Chapter 12
Utilising NIRS for Qualitative and Non-destructive Identification of Seed Mutants in Large Populations

Johann Vollmann and Joanna Jankowicz-Cieslak

Abstract Phenotyping of large plant populations for genetic research or plant breeding is often time-consuming and expensive. Seed composition is a primary breeding objective as this determines quality for various markets, e.g. food, fodder and industrial processing. Near-infrared reflectance spectroscopy (NIRS) is a fast developing analytical tool for seed composition screening. For example, it is utilised in plant breeding programmes to predict compositional concentrations in various samples. NIRS can be used to detect variation between seed lots and between individual seeds and can be used to identify and isolate new phenotypes including mutants based on spectroscopic sample properties. Spectral data of seed samples may be subjected to principal component analysis to separate groups and individuals with distinct compositional properties. Spectroscopic outliers such as mutants with novel seed quality alleles may then be selected based on principal component scores. Outliers represent a small subset of the entire population, and these may be subject to more rigorous analyses (chemical, physiological and genetic). In determining their potential exploitation, NIRS is a high-throughput phenotyping platform that can be used to reduce large sample sizes, e.g. a mutant population to manageable numbers.

Keywords Near-infrared reflectance spectroscopy • Spectroscopic fingerprinting • Outlier identification • Mutant preselection • High-throughput phenotyping

J. Vollmann (✉)
Plant Breeding Division, Department of Crop Science, University of Natural Resources and Life Sciences Vienna (BOKU), Tulln an der Donau, Austria
e-mail: johann.vollmann@boku.ac.at

J. Jankowicz-Cieslak
Plant Breeding and Genetics Laboratory, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, IAEA Laboratories Seibersdorf, International Atomic Energy Agency, Vienna International Centre, PO Box 100, 1400 Vienna, Austria

12.1 Introduction

Field, glasshouse and laboratory-based phenotyping of large and genetically diverse populations are often time-consuming and costly and are considered a major bottleneck in plant genetic improvement (Araus and Cairns 2014). Rapid developments in genotyping have generated massive data banks, but this has not been matched by phenotyping; this has led to the so-called phenotype gap which limits our ability to link genes to traits. Near-infrared reflectance spectroscopy (NIRS) is a fast analytical method based on reflection/absorption from functional chemical groups in the near-infrared region of light (about 800–2500 nm wavelength range). This is suitable to predict the concentration of major organic components in seed or other biological materials without requiring extensive sample preparation. Non-destructive NIRS methods enable rapid prescreening that may be performed on seed prior to field or glasshouse propagation. While most NIRS applications are implemented in agriculture, food and environmental industries, it is also becoming popular in pharmaceutical and medical fields due to improvements in instrumentation and in statistical analysis (Bosco 2010). Owing to the low per-sample cost and speed of measurement and analysis, NIRS has the potential to serve as a component of high-throughput phenotyping platforms in various crop improvement and quantitative genetics studies.

In plant breeding, NIRS has been utilised in two ways. Quantitative analysis involves the prediction of concentrations of identified components. In this approach, calibration standards need to be developed to assign measured spectra to specific components accurately (e.g. Sato et al. 2012; Xie et al. 2014). The second major approach is qualitative analysis. This involves the classification of samples according to their spectroscopic properties. Samples are clustered according to their similarity/dissimilarity (Munck 2007). This is based on "spectroscopic fingerprints" of each sample. Different statistical criteria can be applied to separate outliers from "normal" samples. This approach is ideally suited to projects involving large populations where outliers are expected at a rare frequency, such as with chemically or irradiated mutant plant populations. Low densities of induced mutations typically mean that many thousands of plants need to be screened to identify rare variants with the desired improved trait (Jankowicz-Cieslak et al. 2011). Moreover, in contrast to quantitative analysis for predicting concentrations of individual analytes, qualitative NIRS analysis does not require reference samples and reference chemistry analysis for the development of calibrations.

Apart from genotype identification and verification based on seed (Turza et al. 1998; Wu et al. 2008), NIRS can be used in the identification of transgenic food materials (Alishahi et al. 2010) or seeds, such as separating RoundupReady® from non-GMO soybeans (Esteve Agelet et al. 2012). NIRS can also be used in mutant identification, for example, barley endosperm mutants such as those expressing high lysine or high and low beta-glucan can be differentiated from normal barley genotypes based on their chemometric patterns (Jacobsen et al. 2005). In wheat, starch mutants carrying non-functional alleles in the amylose

synthesis gene resulting in the "waxy" phenotype can be detected by NIRS (Delwiche et al. 2011). A major step forward in NIRS is the analysis of single seeds, e.g. NIRS analysis can distinguish individual amylose-free seeds in both hexaploid and tetraploid wheat (Dowell et al. 2009). Thus, waxy wheat seeds could be selected in segregating populations, for purifying advanced breeding lines or for mutant identification and isolation. In maize, prediction of constituents such as starch, oil or protein concentration and seed weight is feasible on individual seeds through classical calibrations based on partial least square regression statistics, while oil, starch and protein mutant phenotypes can be identified from principal component analysis of spectral data, implemented using single seed glass tube NIR spectrometer. The design of the instrument enables high-throughput data collection and is of great interest for single seed-based selection, genetic screening and seed phenomics (Spielbauer et al. 2009).

Evidence from different crop species demonstrates the potential of utilising NIRS spectral data for classifying seed samples according to spectral similarity. A major concern in identifying mutants (traditionally mutagenised by seed treatments with chemical or physical mutagens) is their low frequency; usually thousands of individuals or lines must be evaluated to identify the rare novel phenotypes of interest. We have developed a high-throughput preselection method utilising qualitative NIRS analysis of rice seed in which rare and novel phenotypes can be identified. Since the outliers are from the same population and share a common and highly homogenous genetic background, any change is easily picked up and is a potential mutant that may be validated by further analyses. A practical, user-friendly method for NIRS-based screening of mutant seed populations is given below for spectroscopic outlier detection.

12.2   Materials

12.2.1   Equipment and Hardware

1. NIRS equipment, e.g. Bruker Matrix-I FT-NIR machine (*see* **Note 1**)
2. Mill, equipped with grid nets for uniform grinding, e.g. CT 1093 Cyclotec Sample Mill (FOSS, Sweden)
3. Funnel
4. Spoon
5. Brush
6. Small scale vacuum cleaner
7. Sample glass cuvettes compatible with NIRS equipment

12.2.2 Software

1. OPUS 7.5 (Bruker, Ettlingen, Germany) or software compatible with NIRS equipment
2. Unscrambler® (Camo Software AS, Oslo, Norway)
3. Standard spreadsheet software for data handling

12.2.3 Plant Materials

1. Mutant populations (*see* **Note 3**).
2. Dry seeds with moisture content below 14 % (*see* **Notes 4 and 5**).

12.3 Methods

12.3.1 Seed Preparation

1. Dry mature seed *per* mutant population or line to have moisture content below 14 %, in a standard desiccator or a dry room under ambient conditions (*see* **Note 5**).
2. Grind seed to a fine powder (*see* **Note 6**).

### 12.3.2 NIRS Analysis

12.3.2.1 Destructive Approach

1. For each accession or a mutant line, grind 3–5 g of dry seeds (*see* **Note 7**).
2. Disassemble the mill and clean it with a vacuum cleaner and brush so that no particles of the previous sample remain (*see* **Note 8**).
3. Fill the sample containers with the milled sample according to manufacturer's instructions.
4. Scan sample in the NIRS instrument to collect and record spectral reflectance characteristics of the samples (*see* **Note 9**).
5. Repeat the measurement 2–4 times, turning the sample cup or mixing the sample each time (*see* **Note 10**).
6. Perform statistical analysis (*see* **Notes 11–15**, Figs. 12.1, 12.2 and 12.3).
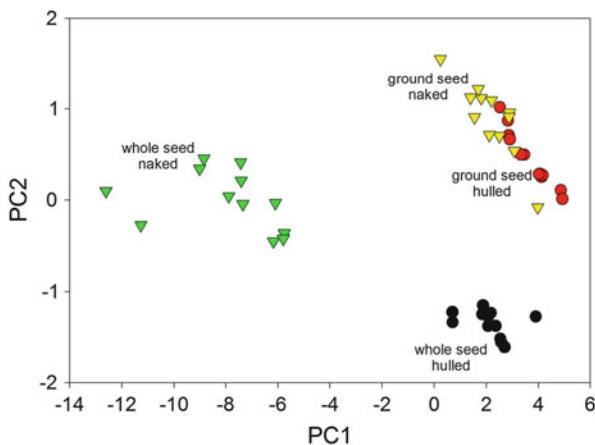
Fig. 12.1 Classification of a test set of 12 rice mutant lines based on principal component analysis of NIRS spectra as influenced by sample pretreatment such as dehulling or grinding of samples
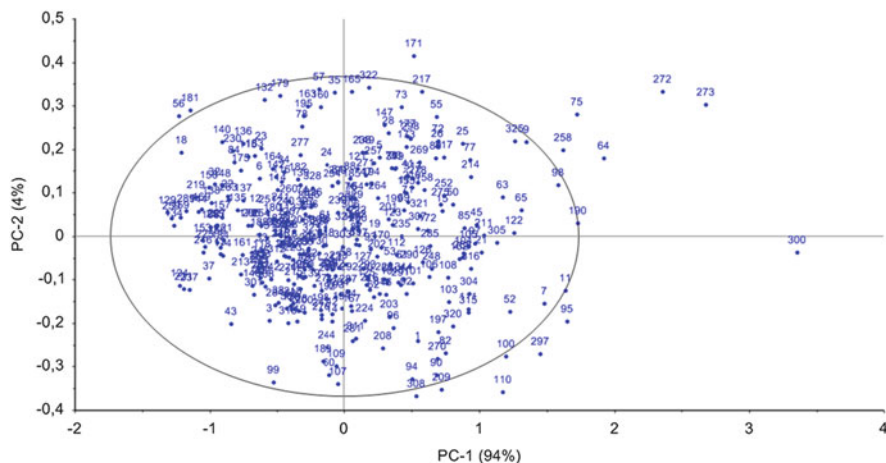

Fig. 12.2 Classification of a mutant population of 329 rice samples using PC analysis of spectroscopic data. Spectroscopic outliers are visualised as samples outside the boundaries of a Hotelling's T2 statistics ellipse

12.3.2.2   Non-destructive Approach

1. For each accession/mutant line prepare 3–5 g of dry seeds.
2. Fill sample containers with milled samples according to manufacturer's instructions.
3. Scan sample in the NIRS instrument to collect and record spectral reflectance characteristics of the samples (*see* **Note 9**).
4. Repeat the measurement 2–4 times, turning the sample cup or mixing the sample prior to each measurement. If the amount of sample material permits, use a new portion for each measurement.

Fig. 12.3 Spectroscopic classification of 329 rices mutant lines with the top 5 % of samples highest in seed protein content highlighted in *red*
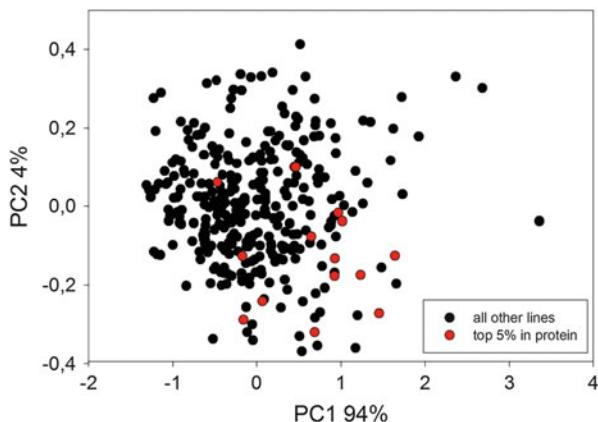
5. Keep the seed for future analysis or for mutant multiplication.
6. Perform statistical analysis (*see* **Notes 11–15,** Fig. 12.1, Fig. 12.2 and Fig. 12.3).

12.3.3   Statistical Analysis

1. Download spectroscopic data from the NIRS computer.
2. Import sample spectra to Unscrambler or other statistical software suitable for multivariate data analysis.
3. Carry out principal component analysis for data reduction and clustering of samples (*see* **Notes 12–15**).
4. Save PCA scores calculated for each sample of the population in a spreadsheet software programme.
5. Use PCA scores to plot *x–y* graphs of the samples to visualise clustering (similar phenotypes) and outliers (unique phenotypes).

12.4   Notes

1. Various types of NIRS instruments with different technologies are available on the market at present. Newer models often have better performance so that more spectral information can be recovered from a sample. Among important properties are resolution, light throughput, wavenumber accuracy, repeatability and signal-to-noise. Beside the near-infrared region (typically 800–2500 nm), some instruments have an extended wavelength range (e.g. 400–2500 nm) including the range of the visible spectrum of light; this may be of interest in measuring colour differences which can be caused by variation in carotenoids, anthocyanins or other components of seed samples. NIRS instruments also differ in

sample presentation methods which have an influence on surface reflection, particle size and other effects; depending on sample presentation modes, seed samples might either need to be finely ground, might be scanned non-destructively or analysed as single intact seeds. Here we provide Bruker Matrix-I as an example of a modern FT-NIR instrument which is very well suited for combined single seed and bulk material analysis.

2. Proprietary software is required for most instruments in order to carry out basic instrument operations such as optical adjustments or collection of spectral data from samples. Programmes are normally available to assist in the development and validation of chemometric calibrations from sample sets with reference data, or they can be used to predict seed composition from external calibrations. These are commercially available for major crop species and include the most important grain components such as for instance starch, protein and pigments.

3. Mutant populations should be developed following established procedures (Lee et al. 2014). It is important to note that the first (M_1) generation after mutagenesis is chimeric, and most mutations will not be heritable, also the M_1 generation suffers from physiological disorders as a result of mutagenic treatments. Therefore, mutation screening should not be carried out until the M_2 or subsequent generations. Seed weight and number are important considerations when selecting the generation to screen. One should select a generation where there is at least 3–5 g of seed *per* line.

4. In quantitative NIRS analysis for predicting particular analytes using calibrations, reference samples need to be collected from different environments (e.g. locations, growing seasons) to cover environmental variation. In qualitative NIRS analysis for classifying samples, spectral variation caused by environmental effects might bias classification results. Therefore, samples to be classified have to be grown in the same environment (field location, greenhouse) to avoid environmental effects. Moreover, repeated controls have to be used to be able to monitor and estimate environmental variation effects. A dry seed with moisture content below 14 % is a suitable material for NIRS analysis. Samples differing in water content cannot be analysed, as very broad water peaks in certain regions of the NIRS spectrum mask useful spectral variation between samples.

5. Generally, seed of cereal crops can be stored at a moisture level of 14 % or below. In dry environments, this level is reached at full maturity; in moist or high-humidity environments, some postharvest drying may be necessary. Moisture level of 14 % or below is reached when single seeds cannot be indented with finger nails.

6. Depending on seed size and the sample presentation method of the NIRS instrument, grinding of seeds into a homogeneous powder or replicated measurements (e.g. mix by inverting the vial 2–4 times) of the sample might be needed for avoiding surface reflection and to obtain representative and reproducible spectral data. Larger amounts of sampling materials measured in higher number of replications might avoid the need for grinding thus enabling non-destructive measurements in nonhomogenous (i.e. large seeded) samples. However, non-destructive measurement vs. grinding or dehulling of samples has a strong effect on spectroscopic sample classification, as shown in Fig. 12.1.

7. The method described here is a destructive method that involves grinding the entire seed (embryo and endosperm) to make the measurement. In advanced lines, induced mutations should segregate in a Mendelian fashion. In order to recover mutations identified in NIRS, one should retain 20 or more seeds *per* line.

8. Cross-over contamination from previous measurements can lead to errors.

9. Depending on the NIRS instrument and the measuring mode used, scanning a sample takes between 30 s and 2 min.

10. For qualitative NIRS analysis to classify samples, general statistical software packages suitable for data reduction and multivariate analysis could be utilised. A machine-independent software package dedicated to spectral data analysis (spectra pretreatment and transformation, calibration development and validation, multivariate classification, etc.) is Unscrambler® (Camo Software AS, Oslo, Norway), which is widely used in multivariate analysis of spectroscopic data.

11. In spectroscopic data analysis for sample classification, specific data pretreatment methods such as multiplicative scatter correction or standard normal variate transformation are used to adjust for radiation-scattering effects. Various other data pretreatment methods such as first or second derivative functions might be applied as well for reducing data noise and enhancing sample signals on a trial-error basis, but data pretreatment might cause information loss thus reducing the discrimination power of spectral data.

12. Spectral data are subject to principal component analysis (PCA), and PCA scores for samples are calculated and further used in score plots to visualise classification results. Either full spectra or specific wavelength regions can be utilised to calculate PCA models. PCA classification models can be validated, and models can be selected based on the level of total variance explained. In mutant populations, spectroscopic outliers can subsequently be detected based on the distance to untreated control genotypes of the same genetic background. In Fig. 12.2, 329 rice samples derived from a mutant population are classified by PCA analysis of spectroscopic data; outliers are visualised as samples outside the boundaries of a Hotelling's T2 statistics ellipse which could be subject to further analysis.

13. In sample populations segregating for qualitative characters such as the waxy wheat trait, samples with known group membership (e.g. waxy vs. normal starch) could be used as anchors to classify unknown samples or for developing discriminant analysis functions based on the PCA scores.

14. If calibrations for quantitative determination of individual analytes are available, they might be combined with spectroscopic classifications to gain additional information about individual samples. In Fig. 12.3, the sample set of Fig. 12.2 (329 rices samples, mutant population) was subject to quantitative analysis of seed protein content: The majority of the 5 % of samples with highest protein content are appearing in the lower right section of the scatter plot as highlighted. Using quantitative calibrations, the concentrations of various other seed components of interest can be calculated and considered for further selection steps, as given in Table 12.1 for the highlighted samples of Fig. 12.3.

Table 12.1 Protein content and other parameters of the top 5 % in protein content of Fig. 12.3 rice mutant population as compared to the control mean

Line	Ash	Fat	Fibre	Protein	Moisture
281	4.58	5.19	3.50	13.45	9.96
52	4.20	5.78	4.89	13.52	8.78
27	4.20	5.74	4.07	13.65	9.24
189	4.00	5.92	4.90	13.78	9.78
97	4.52	5.79	5.94	13.80	9.33
315	4.08	5.73	4.71	14.12	8.29
241	4.90	5.52	4.69	14.24	9.68
11	4.55	6.45	6.18	14.41	7.78
297	4.61	5.39	4.34	14.60	9.92
21	4.31	5.88	5.73	14.63	9.43
304	3.82	5.97	4.59	14.76	10.29
248	4.35	5.96	4.68	15.30	10.13
90	4.75	5.96	3.73	16.34	9.15
17	4.80	6.10	5.45	16.55	9.45
Control mean	3.75	4.94	5.51	11.25	4.65

15. False detections: In addition to genotype, the composition of a seed is influenced by the environment in which it develops. Environmental effects could be at the scale of the plant environment or the micro-environment in which the seed develops. For example seeds that develop at the extremities of an inflorescence may not receive the same nutrient supplies as those that develop in central locations and have better links to the vascular system; seed composition may also be affected by pests and diseases. Such affected seeds may be detected as having abnormal NIR spectra. Measures can be taken to discard such "off-type" seeds, e.g. they are often small or deformed and removed by sieving. Or they can be included in the preselections and retested in subsequent generations.

References

Alishahi A, Farahmand H, Prieto N, Cozzolino D (2010) Identification of transgenic foods using NIR spectroscopy: a review. Spectrochim Acta Part A 75:1–7

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19:52–61

Bosco GL (2010) James L. Waters Symposium 2009 on near-infrared spectroscopy. Trends Anal Chem 29:197–208

Delwiche SR, Graybosch RA, Amand PS, Bai G (2011) Starch waxiness in hexaploid wheat (*Triticum aestivum* L.) by NIR reflectance spectroscopy. J Agric Food Chem 59:4002–4008

Dowell FE, Maghirang EB, Graybosch RA, Berzonsky WA, Delwiche SR (2009) Selecting and sorting waxy wheat kernels using near-infrared spectroscopy. Cereal Chem 86:251–255

Esteve Agelet L, Gowen AA, Hurburgh CR, O'Donell CP (2012) Feasibility of conventional and roundup ready (R) soybeans discrimination by different near infrared reflectance technologies. Food Chem 134:1165–1172

Jacobsen S, Søndergaard I, Møller B, Desler T, Munck L (2005) A chemometric evaluation of the underlying physical and chemical patterns that support near infrared spectroscopy of barley seeds as a tool for explorative classification of endosperm genes and gene combinations. J Cereal Sci 42:281–299

Jankowicz-Cieslak J, Huynh OA, Bado S, Matijevic M, Till BJ (2011) Reverse-genetics by TILLING expands through the plant kingdom. Emirates J Food Agric 23:290–300

Lee LS, Till BJ, Hill H, Huynh OA, Jankowicz-Cieslak J (2014) Mutation and mutation screening. In: Robert JH, Furtado A (eds) Cereal genomics: methods and protocols, methods in molecular biology, vol 1099. Springer, New York, pp. 77–95

Munck L (2007) A new holistic exploratory approach to systems biology by near infrared spectroscopy evaluated by chemometrics and data inspection. J Chemometr 21:406–426

Sato T, Zahlner V, Berghofer E, Losak T, Vollmann J (2012) Near-infrared reflectance calibrations for determining sucrose content in soybean breeding using artificial reference samples. Plant Breed 131:531–534

Spielbauer G, Armstrong P, Baier JW, Allen WB, Richardson K, Shen B, Settles AM (2009) High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. Cereal Chem 85:556–564

Turza S, Tóth ÁI, Váradi M (1998) Multivariate classification of different soyabean varieties. J. Near Infrared Spectrosc 6:183–187

Wu D, Feng L, He Y, Bao Y (2008) Variety identification of Chinese cabbage seeds using visible and near-infrared spectroscopy. Trans ASABE 51:2193–2199

Xie LH, Tang SQ, Chen N, Luo J, Jiao GA, Shao GN, Wei XJ, Hu PS (2014) Optimisation of near-infrared reflectance model in measuring protein and amylose content of rice flour. Food Chem 142:92–100