Received: 5 May 2021 | Accepted: 4 February 2022

DOI: 10.1002/ppj2.20040

The Plant Phenome Journal OPEN ACCESS

SPECIAL SECTION: BELOWGROUND PHENOTYPING

Low-cost, handheld near-infrared spectroscopy for root dry matter content prediction in cassava

Jenna Hershberger[1] | **Edwige Gaby Nkouaya Mbanjo**[2] | **Prasad Peteti**[2] | **Andrew Ikpan**[2] | **Kayode Ogunpaimo**[2] | **Kehinde Nafiu**[2] | **Ismail Y. Rabbi**[2] | **Michael A. Gore**[1]

[1] Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell Univ., Ithaca, NY 14853, USA

[2] International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria

Correspondence
Jenna Hershberger, Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell Univ., 353 Plant Science Building, Ithaca, NY 14853, USA.
Michael A. Gore, Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell Univ., 358 Plant Science Building, Ithaca, NY 14853, USA.
Email: jmh579@cornell.edu; mag87@cornell.edu

Assigned to Associate Editor Saoirse Tracy.

Funding information
United States Agency for International Development Cooperative Agreement No. 7200AA-19LE-0 Foreign, Commonwealth and Development Office NextGen Cassava Bill and Melinda Gates Foundation NextGen Cassava INV-007637 National Institute of Food and Agriculture EWD Predoctoral Fellowship 2019-67011-29606 National Science Foundation BREAD IOS-1543958 Cornell Institute for Digital Agriculture, Cornell University (CIDA)

Abstract

Over 800 million people across the tropics rely on cassava (*Manihot esculenta* Crantz) as a major source of calories. While the root dry matter content (RDMC) of this starchy root crop is important for both producers and consumers, characterization of RDMC by traditional methods is time-consuming and laborious for breeding programs. Alternate phenotyping methods have been proposed but lack the accuracy, cost, or speed ultimately needed for cassava breeding programs. For this reason, we investigated the use of a low-cost, handheld near-infrared spectrometer (740–1070 nm) for field-based RDMC prediction in cassava. Oven-dried measurements of RDMC were paired with 21,044 scans of roots of 376 diverse genotypes from 10 field trials in Nigeria and grouped into training and test sets based on cross-validation schemes relevant to plant breeding programs. Mean partial least squares regression model performance ranged from R^2_{P} = 0.62–0.89 for within-trial predictions, which is within the range achieved with laboratory-grade spectrometers in previous studies. Relative to other factors, model performance was highly affected by the inclusion of samples from the same environment in both the training and test sets. With appropriate model calibration, the tested spectrometer will allow for field-based collection of spectral data with a smartphone for accurate RDMC prediction and potentially other quality traits, a step that could be easily integrated into existing harvesting workflows of cassava breeding programs.

Abbreviations: CV, cross-validation; IITA, International Institute of Tropical Agriculture; NIRS, near-infrared spectroscopy; PLSR, partial least squares regression; R^2_{CV}, coefficient of multiple determination of cross-validation; R^2_{P}, squared Pearson's correlation between predicted and observed values in a test set; RDMC, root dry matter content; RF, random forest; $\mathrm{RMSE}_{\mathrm{CV}}$, root mean squared error of cross-validation; $\mathrm{RMSE}_{\mathrm{P}}$, root mean squared error of prediction; SVM, support vector machine.

1 | INTRODUCTION

Cassava (*Manihot esculenta* Crantz) is a starch-rich industrial root crop and staple food for hundreds of millions of people throughout the tropics (Howeler et al., 2013). Although cassava is especially valued for its role as a subsistence crop, as it is able to withstand environmental conditions that other crops cannot and can be relied upon in times of drought and food scarcity (El-Sharkawy, 1993, 2004), its role in commercial farming and industrial processing is increasing globally (Parmar et al., 2017). Though more than half of global cassava production occurs in Africa, yields on the continent remain far below the global average (FAO, 2021). While a portion of this yield gap is likely due to management factors such as differences in fertilizer applications, low rates of improved germplasm adoption also contribute to this gap (Fermont et al., 2009; Njine, 2010; Oparinde et al., 2016; Owusu & Donkor, 2012).

As is true for other crops, this lack of improved cultivar adoption is largely attributed to the need for breeding objectives to place more weight on selection for end user preferred traits (Acheampong et al., 2018; Asrat et al., 2010). Early breeding objectives focused heavily on the development of cultivars resistant to diseases that threatened to wipe out cassava production across the African continent (Hahn et al., 1980; International Institute of Tropical Agriculture [IITA], 1990; Jennings, 1957), but following success in those efforts, emphasis has since shifted to breeding for yield and culturally relevant quality traits. Root dry matter content (RDMC), a major component of both dry yield and food quality that is of high importance along the entire value chain (Okechukwu & Dixon, 2008; Teeken et al., 2018, 2020), is now directly included in the selection indices used by major cassava breeding programs including those implementing genomic selection (Ceballos et al., 2012; Kawano, 2003; Kawuki et al., 2011; Wolfe et al., 2016).

Traditional methods of cassava RDMC characterization are time-consuming and laborious, and therefore unsuitable for phenotyping the large number of plots at the early stages of selection. Oven drying, the gold standard phenotyping method for RDMC, is not only tedious but also requires a cheap and stable source of heat energy. However, this is not available in the majority of off-station testing locations, thereby necessitating transportation of roots to centralized laboratories (Safo-Kantanka & Owusu-Nipah, 1992; Teles et al., 1993; Teye et al., 2011). To work around these obstacles, researchers introduced a simple linear regression equation to relate specific gravity to RDMC (Kawano et al., 1987), a method that can be performed directly in the field with just a scale and basin of water. However, due to the relatively thick peels of cassava roots and an inability to wash them thoroughly in the field, this method is often inaccurate (Ikeogu et al., 2017; Pérez et al., 2011), limiting its potential benefits.

Core Ideas

- A low-cost, handheld near-infrared spectrometer was tested for phenotyping of cassava roots.
- Plant breeding-relevant cross-validation schemes were used for predictions.
- High prediction accuracies were achieved for cassava root dry matter content.

More recently, near-infrared spectroscopy (NIRS) was introduced to cassava breeding as an alternative method for RDMC phenotyping (Sánchez et al., 2014). With the correct model calibration, NIRS offers highly accurate estimates of RDMC in a fraction of the time of oven drying, potentially increasing the throughput without proportional rises in cost. Sánchez et al. found the Foss 6500, a lab-grade, benchtop visible and NIR spectrometer, to be highly predictive (squared Pearson's correlation between predicted and observed values in a test set [R^2_{P}] = 0.79–0.95) of RDMC in cassava over several growing seasons. While the use of a benchtop spectrometer saves effort compared with oven drying, it still requires root transport and extensive sample preparation, ultimately resulting in an equivalent amount of labor to the oven method of RDMC determination. This study was followed by an evaluation of the ASD QualitySpec Trek (Malvern Panalytical), a mobile visible and NIR spectrometer, that found it to be sufficiently accurate for RDMC prediction in breeding programs (Ikeogu et al., 2017). Although this instrument can be taken to the field and does not require samples to be shredded or blended before scanning, its high cost limits accessibility for breeding programs with more rigid budget constraints.

To truly meet the needs of cassava breeding programs, there is a need for a spectrometer that is accurate, field-based, and low-cost. Recent developments in NIRS technology have resulted in smaller, light-weight devices that require minimal sample preparation, effectively allowing the user to bring the laboratory to the sample (see Table 1 in Teixeira Dos Santos et al., 2013). Many of these mobile spectrometers (e.g., ConsumerPhysics: SCiO, 740–1070 nm; Stratio: LinkSquare, 400–1000 nm; Tellspec Inc.: Tellspec, 900–1,700 nm; Felix Instruments Inc.: F-750 Produce Quality Meter, 310–1,100 nm; Allied Scientific Pro: NIRVascan, 900–1,700 nm) are also considerably less expensive than their predecessors, which could facilitate their adoption by breeding programs with limited funding. Scans from one such mobile spectrometer, the SCiO, have been found to be highly predictive of dry matter content and other quality traits

TABLE 1 Metadata for 10 International Institute of Tropical Agriculture cassava field trials as planted

Abbreviated trial name	Cassavabase trial name	Trial type[a]	Planting date	Harvest date	Experimental design[b]	# Replicates	Total # unique genotypes	# Check genotypes	# Plants per plot	Plot dimensions (m)
A-17IB	17.CASS.PYT.49.setA.IB	PYT	21 Apr. 2017	9 May 2018	RCBD	2	49	4	10	2 × 4
B-17IB	17.GS.C3.PYT.80.IB	PYT	11 May 2017	17 May 2018	Alpha-lattice	2	80	6	10	2 × 4
C-18IB	18.CASS.PYT.52.IB	PYT	5 July 2018	5 July 2019	RCBD	2	52	4	36	6 × 4.8
D-18IB	18.GS.C2.setA.UYT.36.IB	UYT	19 July 2018	5 July 2019	Alpha-lattice	3	36	5	42	6 × 5.6
E-18IB	18.GS.C2.setB.UYT.36.IB	UYT	21 Jan. 2018	19 July 2019	Alpha-lattice	3	36	5	42	6 × 5.6
F-19IB	19.CMSSurveyVarieties.AYT.33.IB	AYT	29 Apr. 2019	20 Apr. 2020	RCBD	2	33	5	20	4 × 4
G-19IB	19.GS.C2.UYT.36.setA.IB	UYT	16 May 2019	20 Apr. 2020	RCBD	2	36	5	20	4 × 4
H-19IB	19.GS.C2.UYT.36.setB.IB	UYT	16 May 2019	28 Apr. 2020	RCBD	2	36	5	20	4 × 4
I-19IK	19.CASS.PYT.52.IK	PYT	25 June 2019	20 July 2020	RCBD	2	52	4	20	4 × 4
J-19IK	19.GS.C4B.PYT.500.IK	PYT	4 Aug. 2019	27 Oct. 2020	Alpha-lattice	2	452	5	9	3 × 2.4

[a] AYT = Advanced yield trial; PYT = Preliminary yield trial; UYT = Uniform yield trial.

[b] RCBD = randomized complete block design.

in several plant systems (Kaur et al., 2017; Kosmowski & Worku, 2018; Wiedemair & Huck, 2018; Li et al., 2018; Subedi & Walsh, 2020), indicating that the SCiO may be useful for the prediction of cassava RDMC. The application of this technology in a cassava breeding context has the potential to boost the throughput of RDMC phenotyping without increasing the cost or time investment per sample, but to our knowledge, the use of lower cost, mobile NIR spectrometers have not been reported for quantification of RDMC in cassava.

In this study, we developed and evaluated a new phenotyping procedure to predict cassava RDMC in an active breeding program. The main objectives of the study were to (a) assess a low-cost, handheld NIR spectrometer for the prediction of cassava RDMC and (b) evaluate the utility of collected NIRS data in the context of a cassava breeding program, developing best practices for routine use.

2 | MATERIALS AND METHODS

2.1 | Plant materials and experimental design

To test spectrometer and model performance within and across populations and environments, we evaluated 10 field experiments (Table 1). These experiments were representative of field trials commonly used to test the genetic potential of cassava clones (hereafter genotypes) for a range of phenotypes, containing genotypes of varying levels of improvement, relatedness, and RDMC. A complete list of check and non-check genotypes and corresponding phenotypic data are available through Cassavabase (www.cassavabase.org).

Eight of the field trials were planted at IITA in Ibadan, Nigeria (A-17IB, B-17IB, C-18IB, D-18IB, E-18IB, F-19IB, G-19IB, H-19IB), while two were planted in Ikenne, Nigeria

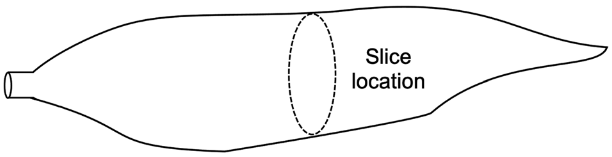

FIGURE 1 Roots from trials A-17IB, B-17IB, and C-18IB were sliced crosswise in the central region as shown and six scans were taken on each side of the cut surface

(I-19IK and J-19IK). All trials were planted with 1-m interrow spacing and 0.8-m alleys at the end of each plot. Phenotypic data were collected on two replicates of each non-check genotype in each trial.

A-17IB, C-18IB, and I-19IK included genotypes selected from the genetic gain germplasm collection, a population of historically important cassava genotypes from IITA, as previously described by Okechukwu and Dixon (2008), Ly et al. (2013), and Wolfe et al. (2016). F-19IB included genotypes sampled from a survey of popular cultivars in Nigeria. Genotypes in all other trials were derived from various cycles of genomic selection originating from the genetic gain population: Trials D-18IB, E-18IB, G-19IB, and H-19IB consisted of genotypes from the second cycle, B-17IB genotypes were selected from the third cycle, and J-19IK included genotypes from the fourth cycle.

2.2 | Dry matter determination and spectral data acquisition

The SCiO molecular sensor (Consumer Physics Inc.) is a handheld, portable NIR spectrometer that connects to the SCiO Lab smartphone application via Bluetooth for spectral data acquisition. Using an active light source, it measures diffuse reflectance from 740 to 1,070 nm, providing values in 1-nm intervals. Each of these measurement scans takes approximately 2 s and is immediately transferred to the SCiO server for data storage. Prior to each use, calibration was performed using the built-in reference standard in the SCiO case. A manufacturer-provided plastic "light shield" was attached to the spectrometer for all scans to block ambient light and to maintain a consistent 9-mm distance from the sensor to the sample surface during the scanning process.

Six commercial-sized roots were collected on a plot basis for all trials and phenotyped on the day of harvest. In A-17IB, B-17IB, and C-18IB, roots were sliced in half and immediately scanned with the SCiO three times on each cut surface for a total of six scans per root and 36 scans per plot (Figure 1, Supplementary Figure S1). Immediately after scanning, individual roots were peeled, shredded, and a single, homogenized sample of 100 g was oven-dried at 80 ˚C until reaching a constant weight (48 h). Percent RDMC was calculated as

$$\text{RDMC} = \frac{\text{dry weight}}{\text{fresh weight}} \times 100\% \tag{1}$$

In the seven remaining trials, scans were taken immediately after grating the root to obtain a homogeneous plot-level sample. Three to ten subsamples of homogenized fresh root tissue were then immediately placed in a quartz glass container and scanned with the SCiO six times. Percent RDMC was obtained using the same oven-drying method as the other three trials. Individual field plots with incomplete spectral or RDMC data were removed from the analysis, at times resulting in a single replicate of a given genotype in the final dataset.

2.3 | Outlier removal and sample aggregation

After data collection, all scans were filtered according to Mahalanobis distance. Samples with Mahalanobis distances greater than a cutoff set by a χ^2-distribution with 331 degrees of freedom ($\alpha = 0.05$) were removed from the analysis (Johnson & Wichern, 2007) using the *waves* R package version 0.1.1 (Hershberger et al., 2021) in R version 3.5.2 (R Core Team, 2018). In total, eight scans were removed through this procedure (Supplemental Table S1). After outlier removal, sample aggregation by means, as typical for NIRS studies (Ikeogu et al., 2017; A. Kaur et al., 2020; Lebot et al., 2009; Sánchez et al., 2014), was performed to enable RDMC prediction on a plot-level basis (i.e., the experimental unit) for all trials.

2.4 | Spectral pretreatment

Twelve combinations of common pretreatment methods including standard normal variate (Barnes et al., 1989), first and second derivatives, and Savitzky-Golay polynomial smoothing (Savitzky & Golay, 1964) were applied using the R package *waves* version 0.1.1 (Hershberger et al., 2021). No clear differences in model performance were found between raw and pretreated data using any of the pretreatment methods for within-trial predictions (Supplemental Figure S3, Supplemental Table S2), thus raw spectral data were used for all subsequent analyses.

2.5 | Cross-validation (CV)

A Monte Carlo CV scheme was developed for within-trial predictions (Xu et al., 2001). In this scheme, samples were separated into training and test sets (70:30) using stratified

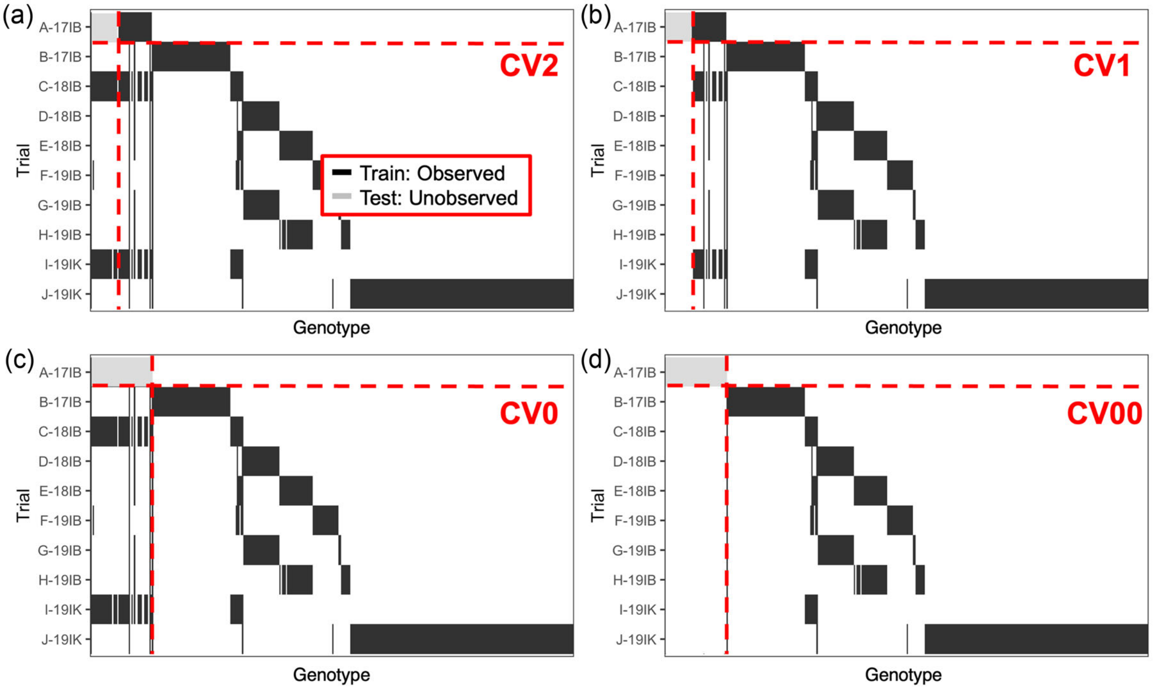

FIGURE 2 Graphic representation of four of the five cross-validation (CV) schemes used in this study. Unique cassava genotypes (clones) are shown along the *x* axis and individual trials are along the *y* axis. Genotypes appearing in test sets are represented in gray, while genotypes in the training sets are black. In this example, various proportions of genotypes from trial A-17IB are included in the test set according to the CV scheme. (a) CV2 represents a case in which observations of the test set genotypes are present in the training set along with observations of genotypes from all environments. (b) CV1 also includes observations from all environments, but no observations from the test set genotypes are included in the training set. (c) CV0 includes observations from all genotypes but includes all genotypes from trial A-17IB in the test set so that environment is not represented in the training set. (d) CV00 omits all observations from the training set that overlap with the test set in genotype or environment. These CV schemes are modeled after Jarquín et al. (2017)

random sampling based on RDMC to ensure that the full range of RDMC values was represented in each set. This scheme did not take non-check or check plots into account, so in some cases different plots from the same non-check or check genotypes were included in both the training and test sets.

Four additional CV schemes representing relevant plant breeding scenarios were applied across the 10 trials with each trial considered an independent environment due to differences in planting and harvest dates even within the same year. These CV schemes are described in detail in Jarquín et al. (2017), but in brief, all pairwise combinations of tested and untested genotypes in tested and untested environments were evaluated in four CV schemes (Figure 2). For CV2 (tested genotypes in tested environments; Figure 2a), a random subset of 30% of the genotypes from a given trial were used as the test set. Data from all remaining genotypes from that trial and all other trials were combined into one training set. This process was repeated with 50 iterations for each test trial, each with a different random sample of genotypes in the test set. CV1 (untested genotypes in tested environments; Figure 2b) was implemented using the same subsetting procedure as CV2. In this scheme, however, genotypes present in the test set were removed from the training set altogether. CV0 (tested genotypes in untested environments; Figure 2c) involved the inclusion of an entire trial as the test set, with all other trials making up the training set whether or not they contained genotypes represented in the test set trial. The implementation of CV00 (untested genotypes in untested environments; Figure 2d) followed the same procedure as CV0, but, as with CV1, all genotypes present in the test set were removed from the training set prior to model training. In all cases, all plots of each genotype within a given trial were grouped, occurring either both in the training or both in the test set. Counts of overlapping genotypes between trials are listed in Supplemental Table S3. All CV schemes were implemented in the *waves* R package version 0.1.1 (Hershberger et al., 2021).

2.6 | Prediction

Prediction models were developed using three different algorithms in the *waves* R package version 0.1.1 (Hershberger et al., 2021): partial least squares regression (PLSR) (H. Wold, 1982; S. Wold et al., 1984), random forest (RF) (Breiman, 2001), and support vector machine (SVM) (Drucker et al., 1997). Partial least squares regression both displayed the best

TABLE 2 Root dry matter content (RDMC) summary statistics for 10 International Institute of Tropical Agriculture cassava field trials. The number of genotypes and plots includes only those with complete spectral and phenotypic data

Abbreviated trial name	# Phenotyped non-check/check genotypes	# Phenotyped plots	Mean RDMC	Maximum RDMC	Minimum RDMC	RDMC standard deviation
A-17IB	44/4	92	24.2	39.8	10.8	6.2
B-17IB	59/6	156	32.5	42.0	21.8	4.1
C-18IB	47/4	97	32.1	43.7	19.3	5.7
D-18IB	30/5	103	40.7	45.9	29.7	3.1
E-18IB	31/5	100	40.1	45.9	30.6	2.8
F-19IB	25/5	52	28.5	39.1	17.1	4.6
G-19IB	30/5	68	33.6	43.2	26.7	3.6
H-19IB	31/5	65	30.9	36.0	21.7	2.7
I-19IK	46/4	98	27.6	43.3	9.8	7.9
J-19IK	175/5	331	37.9	47.1	27.0	4.2

performance for within-trial predictions and required the least time-intensive training (Supplemental Table S2), so it was used for all subsequent prediction analyses in this study.

Partial least squares regression (PLSR) is a popular method used in NIRS analysis (reviewed in Roggo et al., 2007). This algorithm uses latent variables to reduce the dimensionality of the spectral dataset, allowing for efficient model training. For each CV scenario, the number of latent variables used in the final model was tuned using 5-fold CV within the training set. The best number of latent variables was chosen based on the lowest root mean squared error of CV ($\mathrm{RMSE}_{\mathrm{CV}}$) and predictions were made on the test set using this final tuned model. This process of sampling and tuning was repeated 50 times for each combination of pretreatment technique and algorithm for a total of 1,950 runs per trial. For CV0 and CV00 scenarios, only a single iteration of hyperparameter tuning and prediction was performed, as no sampling was performed in these cases.

This pipeline was also used to evaluate the optimal number of scanned samples (homogenized subsamples or sliced root subsets) per plot. For trials in which root slices were scanned, A-17IB, B-17IB, and C-18IB, a sample was defined as a single root. For all remaining trials, each 100 g homogenized subsample within a plot was treated as an individual sample. Within each trial, n versions of the RDMC and spectral dataset were created, each containing a different number of samples per plot with n representing the maximum number of samples available for that trial. For the versions with one to n-1 samples per plot, where a different individual or combination of individual samples may be selected from each plot, 50 iterations of random statistical subsampling were performed to get a representative sample of the possible combinations of homogenized subsamples or sliced root subsets. Each version of the dataset was then run through the *waves* within-trial tuning and prediction pipeline 10 times using PLSR for model performance comparison. All plots for this and other analyses were prepared with ggplot2 version 3.3.3 (Wickham, 2016).

3 | RESULTS AND DISCUSSION

Mobile NIRS has the potential to provide rapid, in-field phenotyping of cassava roots for dry matter content, but validation is required to ensure that realistic expectations can be set regarding the accuracy of prediction with a given spectrometer and statistical model. We explored the use of an inexpensive, handheld spectrometer that captures a subset (740–1070 nm) of the full NIR range for the prediction of cassava RDMC in the context of an active breeding program. The results presented here demonstrate the feasibility of RDMC prediction with low-cost, mobile spectrometers such as the SCiO used in our study and inform its use in routine breeding decisions.

3.1 | Variation for dry matter content of cassava roots

We scanned roots with the SCiO and measured the oven-based RDMC of 376 cassava genotypes evaluated in one or more of 10 field trials in Nigeria (Supplemental Figure S2). Summary statistics of plot averages of raw RDMC values as measured with the oven method for the 10 trials in this study are shown in Table 2, with plot mean RDMC distributions shown in Figure 3. Tukey's honest significant difference test identified significant differences (α = 0.05) between the mean values of almost all trials. I-19-IK contained the largest range of RDMC, spanning from 9.75 to 43.30%, resulting in a standard deviation of 7.86. The two other trials with this same set of germplasm, A-17IB and C-18IB, had the second (6.17) and

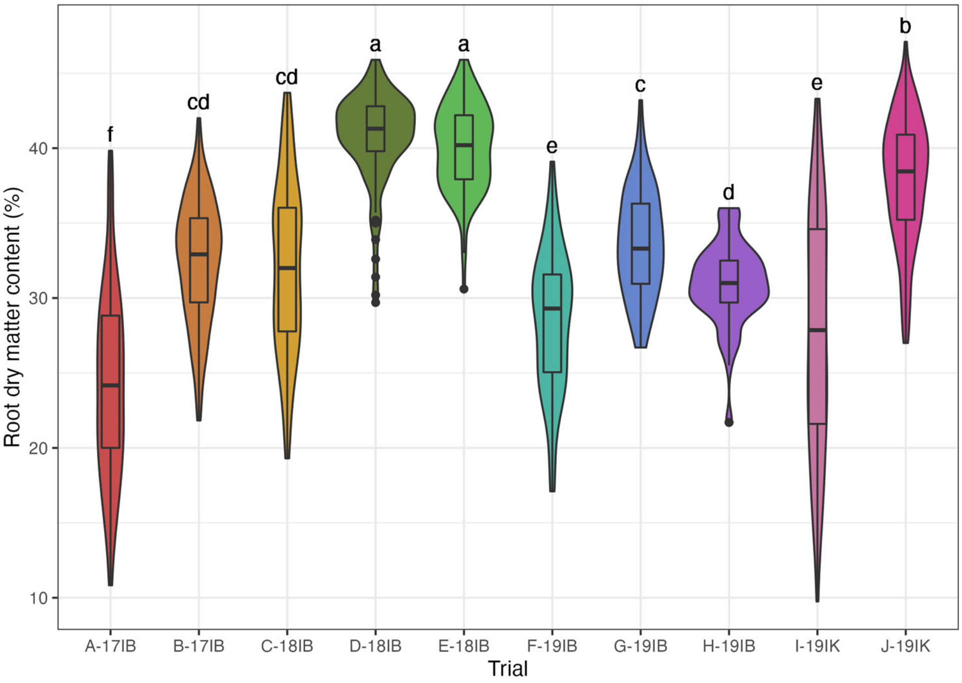

FIGURE 3 F Plot mean root dry matter content distributions are shown for each trial. Letters above the violin and boxplots represent significantly different groups as identified by Tukey's honestly significant difference test at α = 0.05

third (5.69) highest standard deviations. Maximum RDMC ranged from 36.0% in H-19IB to 47.1% in J-19IK, with maximums from the other trials falling in between. Minimum RDMC ranged from 9.75% (I-19IK) to 30.6% (E-18IB). The range of RDMC across the 10 trials (9.75–47.10%) is representative of the typical RDMC range in cassava germplasm, with adequate variation for improvement through breeding (Sánchez et al., 2009).

3.2 | Within-trial prediction

To assess within-trial PLSR prediction accuracy of RDMC, we used a stratified random sample of 70% of the plots in a given trial as the training set for hyperparameter tuning and testing model performance on the remaining 30% of plots. Mean prediction performance over 50 iterations of subsampling varied across years, locations, populations, and sample preparation methods (Table 3 and Figure 4), with the highest R^2_P from I-19IK at 0.89 and the lowest root mean squared error of prediction ($RMSE_P$) from H-19IB at 1.31. The poorest performing trials in terms of R^2_P were C-18IB and D-18IB, both with an R^2_P of 0.63. C-18IB also had the least favorable $RMSE_P$ at 3.55. Squared Spearman's rank correlation values (0.57–0.89) were similar to R^2_P (0.62–0.89) in most trials. The optimal number of latent variables in the most iterations of subsampling ranged from 2–10, again showing no identifiable pattern according to year, location, population, or sample preparation method (Table 3). Results showing the same general trends with 12 different pretreatment technique combinations for PLSR as well as RF and SVM can be found in Supplemental Table S2.

Statistics reported in terms of prediction of a holdout test set (e.g., R^2_P and $RMSE_P$) show weaker model performance as compared with those reported in terms of CV (coefficient of multiple determination of CV [R^2_{CV}] and $RMSE_{CV}$) across all trials in this study. In previous studies of NIRS for RDMC prediction in cassava, only leave-one-out CV statistics are discussed, though R^2_P statistics are also reported (Ikeogu et al., 2017; Sánchez et al., 2014). This can lead to misleading conclusions, as we can expect the predictions of future samples to perform more similarly to the R^2_P results rather than those of R^2_{CV} due to the new samples not being present in the training set. When we directly compare our R^2_P values to those found in previous studies, the performance of the laboratory-grade FOSS 6500 ($R^2_P = 0.79–0.95$) and QualitySpec Trek ($R^2_P = 0.84$) spectrometers are comparable to our R^2_p results with the SCiO (within-trial PLSR $R^2_P = 0.62–0.89$), indicating that the SCiO may be a suitable alternative for either of these more expensive spectrometers for RDMC phenotyping, especially when factoring in cost and throughput.

3.3 | Algorithms and pretreatment

Many spectral pretreatment methods have been explored (reviewed in Rinnan et al., 2009) to reduce noise, but the ben-

TABLE 3 Summary statistics for within-trial cassava root dry matter content partial least squares regression predictions with 50 iterations of subsampling

Trial	RMSE_P	R^2_P	RPD	RPIQ	CCC	Bias	SEP	RMSE_CV	R^2_CV	R^2_SP	Best number of latent variables
A-17IB	2.88 (0.46)	0.79 (0.08)	2.24 (0.41)	2.90 (0.60)	0.88 (0.04)	0.09 (0.57)	2.94 (0.46)	2.22 (0.34)	0.87 (0.04)	0.76 (0.09)	5 (2.79)
B-17IB	2.58 (0.26)	0.62 (0.07)	1.61 (0.16)	2.19 (0.23)	0.77 (0.05)	0.00 (0.38)	2.61 (0.26)	2.05 (0.20)	0.74 (0.05)	0.57 (0.07)	10 (2.01)
C-18IB	3.55 (0.40)	0.63 (0.10)	1.60 (0.21)	2.26 (0.26)	0.77 (0.06)	0.17 (0.81)	3.61 (0.41)	2.69 (0.34)	0.77 (0.06)	0.65 (0.11)	6 (2.89)
D-18IB	1.95 (0.48)	0.63 (0.17)	1.67 (0.39)	1.56 (0.42)	0.74 (0.12)	0.14 (0.43)	1.99 (0.48)	1.70 (0.18)	0.69 (0.06)	0.61 (0.11)	5 (1.03)
E-18IB	1.56 (0.31)	0.71 (0.12)	1.88 (0.34)	2.73 (0.54)	0.82 (0.07)	0.04 (0.30)	1.59 (0.32)	1.21 (0.25)	0.81 (0.07)	0.71 (0.08)	7 (2.48)
F-19IB	2.03 (0.39)	0.82 (0.06)	2.30 (0.44)	2.67 (0.70)	0.89 (0.04)	0.04 (0.64)	2.12 (0.41)	1.51 (0.34)	0.89 (0.04)	0.76 (0.11)	2 (3.79)
G-19IB	1.61 (0.23)	0.80 (0.07)	2.23 (0.42)	2.87 (0.55)	0.88 (0.04)	0.03 (0.39)	1.65 (0.24)	1.00 (0.15)	0.92 (0.02)	0.79 (0.07)	10 (1.84)
H-19IB	1.31 (0.24)	0.74 (0.12)	2.06 (0.53)	2.22 (0.50)	0.84 (0.09)	0.06 (0.34)	1.35 (0.24)	1.04 (0.11)	0.84 (0.04)	0.65 (0.16)	7 (1.32)
I-19IK	2.66 (0.48)	0.89 (0.04)	3.05 (0.55)	4.77 (0.98)	0.94 (0.02)	0.04 (0.63)	2.71 (0.49)	2.21 (0.40)	0.92 (0.03)	0.89 (0.04)	3 (3.40)
J-19IK	2.28 (0.16)	0.70 (0.04)	1.83 (0.13)	2.50 (0.18)	0.83 (0.02)	−0.04 (0.20)	2.29 (0.16)	2.07 (0.10)	0.75 (0.02)	0.70 (0.05)	7 (2.27)

Note. The mean of each statistic over all iterations is shown for raw data post-outlier removal but without pretreatment except for the best number of latent variables, in which the mode is displayed. The standard deviation for each statistic is shown in parentheses. Model performance statistics include Lin's concordance correlation coefficient (CCC), coefficient of multiple determination of cross-validation (R^2_CV), squared Pearson's correlation of predicted and observed values in a test set (R^2_P), squared Spearman's correlation of predicted and observed values in a test set (R^2_SP), root mean squared error of cross-validation (RMSE_CV), root mean squared error of prediction (RMSE_P), residual predictive deviation (RPD), ratio of performance to interquartile distance (RPIQ), and standard error of prediction (SEP).

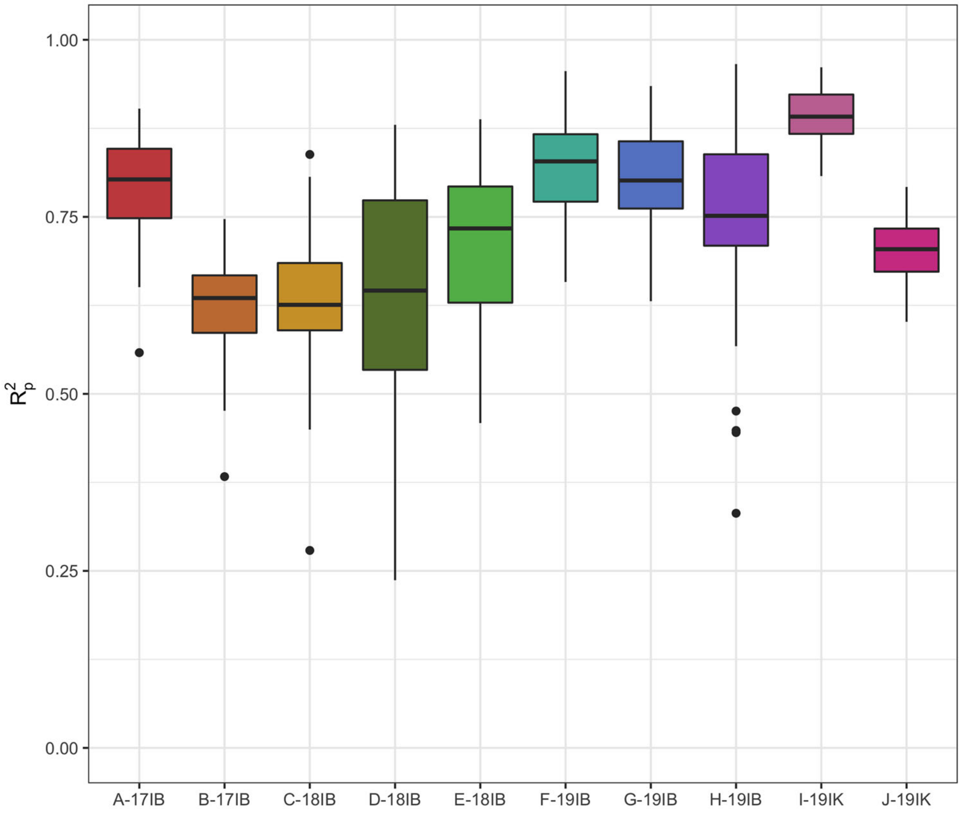

FIGURE 4 Within-trial predictions of cassava root dry matter content were performed on a plot basis for each trial. The squared Pearson's correlation between predicted and observed root dry matter content (R^2_P) is shown for 50 iterations of the *waves* prediction pipeline with no spectral pretreatment

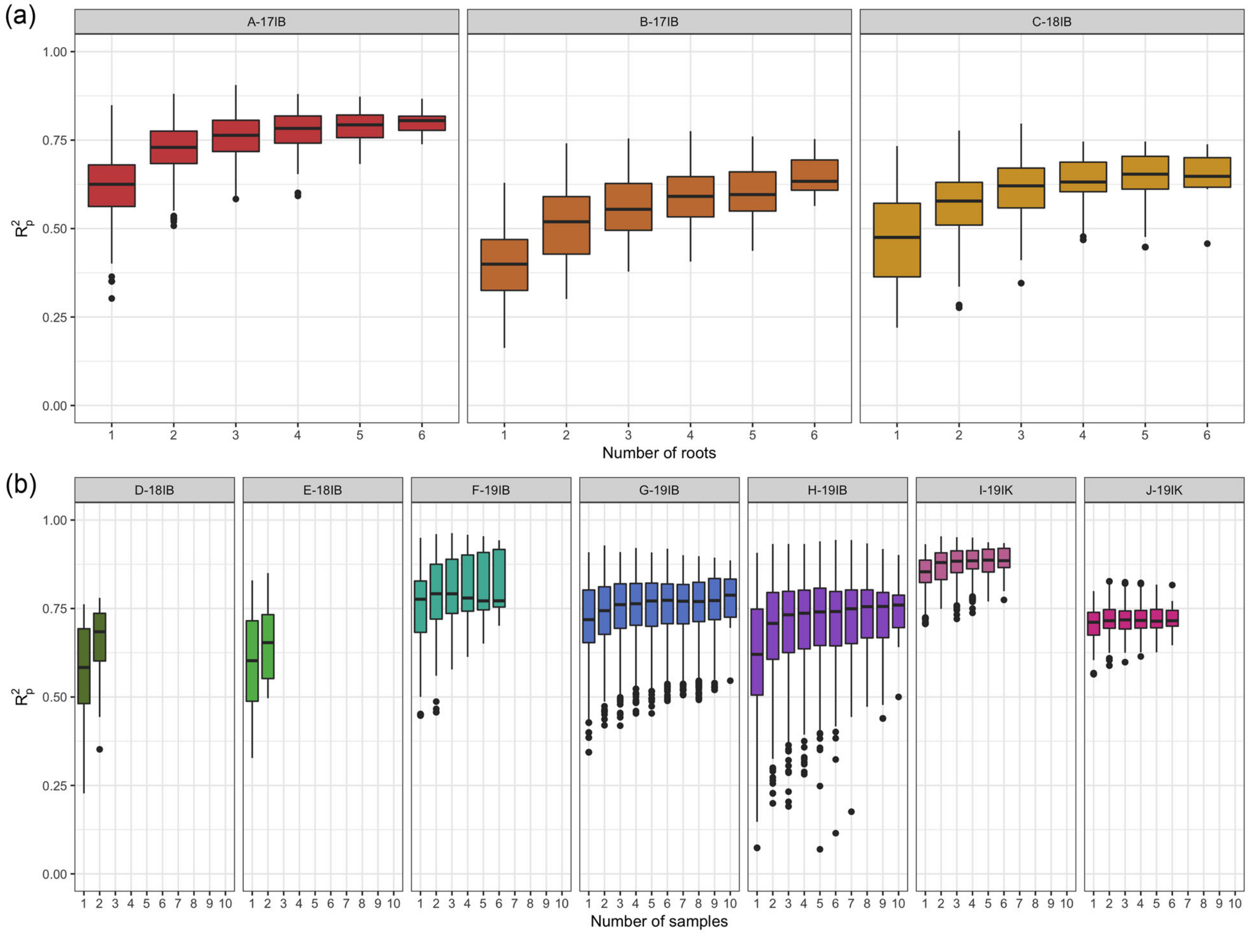

FIGURE 5 Within-trial predictions of cassava root dry matter content were performed with either subsets of roots (a) or homogenized subsamples (b) for each trial. The squared Pearson's correlation between predicted and observed plot mean root dry matter content (R^2_{P}) is shown for 50 iterations of subsampling of scans within each plot followed by 10 iterations of the *waves* prediction pipeline. The plot-level mean of all scans was taken prior to model development in all cases

efit of individual methods varies depending on the trait of interest, sample set, and model algorithm (Ikeogu et al., 2017; Kosmowski & Worku, 2018; Pizarro et al., 2004). Because the SCiO had not yet been tested for use in the prediction of cassava RDMC, we tested 12 combinations of pretreatment techniques and three model algorithms to identify the combination that provided the best model performance for this spectrometer. Overall, we found no additional benefit to the RF and SVM algorithms as opposed to the more traditional PLSR method (Supplemental Table S2). Though working with group discrimination rather than regression, Kosmowski and Worku (2018) also found that RF and SVM were not superior to PLS-discriminant analysis when investigating SCiO performance. Interestingly, they also found that pretreatment techniques did not affect model performance with PLS or SVM but that data pretreatment was necessary to improve performance of RF models. Studies exploring the effect of these factors on model performance with other spectrometers have not followed this same pattern (Ikeogu et al., 2017; Sampaio et al., 2018), indicating that the ideal combination of spectral pretreatment technique and model algorithm may be spectrometer specific.

3.4 | Sample preparation

Cassava RDMC can vary significantly within roots from proximal to distal end (Chávez et al., 2008), so root tissue is traditionally shredded and mixed to get a representative sample for analysis. Because this method requires the time-consuming transport of bulky roots to a laboratory environment, our study included scans of both shredded tissue and roots prepared for scanning with an alternate sliced method that can be performed directly in the field in order to identify potential tradeoffs in terms of predictive ability. Comparing results from plot-mean scans, we observed slightly more favorable model performance statistics from trials in which roots were shredded prior to scanning; the mean within-trial R^2_{P} ranged from

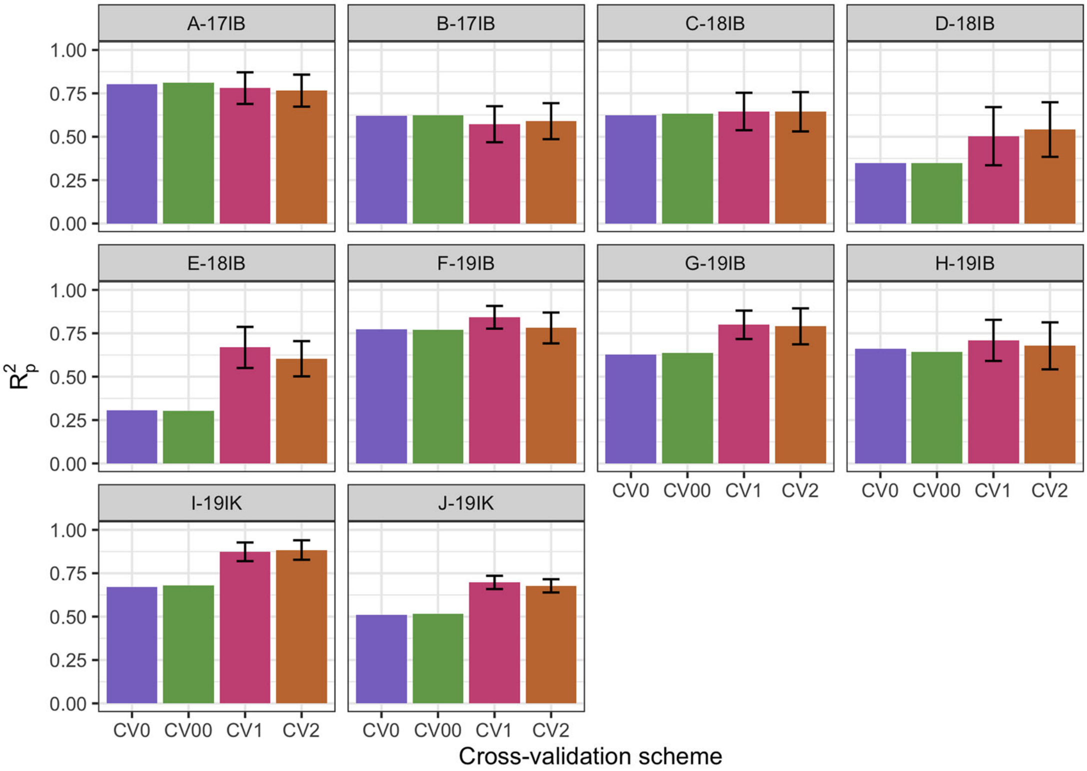

FIGURE 6 Partial least squares regression prediction of cassava root dry matter content performed on a plot-mean basis for each trial. Cross-validation (CV) scheme results are displayed as the $R^2{}_{\mathrm{P}}$ (squared Pearson's correlation between predicted and observed values) for 50 iterations of the *waves* prediction pipeline with no spectral pretreatment. CV0 indicates leave-one-trial-out CV, CV00 indicates that there was no overlap between genotypes and environments in the training and test sets, CV1 indicates overlap in environment but not genotypes between the training and test sets, and CV2 indicates overlap of both genotypes and environments in the training and test sets, though in all cases genotypes with multiple replicates within a trial were sorted together. Error bars show standard deviation for schemes with subsampling (CV1 and CV2). As no subsampling occurred in either the CV0 or CV00 schemes, standard deviation was not calculated and therefore no error bars are displayed

0.62 to 0.79 for the three sliced root trials and 0.63 to 0.89 for the seven shredded root trials (Table 3). In agreement with expectations that homogenized samples better capture whole-root variability for RDMC, our findings align with those of Ikeogu et al. (2017), as they also found scans of shredded root samples to slightly outperform those of sliced roots in terms of cassava RDMC predictions.

The small sample size in this study limits our ability to confidently draw conclusions based on plot-mean scan model performance alone, so we also sought to determine whether our samples captured the plot-level RDMC within each trial. By varying the number of homogenized subsamples or subsets of sliced roots (hereafter samples) included in the plot mean scan, we were able to identify the point at which within-trial prediction model performance began to level off, indicating that additional samples would not further increase performance and therefore within-plot variation had been adequately captured. Despite differences in the minimum number of homogenized subsamples and sets of sliced roots required to stabilize model performance, five samples were sufficient to capture variation in all trials, regardless of sample preparation method (Figure 5). In seven of the 10 trials in this study, a boost in predictive performance occurred as additional samples were included in the mean scan used for model development up to the leveling point. This increase in performance with additional samples is not surprising, as we expect the addition of technical replicates to stabilize measurements within a given experimental unit by controlling measurement error (Blainey et al., 2014). These results can also be used to guide the development of a refined slicing protocol that would potentially eliminate the need for shredding.

3.5 | Application of low-cost NIRS in a cassava breeding program

Optimal NIRS model calibration requires the training set to be representative of the greater population of samples that will be predicted. Because the RDMC of cassava is affected by the environment in which it is grown, we chose to include trials from a range of years, populations, and locations, and to investigate the influence of these factors on model performance using CV schemes designed for plant breeding programs. Overall mean model performance using these schemes ranged from $R^2{}_{\mathrm{P}} = 0.60$ with CV0 and CV00 to $R^2{}_{\mathrm{P}} = 0.71$ for CV1 (Supplemental Table 4). In general, we found that schemes in which the test set environment was represented in the training set (CV1 and CV2) outperformed schemes in which there was no overlap in environments between the two sets (CV0 and CV00), while the differences in performance within each of these groups was minimal (Figure 6, Supplemental Table 4). This indicates that sharing an environment between the training and test sets has more of an effect on model performance than sharing a set of genotypes. We therefore recommend that training sets be updated to include a subset of oven-phenotyped plots from each new trial to maximize model performance, as is done in the CV1 and CV2 schemes. However, as the addition of oven-dried phenotyping represents a tradeoff between model performance and labor, each breeding program will need to evaluate whether the additional performance boost is worthwhile. We also observed differences in overall model performance between trials, following the same pattern as our within-trial prediction results. Given the mix of overall trial performance amongst locations, years, populations, and scanning protocols, it is not possible to determine the source of this pattern with the limited sample size of this study.

4 | CONCLUSIONS

The present study shows that miniature NIR spectrometers such as the SCiO have the potential to enable cheap and high-throughput quality assessments particularly for breeding programs that cannot afford expensive, laboratory-based NIR equipment. The results of this study support the application of the SCiO for the phenotyping of cassava RDMC directly in the field, with model performance matching that of laboratory-grade spectrometers in many trials. The low per-unit cost of this new generation of consumer-grade spectrometers will contribute not only to the efficiency with which cassava breeders are able to phenotype quality traits but also for other starchy root and tuber crops.

Our results indicate that scans taken on sliced roots may match the performance of scans taken on homogenized root subsamples, but additional studies directly comparing these two methods within the same trials would help to clarify this relationship. The ability to accurately predict RDMC with sliced roots represents a large savings in terms of time and labor on the part of breeding programs, allowing maximum benefit from the use of handheld spectrometers. Further, we observed diminishing returns with the inclusion of additional samples per plot, indicating that scanning a subset of five or more roots or alternatively two or more homogenized subsamples should be sufficient for capturing within-plot scan variation with the SCiO. Overall, we found prediction with the SCiO, a consumer-grade NIR spectrometer, to be a highly accurate, field-based, and low-cost method for cassava RDMC phenotyping. The scanning and analysis protocols explored in this study are readily applicable in the phenotyping of other quality traits in cassava and beyond.

DATA AVAILABILITY STATEMENT

Raw RDMC and spectral data are available for download on Cyverse at https://datacommons.cyverse.org/browse/iplant/home/shared/GoreLab/dataFromPubs/Hershberger_CassavaNIRS_2021/. Analysis R code is available on Github at http://www.github.com/GoreLab/CassavaNIRS.

ACKNOWLEDGMENTS

We are grateful to the Next Generation Cassava project and in particular the IITA cassava breeding team for their interest in and support of this work. This study would not have been possible without their efforts. We also thank the PhenoApps team, Kelly Robbins, Jean-Luc Jannink, and the Gore research group for valuable feedback and productive discussions. This work has been supported by USDA NIFA AFRI EWD Predoctoral Fellowship 2019-67011-29606 (J.H.) and NSF BREAD IOS-1543958 (M.A.G.). Work at IITA was supported by the United Kingdom Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation (Grant INV-007637). This work also benefited from the support of the Cornell Institute for Digital Agriculture (CIDA). This study was also made possible by the support of the American People provided to the Feed the Future Innovation Lab for Crop Improvement through the United States Agency for International Development (USAID). The contents are the sole responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government. Program activities are funded by the United States Agency for International Development (USAID) under Cooperative Agreement No. 7200AA-19LE-00005.

AUTHOR CONTRIBUTIONS

Jenna Hershberger: Conceptualization; Data curation; Formal analysis; Methodology; Writing-original draft; Writing-review & editing. Edwige Gaby Nkouaya Mbanjo: Investigation; Writing-review & editing. Prasad Peteti: Data curation; Investigation; Methodology. Andrew Ikpan: Investigation; Methodology. Kayode Ogunpaimo: Investigation; Methodology. Kehinde Nafiu: Investigation; Methodology. Ismail Y. Rabbi: Conceptualization; Funding acquisition; Methodology; Supervision; Writing-review & editing. Michael A.

Gore: Conceptualization; Funding acquisition; Supervision; Writing-review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Jenna Hershberger https://orcid.org/0000-0002-3147-6867
Edwige Gaby Nkouaya Mbanjo https://orcid.org/0000-0002-9982-1137
Prasad Peteti https://orcid.org/0000-0002-6013-8947
Michael A. Gore https://orcid.org/0000-0001-6896-8024

REFERENCES

Acheampong, P. P., Owusu, V., & Nurah, G. (2018). How does farmer preference matter in crop variety adoption? The case of improved cassava varieties' adoption in Ghana. *Open Agriculture*, *3*(1), 466–477. https://doi.org/10.1515/opag-2018-0052

Asrat, S., Yesuf, M., Carlsson, F., & Wale, E. (2010). Farmers' preferences for crop variety traits: Lessons for on-farm conservation and technology adoption. *Ecological Economics: The Journal of the International Society for Ecological Economics*, *69*(12), 2394–2401. https://doi.org/10.1016/j.ecolecon.2010.07.006

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, *43*(5), 772–777. https://doi.org/10.1366/0003702894202201

Blainey, P., Krzywinski, M., & Altman, N. (2014). Points of significance: Replication. *Nature Methods*, *11*(9), 879–880. https://doi.org/10.1038/nmeth.3091

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Ceballos, H., Hershey, C., & Becerra Lopez-lavalle, L. A. (2012). New approaches to cassava breeding. In J. Janick (Ed.), *Plant breeding reviews* (Vol. 36, pp. 427–504). Wiley-Blackwell.

Chávez, A. L., Ceballos, H., Rodriguez-Amaya, D. B., Perez, J. C., Sanchez, T., Calle, F., & Morante, N. (2008). Sampling variation for carotenoids and dry matter content in cassava roots. *African Journal of Root and Tuber Crops*, *34*(1), 43–49.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). *Support vector regression machines* (pp. 155–161).

Dos Santos, C. A. T., Lopo, M., Páscoa, R. N. M. J., & Lopes, J. A. (2013). A review on the applications of portable near-infrared spectrometers in the agro-food industry. *Applied Spectroscopy*, *67*(11), 1215–1233. https://doi.org/10.1366/13-07228

El-Sharkawy, M. A. (1993). Drought-tolerant cassava for Africa, Asia, and Latin America. *Bioscience*, *43*(7), 441–451. https://doi.org/10.2307/1311903

El-Sharkawy, M. A. (2004). Cassava biology and physiology. *Plant Molecular Biology*, *56*(4), 481–501. https://doi.org/10.1007/s11103-005-2270-7

Fermont, A. M., Van Asten, P. J. A., Tittonell, P., Van Wijk, M. T., & Giller, K. E. (2009). Closing the cassava yield gap: An analysis from smallholder farms in East Africa. *Field Crops Research*, *112*(1), 24–36. https://doi.org/10.1016/j.fcr.2009.01.009

Food and Agriculture Organization of the United Nations. (2021). *FAOSTAT statistical database*. www.faostat.fao.org

Hahn, S. K., Terry, E. R., & Leuschner, K. (1980). Breeding cassava for resistance to cassava mosaic disease. *Euphytica*, *29*(3), 673–683. https://doi.org/10.1007/BF00023215

Hershberger, J., Morales, N., Simoes, C. C., Ellerbrock, B., Bauchet, G., Mueller, L. A., & Gore, M. A. (2021). Making waves in Breedbase: An integrated spectral data storage and analysis pipeline for plant breeding programs. *The Plant Phenome Journal*, *4*(1), e20012. https://doi.org/10.1002/ppj2.20012

Howeler, R., Lutaladio, N., & Thomas, G. (2013). *Save and grow: Cassava a guide to sustainable production intensification*. FAO.

IITA. (1990). *Cassava in tropical Africa. A reference manual*. IITA.

Ikeogu, U. N., Davrieux, F., Dufour, D., Ceballos, H., Egesi, C. N., & Jannink, J.-L. (2017). Rapid analyses of dry matter content and carotenoids in fresh cassava roots using a portable visible and near infrared spectrometer (Vis/NIRS). *Plos One*, *12*(12), 1–17. https://doi.org/10.1371/journal.pone.0188918

Jarquín, D., Lemes Da Silva, C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., Battenfield, S., & Crossa, J. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype × environment interactions in Kansas wheat. *The Plant Genome*, *10*(2). https://doi.org/10.3835/plantgenome2016.12.0130

Jennings, D. L. (1957). Further studies in breeding cassava for virus resistance. *East African Agricultural and Forestry Journal*, *22*(4), 213–219. https://doi.org/10.1080/03670074.1957.11665107

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson.

Kaur, A., Donis-Gonzalez, I. R., & St Clair, D. A. (2020). Evaluation of a hand-held spectrophotometer as an in-field phenotyping tool for tomato and pepper fruit quality. *The Plant Phenome Journal*, *3*(1), e.20008. https://doi.org/10.1002/ppj2.20008

Kaur, H., Künnemeyer, R., & Mcglone, A. (2017). Comparison of hand-held near infrared spectrophotometers for fruit dry matter assessment. *Journal of Near Infrared Spectroscopy*, *25*(4), 267–277. https://doi.org/10.1177/0967033517725530

Kawano, K. (2003). Thirty years of cassava breeding for productivity—biological and social factors for success. *Crop Science*, *43*(4), 1325–1335. https://doi.org/10.2135/cropsci2003.1325

Kawano, K., Fukuda, W. M. G., & Cenpukdee, U. (1987). Genetic and environmental effects on dry matter content of cassava root. *Crop Science*, *27*(1), 69–74. https://doi.org/10.2135/cropsci1987.0011183X002700010018x

Kawuki, R. S., Pariyo, A., Amuge, T., Nuwamanya, E., Ssemakula, G., Tumwesigye, S., Bua, A., Baguma, Y., Omongo, C., Alicai, T., & Orone, J. (2011). A breeding scheme for local adoption of cassava (*Manihot esculenta* Crantz). *Journal of Plant Breeding and Crop Science*, *3*(7), 120–130.

Kosmowski, F., & Worku, T. (2018). Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia. *Plos One*, *13*(3), e0193620. https://doi.org/10.1371/journal.pone.0193620

Lebot, V., Champagne, A., Malapa, R., & Shiley, D. (2009). NIR determination of major constituents in tropical root and tuber crop flours. *Journal of Agricultural and Food Chemistry*, *57*(22), 10539–10547. https://doi.org/10.1021/jf902675n

Li, M.o, Qian, Z., Shi, B., Medlicott, J., & East, A. (2018). Evaluating the performance of a consumer scale SCiO molecular sensor to predict quality of horticultural products. *Postharvest Biology and Technology*, *145*(7), 183–192. https://doi.org/10.1016/j.postharvbio.2018.07.009

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H. G., Okechukwu, R., Dixon, A. G. O., Kulakow, P., & Jannink, J.-L. (2013). Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. *Crop Science*, *53*(4), 1312. https://doi.org/10.2135/cropsci2012.11.0653

Njine, M. (2010). Social and economical factors hindering adoption of improved cassava varieties in Kiganjo Location, Nyeri Municipality Division, Kenya. *Journal of Developments in Sustainable Agriculture*, *5*(2), 178–190.

Okechukwu, R. U., & Dixon, A. G. O. (2008). Genetic gains from 30 years of cassava breeding in Nigeria for storage root yield and disease resistance in elite cassava genotypes. *Journal of Crop Improvement*, *22*(2), 181–208. https://doi.org/10.1080/15427520802212506

Oparinde, A., Abdoulaye, T., Manyong, V., Birol, E., Asare-Marfo, D., Kulakow, P., & Ilona, P. (2016). A technical review of modern cassava technology adoption in Nigeria (1985–2013): Trends, challenges, and opportunities. *HarverstPlus Working Paper*, (23), 1–26.

Owusu, V., & Donkor, E. (2012). Adoption of improved cassava varieties in Ghana. *Agricultural Journal*, *7*(2), 146–151. https://doi.org/10.3923/aj.2012.146.151

Parmar, A., Sturm, B., & Hensel, O. (2017). Crops that feed the world: Production and improvement of cassava for food, feed, and industrial uses. *Food Security*, *9*(5), 907–927. https://doi.org/10.1007/s12571-017-0717-8

Pérez, J. C., Lenis, J. I., Calle, F., Morante, N., Sánchez, T., Debouck, D., & Ceballos, H. (2011). Genetic variability of root peel thickness and its influence in extractable starch from cassava (*Manihot esculenta* Crantz) roots. *Plant Breeding*, *130*(6), 688–693. https://doi.org/10.1111/j.1439-0523.2011.01873.x

Pizarro, C., Esteban-Dıéz, I., Nistal, A.-J., & González-Sáiz, J.-M.arıá (2004). Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta*, *509*(2), 217–227. https://doi.org/10.1016/j.aca.2003.11.008

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rinnan, Å., Berg, F. V. D., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, *44*(3), 683–700. https://doi.org/10.1016/j.jpba.2007.03.023

Safo-Kantanka, O., & Owusu-Nipah, J. (1992). Cassava varietal screening for cooking quality: Relationship between dry matter, starch content, mealiness and certain microscopic observations of the raw and cooked tuber. *Journal of the Science of Food and Agriculture*, *60*(1), 99–104. https://doi.org/10.1002/jsfa.2740600116

Sampaio, P. S., Soares, A., Castanho, A., Almeida, A. S., Oliveira, J., & Brites, C. (2018). Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. *Food Chemistry*, *242*, 196–204. https://doi.org/10.1016/j.foodchem.2017.09.058

Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., Zum Felde, T., Domínguez, M., & Davrieux, F. (2014). Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. *Food Chemistry*, *151*, 444–451. https://doi.org/10.1016/j.foodchem.2013.11.081

Sánchez, T., Salcedo, E., Ceballos, H., Dufour, D., Mafla, G., Morante, N., Calle, F., Pérez, J. C., Debouck, D., Jaramillo, G., & Moreno, I. X. (2009). Screening of starch quality traits in cassava (*Manihot esculenta* Crantz). *Starch*, *61*(5), 310–310. https://doi.org/10.1002/star.200990027

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Subedi, P. P., & Walsh, K. B. (2020). Assessment of avocado fruit dry matter content using portable near infrared spectroscopy: Method and instrumentation optimisation. *Postharvest Biology and Technology*, *161*, 111078. https://doi.org/10.1016/j.postharvbio.2019.111078

Teeken, B., Agbona, A., Bello, A., Olaosebikan, O., Alamu, E., Adesokan, M., Awoyale, W., Madu, T., Okoye, B., Chijioke, U., Owoade, D., Okoro, M., Bouniol, A., Dufour, D., Hershey, C., Rabbi, I., Maziya-Dixon, B., Egesi, C., Tufan, H., & Kulakow, P. (2021). Understanding cassava varietal preferences through pairwise ranking of gari-eba and fufu prepared by local farmer–processors. *International Journal of Food Science & Technology*, *56*(3), 1258.

Teeken, B., Olaosebikan, O., Haleegoah, J., Oladejo, E., Madu, T., Bello, A., Parkes, E., Egesi, C., Kulakow, P., Kirscht, H., & Tufan, H. A. (2018). Cassava trait preferences of men and women farmers in Nigeria: Implications for breeding. *Economic Botany*, *72*(3), 263–277. https://doi.org/10.1007/s12231-018-9421-7

Teles, F. (1993). An easy technique for rapid determination of dry-matter content in cassava roots (*Manihot esculenta* Crantz). *Food Chemistry*, *47*(4), 375–377. https://doi.org/10.1016/0308-8146(93)90180-N

Teye, E., Asare, A. P., Amoah, R. S., & Tetteh, J. P. (2011). Determination of the dry matter content of cassava (*Manihot esculenta*, Crantz) tubers using specific gravity method. *ARPN Journal of Agricultural and Biological Science*, *6*(11), 23–28.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wiedemair, V., & Huck, C. W. (2018). Evaluation of the performance of three hand-held near-infrared spectrometer through investigation of total antioxidant capacity in gluten-free grains. *Talanta*, *189*, 233–240. https://doi.org/10.1016/j.talanta.2018.06.056

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Joreskog & H. O. A. Wold (Eds.), *Systems under indirect observation* (pp. 1–54). Elsevier.

Wold, S., Ruhe, A., Wold, H., & Dunn Iii, W. J. (1984). The collinearity problem in linear regression. The partial least squares

(PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743. https://doi.org/10.1137/0905052

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., Lozano, R., Carpio, D. P. D., Ramu, P., & Jannink, J.-L. (2016). Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *The Plant Genome*, *9*(2), 1–13. https://doi.org/10.3835/plantgenome2015.11.0118

Xu, Q.-S., & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Hershberger, J., Mbanjo, E. G. N., Peteti, P., Ikpan, A., Ogunpaimo, K., Nafiu, K., Rabbi, I. Y., & Gore, M. A. (2022). Low-cost, handheld near-infrared spectroscopy for root dry matter content prediction in cassava. *The Plant Phenome Journal*, *5,* e20040. https://doi.org/10.1002/ppj2.20040