 2011-0225680 A1 (COPE, JASON) 15 September 2011 See paragraphs [0172]-[0208] and figures 1, 4A-4D.	1-28
A	US 2011-0297589 A1 (BECKER et al.) 08 December 2011 See paragraphs [0040]-[0079] and figures 3-16.	1-28



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

12 June 2017 (12.06.2017)

Date of mailing of the international search report

12 June 2017 (12.06.2017)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea



Facsimile No. +82-42-481-8578

Authorized officer

LEE, Hun Gil

Telephone No. +82-42-481-8525



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/022310

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 02-48687 A2	20/06/2002	AR 031284 A1	17/09/2003
		AU 2002-043407 A1	24/06/2002
		BG 107842 A	30/01/2004
		BR 0115054 A	30/09/2003
		CA 2426933 A1	20/06/2002
		CA 2426933 C	24/01/2012
		CN 1484761 A	24/03/2004
		CN 1484761 C	04/06/2008
		EP 1332354 A2	06/08/2003
		EP 1332354 B1	21/08/2013
		ES 2431955 T3	28/11/2013
		HU 0302495 A2	28/10/2003
		HU 0302495 A3	28/02/2006
		JP 2004-515778 A	27/05/2004
		MX PA03003819 A	20/04/2004
		NO 20031925 A	26/06/2003
		PL 362509 A1	02/11/2004
		RU 2003-116064 A	20/02/2005
		RU 2288461 C2	27/11/2006
		UA 79083 C2	25/05/2007
		US 6646264 B1	11/11/2003
		WO 02-48687 A3	25/07/2002
		ZA 200303298 A	29/07/2004
EP 2140749 A1	06/01/2010	WO 2010-000266 A1	07/01/2010
US 2008-0035532 A1	14/02/2008	US 2002-0144458 A1	10/10/2002
		US 2004-0118754 A1	24/06/2004
		US 2008-0034652 A1	14/02/2008
		US 2008-0179226 A1	31/07/2008
		US 6706989 B2	16/03/2004
		US 7290665 B2	06/11/2007
		US 7588151 B2	15/09/2009
		US 7591374 B2	22/09/2009
		US 7905050 B2	15/03/2011
US 2011-0225680 A1	15/09/2011	BR PI0916984 A2	24/11/2015
		BR PI0917251 A2	10/11/2015
		CA 2732783 A1	25/02/2010
		CA 2735045 A1	25/02/2010
		CN 102165320 A	24/08/2011
		CN 102197309 A	21/09/2011
		EP 2316037 A2	04/05/2011
		EP 2316038 A2	04/05/2011
		US 2010-0044356 A1	25/02/2010
		US 2010-0050300 A1	25/02/2010
		US 2013-0302805 A1	14/11/2013
		US 2013-0309764 A1	21/11/2013
		US 7968282 B2	28/06/2011

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2017/022310

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011-0297589 A1	08/12/2011	US 8519297 B2	27/08/2013
		US 8535877 B2	17/09/2013
		US 8815500 B2	26/08/2014
		US 8907245 B2	09/12/2014
		WO 2010-022289 A2	25/02/2010
		WO 2010-022289 A3	27/05/2010
		ZA 201100680 B	27/06/2012
		ZA 201100681 B	27/06/2012
		BR 112012033133 A2	29/11/2016
		CN 102918377 A	06/02/2013
		CN 102918377 B	20/05/2015
		EP 2580568 A1	17/04/2013
		US 8833565 B2	16/09/2014
		WO 2011-156184 A1	15/12/2011
		ZA 201208552 B	31/07/2013