frontiers | Research Topics

Advances in cassava genomics, genetics and breeding

Edited by
Xiaofei Zhang, Robert Kawuki, Ismail Rabbi and
Eder Jorge Oliveira

Published in
Frontiers in Plant Science

ISSN 1664-8714
ISBN 978-2-8325-2589-0
DOI 10.3389/978-2-8325-2589-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Advances in cassava genomics, genetics and breeding

Topic editors

Xiaofei Zhang — International Center for Tropical Agriculture (CIAT), Colombia
Robert Kawuki — National Crops Resources Research Institute (NaCRRI), Uganda
Ismail Rabbi — International Institute of Tropical Agriculture (IITA), Nigeria
Eder Jorge Oliveira — Embrapa Mandioca e Fruticultura, Brazil

Citation

Zhang, X., Kawuki, R., Rabbi, I., Oliveira, E. J., eds. (2023). *Advances in cassava genomics, genetics and breeding*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-2589-0

Table of contents

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 16 September 2022
DOI 10.3389/fpls.2022.973206

Long-day photoperiod and cool temperature induce flowering in cassava: Expression of signaling genes

Peter T. Hyde and Tim L. Setter*

Section of Soil and Crop Sciences, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States

Check for updates

OPEN ACCESS

EDITED BY
Eder Jorge Oliveira,
Embrapa Mandioca e Fruticultura,
Brazil

REVIEWED BY
Paul Gibson,
Makerere University, Uganda
John Saviour Yaw Eleblu,
University of Ghana, Ghana
Ueric José Borges de Souza,
Federal University of Tocantins, Brazil

*CORRESPONDENCE
Tim L. Setter
TLS1@cornell.edu

SPECIALTY SECTION
This article was submitted to
Crop and Product Physiology,
a section of the journal
Frontiers in Plant Science

RECEIVED 19 June 2022
ACCEPTED 23 August 2022
PUBLISHED 16 September 2022

CITATION
Hyde PT and Setter TL (2022)
Long-day photoperiod and cool temperature induce flowering in cassava: Expression of signaling genes.
Front. Plant Sci. 13:973206.
doi: 10.3389/fpls.2022.973206

Cassava is a staple food crop in the tropics, and is of particular importance in Africa. Recent development of genomic selection technology have improved the speed of cassava breeding; however, cassava flower initiation and development remains a bottleneck. The objectives of the current studies were to elucidate the effect of photoperiod, temperature and their interactions on the time of flowering and flower development in controlled environments, and to use RNA-sequencing to identify transcriptome expression underlying these environmental responses. Compared to a normal tropical day-length of 12 h, increasing the photoperiod by 4 h or decreasing the air temperature from 34/31 to 22°/19°C (day/night) substantially hastened the time to flowering. For both photoperiod and temperature, the environment most favorable for flowering was opposite the one for storage root harvest index. There was a pronounced treatment interaction: at warm day-time temperatures, percent flowering was low, and photoperiod had little effect. In contrast, at cooler temperatures, percent flowering increased, and long-day (LD) photoperiod had a strong effect in hastening flowering. In response to temperature, many differentially expressed genes in the sugar, phase-change, and flowering-time-integrator pathways had expression/flowering patterns in the same direction as in Arabidopsis (positive or negative) even though the effect of temperature on flowering operates in the reverse direction in cassava compared to Arabidopsis. Three trehalose-6-phosphate-synthase-1 (TPS1) genes and four members of the SPL gene family had significantly increased expression at cool temperature, suggesting sugar signaling roles in flower induction. In response to LD photoperiod, regulatory genes were expressed as in Arabidopsis and other LD flowering plants. Several hormone-related genes were expressed in response to both photoperiod and temperature. In summary, these findings provide insight on photoperiod and temperature responses and underlying gene expression that may assist breeding programs to manipulate flowering for more rapid crop improvement.

KEYWORDS

flower induction, flower initiation, photoperiod, temperature, transcriptome expression, RNA-seq, ambient temperature, *Manihot esculenta*

Introduction

Cassava (*Manihot esculenta,* Crantz) is a tropical crop grown as a source of food and specialty starch products. More than half of the worldwide production occurs in Africa where hundreds of millions of people depend on cassava as a staple food (Jarvis et al., 2012; Parmar et al., 2017). The multiple uses of cassava include food from the storage roots and leaves, tapioca and other processed starch products, and livestock feed (Parmar et al., 2017). The ability of cassava to produce appreciable yields under sub-optimal conditions has led to its wide adoption by both small- and large-holder farmers. Furthermore, with respect to climate change, given its relative tolerance of drought and high optimal temperature for growth, it is predicted to be one of the least adversely affected staple-food crops in sub-Saharan Africa (SSA) (Jarvis et al., 2012), further increasing the need for research and development of this vital crop.

Breeding is needed to develop cultivars for the diversity of farmer's preferences including high yield, disease resistance and consumer-preferred quality traits. Village surveys have determined that small holder farmers grow multiple cultivars in regions such as Uganda (Iragaba et al., 2020) and Ghana (Rabbi et al., 2015). These different varieties serve multiple needs including risk aversion and market demands (Nakabonge et al., 2018). Furthermore, consumer-preferred traits vary by gender and locale (Teeken et al., 2018; Iragaba et al., 2021) such that including these varied traits in breeding schemes can improve adoption of improved varieties (Iragaba et al., 2021). Breeding for resistance to newly emerging diseases such as cassava brown streak disease (CBSD) is needed to reduce devastating outbreaks (Kawuki et al., 2016). These reports emphasize the need to breed a large diversity of improved cultivars that can be made available to farmers.

Recent developments of genomic selection technology have improved the speed of cassava breeding (Wolfe et al., 2017; Andrade et al., 2019). However, reliable and prompt flower initiation and development, which are essential for conventional breeding and genomic selection, remains a bottleneck (Alves, 2002; Ceballos et al., 2012; Andrade et al., 2019). The timing of floral initiation of the apical meristem is exceptionally variable, with some varieties first developing an inflorescence as early as 2 months after planting and others almost never produce an inflorescence (Ceballos et al., 2004; Adeyemo et al., 2019; Pineda et al., 2020a; Tokunaga et al., 2020). Even if a plant initiates an inflorescence it often produces few or no viable female flowers, further hindering breeding efforts (Adeyemo et al., 2019; Hyde et al., 2020; Pineda et al., 2020b).

Flowering in cassava has long been thought to be induced by long-day photoperiods (Keating et al., 1982; Pineda et al., 2020a; Souza et al., 2020; Tokunaga et al., 2020); however, these studies involved naturally occurring environmental conditions in the field where interpretation of the response to one environmental property is confounded by variation in several others. For example, Keating et al. (1982) planted cassava over several months in a location at 27°S latitude and observed that time-to-flowering was earlier in mid-summer, leading them to suggest that cassava flowers in response to long days, even though there were other co-variate factors, such as temperature, which were not evaluated. Similarly, studies that involved evaluating flowering over a 1-year time-frame have led researchers to conclude that cassava responds to long days (Souza et al., 2020). More recently, studies in controlled environments have shown that cassava flowering is indeed induced in long days, but also by cool temperatures (22°C) (Adeyemo et al., 2019; Oluwasanya D. N. et al., 2021).

In the model species Arabidopsis, signaling pathways that regulate the transition from vegetative growth phase to reproductive phase have been identified and characterized in detail. These pathways include circadian, hormone, autonomous, age, sugar, vernalization, ambient temperature, and photoperiod (Fornara et al., 2010; Blümel et al., 2015). There is a convergence of these signaling pathways toward a small number of floral integrator genes, notably FLOWERING LOCUS T (FT). FT encodes a phloem-mobile signaling protein which travels from the leaves to the apical meristem where it promotes floral induction. Several genes in the photoperiod pathway have been identified in cassava as homologs of their counterparts in Arabidopsis and other species, including MeFT1 and MeFT2, which are homologs of Arabidopsis FT (Adeyemo et al., 2019; Behnam et al., 2021). A large number of other photoperiod signaling homologs have been identified in cassava and their expression detected by transcriptome analysis, including TERMINAL FLOWER1 (TFL1), CONSTANS (CO), GIGANTEA (GI), and TEMPRANILLO1 (TEM1) (Adeyemo et al., 2019; Tokunaga et al., 2020; Behnam et al., 2021; Oluwasanya D. et al., 2021). Additionally, overexpression of Arabidopsis FT in cassava induces both flower initiation (Adeyemo et al., 2017; Bull et al., 2017; Odipio et al., 2020) and flower proliferation (Adeyemo et al., 2017). Together, this shows that many of the genes regulating flowering in model species have homologs in cassava and behave in a similar manner.

In the current study we elucidated the effect of photoperiod, temperature and their interactions on the time-to-flowering and flower development in cassava genotypes. To further our understanding of the mechanisms behind these environmental effects we investigated the transcriptome of plants grown under controlled environments of temperature and photoperiod. For several flowering regulatory pathways, we compared the transcriptome expression in cassava with that seen in Arabidopsis and other species, and obtained insight that may assist breeding programs in manipulating flowering for more rapid crop improvement.

Materials and methods

Plant material

Five cassava genotypes were used in the photoperiod and temperature experiments. TMS-IBA-980002 (also known as TMSI980002) and TMEB419 were obtained from the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria; Nase14, Nase3 (also known as TMS30572) and TME204 were obtained from the National Crop Resources Research Institute (NaCRRI), Namulonge, Uganda.

Growth conditions

Stem cuttings (stakes) were cut to about 15 cm length from the bottom 1 m of previously grown plants. Stakes were planted into 11-L pots (Polytainer #3; Nursery Supplies Inc., Chambersburg, PA, United States). Rooting media was a mixture of peat:vermiculite:perlite (62:22:11; v:v) with added dolomitic limestone and 2.2% (w:v) of fertilizer (10-5-10 Jacks Pro Media mix plus III; J.R. Peters, Inc., Allentown, PA, United States), as previously described (Hyde et al., 2020).

Photoperiod experiment 1 and 2

Two photoperiod experiments were conducted in a pair of matched growth chambers (Sherer, model CEL 511-38 walk-in room, 130 cm × 260 cm × 200 cm [depth × width × ht.], Sherer Inc., Marshall, MI, United States) with illumination by Philips cool white (4100 K) fluorescent lamps (Amsterdam, Netherlands) which provided 400 μmol m^{-2} s^{-1} (400–700 nm) photon flux density. In both Photoperiod Exp. 1 and 2, treatments were short day (SD), with 10 h of illumination from 6:00 until 16:00, and long day (LD) with 10 h of full illumination from 06:00 until 16:00 and with an additional 4 h of illumination with 10 μmol m^{-2} s^{-1} (dim light extension) from 16:00–20:00 provided by red light-emitting-diode (LED) lamps (spectral peak at 660 nm). In Photoperiod Exp. 1, temperature was 25°C (±0.38°C STD) starting at the beginning of the light period and extending 12 h, and 20°C for the second 12 h of a 24-h period. In Photoperiod Exp. 2, temperature was 30°C (±0.32°C SD) starting at the beginning of the light period and extending 12 h, and 25°C for the second 12 h of a 24-h period.

Each of the photoperiod experiments had two sequential batches of plants, each of which had 3 replicates of each genotype (TMSI980002, Nase 3, Nase 14 and TME 419) in a randomized block design, where batches (blocks) were considered a random effect. Photoperiod experiments were terminated when plants out-grew the height of the growth chamber; this averaged 182 d after planting (DAP).

Photoperiod x temperature experiment

The Photoperiod × Temperature Experiment was conducted in four matched growth chambers (Model CEL-63-10, Sherer Inc., Marshall, MI, United States) which had interior dimension of 112 cm × 74 cm × 83 cm (width × depth × ht.) and 400 μmol photons of photosynthetically active radiation (400–700 nm) m^{-2} s^{-1} at the top of the canopy, supplied by fluorescent lamps (Philips F48T8/TL841/HO). The two temperature treatments were (1) warm, with 35°C (±0.18 SD) from 6:00 until 18:00 (day) and 30°C (±0.12) from 18:00 until 06:00 (night), and (2) cool, with 25°C (±0.22) from 6:00 until 18:00 (day) and 20°C (±0.13) from 18:00 until 06:00 (night). Photoperiod treatments were short day (illumination of 400 μmol m^{-2} s^{-1} from 6:00 until 18:00) and long day (illumination of 400 μmol m^{-2} s^{-1} from 6:00 until 18:00 with 10 μmol m^{-2} s^{-1} dim light extension from 18:00–22:00). Lighting for dim light extension was provided by Philips Decorative Twister (4100K) lamps. The experiment had a 2 × 2 factorial arrangement in a randomized complete block design of two temperatures and two photoperiods (four treatment combinations). The experiment was run in four sequential batches (blocks), each containing a complete representation of the four genotypes and four temperature × photoperiod treatments. Treatments where imposed for an average of 114 days and experiments ended when plants out-grew the height of the growth chambers.

Temperature experiment

The Temperature Experiment was conducted in three matched growth chambers (Conviron Controlled Environments Ltd., Winnipeg, MB, Canada) (135 cm × 245 cm × 180 cm [depth × width × ht.]) with ten 400 W high pressure sodium and ten 400 W metal halide lamps, providing 600 μmol photons (400–700 nm) m^{-2} s^{-1} with a 12 h photoperiod. The daytime temperatures were 22, 28, and 34°C and night temperatures were 3°C lower than the day. Two sequential batches of plants were run, each including all three temperatures. While the purpose of the study was to elucidate temperature effects, we included a range of genotypes: the first batch had four replicate plants of TMSI980002 and three replicates of Nase 3, Nase 14 and TME 419, and the second batch had five replications of TMSI980002 and two replications of Nase 3, Nase 14 and TME 419. The study was an unbalanced randomized block (batches) design with blocks considered a random effect and temperature and genotype fixed effects.

Root zone temperature experiment

The Root-zone Temperature Experiment was conducted in the chambers described above for the photoperiod experiments. Plants were grown with 12-h daylength and chamber temperature of 30/25°C (day/night). Four root-zone temperature treatments were imposed: 15, 20, 30, and 40°C. Root zone temperatures were constant throughout night and day, and were obtained by installing about 1 m of copper tubing (9.5 mm outside dia.), which was coiled four turns such that the coils were about four cm from the periphery of the pot. Pots were insulated with 6-mm thick reflective bubble wrap insulation (Everbuilt Double Reflective Insulation, Home Depot Product Authority, Atlanta, GA, United States). Water was pumped through the coils, with the water temperatures thermostatically regulated by circulating thermo-controllers (Allied, Model 900, Fisher Scientific, Pittsburg, PA, United States), and soil temperature at the center of the pot was monitored and adjusted as needed. The duration of the experiment was 180 d. The experiment was a factorial arrangement of two genotypes (TMSI980002 and Nase 14) and four root-zone temperatures; two batches with two replicate plants for each treatment combination were run.

Gene expression in response to temperature

An additional set of plants grown in two growth chambers as described for the Temperature Experiment (above) was used to evaluate temperature effects on gene expression. The daytime temperature treatments were 22 and 34°C with night temperatures 3°C lower than the day. TMSI980002 and TMEB419 were used with 17 replicate plants at 22°C and 14 replicate plants at 34°C. Leaves were sampled as described below.

Gene expression in response to photoperiod

Leaves were sampled for analysis of gene expression from genotype TMEB 419 and Nase14 from Photoperiod Exp. 1 (described above). The experiment had photoperiods of 10 h (SD) and 14 h (LD) and day/night temperatures of 25/20°C. As described above, the LD treatment had 10 h of light at full flux density, followed by 4 h of dim light. Three replicate plants were used for each treatment × genotype combination. Leaf tissue was sampled 15 min before the end of the photoperiod at 69, 104, and 132 DAP, as described below.

Flower terminology and data collection

Flower induction in cassava occurs when the shoot apical meristem transitions to an inflorescence meristem, which is accompanied by the growth of two to four axillary buds directly below the inflorescence (Perera et al., 2013; Pineda et al., 2020a). Shoot growth from these buds forms a fork, which is indicative of the floral induction event. Subsequent transitions of the shoot apices on each of the fork branches to inflorescences is described as second tier forking. The identification of a developing fork was used to determine the timing of flowering. Although the inflorescences of cassava are technically (botanically) cyathia (Perera et al., 2013), we will refer to the entire structure of petal-like bracts and associated pistils or anthers/stamens as female or male flowers, respectively. We will refer to the entire reproductive stalk with multiple female and male flowers as an inflorescence.

Weekly counts of the number of flower buds greater than 2 mm diameter and mature flowers were used to calculate (1) the maximum number of flowers on a given week on an individual plant (maximum flower count), (2) the number of days that an individual plant had non-senesced flowers (flower retention), and (3) the sum of all the weekly flower counts (flower integral). At the final harvest, storage roots were counted, and above-ground and storage-root plant material was separated, dried and weighed.

Statistical analysis

Each experiment had a randomized complete block design. Mixed-model ANOVA was used with the modeled fixed effects including treatment (photoperiod and/or temperature), genotype, and genotype by treatment interaction. Batches of plants (complete blocks with all treatments and genotypes of a given experiment represented) were modeled as random effects, which accounted for batch-to-batch variation when the experiment was repeated in the same set of growth chambers over time. Linear Models and ANOVA were calculated using the lm and anova function of the "stats" package conducted in R studio (R Core Team, 2017). The emmeans package (Lenth, 2019) was used for mean comparisons both pairwise with *t*-tests and with multiple tests using Tukey–Kramer honest significant difference tests. The time to flowering or termination of experiment and the proportion of plants that flowered during the experiment were analyzed using the Cox proportional hazard test (Cox, 1972) in the *R* package "Survival" (Therneau, 2015). The loess curve-smoothing regression function (span = 2, degree = 2) of the "stats" package (R Core Team, 2017) was used to calculate the matrix used for the 3-dimensional image of photoperiod × temperature × percent flowering.

Analysis of gene expression with RNA-sequencing

Tissue sampling

Three leaf lobes were sampled from the youngest fully developed leaves on the upper nodes of plants of the temperature and photoperiod experiments from the youngest fully expanded, mature leaf on each plant. Samples were excised approximately 15 min prior to the dark period, enclosed in porous polyester tea bags and immediately submerged in liquid N_2 and subsequently transferred to a −80°C freezer awaiting RNA extraction.

RNA extraction

Total RNA was extracted by a modified CTAB protocol, and purified on silica RNA columns as previously described (Oluwasanya D. et al., 2021). Samples were ground in a mortar and pestle chilled with liquid N_2; about 0.5 g of the powder was vortexed for 5 min with 1 mL of extraction buffer containing 1% [w/v] CTAB detergent, 100 mM Tris-HCl [pH 8.0], 1.4 M NaCl, 20 mM EDTA, and 2% [v/v] 2-mercaptoethanol followed by 0.2 mL of chloroform; the suspension was mixed for 1 min, tubes were centrifuged and the top layer was moved to a new tube and 700 μL of a buffer containing 4 M guanidine thiocyanate, 10 mM MOPS (pH 6.7) and 500 μL of 100% ethanol (100%) was added and mixed. This mixture was applied to silica RNA columns (RNA mini spin column, Epoch Life Science, Missouri City, TX, United States), then washed sequentially with 750 μL each of 10 mM MOPS-HCl [pH 6.7] with 1 mM EDTA, containing 80% [v/v] ethanol, then 80% ethanol (twice), and to elute the RNA, 20 μL RNAase-free water. The RNA quality was evaluated with a gel system (TapeStation 2200, Agilent Technologies, Santa Clara, CA, United States).

3′RNA sequencing

The 3′RNA-seq libraries were prepared from ∼500 ng total RNA at the Cornell Genomics facility[1] using the Lexogen QuantSeq 3′ mRNA-Seq Library Prep Kit FWD for Illumina (Greenland, NH, United States). For each experiment (temperature and photoperiod), the pool was sequenced on one lane of an Illumina NextSeq500 sequencer using Illumina bcl2fastq2 software. Illumina adapters were removed from the de-multiplexed fastq files using Trimmomatic (version 0.36; Bolger et al., 2014). Poly-A tails and poly-G stretches of at least 10 bases in length were then removed using the BBDuk program in the package BBMap[2] (version 37.50), keeping reads at least 18 bases in length after trimming. Poly-G stretches result from sequencing past the ends of short fragments (G = no signal). The trimmed reads were aligned to the *Manihot esculenta* genome assembly 520_v7 (Mesculenta_520_v7.fa[3]) using the STAR aligner (version 2.7.0f; Dobin et al., 2012) allowing a read to map in at most 10 locations -outFilterMultimapNmax 10 with at most 6% mismatches (-outFilterMismatchNoverLmax 0.06), while filtering out all non-canonical intron motifs (-outFilterIntronMotifs RemoveNoncanonicalUnannotated). For the STAR indexing step, the number of reads overlapping each gene in the forward strand were counted using HTSeq-count [version 0.6.1 (Anders et al., 2015)].

Differential expression analysis

Analysis of differential gene expression was accomplished using the DESeq2 package (Love et al., 2014), which adjusts *P*-values for multiple testing due to the large number of tests. Manihot esculenta genes and their homologs from Arabidopsis thaliana and functional annotations were sourced from Phytozome13 (Bredeson et al., 2016). Arabidopsis flowering time genes, their pathways and expected effects on flowering were acquired from the Flowering Interactive Database FLOR-ID (Bouché et al., 2015). A list of 498 Manihot esculenta genes (version 7.1) and their annotations was created by matching them in Phytozome13[4] (accessed 2021.08.08) with corresponding flowering time genes from FLOR-ID database for Arabidopsis. Also, 30 cassava homologs of flowering genes that were not included in auto-annotation, were added to the list.

Results

Photoperiod and temperature effects on flowering age and abundance

A series of tests showed that both photoperiod and air temperature influence the timing of cassava flower initiation (**Table 1**). In experiment Photoperiod 1 (**Table 1A**), with 25/20°C, the extended photoperiod (10 + 4 h) treatment hastened the days-to-flower at flowering tier 1 for the four genotypes by 22 d and increased the percentage of plants that flowered during the experimental period to 75% in the 10 + 4-h treatment, compared to 33% in the 10-h treatment. In experiment Photoperiod 2 (**Table 1B**), extended photoperiod decreased the average days-to-flower by 44 d and increased the percentage flowering from 0 to 58%. These effects were statistically significant ($P \leq 0.05$) according to a Cox Proportional Hazard model, which considers both

[1] http://www.biotech.cornell.edu/brc/genomics-facility

[2] https://sourceforge.net/projects/bbmap/

[3] https://genome.jgi.doe.gov

[4] https://phytozome-next.jgi.doe.gov/

TABLE 1 Response of cassava to varying photoperiods and temperatures on time-to-flower and per-cent of flowering, in the first and second tier.

						Tier 1			Tier 2			Whole plant		
Experiment (table section)	Air Temperature (Day/Night)	Photo-period (h)	Root Zone Temperature (constant)	No. Genotypes	No. Plants	Days to flowering or termination	Percent flowering	Cox Proportional Hazard[†]	Days to flowering or termination	Percent flowering	Cox Proportional Hazard	Dried Weight (g)[¶]	Number of Roots	HI[§]
(A) Photoperiod-1	25/20°C	10		4	26	134	33%	a	169	8%	a	167 a	15 a	0.62 a
		10 + 4*				112	75%	b	145	58%	b	180 b	14 a	0.48 b
(B) Photoperiod-2	30/25°C	10		4	24	207	0%	a	207	0%	a	262 a	9 a	0.39 a
		10 + 4				163	58%	b	198	25%	b	234 a	7 a	0.27 b
(C) Temperature	22/19°C	12		4	56	91	96%	a	137	75%	a	271 a	13 a	0.32 a
	28/25°C					116	75%	a	145	63%	a	557 b	12 ab	0.43 b
	34/31°C					151	21%	b	169	4%	b	531 b	9 b	0.42 b
(D) Root Zone Temperature	25/20°C	12	20°C	2	8	91	100%	a	NA			204 a	16 a	0.40 a
			30°C			81	100%	a				333 b	11 b	0.65 b
(E) Photoperiod × Temperature	2520°C	12 + 4		4	64	112	100%	a	NA			68 a	12 a	0.46 a
		12				132	50%	b				69 a	13 a	0.47 a
	35/30°C	12 + 4				154	0%	c				127 b	10 b	0.64 b
		12				154	0%	c				128 b	10 ab	0.70 b

Plant growth was evaluated on a whole plant basis including whole plant dried biomass (Dried Weight), number of storage roots (Roots), and harvest index (HI). Shown are averages across genotypes and summary data for the five experiments; further details on each genotype are shown in Supplementary Tables. *Photoperiod +4 indicates a photoperiod extension with 4 h of dim light ca. 10 µmol m^{-2} s^{-1}. [†]Treatments within individual experiments not labelled with the same letter are significantly different using the Cox Proportional Hazard test with Bonferroni correction for multiple comparisons. This test uses a model that considers both the time that passes before a flower initiation event occurs and the fraction of plants that flower within the period of observation. [¶]Treatments within an individual experiment not connected by same letter are significantly different using the t test for paired comparisons and Tukey HSD for multiple comparisons. [§]HI (harvest index) is calculated as the proportion of the whole plant dried biomass that is dried storage root.

the proportion of plants that flower and the time that passes before a flower initiation event occurs (Cox, 1972). Similar treatment effects were observed at flowering tier 2. Comparing across the two experiments, flowering was earlier when plants were grown at cooler temperatures of 25/20°C (Photoperiod 1) than at warmer temperatures of 30/25°C (Photoperiod 2). Genotypic differences were seen in terms of the magnitude of treatment effect, but not a crossover interaction (Supplementary Tables 1, 2).

In the Temperature Experiment (Table 1C), plants grown at the coolest temperature of 22/19°C (day/night) flowered earliest (91 DAP), and had the highest percentage of plants flowering during the observation period (96%). When grown at the moderate temperature of 28/25°C, the mean age of flowering and percent flowering was intermediate (116 DAP, 75%), and at the highest temperature of 34/31°C the average age of flowering was the latest (151 DAP) and only 21% of plants flowered. Treatment comparisons were statistically significant ($P \leq 0.05$) between the two lower temperatures and the high temperature using the Cox Proportional Hazard model to evaluate days-to-flowering and percent flowering. Flowering at the second tier had a similar pattern: plants at the two lower temperatures (daytime: 22 and 28°C) flowered earlier and a higher percentage flowered than at 34°C. Different genotypes had different responses to temperature in terms of the magnitude of the effect; however, there was not a crossover interaction and all genotypes had earlier flowering and a greater percentage of flowering at lower temperatures (Supplementary Table 3).

In the root zone temperature experiment (Table 1D), varying the root zone temperature 5°C above or below the air temperature of 25°C did not significantly affect days-to-flower and percent of plants that flowered, indicating that temperature response is likely due to above-ground processes.

To evaluate photoperiod × temperature interaction, a study with two temperatures (25/20°C and 35/30°C; day/night) and two photoperiods (12 h daylength and 12 + 4 h) was conducted (Table 1E, Photoperiod × Temperature Exp.). At 35/30°C, flowering was not observed during the experimental period in either photoperiod. However, at 25/20°C, flowering occurred in both photoperiods; among them, long daylength (LD) induced earlier days-to-flower and a higher percent flowering compared to short daylength (SD).

The response of each genotype to photoperiod × temperature indicated that in all genotypes except, TMSI980002, flowering was earliest at LD and cool temperature (Figure 1). Differences in flowering were indicative of a significant genotype by treatment interaction detected by ANOVA and Cox Proportional Hazard test (Supplementary Table 4). In TMSI980002, which began flowering much earlier than the others, plants in LD and SD flowered similarly at 25/20°C. In

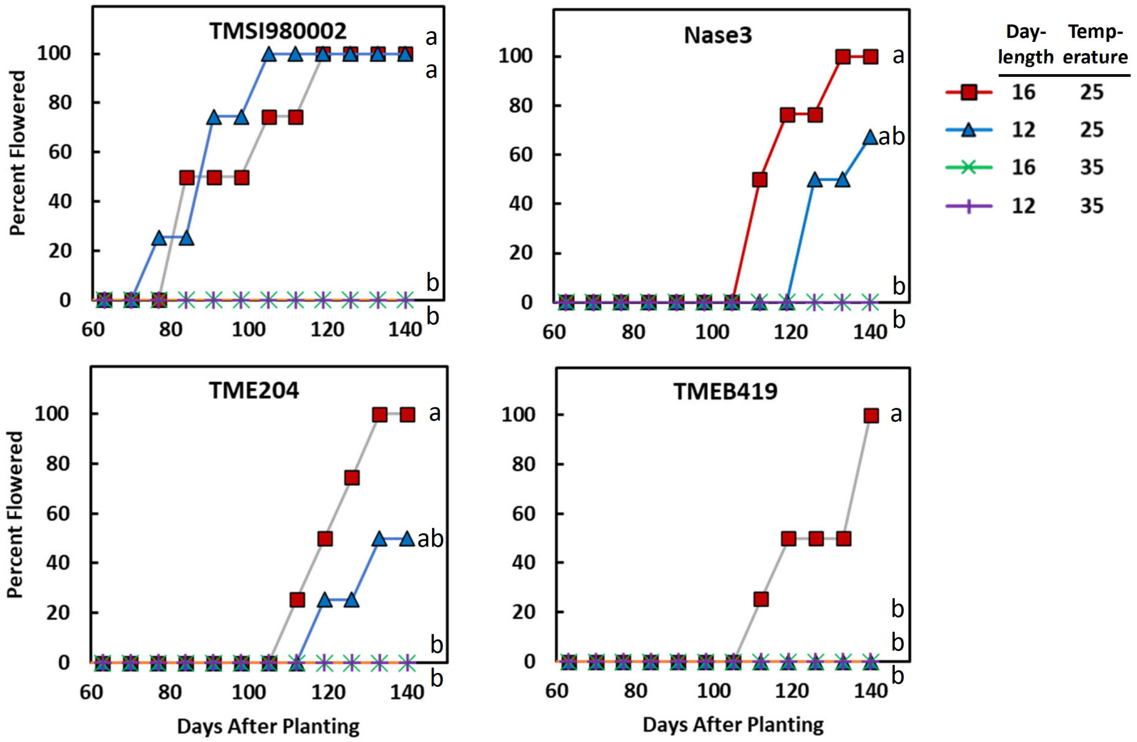


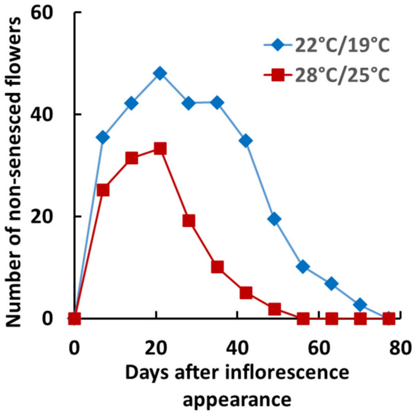
FIGURE 1
The effect of photoperiod and temperature on percentage of plants flowering as a function of time in the genotypes TMSI980002, Nase 3, TME 204, and TMEB 419. Day-lengths were 16 h (12 h full light + 4 h dim light) and 12 h full light. Temperatures were 25/20°C (day/night) and 35/30°C. Data shown represent the average of 4 replicate plants for each photoperiod × temperature × genotype combination. Curves labeled with different letters differed ($P \leq 0.05$) according to Cox Proportional Hazard tests, which considers both the proportion of plants that flower and the time-to-flowering.

the genotype with latest flowering, TMEB419, flowering only occurred in LD and cool temperature.

Partitioning of carbon for the growth of alternative plant parts was also affected by photoperiod and temperature, but differently than flowering (Table 1, columns on right). At both 25/20 and 30/25°C, in Photoperiod Exp. 1 and 2, respectively, plants in the 10-h photoperiod were not significantly different from those in 14-h photoperiod in their total plant weight or root count; however, HI was significantly greater with 10-h daylength. When cassava was grown with 12-h daylength at three temperatures, both total plant dry weight and HI were higher at 28/25 and 34/31°C than at 22/19°C, though the number of storage roots were fewer at the warm temperatures. The photoperiod by temperature experiment also showed that when grown at 35/30°C, plants had a greater total weight and HI than plants grown at 25/20°C. Hence, for both photoperiod and temperature, the environment most favorable for flowering was opposite of the one for storage root HI: flowering was favored in LD and cool environments; whereas storage root HI was favored in SD and warm environments.

To determine the extent to which carbon partitioning responds to root temperature, we grew plants at a common above-ground temperature and subjected root-zones to two temperatures (Table 1D, Root Zone Exp.). At a warm root-zone temperature of 30°C compared to 20°C, flowering was not affected, while both total plant dry weight and storage root HI were higher at a root-zone temperature of 30°C than 20°C. Thus, whereas the temperature response of flowering was apparently due to above-ground temperature, carbon partitioning attributes were affected by both root-zone and whole-plant temperature.

Temperature also affected flower prolificacy, i.e., the number of flowers produced per plant, and inflorescence longevity, the time-frame over which flowers were produced and remained viable/non-senescent, as illustrated by the graph in Figure 2. For this experiment the late flowering lines (TME204 and TMEB419) were not included and an early line (Nase14) was substituted. At 22°C (day-time), the count number of non-senesced flowers averaged across the three genotypes reached a maximum of 49 flowers, whereas at 28°C, the average maximum was 37 flowers (Figure 2, embedded table). Longevity was also greater at the cooler temperature: at 22°C the average days of flower retention was 33 d, whereas at 28°C it was 17 d. The integral of flower counts over time (area under the



Genotype	Air Temperature (Day/Night)	Maximum Flower Count	Days of Flower Retention	Flower Integral
Nase 14	22°C/19°C	12 a	13 a	64 a
	28°C/25°C	0 b	0 b	0 b
Nase 3	22°C/19°C	44 a	49 a	205 a
	28°C/25°C	7 b	9 b	13 b
TMS I980002	22°C/19°C	72 a	37 a	247 a
	28°C/25°C	59 a	25 a	154 a
Combined Genotypes	22°C/19°C	49 a	33 a	187 a
	28°C/25°C	37 b	17 b	95 b
ANOVA †	source		P-value	
Treatment Effect		0.4808	0.0225	0.1591
Genotype Effect		<0.0001	<0.0001	<0.0001
Genotype X Treatment		0.1302	0.0308	0.0617
Block		0.0527	<0.0001	0.0022

*Comparisons between treatments within each genotype which do not have the same letter are significantly (P ≤ 0.05) different using a t test; based on log transformation of the data.

†ANOVA based on a model with temperature Treatment, Genotype, Genotype X Treatment interaction, and block effects; shown are probabilities that the effect was not a significant source of variation.

FIGURE 2

Floral development in genotypes Nase 14, Nase 3, and TMSI980002 at tier 2, grown at day/night temperatures of 22/19 and 28/25°C. Graph represents weekly mean flower counts averaged across all three genotypes.

curve) was 187 flower-days at 22°C and 95 flower-days at 28°C. Genotypes differed in these properties: In Nase14 and Nase 3, maximum flower counts, retention, and flowering integral were substantially greater at 22°C than 28°C ($P < 0.05$). In TMSI980002 flowering was abundant by all measures, though the effects of temperature on flowering were not significant ($P \leq 0.05$).

Gene expression in response to temperature and photoperiod

We conducted two studies to determine gene expression of the transcriptome in response to environmental factors: (1) a comparison of plants grown at cool temperatures (22/19°C day/night) vs. warm temperatures (34/31°C day/night), and (2) a comparison of plants grown in long-day (14 h) vs. short-day (10 h) photoperiods. In both cases, mature leaf tissue was sampled 15 min before the end of the photoperiod. Using an experiment-wise adjusted *P*-value (P_{adj}) of 5% and genome-wide statistical analyses of differentially expressed genes, we identified 7946 genes that differed in the comparison of plants grown at 22 vs. 34°C (Supplementary Table 5), and 6616 genes that differed in the LD vs. SD comparison (Supplementary Table 6). To provide an overview of the types of genes that were affected by these environments, we performed enrichment analysis with ShinyGO (Ge et al., 2019) on each group of genes differentially expressed in response to treatments. This analysis identified gene ontology (GO) categories which were represented in greater proportion than what would be expected by random chance. In the temperature experiment, among genes that had significantly higher expression at 34°C than 22°C, genes upregulated by 34°C were enriched, based on the statistical false discovery rate (FDR) in the GO categories "response to stress" (FDR 1.4×10^{-25}; 435 genes), "response to heat" (FDR 3.3×10^{-17}; 61 genes) and "response to temperature stimulus" (FDR 5.2×10^{-11}; 95 genes) (Supplementary Table 7). Collectively, among genes upregulated by 34°C, there were 593 genes in stress- and heat-related categories, whereas among genes downregulated by 34°C, there were only 102 genes. Hence, the high temperature treatment probably induced expression of a large number of genes not directly related the regulation of flowering, such as those associated with high temperature.

Enrichment analysis for the photoperiod experiment indicated that among genes downregulated in the LD treatment, GO categories related to photosynthesis and response to light were enriched relative to what could be attributed to random chance (Supplementary Table 7). In GO categories related to photosynthesis and response to light, there were 11 genes upregulated by LD relative to SD, and 166 downregulated in LD (i.e., upregulated in SD). This outcome is consistent with the light levels that existed when leaves were sampled for transcript analysis, which was about 15 min before the beginning of the dark period in each case. In the SD treatment, leaves were sampled at the end of the photosynthetic period when photon flux was high (400 μmol m^{-2} s^{-1}) and photosynthesis was active. In the LD treatment, leaves were sampled during the dim-light extension of the photoperiod when photon flux was low (10 μmol m^{-2} s^{-1}) and photosynthesis was minimal. Thus the expression profiles for both the temperature experiment and the photoperiod experiment likely included a large number of genes not directly related to regulation of flowering.

To assess expression of genes involved in flowering, we focused on cassava homologs of genes identified in the FLOR-ID database of flowering-related genes, a database that includes genes with both direct and indirect roles in signaling/regulating flower development (Bouché et al., 2015). From this database of Arabidopsis flowering-related genes, we obtained a list of 498 cassava homologs. Also, we added 30 cassava homologs of flowering-related genes that were not in the initial list (**Supplementary Table 8**). Among these genes, in the temperature experiment, plants grown at 22 vs. 34°C had 198 flowering-related differentially expressed genes (DEG) using an experiment-wise adjusted *P*-value of 5% ($P_{adj} \leq 0.05$) (**Supplementary Table 8**). Based on the FLOR-ID database, we classified these genes into individual flower-signaling and regulatory pathways (Bouché et al., 2015).

Temperature

Several differentially expressed cassava genes ($P_{adj} \leq 0.05$) that were categorized into the ambient temperature pathway were expressed at lower levels in cool than warm temperatures (**Table 2**). In cooler growth conditions, expression of cassava homologs of the Arabidopsis PIF family and FCA genes were decreased whereas SVP was increased. PIF-family genes are of particular interest because in Arabidopsis, PIF4 has a key role in the ambient temperature pathway and in regulation of flowering (Kumar et al., 2012; Proveniers and van Zanten, 2013). A comparison of amino acid sequences of the PIF homologs in cassava indicates that the cassava PIF that was significantly expressed in response to temperature, Manes.13G043000, had close similarity to Arabidopsis PIF4 and PIF3 (**Table 2**; **Supplementary Figure 1**). Given that cassava flowering was enhanced by cooler temperatures (**Table 1**), and that expression of cassava PIF4 and FCA homologs were significantly lower at cool temperature (**Table 2**), expression of these genes was negatively correlated with flowering. SVP was positively correlated with flowering. However, these expression patterns were opposite that found in Arabidopsis. While the temperature effects on expression were in the same direction in both species (lower expression at cool than warm temperature), in Arabidopsis warm temperature promotes flowering so expression of PIF and FCA is positively correlated with flowering and SVP expression is negatively correlated with flowering. Based on these findings we suggest that for these genes the mechanism by which temperature affects flowering in cassava differs from that in Arabidopsis, and may depend on additional factors or have opposite effects.

Several cassava genes which responded to temperature ($P_{\text{adj}} \leq 0.05$) were classified into the sugar and aging pathways. The sugar pathway relates to the enhancement of flowering in response to abundant photosynthetic activity, while the "aging" pathway refers to developmental regulation of phase changes, notably the phase transitions from juvenile to adult and from vegetative to reproductive development. A high fraction of these genes (80% in sugar pathway and 100% in aging pathway) had correlations between gene expression and flowering that agreed with the direction of the response (positive vs. negative) in Arabidopsis (**Table 3**). Four genes in the sugar signaling and response pathway (TPS1, PGM1, SUS4, and ADG1) had higher expression in cassava plants grown at flower-enhancing 22 vs. 34°C. This is an expression pattern with the same relationship to flowering as in Arabidopsis (positive correlation). While several of the genes in this pathway could be viewed as enzymes whose role is metabolism rather than signaling (PGM1, SUS4, and ADG1), TPS1 (trehalose-6-phosphate synthase-1) is considered to have roles in signaling sugar status and regulating metabolic and developmental processes (Wahl et al., 2013; Ponnu et al., 2020). A comparison of amino acid sequences of the TPS1 homologs in cassava indicates that the cassava TPS1 genes that were significantly expressed in response to temperature (Manes.15G116000, Manes.15G098800, and Manes.17G062500), had close similarity to Arabidopsis TPS1 (**Table 3**; **Supplementary Figure 2**).

In the developmental age pathway (regulation of developmental phase transitions), SPL genes (SPL3, SPL5, and SPL9) had higher expression in cassava plants which had enhanced flowering at 22°C (**Table 3**), matching their positive correlation with flowering in Arabidopsis. Also, in the aging pathway four different TPL homologs, and one TOE1 homolog had expression that matched the Arabidopsis direction of relationship to flowering, though in this case the magnitude of expression was negatively correlated with tendency for

TABLE 2 Significant differentially expressed genes (DEGs) and log2 Fold Change of core downstream genes in the ambient temperature pathway.

Arabidopsis gene name	DESeq2 results: expression at 22 vs. 34°C		Correlation between gene expression and flowering			
	Log2 Fold Change	P_{adj}	Cassava	Arabidopsis	*M. esculenta* Gene	*A. thaliana* Gene
PIF	−1.34	0.0001	Negative	Positive	Manes.13G043000	AT2G20180.2
FCA	−0.60	0.0003	Negative	Positive	Manes.03G206500	AT4G16280.2
FCA	−0.54	0.0126	Negative	Positive	Manes.01G230100	AT4G16280.4
SVP	0.64	0.0074	Positive	Negative	Manes.10G099000	AT2G22540.1

flowering at 22°C. Furthermore, three flower development and identity DEGs homologous with FUL and AP2, and two homologs of the flower time integrator gene, SOC1, were expressed in the same manner as Arabidopsis. These findings indicate that many of the DEGs in the sugar, aging, and flowering-time-integrator pathways had expression/flowering patterns in the same direction as in Arabidopsis (positive or negative) even though the effect of temperature on flowering operates in the reverse direction in cassava compared to its direction in Arabidopsis.

Photoperiod

To evaluate transcript expression in response to photoperiod, we compared plants grown in a short-day (SD) 10 h photoperiod vs. in a long-day (LD) extended photoperiod (10 h + 4 h). The sampling time was about 15 min before the dark period in each case, thus targeting differential expression caused by photoperiod signaling systems. Due to this sampling method, we have avoided placing our focus on genes whose expression is strictly related to time-of-day and the circadian cycle, and instead have focused on photoperiod-related genes. Two phytochrome genes PHYA and PHYB were both differentially expressed between long and short days (Table 4). PHYA was upregulated in long days whereas PHYB was down regulated, which indicates PHYA expression was positively correlated with flowering whereas PHYB was negatively correlated with flowering. These findings in cassava match the relationship of these genes to flowering in Arabidopsis. Two genes belonging to the family of phytochrome interacting factors (PIF) were also up-regulated by long days, matching the positive relationship between gene expression and flowering that is found in Arabidopsis. The cryptochromes

TABLE 3 Significant differentially expressed genes (DEGs) and log2 Fold Change from flowering pathways showing the most similarity to flower regulation in A thaliana and relevant down stream genes part of the flower development and meristem identity and flowering time integrator pathways.

		DESeq2 Results: expression at 22 vs. 34°C		Correlation between gene expression and flowering			
Pathway	**Arabidopsis gene name**	**log2 Fold Change**	**Padj**	**Cassava**	**Arabidopsis**	***M. esculenta* Gene**	***A. thaliana* Gene**
Sugar	ADG1, APS1	1.03	<0.0001	Positive	Positive	Manes.12G067900	AT5G48300.1
	ADG1, APS1	1.22	<0.0001	Positive	Positive	Manes.13G058900	AT5G48300.1
	AKIN10, SNRK1.1	0.42	0.0104	Positive	Negative	Manes.02G049300	AT3G01090.2
	HXK1, GIN2	−1.93	<0.0001	Negative	Positive	Manes.03G026700	AT4G29130.1
	PGM1	0.82	0.0094	Positive	Positive	Manes.06G141300	AT5G51820.1
	PGM1	0.67	0.0795	Positive	Positive	Manes.14G031100	AT5G51820.1
	SUS4	0.98	<0.0001	Positive	Positive	Manes.03G044400	AT3G43190.1
	SUS4	1.05	<0.0001	Positive	Positive	Manes.16G090600	AT3G43190.1
	TPS1	2.95	<0.0001	Positive	Positive	Manes.15G116000	AT1G78580.1
	TPS1	0.54	0.0003	Positive	Positive	Manes.15G098800	AT1G78580.1
	TPS1	1.63	0.0381	Positive	Positive	Manes.17G062500	AT1G78580.1
Developmental Age	SPL3	1.34	<0.0001	Positive	Positive	Manes.17G047500	AT2G33810.1
	SPL3	2.41	<0.0001	Positive	Positive	Manes.16G029900	AT2G33810.1
	SPL5	1.28	0.0118	Positive	Positive	Manes.03G106900	AT3G15270.1
	SPL9	1.80	<0.0001	Positive	Positive	Manes.09G032800	AT2G42200.1
	TOE1, RAP2.7	−1.06	0.0001	Negative	Negative	Manes.10G041100	AT2G28550.3
	TPL	−0.52	<0.0001	Negative	Negative	Manes.09G124300	AT1G15750.1
	TPL	−1.83	<0.0001	Negative	Negative	Manes.16G124800	AT1G15750.3
	TPL	−0.64	0.0131	Negative	Negative	Manes.04G108600	AT1G15750.4
	TPL	−0.27	0.0829	Negative	Negative	Manes.08G164000	AT1G15750.4
Flower development and meristem identity	AP2	−0.43	0.1422	Negative	Negative	Manes.12G106400	AT4G36920.1
	FUL, AGL8	1.67	<0.0001	Positive	Positive	Manes.14G088500	AT5G60910.1
	FUL, AGL8	0.70	0.0064	Positive	Positive	Manes.02G059300	AT5G60910.1
Flowering time integrator	SOC1, AGL20	0.29	0.1225	Positive	Positive	Manes.01G263500	AT2G45660.1
	SOC1, AGL20	0.22	0.4651	Positive	Positive	Manes.05G041900	AT2G45660.1

The reference condition was 34°C.

TABLE 4 Significant differentially expressed genes (DEGs) and log2 Fold Change from flowering pathways showing the most similarity to flower regulation in A thaliana and relevant down stream genes part of the phytochrome and cryptochrome light signaling pathways.

		DESeq2 Results: expression at Long-day vs. Short-day		Correlation between gene expression and flowering			
Pathway	Arabidopsis gene name	log2 Fold Change	Padj	Cassava	Arabidopsis	*M. esculenta* Gene	*A. thaliana* Gene
Phytochrome	PHYA	0.41	<0.0001	Positive	Positive	Manes.09G182500	AT1G09570.1
	PHYB	−0.48	<0.0001	Negative	Negative	Manes.03G205100	AT2G18790.1
	PIF	0.64	0.0008	Positive	Positive	Manes.12G044000	AT2G20180.3
	PIF	0.70	0.0191	Positive	Positive	Manes.13G043000	AT2G20180.2
	ATCOL2,COL2	2.11	<0.0001	Positive	Positive	Manes.01G106200	AT3G02380.1
	ATCOL2,COL2	1.54	<0.0001	Positive	Positive	Manes.02G062700	AT3G02380.1
	FBH3, AKS1, BHLH122	0.96	0.0001	Positive	Positive	Manes.06G167700	AT1G51140.1
	FBH4, AKS3	1.69	<0.0001	Positive	Positive	Manes.10G146400	AT2G42280.1
	FBH4, AKS3	−0.44	0.0010	Negative	Positive	Manes.09G031000	AT2G42280.1
Crypto-chrome	CRY1	0.84	<0.0001	Positive	Positive	Manes.02G152400	AT4G08920.1
	CRY1	0.48	<0.0001	Positive	Positive	Manes.18G067300	AT4G08920.1
	CRY2	0.31	0.0115	Positive	Positive	Manes.15G040500	AT1G04400.1
	COP1	−0.21	0.0108	Negative	Negative	Manes.12G068900	AT2G32950.1
	SPA1	−0.98	<0.0001	Negative	Negative	Manes.01G248600	AT2G46340.1

Short-day was the reference condition.

CRY1 and CRY2 also had higher expression in long days, with a relationship between expression and flowering that matched that in Arabidopsis, and two negative factors in the cryptochrome pathway, COP1 and SPA1, were expressed at lower levels in LD, a pattern that is in the same direction as in Arabidopsis. Genes downstream of the phytochrome and cryptochrome pathways, in the CONSTANS-like (COL) gene family, also had expression in the same positive direction in cassava and Arabidopsis. Three members of the FBH family of CONSTANS-interacting factors had significant differential expression in response to photoperiod, and two of these had expression with a positive correlation with flowering, as in Arabidopsis, while one was negative.

Significant differentially expressed genes found in both the temperature and photoperiod experiments

Given that both LD photoperiod and cool temperature induced flowering, we explored which genes were differentially expressed in response to both treatments. Eighteen differentially expressed genes were identified as significant in both the photoperiod and temperature experiments (Table 5). As expected, many of these genes appear to be in pathways downstream of initial environmental perception at points that integrate multiple signaling pathways to determine flowering time. For example, genes in the pathways for aging (developmental phase change), CONSTANS-like (COL), flower development, and meristem identity were among the genes that were significantly up-regulated in response to both LD and cool temperature. A circadian clock LUX homolog was down-regulated in response to both flower-inducing treatments. There were also several genes in various general pathways, and some related to hormone signaling which were significantly affected by both photoperiod and temperature. Among the hormone-related genes were those involving gibberellins (GID1C, GA2), cytokinin (RR2), and auxin (IAA7). These findings indicate that there are numerous regulatory pathways that operate in response to both environmental factors, thus helping us distinguish pathways that may interact or are additive from those that may operate differently in each environmental response.

Discussion

Effect of temperature and photoperiod on flower induction

The current studies were conducted with the overarching goal of facilitating cassava breeding of genotypes which are poor flowering but have desirable agronomic traits. To this end, we evaluated the effect of photoperiod and temperature, two environmental factors that are known to affect flowering in many flowering plant species (Yan and Wallace, 1996). Our

TABLE 5 Significant differentially expressed genes (DEGs) and log2 Fold Change from flowering pathways that where identified in both photoperiod and temperature induction of flowering in cassava.

		DESeq2 Results: expression at Long-day vs. Short-day		DESeq2 Results: expression at 22 vs. 34°C			
Pathway	**Arabidopsis gene name**	**log2 Fold Change**	P_{adj}	**log2 Fold Change**	P_{adj}	***M. esculenta* Gene**	***A. thaliana* Gene**
Aging	SPL3	0.71	<0.0001	1.34	<0.0001	Manes.17G047500	AT2G33810.1
Circadian Clock	LUX, PCL1	−1.12	<0.0001	−1.08	<0.0001	Manes.01G170000	AT3G46640.1
Flower development and meristem identity	FUL, AGL8	1.06	<0.0001	1.67	<0.0001	Manes.14G088500	AT5G60910.1
	AGL1,SHP1	0.76	<0.0001	2.46	<0.0001	Manes.02G085501	AT3G58780.1
	AGL4,SEP2	0.76	0.0028	1.23	0.0009	Manes.01G103100	AT3G02310.1
	AGL1,SHP1	0.41	0.0321	2.21	<0.0001	Manes.01G128500	AT3G58780.1
General	MRG1	−0.54	0.0166	−1.36	<0.0001	Manes.05G178100	AT4G37280.1
	OTS1, ULP1D	−0.33	0.0753	−0.51	0.0169	Manes.12G029500	AT1G60220.1
	HTA9	−0.95	<0.0001	−0.21	0.0924	Manes.03G018100	AT1G52740.1
	UBC2	0.28	0.0027	0.50	0.0107	Manes.08G154200	AT2G02760.1
	MYB30	0.45	0.0177	0.80	0.0020	Manes.09G135700	AT3G28910.1
Hormones	GID1C	−1.07	<0.0001	−0.62	0.0096	Manes.09G161600	AT5G27320.1
	RR2, ARR2	−1.69	0.0039	−1.38	0.0279	Manes.01G262600	AT4G16110.1
	GA2, ATKS1	−0.81	0.0751	−2.88	<0.0001	Manes.16G068951	AT1G79460.1
	MYB33	0.89	<0.0001	1.23	0.0007	Manes.11G009900	AT5G06100.2
	IAA7, AXR2	0.48	0.0038	1.15	<0.0001	Manes.03G169700	AT3G23050.1
Photoperiodism, light perception and signaling	ATCOL2,COL2	2.11	<0.0001	1.31	<0.0001	Manes.01G106200	AT3G02380.1
	ATCOL2,COL2	1.54	<0.0001	1.16	<0.0001	Manes.02G062700	AT3G02380.1

results show that compared to a normal tropical day-length of 12 h, increasing the photoperiod by 4 h decreased the time to flowering and increased the percentage of plants that flowered. Decreasing the air temperature from 34/31 to 22/19°C (day/night) greatly hastened the time to flowering on both the first and second tier of plant branching/flowering.

Temperature and photoperiod effects interacted (**Table 1E**). At the warm temperature of 35°C, extended photoperiod did not hasten flowering, whereas at cooler temperatures, genotypes responded to photoperiod and flowering was hastened by LD. A 3-dimensional summary of the response of percent flowering to photoperiod × temperature is shown in **Figure 3**. The response surface was calculated based on data from the all of the photoperiod and temperature experiments shown in **Table 1**. The response surface indicates that at warm day-time temperatures, percent flowering was low, and photoperiod had little effect. In contrast, at cooler temperatures, percent flowering increased, and a pronounced interaction with photoperiod induced the highest percent flowering with the combination of LD and cool temperatures.

Previous anecdotal evidence suggests that extended photoperiod is more effective at cooler temperatures (Pineda et al., 2020a), and previous growth chamber trials indicated cool temperature is favorable for flowering (Adeyemo et al., 2019; Oluwasanya D. N. et al., 2021). Our results substantiate that cassava breeders aiming to utilize extended photoperiod to produce more flowers would benefit from utilizing locations with cooler temperatures.

Temperature effect on transcriptome

Due to global climate change, interest in understanding temperature effects on all aspects of plant development including flowering has increased (Lee et al., 2013; Jin and Ahn, 2021). Recent work in Arabidopsis has elucidated several regulatory factors by which temperature affects time-to-flowering (Capovilla et al., 2014; Susila et al., 2018; Jin and Ahn, 2021). However, in contrast to our findings in cassava where flowering was enhanced by cooler temperature (**Table 1** and

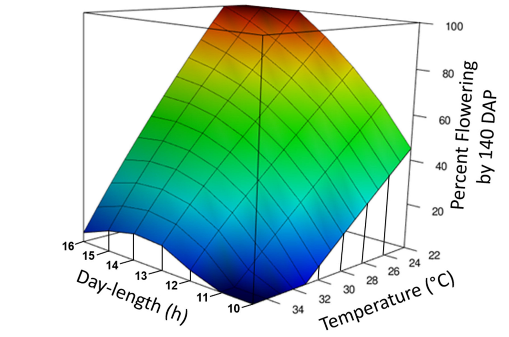


FIGURE 3
Graphical representation of photoperiod and temperature effect on percentage of plants induced to flower. The response surface was calculated by loess regression based on data from experiments Photoperiod 1, Photoperiod 2, Temperature, and Photoperiod × Temperature. All experiments were standardized by using the data for percent forking at 140 days after planting (DAP) and genotypes TMS98I0002, Nase3, and TMEB419. Day-time temperatures are shown.

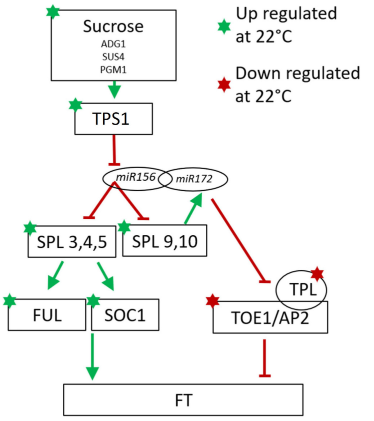


FIGURE 4
Minimal model of the of Arabidopsis sugar signaling and developmental phase change pathway for regulation of flowering with current findings for cassava gene expression indicated with stars for genes significantly up-(green) or down-(red) regulated at cool temperatures. Involvement of the non-coding microRNAs miR156 and miR172 are also shown. Arrows indicate promotive effects; transverse bars indicate inhibitory effects. Model adapted from: Wu et al. (2009), Wahl et al. (2013), Wang (2014), Ponnu et al. (2020).

Figures 1, **3**), in Arabidopsis, flowering is promoted by warmer temperature. Accordingly, in Arabidopsis, expression of the transcription factor PIF4 is increased at warmer temperatures and it binds to the promoter region of the key flower-inducing gene FT, thereby increasing its expression (Kumar et al., 2012). In the current work, expression of FT was below the detectable threshold of our RNA-seq method, so it was not possible to confirm that FT expression was higher at cool temperatures, as would be expected from our flowering results. Our findings indicate that expression of a member of the PIF family in cassava with close homology with AtPIF3 and AtPIF4 was also increased by warmer temperature (**Table 2** and **Supplementary Figure 2**); however, warmer temperature decreased flowering in cassava. It is plausible that the cassava PIF homolog that was up-regulated at warm temperature, Manes.13G043000, is operating similarly to Arabidopsis PIF3 which has been shown to inhibit flowering, and knockdown of its expression results in earlier flowering (Oda et al., 2004). Alternatively, it is possible that other factors are operational in cassava's response to temperature.

Other potential contributors to ambient temperature response are the transcription factor SVP, which in Arabidopsis interacts with the FT promoter and negatively regulates flowering (Lee et al., 2013), and FCA, which promotes flowering at higher temperatures through the induction of FT (Jung et al., 2012). In cassava these components of the thermosensory pathway were differentially expressed in response to temperature; however, their expression was correlated with temperature effects on flowering in the opposite direction (**Table 2**). In cassava flowering was enhanced at low temperature whereas in Arabidopsis flowering is enhanced by warmer temperature. These results indicate that these components of the ambient temperature pathway may operate differently in cassava, or interact with additional components.

In contrast to the lack of agreement between cassava and Arabidopsis in the correlation between temperature-regulated expression of the genes described in **Table 2**, a high proportion of flowering-related genes in the age signaling pathway (developmental phase change) and sugar signaling/response pathway were regulated in a direction similar to that seen in the model species Arabidopsis (**Table 3**). Phase change from juvenile to adult is needed for competence to flower (Poethig et al., 2013). A minimal model of these modules in Arabidopsis are represented in **Figure 4** with the expression direction of homologs from the current study overlayed to assist in interpreting the differential expression we have found in cassava (Wu et al., 2009; Huijser and Schmid, 2011).

Consistent with the model in **Figure 4**, in cassava many of the genes in the SPL family including SPL3, 4, and 5, which promote flowering, and SPL 9 which promotes the juvenile to adult phase transition, had higher expression in when grown under cool temperatures which also induced flowering. In Arabidopsis, SPL9 interacts with miR172 to block the floral repressors TOE1 and AP2, and their activator, TPL (Wu et al., 2009; Huijser and Schmid, 2011; Wang, 2014). Consistent with their expected effects on flowering, in cassava, expression of genes homologous to TOE1, AP2 and TPL was significantly decreased when grown under cool flower inductive conditions.

SPLs in Arabidopsis are direct regulators of SOC1 and FUL which influence FT expression in the leaves (Wang et al., 2009; Yamaguchi et al., 2009; Wang, 2014). Our transcriptome analysis shows a concurrent up regulation of SPL3, SPL4, SPL5, SOC1, and FUL in the same manner as seen in Arabidopsis. The amino acid similarity of cassava homologs of the Arabidopsis SPL family are shown in **Supplementary Figure 3**. It has been shown in some woody perennial species including Citrus and Jatropha, that homologs of SPL3 and SPL5 are up-regulated in relation to flower initiation (Shalom et al., 2015; Yu et al., 2020). In Jatropha, a close relative of cassava, 15 SPL homologs were identified with most being highly conserved (Yu et al., 2020). Nine of the Jatropha SPL genes are regulated by miR156 (Yu et al., 2020). Jatropha JcSPL3 has increased expression in the leaves of plants entering the flowering stage of development and expression of JcSPL3 in Arabidopsis triggers earlier flowering (Yu et al., 2020). Together, based on this evidence, we suggest that cassava may operate similarly to Jatropha where JcSPL3 is responsible for vegetative phase transition (Yu et al., 2020).

In Arabidopsis many other signaling molecules are known to interact with the miR/SPL module. Trehalose-6-phosphate (T6P) signals carbohydrate status of the plant to regulate flowering (Wahl et al., 2013) and T6P-synthase-1 (TPS1) activity in the leaves is necessary for induction of FT. Recent work has shown that T6P regulates the juvenile-to-adult vegetative phase change by interactions with miR156/SPL module (Ponnu et al., 2020). In cassava, a high proportion of genes in the sugar-related pathway were expressed in response to temperature with the same relationship to flowering (positive vs. negative) as seen to induce flowering in Arabidopsis (**Table 3**). Among these are 3 TPS1 homologs (**Table 3** and **Figure 4**). Based on this evidence, we suggest that in cassava, these genes might be involved in signaling carbohydrate status of the leaves to regulate flowering.

Several other regulatory networks have been shown to interact with the SPL/miR172 module and affect the time to flowering, including those involving the plant hormone GA, GI (the core regulatory component of the circadian pathway) (Wang, 2014) and TEMs (Aguilar-Jaramillo et al., 2019). In both field and greenhouse experiments involving temperature, Oluwasanya D. N. et al. (2021) found that expression of TEM2 in cassava was consistent with a role in regulating flowering. Our results did not show significant differentially expressed genes related to these pathways but they cannot be ruled out.

Photoperiod effect on transcriptome

The external coincidence model for photoperiod regulation of flowering has many layers (Song et al., 2015). In general, our findings in cassava indicate that expression of genes in cassava follows this model, though as in other species, there may be some variation in these pathways (Hayama and Coupland, 2003; Song et al., 2015). As predicted by the model, whereby the circadian clock entrains accumulation of CONSTANS transcripts in the evening, in cassava, we found that two CO-like genes have significantly higher expression in the end-of-day (pre-dark) period in LD compared to SD (**Table 4**). The model predicts that this accumulation is followed by multiple layers of posttranslational regulation of the CO protein. PHYA stabilizes the CO protein (Hayama and Coupland, 2003; Song et al., 2015) and PHYB promotes the degradation of the CO protein (Hayama and Coupland, 2003). Blue light stabilizes CO protein through CRY and COP1. COP1 is a negative regulator of flowering, reducing CO abundance, and CRY negatively regulates COP1 (Liu et al., 2008). Our findings in cassava show evidence that expression of homologs for all of these components follow the expression predicted by the model (**Table 4**).

Studies of cassava conducted by Behnam et al. (2021) indicated that CO-like homologs COL2, COL3, and COL4 were expressed in leaves at young stages of plant development, whereas COL5, COL6, and COL7 were expressed at mature stages from 4 months after transplanting when flowering was taking place. These studies involved sampling at the middle of the day when CO expression might not be fully reflective of photoperiod or other environmental regulatory effects. Our studies, which involved sampling in the last 15 min of the light period, only found COL2 as differentially expressed in response to temperature and photoperiod (**Tables 4**, **5**). Hence it is possible that only this homolog is involved in end-of-day expression associated with flower induction by these treatments, or differential expression of other CO homologs was below our limit of detection.

Other CO transcriptional promotors are also known to regulate the amplitude of CO transcripts in addition to the circadian cycling of expression. One such family are FBH1, 2,3,4 basic helix-loop-helix type transcription factors which bind directly to the CO promotor and likely function with multiple redundancies (Ito et al., 2012). In our study of cassava, two members of this family had higher expression in LD than SD, consistent with Arabidopsis, whereas a third member had the opposite pattern of expression in LD vs. SD (**Table 4**). The core circadian pathway gene GI, which has a strong circadian cyclical expression pattern that decreases in the evening (James et al., 2008; Bouché et al., 2015), was expressed at lower levels in long day plants sampled later in the evening, as expected (**Supplementary Table 8**). Behnam et al. (2021), sampling leaves at mid-day, found that in cassava GI followed the same pattern as FT with both genes expressed at low levels in leaves of young plants but higher in plants at the flowering stage of 4 months or older. Overall, the photoperiod regulatory system in cassava, in which flowering is promoted in long days, appears similar to the photoperiod regulatory mechanism in other long day flowering plants such as Arabidopsis.

Temperature and photoperiod effect on transcriptome

Photoperiod and temperature elicited many of the same differentially expressed genes (**Table 5**). Some of these genes, such as CO and SPL may represent nodes that are targets of multiple upstream signaling pathways. These may help explain the interaction observed on the flowering response with the combination of LD and cool temperature (**Figures 1**, **3**). Others such as flower development and identity genes may represent downstream effects that are part of later flower development. Because they have been identified to respond to both photoperiod and temperature it is likely they play a critical role in the induction of flowering, though at this point we do not have enough information to elucidate their exact function. Of particular interest are several hormone-related genes that were differentially expressed in response to both temperature and photoperiod treatments. In previous study, plant growth regulator (PGR) treatments with the anti-ethylene silver thiosulfate (STS) and cytokinin were found to be effective tools in stimulating flower proliferation and feminization of flower development in cassava (Hyde et al., 2020; Oluwasanya D. et al., 2021) and cytokinin stimulates female flower proliferation in Jatropha (Chen et al., 2014, 2019; Froeschle et al., 2017). A previous study of the effect of temperature on flowering in cassava identified differentially expressed genes in several hormone pathways (Oluwasanya D. N. et al., 2021). Collectively, these studies have revealed that application of these hormones stimulates expression in a wide range hormone pathways. It may be possible to use this information to further improve PGR treatment protocols or breeding efforts to improve flowering.

Conclusion

Environmental effects on flowering

These investigations showed that cooler air temperature and extended photoperiod stimulate earlier flowering, whereas these conditions were unfavorable to storage-root growth and harvest index. Warmer temperatures limited the benefit of extended photoperiod. Considering this, we conclude that the most favorable conditions for stimulating earlier and more prolific flowering in cassava are long-day photoperiods with relatively cool day temperatures of approximately 22°C.

Transcriptome data for cassava suggested that the regulatory system for photoperiod and downstream pathways leading to flower induction operate similarly to those in Arabidopsis. Many of the known genes involved in photoperiodism regulating flowering were differentially expressed in the same direction as in Arabidopsis, a model LD plant, confirming that cassava is also a long day plant. In contrast, Arabidopsis and cassava respond to ambient temperature in opposite directions – Arabidopsis flowers earlier in warm temperature whereas cassava flowers earlier in cool temperatures – yet expression of temperature-responsive flowering genes was similarly affected by temperature. These results indicate that the ambient temperature regulatory pathway as described based on studies of Arabidopsis does not function similarly in cassava flower induction. The sugar and developmental age pathways in Arabidopsis and cassava have similar gene expression patterns and correlation to flowering, therefore, it is likely these pathways are involved in the induction of flowering in cassava.

Implication for breeding

The current findings provide an improved understanding of cassava flowering in response to photoperiod and temperature. Cassava breeders can utilize this information to provide guidance on the extent to which there may be benefit in locating their crossing nurseries where the climate is relatively cool and employing lights to extend the photoperiod. This is of immediate interest as global climate change is likely to increase average temperatures. This will be useful for all types of breeding designs that require making crosses, from genomic to mass selection approaches. Furthermore, breeders interested in developing non-branching cultivars can utilize this transcriptome information to target genes that may regulate flowering and in turn branching.

Data availability statement

The original contributions presented in this study are publicly available. This data can be found here: https://cassavabase.org/ftp/manuscripts/Hyde_et_al_2022/.

Author contributions

PH conducted the plant growth and laboratory work. PH and TS supervised the work, analyzed the data, wrote the manuscript, revised the article, and approved the submitted version. TS conceived the project and obtained funding. Both authors contributed to the article and approved the submitted version.

Funding

Funding for this work was obtained from the NextGen Cassava Breeding Project, managed by Cornell University with funds from the Bill & Melinda Gates Foundation (Agreement No. OPPID INV-007637) and the Department for International Development of the United Kingdom.

Acknowledgments

We thank J. Glaubitz for bioinformatic technical assistance, and the Guterman greenhouse crew for plant care. We also thank IITA and NaCRRI for providing the cassava cultivars for these studies.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.973206/full#supplementary-material

References

Adeyemo, O. S., Chavarriaga, P., Tohme, J., Fregene, M., Davis, S. J., and Setter, T. L. (2017). Overexpression of *Arabidopsis* flowering locus T (FT) gene improves floral development in cassava (*Manihot esculenta, Crantz*). *PLoS One* 12:e0181460. doi: 10.1371/journal.pone.0181460

Adeyemo, O. S., Hyde, P. T., and Setter, T. L. (2019). Identification of FT family genes that respond to photoperiod, temperature and genotype in relation to flowering in cassava *(Manihot esculenta, Crantz*). *Plant Reprod.* 32, 181–191. doi: 10.1007/s00497-018-00354-5

Aguilar-Jaramillo, A. E., Marín-González, E., Matías-Hernández, L., Osnato, M., Pelaz, S., and Suárez-López, P. (2019). Tempranillo is a direct repressor of the microRNA miR172. *Plant J.* 100, 522–535. doi: 10.1111/tpj.14455

Alves, A. A. C. (2002). "Charpter 5. Cassava botany and physiology," in *Cassava biology, production and utilization*, eds R. J. Hillocks, J. M. Thresh, and A. C. Bellotti (Wallingford: CABI), 67–89.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu638

Andrade, L. R. B. D., Sousa, M. B. E., Oliveira, E. J., Resende, M. D. V. D., and Azevedo, C. F. (2019). Cassava yield traits predicted by genomic selection methods. *PLoS One* 14:e0224920. doi: 10.1371/journal.pone.0224920

Behnam, B., Higo, A., Yamaguchi, K., Tokunaga, H., Utsumi, Y., Selvaraj, M. G., et al. (2021). Field-transcriptome analyses reveal developmental transitions during flowering in cassava (*Manihot esculenta Crantz*). *Plant Mol. Biol.* 106, 285–296. doi: 10.1007/s11103-021-01149-5

Blümel, M., Dally, N., and Jung, C. (2015). Flowering time regulation in crops—what did we learn from *Arabidopsis*? *Curr. Opin. Biotechnol.* 32, 121–129. doi: 10.1016/j.copbio.2014.11.023

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2015). FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res.* 44, D1167–D1171. doi: 10.1093/nar/gkv1054

Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., et al. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* 34, 562–570. doi: 10.1038/nbt.3535

Bull, S., Alder, A., Barsan, C., Kohler, M., Hennig, L., Gruissem, W., et al. (2017). Flowering Locus T triggers early and fertile flowering in glasshouse cassava (*Manihot esculenta Crantz*). *Plants* 6:22. doi: 10.3390/plants6020022

Capovilla, G., Schmid, M., and Posé, D. (2014). Control of flowering by ambient temperature. *J. Exp. Bot.* 66, 59–69. doi: 10.1093/jxb/eru416

Ceballos, H., Iglesias, C. A., Perez, J. C., and Dixon, A. G. O. (2004). Cassava breeding: Opportunities and challenges. *Plant Mol. Biol.* 56, 503–516.

Ceballos, H., Kulakow, P., and Hershey, C. (2012). Cassava breeding: Current status, bottlenecks and the potential of biotechnology tools. *Trop. Plant Biol.* 5, 73–87.

Chen, M.-S., Pan, B.-Z., Wang, G.-J., Ni, J., Niu, L., and Xu, Z.-F. (2014). Analysis of the transcriptional responses in inflorescence buds of *Jatropha curcas* exposed to cytokinin treatment. *BMC Plant Biol.* 14:318. doi: 10.1186/s12870-014-0318-z

Chen, M.-S., Zhao, M.-L., Wang, G.-J., He, H.-Y., Bai, X., Pan, B.-Z., et al. (2019). Transcriptome analysis of two inflorescence branching mutants reveals cytokinin is an important regulator in controlling inflorescence architecture in the woody plant *Jatropha curcas*. *BMC Plant Biol.* 19:468. doi: 10.1186/s12870-019-2069-3

Cox, D. R. (1972). Regression models and life-tables. *J. Royal Stat. Soc. Ser B (Methodol).* 34, 187–220.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Fornara, F., de Montaigu, A., and Coupland, G. (2010). SnapShot: Control of flowering in *Arabidopsis*. *Cell* 141, 550–550.

Froeschle, M., Horn, H., and Spring, O. (2017). Effects of the cytokinins 6-benzyladenine and forchlorfenuron on fruit-, seed- and yield parameters according to developmental stages of flowers of the biofuel plant *Jatropha curcas* L. (Euphorbiaceae). *Plant Growth Regul.* 81, 293–303.

Ge, S. X., Jung, D., and Yao, R. (2019). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* 36, 2628–2629. doi: 10.1093/bioinformatics/btz931

Hayama, R., and Coupland, G. (2003). Shedding light on the circadian clock and the photoperiodic control of flowering. *Curr. Opin. Plant Biol.* 6, 13–19. doi: 10.1016/s1369-5266(02)00011-0

Huijser, P., and Schmid, M. (2011). The control of developmental phase transitions in plants. *Development* 138, 4117–4129. doi: 10.1242/dev.063511

Hyde, P. T., Guan, X., Abreu, V., and Setter, T. L. (2020). The anti-ethylene growth regulator silver thiosulfate (STS) increases flower production and longevity in cassava (*Manihot esculenta Crantz*). *Plant Growth Regul.* 90, 441–453. doi: 10.1007/s10725-019-00542-x

Iragaba, P., Hamba, S., Nuwamanya, E., Kanaabi, M., Nanyonjo, R. A., Mpamire, D., et al. (2021). Identification of cassava quality attributes preferred by Ugandan users along the food chain. *Int. J. Food Sci. Technol.* 56, 1184–1192. doi: 10.1111/ijfs.14878

Iragaba, P., Kawuki, R. S., Bauchet, G., Ramu, P., Tufan, H. A., Earle, E. D., et al. (2020). Genomic characterization of Ugandan smallholder farmer-preferred cassava varieties. *Crop Sci.* 60, 1450–1461. doi: 10.1002/csc2.20152

Ito, S., Song, Y. H., Josephson-Day, A. R., Miller, R. J., Breton, G., Olmstead, R. G., et al. (2012). Flowering Bhlh transcriptional activators control expression of the photoperiodic flowering regulator constans in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3582–3587. doi: 10.1073/pnas.1118876109

James, A. B., Monreal, J. A., Nimmo, G. A., Kelly, C. L., Herzyk, P., Jenkins, G. I., et al. (2008). The circadian clock in *Arabidopsis* roots is a simplified slave version of the clock in shoots. *Science* 322, 1832–1835. doi: 10.1126/science.1161403

Jarvis, A., Ramirez-Villegas, J., Herrera Campo, B., and Navarro-Racines, C. (2012). Is cassava the answer to African climate change adaptation? *Trop. Plant Biol.* 5, 9–29.

Jin, S., and Ahn, J. H. (2021). Regulation of flowering time by ambient temperature: Repressing the repressors and activating the activators. *New Phytol.* 230, 938–942. doi: 10.1111/nph.17217

Jung, J.-H., Seo, P. J., Ahn, J. H., and Park, C.-M. (2012). Arabidopsis RNA-binding protein FCA regulates MicroRNA172 processing in thermosensory flowering*. *J. Biol. Chem.* 287, 16007–16016. doi: 10.1074/jbc.M111.337485

Kawuki, R. S., Kaweesi, T., Esuma, W., Pariyo, A., Kayondo, I. S., Ozimati, A., et al. (2016). Eleven years of breeding efforts to combat cassava brown streak disease. *Breed. Sci.* 66, 560–571. doi: 10.1270/jsbbs.16005

Keating, B. A., Evenson, J. P., and Fukai, S. (1982). Environmental effects on growth and development of cassava *Manihot esculenta* 1. Crop development. *Field Crops Res.* 5, 271–282.

Kumar, S. V., Lucyshyn, D., Jaeger, K. E., Alós, E., Alvey, E., Harberd, N. P., et al. (2012). Transcription factor PIF4 controls the thermosensory activation of flowering. *Nature* 484, 242–245. doi: 10.1038/nature10928

Lee, J. H., Ryu, H.-S., Chung, K. S., Posé, D., Kim, S., Schmid, M., et al. (2013). Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. *Science* 342, 628–632. doi: 10.1126/science.1241097

Lenth, R. (2019). *emmeans: Estimated marginal means, aka least-squares means. R package version 1.4. 3.01*. Available online at: https://CRAN.R-project.org/package=emmeans (accessed February 21, 2021).

Liu, L.-J., Zhang, Y.-C., Li, Q.-H., Sang, Y., Mao, J., Lian, H.-L., et al. (2008). COP1-mediated ubiquitination of constans is implicated in cryptochrome regulation of flowering in *Arabidopsis*. *Plant Cell* 20, 292–306. doi: 10.1105/tpc.107.057281

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Nakabonge, G., Samukoya, C., and Baguma, Y. (2018). Local varieties of cassava: Conservation, cultivation and use in Uganda. *Environ. Dev. Sustain.* 20, 2427–2445.

Oda, A., Fujiwara, S., Kamada, H., Coupland, G., and Mizoguchi, T. (2004). Antisense suppression of the *Arabidopsis* PIF3 gene does not affect circadian rhythms but causes early flowering and increases FT expression. *FEBS Lett.* 557, 259–264. doi: 10.1016/s0014-5793(03)01470-4

Odipio, J., Getu, B., Chauhan, R. D., Alicai, T., Bart, R., Nusinow, D. A., et al. (2020). Transgenic overexpression of endogenous Flowering Locus T-like gene MeFT1 produces early flowering in cassava. *PLoS One* 15:e0227199. doi: 10.1371/journal.pone.0227199

Oluwasanya, D., Esan, O., Hyde, P. T., Kulakow, P., and Setter, T. L. (2021). Flower development in cassava is feminized by cytokinin, while proliferation is stimulated by anti-ethylene and pruning: Transcriptome responses. *Front. Plant Sci.* 12:666266. doi: 10.3389/fpls.2021.666266

Oluwasanya, D. N., Gisel, A., Stavolone, L., and Setter, T. L. (2021). Environmental responsiveness of flowering time in cassava genotypes and associated transcriptome changes. *PLoS One* 16:e0253555. doi: 10.1371/journal.pone.0253555

Parmar, A., Sturm, B., and Hensel, O. (2017). Crops that feed the world: Production and improvement of cassava for food, feed, and industrial uses. *Food Security* 9, 907–927.

Perera, P. I. P., Quintero, M., Dedicova, B., Kularatne, J. D. J. S., and Ceballos, H. (2013). Comparative morphology, biology and histology of reproductive development in three lines of *Manihot esculenta Crantz* (Euphorbiaceae: Crotonoideae). *AoB Plants* 5:46. doi: 10.1093/aobpla/pls046

Pineda, M., Morante, N., Salazar, S., Cuásquer, J., Hyde, P. T., Setter, T. L., et al. (2020a). Induction of earlier flowering in cassava through extended photoperiod. *Agronomy* 10:1273.

Pineda, M., Yu, B., Tian, Y., Morante, N., Salazar, S., Hyde, P. T., et al. (2020b). Effect of pruning young branches on fruit and seed set in cassava. *Front. Plant Sci.* 11:1107. doi: 10.3389/fpls.2020.01107

Poethig, R. S., Rougvie, A. E., and O'Connor, M. B. (2013). "Chapter Five - Vegetative phase change and shoot maturation in plants," in *Current topics in developmental biology*, ed. G. Schatten (New York, NY: Academic Press), 125–152. doi: 10.1016/B978-0-12-396968-2.00005-1

Ponnu, J., Schlereth, A., Zacharaki, V., Działo, M. A., Abel, C., Feil, R., et al. (2020). The trehalose 6-phosphate pathway impacts vegetative phase change in *Arabidopsis thaliana*. *Plant J.* 104, 768–780. doi: 10.1111/tpj.14965

Proveniers, M. C. G., and van Zanten, M. (2013). High temperature acclimation through PIF4 signaling. *Trends Plant Sci.* 18, 59–64.

R Core Team (2017). *R: A language and environment for statistical computing [Online]*. Vienna: R Foundation for Statistical Computing.

Rabbi, I. Y., Kulakow, P. A., Manu-Aduening, J. A., Dankyi, A. A., Asibuo, J. Y., Parkes, E. Y., et al. (2015). Tracking crop varieties using genotyping-by-sequencing markers: A case study using cassava (*Manihot esculenta Crantz*). *BMC Genetics* 16:115. doi: 10.1186/s12863-015-0273-1

Shalom, L., Shlizerman, L., Zur, N., Doron-Faigenboim, A., Blumwald, E., and Sadka, A. (2015). Molecular characterization of Squamosa Promoter Binding Protein-Like (SPL) gene family from citrus and the effect of fruit load on their expression. *Front. Plant Sci.* 6:389. doi: 10.3389/fpls.2015.00389

Song, Y. H., Shim, J. S., Kinmonth-Schultz, H. A., and Imaizumi, T. (2015). Photoperiodic flowering: Time measurement mechanisms in leaves. *Annu. Rev. Plant Biol.* 66, 441–464. doi: 10.1146/annurev-arplant-043014-115555

Souza, L. S., Alves, A. A. C., and Oliveira, E. J. (2020). Phenological diversity of flowering and fruiting in cassava germplasm. *Sci. Horticult.* 265:109253.

Susila, H., Nasim, Z., and Ahn, J. H. (2018). Ambient temperature-responsive mechanisms coordinate regulation of flowering time. *Int. J. Mol. Sci.* 19:3196. doi: 10.3390/ijms19103196

Teeken, B., Olaosebikan, O., Haleegoah, J., Oladejo, E., Madu, T., Bello, A., et al. (2018). Cassava trait preferences of men and women farmers in Nigeria: Implications for breeding. *Econ. Bot.* 72, 263–277. doi: 10.1007/s12231-018-9421-7

Therneau, T. M. (2015). *A package for survival analysis in S [Online]. R foundation for statistical computing*. Available online at: https://CRAN.R-project.org/package=survival (Accessed May 2021).

Tokunaga, H., Quynh, D. T. N., Anh, N. H., Nhan, P. T., Matsui, A., Takahashi, S., et al. (2020). Field transcriptome analysis reveals a molecular mechanism for cassava-flowering in a mountainous environment in Southeast Asia. *Plant Mol. Biol.* *VP, doi: 10.1007/s11103-020-01057-0

Wahl, V., Ponnu, J., Schlereth, A., Arrivault, S., Langenecker, T., Franke, A., et al. (2013). Regulation of flowering by trehalose-6-phosphate signaling in *Arabidopsis thaliana*. *Science* 339, 704–707. doi: 10.1126/science.1230406

Wang, J.-W. (2014). Regulation of flowering time by the miR156-mediated age pathway. *J. Exp. Bot.* 65, 4723–4730. doi: 10.1093/jxb/eru246

Wang, J.-W., Czech, B., and Weigel, D. (2009). miR156-Regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. *Cell* 138, 738–749. doi: 10.1016/j.cell.2009.06.014

Wolfe, M. D., Del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., et al. (2017). Prospects for genomic selection in cassava breeding. *Plant Genome* 10:3. doi: 10.3835/plantgenome2017.03.0015

Wu, G., Park, M. Y., Conway, S. R., Wang, J.-W., Weigel, D., and Poethig, R. S. (2009). The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. *Cell* 138, 750–759. doi: 10.1016/j.cell.2009.06.031

Yamaguchi, A., Wu, M.-F., Yang, L., Wu, G., Poethig, R. S., and Wagner, D. (2009). The MicroRNA-regulated SBP-box transcription factor SPL3 is a direct upstream activator of leafy, fruitfull, and apetala1. *Dev. Cell* 17, 268–278. doi: 10.1016/j.devcel.2009.06.007

Yan, W., and Wallace, D. H. (1996). A model of photoperiod × temperature interaction effects on plant development. *Crit. Rev. Plant Sci.* 15, 63–96.

Yu, N., Yang, J.-C., Yin, G.-T., Li, R.-S., and Zou, W.-T. (2020). Genome-wide characterization of the SPL gene family involved in the age development of *Jatropha curcas*. *BMC Genomics* 21:368. doi: 10.1186/s12864-020-06776-8

TYPE Original Research
PUBLISHED 21 September 2022
DOI 10.3389/fpls.2022.978248

Check for updates

OPEN ACCESS

EDITED BY
Muhammad Yasir,
Zhejiang Agriculture and Forestry University, China

REVIEWED BY
Brian Cullis,
University of Wollongong, Australia
Hans-Peter Piepho,
University of Hohenheim, Germany

*CORRESPONDENCE
Jean-Luc Jannink
jeanluc.work@gmail.com

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 25 June 2022
ACCEPTED 08 August 2022
PUBLISHED 21 September 2022

CITATION
Bakare MA, Kayondo SI, Aghogho CI, Wolfe MD, Parkes EY, Kulakow P, Egesi C, Jannink J-L and Rabbi IY (2022) Parsimonious genotype by environment interaction covariance models for cassava (*Manihot esculenta*). *Front. Plant Sci.* 13:978248.
doi: 10.3389/fpls.2022.978248

Parsimonious genotype by environment interaction covariance models for cassava (*Manihot esculenta*)

Moshood A. Bakare[1,2], Siraj Ismail Kayondo[2], Cynthia I. Aghogho[2,3], Marnin D. Wolfe[1,4], Elizabeth Y. Parkes[2], Peter Kulakow[2], Chiedozie Egesi[1,2,5], Jean-Luc Jannink[1,6]* and Ismail Yusuf Rabbi[2]

[1]Plant Breeding and Genetics Section, School of Integrative Plant Science, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, United States, [2]International Institute of Tropical Agriculture, Ibadan, Nigeria, [3]West Africa Centre for Crop Improvement, University of Ghana, Legon, Ghana, [4]Department of Crop, Soil and Environmental Sciences, College of Agriculture, Auburn University, Auburn, AL, United States, [5]National Root Crops Research Institute (NRCRI), Umudike, Umuahia, Nigeria, [6]USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, United States

The assessment of cassava clones across multiple environments is often carried out at the uniform yield trial, a late evaluation stage, before variety release. This is to assess the differential response of the varieties across the testing environments, a phenomenon referred to as genotype-by-environment interaction (GEI). This phenomenon is considered a critical challenge confronted by plant breeders in developing crop varieties. This study used the data from variety trials established as randomized complete block design (RCBD) in three replicates across 11 locations in different agro-ecological zones in Nigeria over four cropping seasons (2016–2017, 2017–2018, 2018–2019, and 2019–2020). We evaluated a total of 96 varieties, including five checks, across 48 trials. We exploited the intricate pattern of GEI by fitting variance–covariance structure models on fresh root yield. The goodness-of-fit statistics revealed that the factor analytic model of order 3 (FA3) is the most parsimonious model based on Akaike Information Criterion (AIC). The three-factor loadings from the FA3 model explained, on average across the 27 environments, 53.5% [FA (1)], 14.0% [FA (2)], and 11.5% [FA (3)] of the genetic effect, and altogether accounted for 79.0% of total genetic variability. The association of factor loadings with weather covariates using partial least squares regression (PLSR) revealed that minimum temperature, precipitation and relative humidity are weather conditions influencing the genotypic response across the testing environments in the southern region and maximum temperature, wind speed, and temperature range for those in the northern region of Nigeria. We conclude that the FA3 model identified the common latent factors to dissect and account for complex interaction in multi-environment field trials, and the PLSR is an effective approach for describing GEI variability in the context of multi-environment trials where external environmental covariables are included in modeling.

KEYWORDS

factor analytic model, genotype-by-environment interaction, partial least squares regression, variance structure, factor loadings, genotypic scores, hybrid relationship matrix, environmental covariables

Introduction

Cassava (*Manihot esculenta* Crantz) is one of the most essential food-security crops in developing countries, particularly in tropical and subtropical regions (Tumuhimbise et al., 2014; Nduwumuremyi et al., 2017). It is a crop grown predominantly by smallholders for subsistence due to its adaptability to survive in drought-prone areas under marginal conditions where other crops may not thrive (Egesi et al., 2007; Sayre et al., 2011). Though cassava grows well in diverse environments, its yield production differs among the genotypes and environments. This difference is due to inbuilt genetic properties, environmental conditions, and genotype-by-environment interaction (Falconer, 1996).

It has long been recognized that phenotypic expression of genotypes is much influenced by environmental conditions (Meyer, 2009). This can result in heterogeneity of variability and different ranking of genotypes performance in different environments, a phenomenon described as genotype-by-environment interaction (GEI). The phenotypic panel for evaluating GEI is often called a multi-environment trial (MET). Traditionally, the resulting empirical data from METs are often analyzed using classical statistical methods (Bakare et al., 2022). These methods include ANOVA, fixed linear bilinear model such as additive main effect and multiplicative interaction (AMMI) model (Gauch and Zobel, 1997; Gauch, 2016) and site regression (SREG) or genotype main effect and genotype-by-environment (GGE) model (Yan et al., 2000), and linear regression type model like Finlay and Wilkinson (1963). These classical analyses are inefficient in handling unbalanced datasets that often arise in METs (Bakare et al., 2022), resulting in unreliable estimates of genetic effects.

Linear mixed models that include fixed and random effects are increasingly used to analyze MET in a plant breeding program (Piepho, 1998b; Smith et al., 2005; Burgueño et al., 2008). These models are centered around a factor analytic (FA; Piepho, 1997, 1998a) form of genetic variance–covariance structure. Factor analytic structures have been reported to be more parsimonious and flexible than other variance–covariance structures (Crossa, 2012), allowing the estimation of a fewer number of parameters in comparison to unstructured (US) variance–covariance model (Smith et al., 2001a,b; Kelly et al., 2007). Graphical tool like heatmaps of estimated genetic correlation across the testing environments (Cullis et al., 2014; Smith et al., 2015) resulting from factor analytic model can be used to make inferences about GEI, adaptability and stability of genotypes (Oliveira et al., 2020). Also, the factor loadings which are environmental effects in the latent factors can be correlated with external environmental covariables such as solar radiation, temperature, precipitation, relative humidity, wind speed and others, to examine the pattern of genotypic response across environments. The measure of these external environmental covariables in different developmental phases of year-long growth period crops such as cassava will result in many predictor variables that are highly correlated. The use of ordinary least squares regression model to quantify the relationship between dependent variable(s) and predictor variables is not adequate due to multicollinearity problem. In this scenario, partial least squares regression (Aastveit and Martens, 1986; Talbot and Wheelwright, 1989; Vargas et al., 1998) can be used to determine which among these environmental covariables influence GEI of fresh root yield.

To date, no implementation of the FA model in the genetic assessment of cassava clones has been reported in Africa nor environmental covariables driving GEI have been explored. However, few studies have been reported to explore GEI in cassava and these studies were conducted in few environments using ANOVA, AMMI (Dixon and Ssemakula, 2007; Jiwuba et al., 2020) and GGE (Akinwale et al., 2011) for analyses. This study examines the utility of variance–covariance structure models and partial least squares regression to: (i) identify optimal variance–covariance structure model that captured GEI and stable genotypes; (ii) identify mega environments, and (iii) identify key environmental covariables that explained GEI for fresh root yield.

Materials and methods

Clonal material and field experimental design

This study used 48 uniform yield trials in three sets named setA, setB, and setC with, respectively, 36, 36, and 34 clones each. A total of 96 clones were evaluated, corresponding to 91 breeding lines and five checks common across sets. These clones were derived from elite X elite crosses as part of a genomic recurrent breeding program. Prior to this field evaluation, they were assessed for susceptibility to cassava mosaic disease (CMD), cassava bacteria blight (CBB), early vigor, and other agronomic traits of interest in earlier evaluation stages. The clones in the UYT were high yielding materials that have passed several stages of field evaluation and selection to eliminate disease susceptible clones. The clones were evaluated in UYT trials in 11 locations across different agro-ecological zones in Nigeria (Figure 1) over four growing seasons (2016–2017, 2017–2018, 2018–2019, and 2019–2020).

Each trial was established as a Randomized Complete Block Design (RCBD) with two or three replicates. The experimental plot consisted of six rows of length 5.6 m with an inter-row spacing of 1 m and intra-row spacing of 0.8 m and only the interior 20 plants (4 m × 4 m) were harvested. Across the full dataset, there were 28 environments (location by year combinations) and a total of 4,575 plots, varying in number across the testing environments from 72 (Onne20) to 318 (Ikenne18 and Mokwa18; Table 1). The trait of interest in this study was fresh root yield (t/ha).

Genotype and pedigree relationship matrices

Following a modified cetyltrimethyl ammonium bromide (CTAB) method, we extracted high-quality genomic DNA from

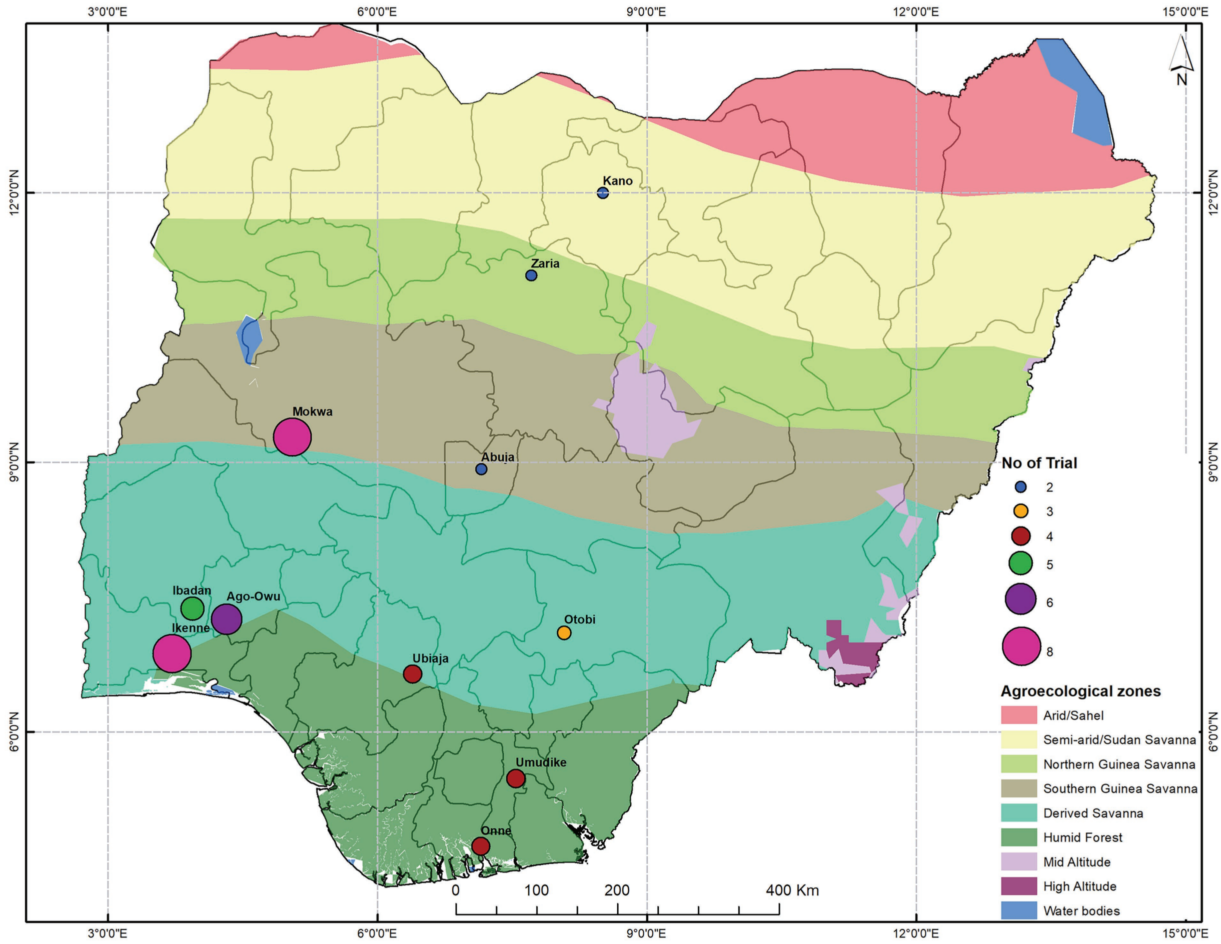


FIGURE 1
A map of Nigeria showing the trial geographical locations across agro-ecological zones.

freeze-dried cassava leaf samples (Dellaporta et al., 1983). The Nanodrop spectrophotometer operating at an absorbance of 260 nm qualified and quantified the extracted DNA before genotyping. The genotyping-by-sequencing (GBS) approach generated a dense genome-wide single nucleotide polymorphism (SNP) dataset as described by Elshire et al. (2011). The ApeKI enzyme reduced genome complexity through restriction digestion, preparing genomic fragments for GBS (Hamblin and Rabbi, 2014). Sequence alignment of the resultant sequence tags was done using the cassava Version 6 genome as a reference (Prochnik et al., 2012). Alignment was followed by the SNP calling step using TASSEL GBS pipeline V4 (Glaubitz et al., 2014). All SNP calls below five reads were masked before imputation using Beagle V4.1 (Browning and Browning, 2016). After imputation, 73,599 biallelic SNP markers with an estimated allelic r-squared value (AR^2) of more than 0.3 were retained for subsequent analyses. Data quality control was carried out on the SNP dataset using the *qc.filtering()* function in the *ASRgenomics* library (Gezan et al., 2021) prior to downstream analyses. The filtering criteria included: (i) removal of SNPs with minor allele frequency (MAF) below 0.05, (ii) removal of individuals whose proportion of missing values was equal or above 20% (call rate 0.2), and (iii) removal of SNPs whose proportion of missing values equal or larger than 20%, retaining 68,279 SNPs in total. However, the available SNP marker data was only available for 81 clones. Thus, we also used pedigree data on 123 individuals out of which 27 individuals were dropped to have a pedigree-based relationship matrix of dimension 96 × 96 for the phenotyped cassava clones. The SNP marker set was used in the derivation of a genomic relationship matrix (GRM) and combined with the pedigree relationship matrix to produce a hybrid relationship matrix (H).

The pedigree-based additive numerator relationship matrix (A-matrix) was constructed following the recursive method presented in Mrode (2014) and was estimated using the *Amatrix()* function of the *AGHmatrix* library (Amadeu et al., 2016). The marker-based relationship matrix (G) and its inverse (G^{-1}) were estimated from SNP marker data using the *G.matrix()* and *G.inverse()* functions of the *ASRgenomics* library (Gezan et al., 2021), respectively.

The H-matrix relates all individuals through the A-matrix but integrates the additional information provided by the G-matrix. The main notion is to replace entries of the A-matrix by the

TABLE 1 Summary of number of trials, cassava clones, plots, blocks and mean fresh root yield (FYLD) per environment.

Environment	Trial	Clones	Plots	Blocks	FYLD (*t*/ha)
Abuja20	2	67	144	4	26.0
Ago-Owu18	2	67	216	6	34.0
Ago-Owu19	2	67	216	6	28.7
Ago-Owu20	2	67	144	4	41.0
Ibadan18	1	33	99	3	36.8
Ibadan19	2	67	216	6	39.9
Ibadan20	2	67	144	4	26.5
Ikenne17	1	34	102	3	37.0
Ikenne18	3	96	318	9	34.1
Ikenne19	2	67	216	6	17.4
Ikenne20	2	67	144	4	41.9
Kano19	2	67	216	6	15.2
Mokwa17	1	34	102	3	22.4
Mokwa18	3	96	318	9	31.7
Mokwa19	2	67	216	6	20.9
Mokwa20	2	67	144	4	18.6
Onne18	1	34	102	3	28.9
Onne19	2	67	216	6	16.9
Onne20	1	36	72	2	13.0
Otobi18	1	34	102	3	25.9
Otobi19	2	67	216	6	41.6
Ubiaja17	1	34	102	3	33.2
Ubiaja18	1	34	102	3	27.6
Ubiaja20	2	67	144	4	15.7
Umudike17	1	34	102	3	24.2
Umudike18	1	34	102	3	21.3
Umudike19	2	67	216	6	31.9
Zaria20	2	67	144	4	13.7

corresponding entries of G-matrix and then adjust the remaining relationships accordingly. Martini et al. (2018) defined matrix H as

$$H = A + \begin{bmatrix} A_{12}A_{22}^{-1}\left(G - A_{22}\right)A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}\left(G - A_{22}\right) \\ \left(G - A_{22}\right)A_{22}^{-1}A_{21} & \left(G - A_{22}\right) \end{bmatrix} \quad (1)$$

where individuals are partitioned into those without (group 1) versus with (group 2) marker data. Therefore, A_{11} contains cells of the A-matrix with relationships within the first group, A_{12} and A_{21} contain cells of the A-matrix with relationships between the individuals of the two groups, and A_{22} contains cell of the A-matrix with relationships within the second group. In this definition of the H-matrix, the inner group pedigree relationship of second group was replaced by the G-matrix indicating that $H_{22} = G$. The term $A_{12}A_{22}^{-1}\left(G - A_{22}\right)$ adapts the relationships within the first group and the relationships between the two groups in accordance to the changed relationships within second group to generate a positive semi-definite and valid covariance structure (Martini et al., 2018).

Since many analyses use the inverse of H that allows for simpler computations, Eq. (1) is often written in terms of its inverse (Misztal et al., 2010; Martini et al., 2018) as

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \left(G^{-1} - A_{22}^{-1}\right) \end{bmatrix} \quad (2)$$

where G^{-1} is the inverse of genomic relationship matrix and A_{22}^{-1} is the inverse of the pedigree-based relationship matrix for genotyped individuals. An approach to combine the A-matrix and G-matrix optimally is implemented by specifying a parameter λ as described by Martini et al. (2018). We used a λ value of 0.9, where λ scales the difference between genomic and pedigree-based information (Misztal et al., 2010), leading to express Eq. (2) as

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \lambda\left(G^{-1} - A_{22}^{-1}\right) \end{bmatrix} \quad (3)$$

The G matrix was derived following (VanRaden, 2008):

$$G = \frac{\left(M - P\right)\left(M - P\right)^{`}}{2\sum_{j=1}^{m} p_j\left(1 - p_j\right)} \quad (4)$$

where M is an allele-sharing matrix with m columns (m = total number of markers) and n rows (n = total number of genotyped individuals), and P is a matrix containing, in each column, the frequency of second allele (p_j) expressed as $2p_j$. M_{ij} was 0 if the genotype of individual i for SNP j was homozygous aa, 1 if heterozygous Aa, or 2 if the genotype was homozygous AA. We note that because all columns of matrix M from which G is constructed are centered, G should not be invertible (contrary to its use in Equation 3). In practice, a number of options are available for matrices that are close to being positive definite (Tier et al., 2015) and we did not encounter difficulty in using the H-matrix described here.

Environmental covariables

Weather data was exploited to identify the potential environmental covariates that influence differential response of the clonal lines across the testing environments. According to each trial's location and growth dates, weather data were collected from the database of the National Aeronautics and Space Administration Prediction of Worldwide Energy Resource (NASA POWER) project.[1] The data included: minimum temperature (°C), maximum temperature (°C), temperature range (°C), precipitation

[1] https://power.larc.nasa.gov/data-access-viewer/

(mm), relative humidity (%), wind speed (m/s), solar radiation (W/m^2), surface soil wetness (%), root zone soil wetness (%), and profile soil moisture (%) for the whole crop growth cycle, i.e., from planting to harvesting of each field trial.

Statistical models

Single trial analysis and data quality control

Before formal analysis, the observed agronomic traits' empirical distribution was visualized across the trials using boxplots and the *ggplot2* package (Wickham, 2016) in R (R Core Team, 2018). The statistical analysis of individual trials was carried out in a linear mixed model framework and the variance components were estimated by restricted maximum likelihood. The univariate linear mixed model fitted was:

$$y = \mu + X_1 r + p\beta + Z_1 g + \epsilon \tag{5}$$

where y is the $(n \times 1)$ vector of observed phenotypic values, in which n is the number of observations in the trial; μ is the intercept (overall mean); r is the $(r \times 1)$ vector of fixed effect of replicates with its associated incidence matrix X_1 of dimension $n \times r$; p denotes the proportion of plant stands harvested as a covariate (e.g., if 28 stands were planted, but only 21 harvested, $p = 0.75$); β is a regression coefficient relating p and y; g is the $(g \times 1)$ vector of random effect of genotype with its associated design matrix Z_1 of dimension $n \times g$, and ϵ is a residual term which is assumed to follow a Gaussian distribution, $\epsilon \sim N\left(0, I_n \sigma_\epsilon^2\right)$.

The quality of each trial was assessed by calculating the coefficient of variation (CV), broad-sense heritability (H^2) on an entry-mean basis, and experimental accuracy (Ac) proposed by Mrode (2014) using the following equations: $CV(\%) = \left(\hat{\sigma}_e / \bar{y}\right) \times 100$,

$$H^2 = \hat{\sigma}_g^2 / \left(\hat{\sigma}_g^2 + \hat{\sigma}_e^2 / r\right), \text{and } Ac = \sqrt{\left(1 - PEV / \hat{\sigma}_g^2\right)}$$

where $\hat{\sigma}_e$ is the estimated residual standard deviation, $\bar{y}$ is the estimate of the overall mean for an agronomic trait; $\hat{\sigma}_g^2$ is the estimated genetic variance, $\hat{\sigma}_e^2$ is the estimated error variance, r is the number of replicates, and PEV is the average of prediction error variance. A trial was removed from a combined analysis based on any of these conditions: The thresholds of CV above 40.5%, H^2 below 0.14 or Ac below 0.40.

Variance–covariance structure models

Before fitting the models, we examined the degree of clone connectivity between pairs of environments (Supplementary Figure 1). This was to have a prior knowledge of the amount of information for estimating a genetic covariance between pairs of environments. Seven variance–covariance structure models were fitted to describe and explore the pattern of GEI. The analysis was carried out using the software ASREML-R version 4.0 (Butler et al., 2017) within the R statistical environment (R Core Team, 2018). This package fits linear mixed models allowing heterogeneity of genetic and error variances across environments where the variance component is estimated using the average information algorithm (Gilmour et al., 1995).

The variance structure models were fit to the data in one-stage analyses using the following linear mixed model:

$$y = \mu + e + set(e) + p\beta + r(set\ e) + g + \epsilon \tag{6}$$

where y, μ, p and β were as defined in the previous equation, e is the $(s \times 1)$ vector of fixed effect of the environment where s is the number of environments; *set(e)* is the fixed effect of the trial set nested within the environment; *r(set e)* is the fixed replicate effect nested with set and environment; g is random effect of genotype nested within environments: $g = \left[g_1^T, g_2^T, \ldots, g_s^T\right]$, where g_j^T is the vector of genotypic effects in environment j with its associated hybrid relationship matrix (H); $g \sim N\left(0, \Sigma\right)$ (see below for the specification of Σ); and ϵ is a residual term that is heterogeneous across the testing environments.

We partitioned the total genetic effects (g) into additive (a) and non-additive (i) components (Oakey et al., 2007) which are assumed to be independent such that $a \sim N\left(0, \sigma_a^2 H\right)$, $i \sim N\left(0, \sigma_i^2 I\right)$ and I is the identity matrix. The non-additive component captures other effects such as dominance, epistasis, and residual additive effects which are not captured by H-matrix. We used an identity matrix to capture that residual after fitting the non-additive effect. This necessitated the scaling of the hybrid matrix associated with additive genetic effect by multiplying main additive genetic and interaction variance matrices by the average of diagonal element of H-matrix which was estimated to be approximately 0.97, closely corresponding to the diagonal element of an identity matrix.

Diagonal variance structure model

We fitted a diagonal variance (DIAG) model as a baseline. This variance–covariance model postulates independence of genetic effects among environments. Being an environment or trial-specific model, if a trial is found to have no genetic variance (variance estimated to be zero), such trial will be excluded from the analysis. The estimates from this model are often used as a starting values when fitting a more complex model like the factor analytic (FA) model. The covariance structure is of the form (assuming four environments):

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix} \otimes \text{H} \tag{7}$$

where the main diagonal elements are the unique genetic variances within environments. For example, σ_1^2 is the genetic variance within an environment 1; and *H* is the hybrid relationship matrix combining pedigree and genomic relationship matrices to account for the relatedness among the cassava clones and same for other models described below.

Compound symmetry model

The compound symmetry (CS) is the most restrictive variance–covariance model. It postulates homogeneity across environments of genetic variance $\left(\sigma_g^2 + \sigma_{ge}^2\right)$ and uniform covariance between any pair of environments $\left(\sigma_g^2\right)$. Note that this variance–covariance model is equivalent to estimating a fixed genotype-by-environment-interaction variance. Its covariance structure is of the form

$$\Sigma = \begin{bmatrix} \sigma_g^2 + \sigma_{ge}^2 & \sigma_g^2 & \sigma_g^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 + \sigma_{ge}^2 & \sigma_g^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 & \sigma_g^2 + \sigma_{ge}^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 & \sigma_g^2 & \sigma_g^2 + \sigma_{ge}^2 \end{bmatrix} \otimes H \quad (8)$$

Compound symmetry heterogeneous model

The compound symmetry heterogeneous (CSH) is an extension of the CS model which postulates a uniform correlation between any pair of environments but heterogeneity across environments of genetic variance and covariance. Its covariance structure is of the form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_1\sigma_4 & \rho\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix} \otimes H \quad (9)$$

where the main diagonal elements were as in Eq. (7), and off-diagonal elements are unique genetic covariances between pairs of environments. For example, $\rho\sigma_1\sigma_2$ is the genetic covariance between environment 1 and 2 in which ρ is the uniform genetic correlation between pairs of environments, and σ_1 and σ_2 are genetic standard deviations of environment 1 and 2, respectively.

Unstructured model

The unstructured (US) model is the least restrictive variance–covariance model, and describes the covariance based on the assumption of heterogeneity of variance within environments and unique covariance between any two environments. As the number of environments (denoted by *s*) increases, it requires a high number of parameters ($p = s(s+1)/2$) resulting in increased computational demand and instability. Therefore, it is rarely used in modeling GEI in the analysis of MET data with a large number of environments. We give this model here for completeness though we were not able to fit it to our data. Its covariance structure is of the form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix} \otimes H \quad (10)$$

where $\sigma_{ij} = \sigma_{ji}$ (the matrix was symmetric), the main diagonal elements were as in Eq. (7), and off-diagonal elements represent unique covariances between pairs of environments.

Factor analytic model

The factor analytic (FA) model is the random effect analogue of AMMI model (Smith and Cullis, 2018) for describing the structure of GEI. It identifies latent (unobserved) common factors that explain GEI while allowing each environment to have a specific variance for effects not explained by the common factors. The FA model provides a parsimonious approximation to the unstructured variance–covariance model (Kelly et al., 2007) but it requires fewer parameters. The model expresses g_{ij}, the random effect of *i*th genotype in the *j*th environment as:

$$g_{ij} = \sum_{k=1}^{t} \lambda_{jk} f_{ik} + \delta_{ij} \quad (11)$$

where λ_{jk} is the loading for latent factor *k* in the *j*th environment (environmental potentiality); f_{ik} is the score or sensitivity of the *i*th genotype (genotypic sensitivity) for latent factor *k* related to the *j*th environment in λ_{jk} ; and δ_{ij} is the residual term representing lack of fit to the model. Thus, the FA model expresses the random effect of *i*th genotype in the *j*th environment as a linear function of latent factors λ_{jk} with random sensitivity f_{ik} for *k* = 1, 2, …, *t* plus an error term δ_{ij} .

The specification of FA model in a covariance form is

$$G = \left(\Lambda\Lambda^T + \psi\right) \otimes H = FA(k) + H \quad (12)$$

where

$$\Lambda\Lambda^T + \psi = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{s1} & \lambda_{s2} & \cdots & \lambda_{sk} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \lambda_{21} & \cdots & \lambda_{s1} \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{s2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{1k} & \lambda_{2k} & \cdots & \lambda_{sk} \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_s \end{bmatrix}$$

where Λ is a $s \times t$ matrix of loadings, with the kth column containing the environment loadings for the kth latent factor ($k = 1, 2, \ldots, t$), and Ψ is an $s \times s$ diagonal matrix with a specific variance for each environment. As above, s is the number of environments.

The FA model can be also taken to be a linear regression of genotype and GEI on environment loadings (λ_{jk}), with each genotype having a distinct slope (genotypic scores, f_{ik}) but a common intercept provided main effect of genotypes are not distinguished from GEI (Crossa, 2012). The genotypic scores measure the genotype's sensitivity to the latent environmental factor represented by the loadings of each environment. Regardless of whether a genotype is evaluated in an environment or not, the FA model provides a predicted genetic effect for each genotype in each testing environment in the dataset.

The number of latent factors is called the order of the model and we use FAk to represent an FA model of order k. We fitted FA1 to FA4 models. The model with the minimum value of AIC was chosen as the most parsimonious model. For FAk models where $k > 1$, the matrix of loadings does not have a unique solution. Therefore, (Cullis et al., 2010) recommends rotating the estimated loadings to their principal component solution *via* singular value decomposition. We use asterisks (*) below to denote rotated loadings and scores.

Assessment of overall performance and stability

We used the factor analytic selection tools proposed by (Smith and Cullis, 2018) to assess and identify the clones with high overall performance and global stability across the testing environments. If λ_1 represents the mean of the loadings for the first factor, then the overall performance (OP) measure for ith genotype is computed as

$$\lambda_1 \tilde{f}^*_{1i} = \frac{1}{s}\sum_{j=1}^{s} \hat{\lambda}^*_{1j} \tilde{f}^*_{1i} \qquad (13)$$

where $\hat{\lambda}^*_{1j}$ is the rotated loading associated with the jth environment in the first latent factor, and $\tilde{f}^*_{1i}$ is the rotated genotypic score of the ith genotype in the first latent factor. The OP measure was based on the first factor loadings because they were all positive and thus represented non-crossover GE interaction (Smith and Cullis, 2018). The OP is on the same scale of measurement as the agronomic trait being analyzed.

The measure of genotype stability is usually based on the higher factors ($k > 1$) which have a mixture of both positive and negative loadings. This practice is justified by the fact that changes in genotype performance due primarily to changes in scale, which are accounted for in the first factor should be eliminated from stability analysis (Smith and Cullis, 2018). The global stability measure for each genotype was obtained as the root mean square deviation (RMSD) from the regression line associated with the first factor. The RMSD for ith genotype is derived as

$$\sqrt{\frac{1}{s}\sum_{j=1}^{s} \tilde{\epsilon}^{*2}_{ij}} \qquad (14)$$

where $\tilde{\epsilon}^{*2}_{ij} = \tilde{\beta}_{ij} + \hat{\lambda}^*_{1j} \tilde{f}^*_{1i}$. The $\tilde{\epsilon}^{*2}_{ij}$ denoted deviations from the first factor prediction in a plot where the x-axis was the first factor loadings and y-axis was the common effects; and $\tilde{\beta}_{ij}$ were the linear combination of factor loadings and genotypic scores. Like OP, RMSD is on the scale of the trait measured. The stability of the genotypes across the environments can be explored in detail by latent regression plot. In this study, we obtained the plot by regressing the predicted breeding value on the factor loading of the FA3 model.

Clustering of target environments and locations

We used the rotated factor loadings resulting from the FA3 model for clustering and delineating the subset of environments and locations into mega-environment using the *hclust()* function in R and the Ward's D2 linkage method. The procedure involved these steps: (i) Computing the Euclidean distance between a pair of environments from the $s \times 3$ factor loadings matrix; (ii) Hierarchical clustering on the derived distance matrix using Ward's minimum variance linkage method (ward.D2) where dissimilarities were squared before clustering; (iii) plotting and visualizing the cluster dendrogram resulting from (ii); and (iv) subjectively determining the number of clusters by imposing a threshold of minimum similarity to be in the same cluster.

To cluster locations (as opposed to environments = location-by-year combination), we computed for each factor separately, the average loadings of the environment that each location was a part of. We then used the approach above to cluster the locations.

We further used an approach proposed by Smith et al. (2021) to group the testing environments into interactive classes (iClasses), a cluster of environments where a negligible crossover GEI exist.

Association of latent factor loadings with environmental covariables

The environmental covariables associated with GEI were identified by correlating each environmental covariable to each of the three latent factor loadings extracted from the FA3 model. Then, we fitted a partial least square regression to describe GEI in terms of differential genotypic responses to environmental covariables. The PLSR is a form of multivariate regression that maximizes covariance between *X* and *Y* data matrices in one single estimation procedure (Vargas et al., 1998). The environment covariables were in a data matrix *X* of dimension 27 × 40 (27 rows representing the testing environments and 40 columns corresponding to the environmental covariables across the developmental phases). The factor loadings were data matrix *Y* of size 27 × 3 (27 rows for testing environments and 3 columns corresponding to the latent factor loadings). Since the PLSR method is variant to the scale of measurement, the columns of *X* and *Y* data matrices were centered (zero mean) and scaled (unit variance).

The PLSR was implemented using the *plsr()* function of *pls* library (Liland et al., 2021) in R (R Core Team, 2018). The underlying multivariate PLSR in a bilinear form is described as

$$X = TP' + \mathrm{E} \tag{15}$$

and

$$Y = UQ' + \mathrm{F} \tag{16}$$

where *T* and *U* are, respectively, $n \times 1$ matrix of projection of *X* (X scores) and projections of *Y* (Y scores); *P* and *Q* denote $m \times 1$ and $p \times 1$ orthogonal loading matrices respectively; *E* and *F* are residual matrices assumed to be independent and identically distributed random normal variables.

We recognize that in this analysis we are using the environmental loadings from FA3 model as if they were observed data as opposed to derived parameters. A better approach would have been to develop a kind of factor analytic model to work directly on the continuous environmental covariables as opposed to using environment labels as a categorical variables. We do not know of a method to do such an analysis, let alone software to fit it. We look forward to the development of such a method.

Results

Single-trial analysis and data quality control

Before formal statistical analysis, the distribution of observed agronomic traits of 96 clones from 48 trials tested in 28 environments revealed that the traits approximated a normal distribution across the testing environments (Supplementary Figure 2) as the mean denoted by blue data point and median represented by a line were approximately the same. The boxplots showed the heterogeneity of variation in the observed traits across the environments. The mean fresh root yield across the 48 trials varied from 0.3 t/ha (18UYT36setAKN, 18UYT36setBKN) in Kano to 83.3 t/ha (18UYT36setAOT) in Otobi with an overall mean of 27.6 t/ha (Supplementary Table 1). The broad-sense heritability on an entry-mean basis (H^2) ranged from 0.06 (19UYT36setAMK) to 0.85 (18UYT36setBIK) across trials. We observed experimental accuracy (Ac) values varying from 0.24 (19UYT36SETAMK) to 0.91 (18UYT36setBIK). The coefficients of variation (CV%) ranged from 14% (17UYT36setAIK) to 42% (18UYT36setAKN). The four trials (17C1UYT34UM, 18UYT36setAKN, 19UYT36setAZA, and 19UYT36setAMK) displayed in red (Supplementary Figures 3a,b) were filtered out from the combined analysis based on threshold defined in the Methods because their error variances were in the range of 17 to 30 fold higher than the genetic variances, which was very unusual in our breeding program. Therefore, subsequent analysis was based on 44 trials across 27 environments.

Variance–covariance structure model

The pair of environments with the least connectivity had five clones in common while Ikenne18 and Mokwa18 had 96 in common (Supplementary Table 2). The low or poor connectivity between some pairs of environments may impact the reliability of estimation of between environment genetic covariances (Smith et al., 2015). Note, however, that because we used an HRM between clones, relationship among clones in a pair of environments helps increase the accuracy of covariance estimation between the pair.

The diagonal variance model revealed that genetic variance within environments ranged from 2.2 (Zaria20) to 82.3 (Ikenne20) under the assumption that genetic correlation between pairs of environments was zero (Supplementary Table 3). The compound symmetry model showed a uniform genetic correlation of 0.42 corresponding to the uniform genetic variance of 22.3 within environments (Supplementary Table 4). The compound symmetry heterogeneous model estimated a uniform genetic correlation of 0.53 but unique genetic variance within environments resulting in different genetic covariances between pair of environments (Supplementary Table 5).

We reported the total number of parameters, the model log-likelihood (Loglik), Akaike information criterion (AIC), Bayesian information criterion (BIC) and percentage of genetic variance captured by factor analytic models (Table 2). The first two ranking models were FA3 and FA4 models having AIC values of 20338.3 and 20339.8, respectively, (Table 2). The FA3 model was chosen as the optimal model because it had the lowest AIC. It required 152 parameters to capture 79.0% of genotypic effect within environments (Table 2).

Pairwise genetic correlations among environments, as estimated by the FA3 model, were predominantly positive (Figure 2), varying from −0.34 (Ago-owu18 vs. Kano19) to 1.00 (Umudike17 vs. Ubiaja17). We report the estimated genetic correlations, variances, and covariances among the environments (Supplementary Table 6). Genetic correlations estimated above 0.70 between any pair of environments were considered high and equivalent to low GEI: the genotypes exhibited similar fresh root yield performance between such environments. In contrast, pairs of environments showing correlations below 0.40, indicated high GEI: the genotypes ranked differently across these pairs of environments.

TABLE 2 Summary of the models fitted to the combined MET data set.

Model	Parameter	LogLik	AIC	BIC	Var (%)
DIAG	89	−10255.7	20689.4	21250.7	
CS	55	−10188.2	20486.3	20833.2	
CSH	79	−10109.0	20377.1	20875.3	
FA1	107	−10078.3	20370.7	21045.5	57.2
FA2	128	−10043.6	20343.2	21150.4	70.8
FA3	152	−10017.3	20338.3	21296.8	79.0
FA4	170	−9999.9	20339.8	21411.9	83.3

Presented is the number of variance–covariance parameters, residual log-likelihood (LogLik), AIC, Akaike information criterion and BIC, Bayesian information criterion, and the mean percentage of variance accounted for. DIAG Diagonal variance model; CS, Compound symmetry model; CSH, Compound symmetry heterogeneous model; and FAk: Factor analytic model of order *k*.

Rotated factor loadings

The first factor loadings after rotation to the principal component solution were all positive, indicating non-crossover GEI, varying from 0.3 to 8.6 with a median of 3.2 and a mean of 3.6 (Table 3). The remaining two factors had ranges extending into negative values indicating crossover GEI. The first-three factors jointly explained 79.0% of the environments' total genetic variability such that the first, second and third factors accounted for 53.5, 14.0, and 11.5% of total genetic variability, respectively, (Table 3). The heritability resulting from genetic and error variances of FA3 model ranged from 0.09 (Kano19) to 0.59 (Ikenne17 and Ikenne18) with an average value of 0.39 across the environments (Table 3).

Assessment of overall performance and stability

The characteristics of first and higher factor loadings can be used to determine the overall performance (OP) and stability of the genotypes. A scatter plot of the OP against the root mean square deviation (RMSD) visualizes genotype performance and its stability (Figure 3). Genotypes in the top left-hand side of the plot

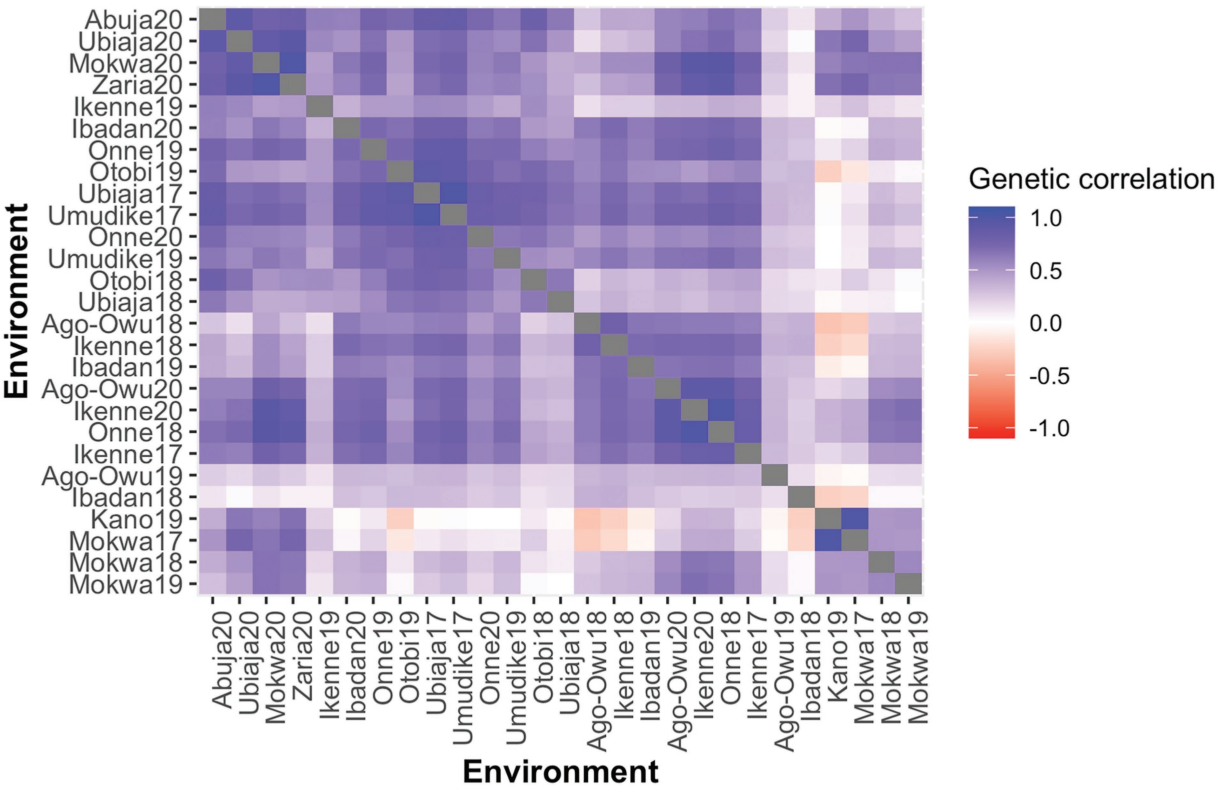


FIGURE 2
A Heatmap of pairwise genetic correlations of fresh root yield estimated the from FA3 model for 27 environments, ordered based on the dendrogram of Ward's D2 linkage method. The color of the square is related to the magnitude of the genetic correlation between environments.

TABLE 3 Summary of the FA3 model in terms of factor loadings, specific variance (Ψ) and genetic variances $\left(\sigma_g^2\right)$, error variances (σ_e^2), heritability (H^2), and interactive classes (iClasses) for environment.

Environment	Factor loadings			Variances			H^2	iClasses
	Factor 1	Factor 2	Factor 3	Ψ	σ_g^2	σ_e^2		
Abuja20	3.5	0.3	−2.8	1.3	21.1	35.1	0.37	ppn
Ago-Owu18	3.4	1.3	2.1	5.4	22.0	47.5	0.32	ppp
Ago-Owu19	1.8	0.4	0.5	2.6	20.5	45.6	0.31	ppp
Ago-Owu20	6.2	−0.9	1.0	8.7	47.5	53.6	0.47	pnp
Ibadan18	2.2	1.8	1.4	0.0	48.4	67.6	0.42	ppp
Ibadan19	5.6	0.9	2.0	11.1	57.7	60.8	0.49	ppp
Ibadan20	3.2	0.7	0.2	4.7	14.7	43.4	0.25	ppp
Ikenne17	7.8	−0.6	0.3	0.0	80.3	54.7	0.59	pnp
Ikenne18	5.5	1.7	2.4	5.6	44.2	30.7	0.59	ppp
Ikenne19	1.7	0.4	−1.7	8.7	14.0	21.4	0.40	ppn
Ikenne20	8.6	−2.9	1.2	0.0	81.8	58.2	0.58	pnp
Kano19	0.3	−1.5	−0.9	0.0	2.8	28.5	0.09	pnn
Mokwa17	1.1	−3.5	−2.7	0.0	19.8	23.7	0.45	pnn
Mokwa18	3.8	−3.4	0.1	1.2	48.4	51.3	0.49	pnp
Mokwa19	2.9	−3.1	0.7	10.8	28.0	33.0	0.46	pnp
Mokwa20	2.9	−1.3	−0.7	0.0	10.0	10.1	0.50	pnn
Onne18	4.7	−1.3	0.3	0.0	23.1	81.4	0.22	pnp
Onne19	3.1	0.5	−0.6	0.0	12.2	15.6	0.44	ppn
Onne20	1.7	0.7	−0.6	1.6	5.0	14.1	0.26	ppn
Otobi18	2.4	1.5	−2.6	0.0	18.5	52.8	0.26	ppn
Otobi19	5.2	4.0	−1.5	5.4	48.7	122.7	0.28	ppn
Ubiaja17	5.0	2.0	−1.4	0.0	29.4	34.5	0.46	ppn
Ubiaja18	2.2	1.6	−1.7	0.0	17.7	23.0	0.43	ppn
Ubiaja20	2.8	−1.1	−2.4	0.6	14.9	15.0	0.50	pnn
Umudike17	3.5	1.1	−0.9	0.0	13.6	36.2	0.27	ppn
Umudike19	4.5	1.0	−0.4	11.2	31.3	61.5	0.34	ppn
Zaria20	1.5	−0.8	−0.6	0.0	3.0	12.9	0.19	pnn
Min	0.3	−3.5	−2.8	0.0	2.8	10.1	0.09	
Max	8.6	4.0	2.4	11.2	81.8	122.7	0.59	
Median	3.2	0.4	−0.6	0.6	21.0	36.2	0.42	
Mean	3.6	0.0	−0.3	2.9	28.8	42.0	0.39	

had high performance and stability while those in the bottom right-hand side had low performance and stability (see also Supplementary Table 7).

Genotype stability may be best viewed using latent regression plots which revealed genotypic responses to each factor loading (Smith et al., 2015). We considered latent regression plots for six clones which included the top two overall performance clones (TMS13F1376P0018 and IITA-TMS-IBA000070), top two stability clones (TMS14F1306P0020 and TMS13F1365P0029), and two clones known for possessing high industrial starch content (TMEB419 and TMS14F1036P0007; Supplementary Figures 4–6). Regression lines have slopes given by the estimated genotype scores for the individual and factor concerned. The regression on the first factor has a maximum impact on the predicted breeding values for explaining the largest percentage (53.5%) of total genetic variation. Since the estimated loadings for this factor are non-negative, large positive regression coefficients for this factor indicate high fresh root yield.

Clustering of target environments and locations

Dendrogram clusters of 27 environments (Figure 4) and 11 locations (Figure 5) using the loadings from the FA3 model reflected how the environments and locations were related. The environments were clustered at a distance of approximately eight while the locations were grouped at distance height of approximately three. This is an indication

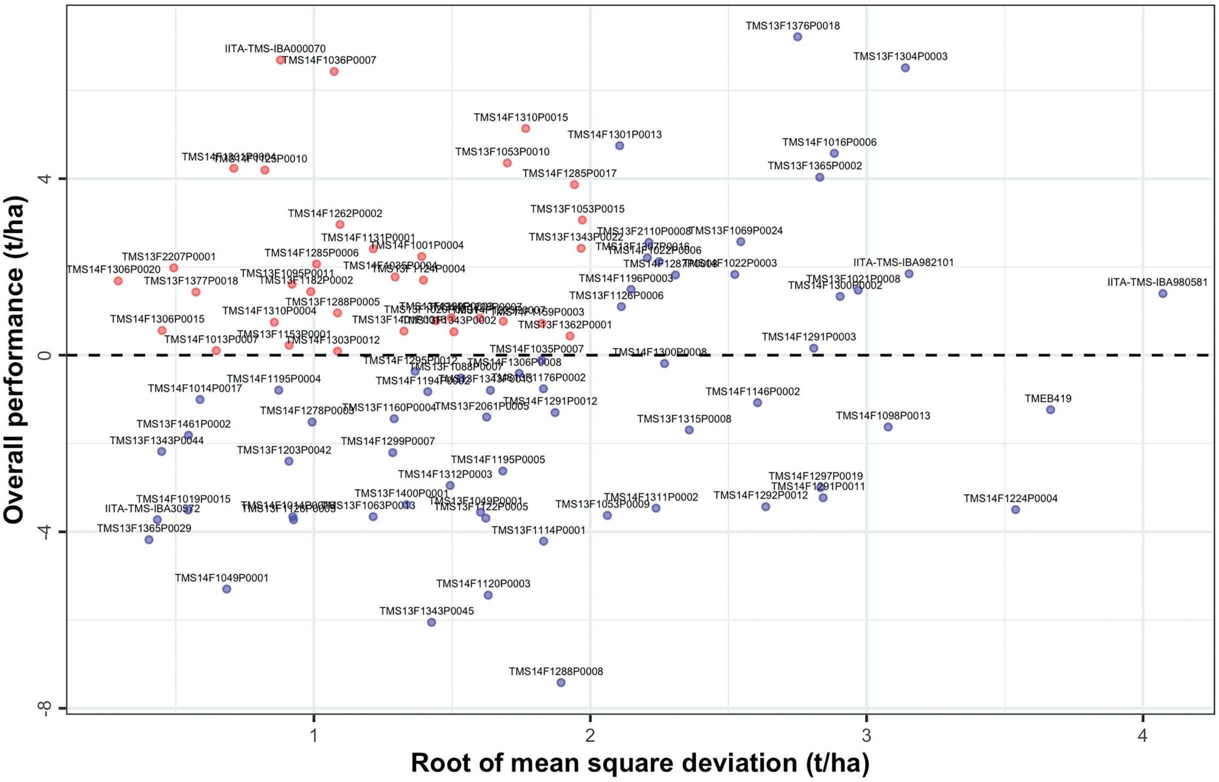


FIGURE 3
Overall performance (OP) vs. stability (Root of mean square deviation, RMSD) for fresh root yield showing all 96 clones evaluated across the environments.

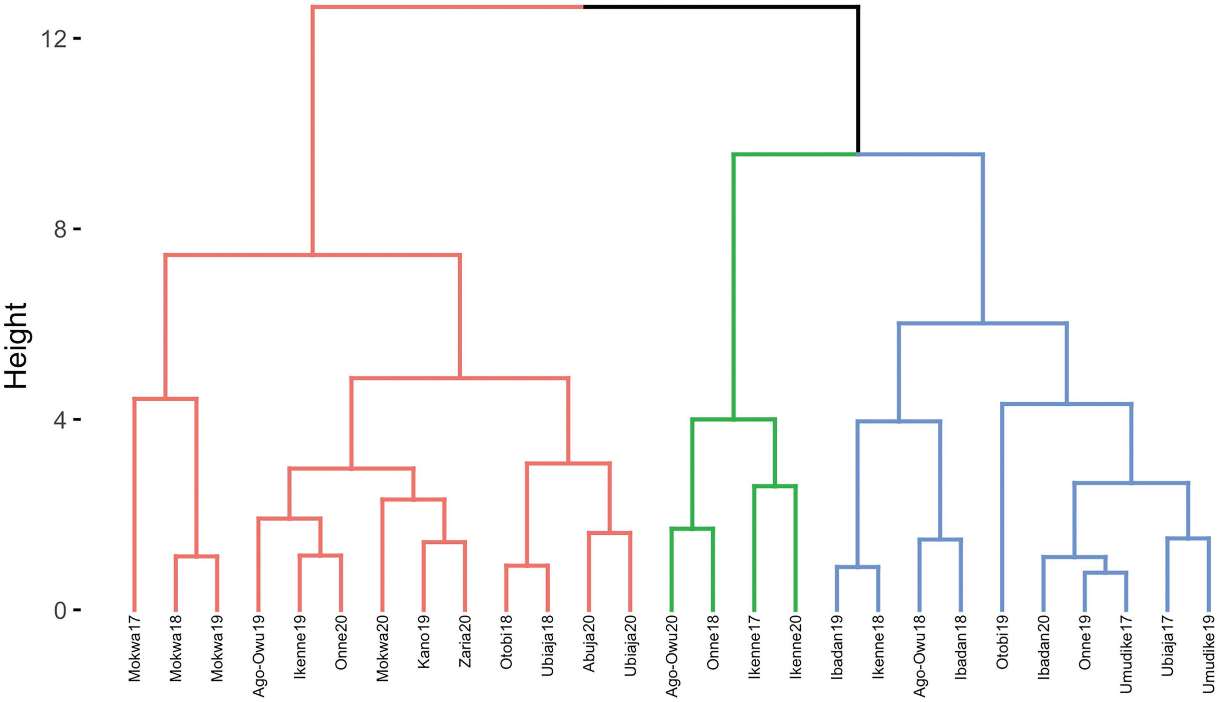


FIGURE 4
Dendrogram of 27 environments based on cassava fresh root yield using rotated factor loadings from FA3 model and Ward's D2 linkage method.

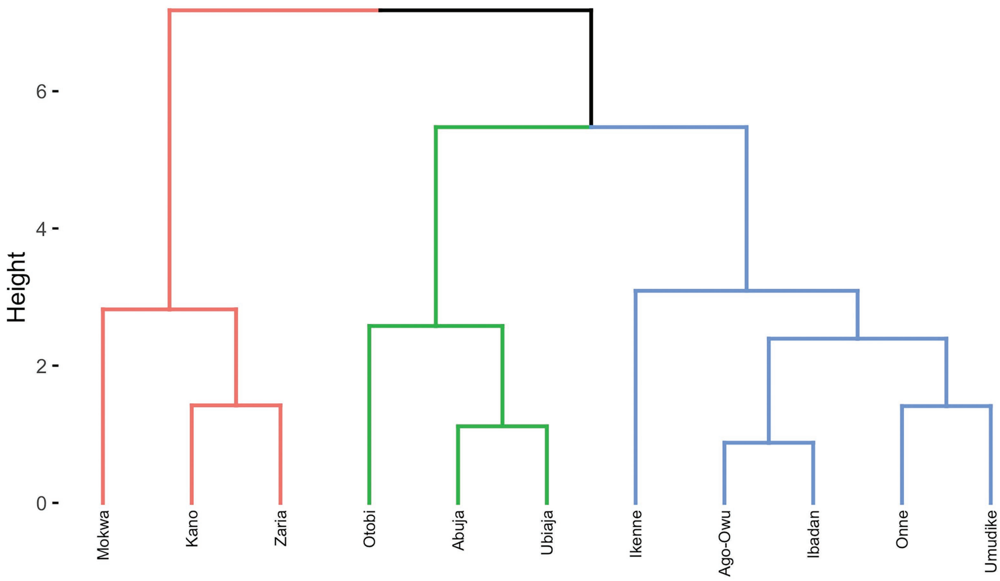


FIGURE 5
Dendrogram of 11 locations based on cassava fresh root yield using average rotated factor loadings from FA3 model and Ward's D2 linkage method.

TABLE 4 Mean factor loadings, number and name of environments within each of four interactive classes (iClasses).

iClass	Factor 1	Factor 2	Factor 3	Number of environment	Environment
pnn	1.7	−1.6	−1.5	5	Kano19, Mokwa17, Mokwa20, Ubiaja20, Zaria20
pnp	5.7	−2.0	0.6	6	Ago-Owu20, Ikenne17, Ikenne20, Mokwa18, Mokwa19, Onne18
ppn	3.3	1.3	−1.4	10	Abuja20, Ikenne19, Onne19, Onne20, Otobi18, Otobi19, Ubiaja17, Ubiaja18, Umudike17, Umudike19
ppp	3.6	1.1	1.4	6	Ago-Owu18, Ago-Owu19, Ibadan18, Ibadan19, Ibadan20, Ikenne18

pnn, positive negative negative; pnp, positive negative positive; ppn, positive positive negative; and ppp, positive positive positive.

that, averaged over years, locations are less differentiated than environments. There was a consistent pattern of Mokwa belonging to the same cluster with environments Kano and Zaria over years (Figures 4, 5). Likewise, the environments associated with Ikenne are in the same cluster except for Ikenne19 which belonged to another cluster. We observed consistent similarity in the environments of Ago-Owu, Onne, Ikenne, and Ibadan, so that these locations were also clustered (Figures 4, 5). The environments Umudike17, Umudike19 shared common characteristics with one out of the three environments in Onne (Figure 4) leading those two locations to be clustered (Figure 5). We identified four interactive classes (pnn, pnp, ppn, and ppp) of the possible $2^3 = 8$ iClasses with 5, 6, 10, and 6 environments each (Table 4). Each of these clusters of environments had a minimal crossover genotype-by-environment interaction and the contrasts between the environments within the same cluster group were eliminated (Smith et al., 2021).

Association of factor loadings with environmental covariables

The first PLSR component had relatively high positive X-loadings for environmental covariables TRAN1, TRAN2, TRAN3, TRAN4, TMAX1, TMAX2, TMAX3, TMAX4, SRAD1, SRAD2, SRAD3, and SRAD4 (Table 5) and showed high negative Pearson's correlation coefficients (*r*) with the first factor loading of the FA3 model. However, these environmental covariables were in contrast to RH1, RH2, RH3, RH4, RZSW1, RZSW2, RZSW3, RZSW4, SM1, SM2, SM3, SM4, SSW1, SSW2, SSW3, SSW4, and TMIN2 showing high negative X-loadings in the first PLSR component and positively correlated to first factor loading. Conversely, the second PLSR component identified WS1, WS2, WS3, and WS4 as environmental covariables that had moderately high negative X-loadings.

Based on Pearson's correlation coefficients, we observed that RH2, RH3, SM1, SM2, and SM3 were weather conditions that had

TABLE 5 X-loadings of the first and second PLSR components of environmental covariables and their Pearson's correlation coefficients sorted in descending order of the first latent factor loadings extracted from the FA3 model.

Environmental covariables	Partial least squares		Factor analytic model		
	Component 1	Component 2	Factor 1	Factor 2	Factor 3
RH3	−0.19	0.07	0.45	0.48	0.45
SM3	−0.19	−0.04	0.42	0.43	0.25
RH2	−0.19	0.03	0.41	0.28	0.35
SM2	−0.20	−0.09	0.40	0.43	0.19
SM1	−0.20	−0.04	0.40	0.49	0.31
TMIN2	−0.16	0.05	0.39	0.10	0.25
RH4	−0.18	−0.17	0.37	0.40	0.17
RH1	−0.18	−0.19	0.36	0.50	0.08
SSW3	−0.20	0.05	0.36	0.41	0.36
RZSW1	−0.20	−0.06	0.36	0.47	0.29
SM4	−0.20	−0.10	0.36	0.48	0.26
RZSW3	−0.19	0.02	0.35	0.42	0.32
SSW1	−0.20	−0.09	0.35	0.51	0.24
SSW2	−0.19	−0.03	0.35	0.37	0.24
RZSW2	−0.19	−0.02	0.33	0.33	0.23
SSW4	−0.19	−0.12	0.28	0.44	0.23
RZSW4	−0.19	−0.12	0.27	0.46	0.23
TMIN3	−0.11	−0.06	0.25	0.34	0.10
PRECIP3	−0.07	0.31	0.14	0.21	0.53
WS1	0.01	−0.43	0.10	0.09	−0.40
TMIN1	−0.08	0.05	0.08	−0.05	0.05
PRECIP4	−0.07	0.07	0.06	0.25	0.30
PRECIP1	−0.11	0.08	0.02	0.20	0.08
TMIN4	0.00	0.04	−0.05	−0.21	−0.18
PRECIP2	−0.10	0.13	−0.05	0.04	0.11
WS4	0.09	−0.41	−0.16	0.03	−0.45
SRAD2	0.16	−0.04	−0.21	−0.11	−0.32
WS3	0.12	−0.41	−0.23	−0.03	−0.48
WS2	0.10	−0.39	−0.24	−0.12	−0.41
TMAX2	0.17	−0.03	−0.24	−0.34	−0.29
SRAD3	0.18	−0.13	−0.27	−0.38	−0.40
TRAN4	0.18	0.14	−0.31	−0.35	−0.13
TMAX4	0.17	0.15	−0.31	−0.42	−0.20
TRAN1	0.18	0.22	−0.31	−0.36	−0.01
SRAD4	0.16	0.06	−0.32	−0.52	−0.25
TMAX1	0.17	0.28	−0.32	−0.44	0.01
TMAX3	0.18	−0.15	−0.32	−0.43	−0.51
SRAD1	0.17	0.07	−0.33	−0.24	−0.08
TRAN2	0.18	−0.05	−0.36	−0.23	−0.30
TRAN3	0.18	−0.07	−0.37	−0.49	−0.40

TMAX, mean maximum temperature; TMIN, mean minimum temperature; TRAN, mean temperature range; PRECIP, total precipitation; RH, mean relative humidity; WS, mean wind speed; SRAD, mean solar radiation; SSW, mean surface soil wetness; RZSW, mean root zone soil wetness; SM, mean soil moisture. The suffixes 1, 2, 3, and 4 denote the covariables measured at first, second, third, and fourth developmental phases of cassava crop, respectively.

high positive association to factor 1 but in contrast to TRAN1, TRAN2, TRAN3, TRAN4, TMAX1, TMAX3, TMAX4, SRAD1, and SRAD4 which revealed high negative correlation (Table 5); SSW1, RH1, RH3, SM1, SM4, RZSW1, RZSW4 were positive highly correlated but in contrast to SRAD4, TRAN3, TMAX1, TMAX3, TMAX4 which showed high negative correlation to factor 2; and PRECIP3, PRECIP4, RH2, RH3, RZSW3, SM1, and SSW3 had positive association but contrary to WS1, WS2, WS3, WS4, SRAD3, TRAN3, and TMAX3 (in factor 3) affected genotypic responses within environments clustered by these three

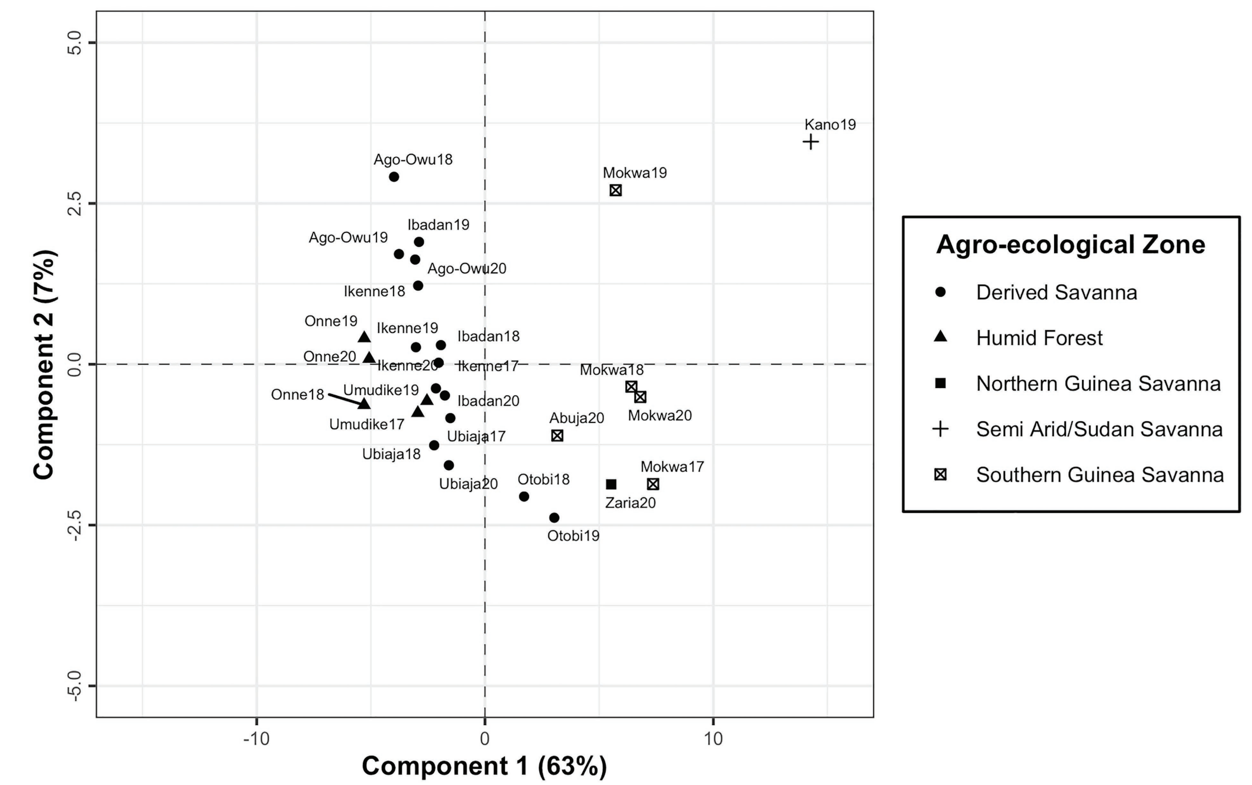


FIGURE 6
A plot of first and second components of X-scores revealing the grouping of the testing environments based on latent factor loadings from FA3 model and environmental covariables.

factors. This analysis provides useful information to understand the environmental covariates' influence on the clonal performance across the environments. It further allows the identification of the most likely environmental conditions affecting GEI in the testing environments.

PLSR was used to maximize covariance between the factor loadings and the environmental covariables at four developmental phases of cassava root crop. It identified the most significant environmental conditions influencing differential genotypic yield response in the testing environments. The first latent component resulting from fitting PLSR model explained 63% of variance in factor loadings. The addition of second component resulted in capturing 70% of total variation, and after the third component which accounted for 78% of variation, no significant improvement in the variance explained in the factor loadings. It was revealed that the first component separated the environments into two clustered groups and conversely its second component did not have a clear interpretation (Figure 6). The environment Kano19 was identified as a leverage point well separated from other environments (Figure 6).

However, a PLSR biplot of *X* and *Y* loadings revealed the association between environmental covariables at different developmental phases and factor loadings. The second component separated the third factor loading (FL3) from the remaining two factor loadings (FL1 and FL2; Figure 7). The environmental covariables found close to each other or in the close vicinity of factor loadings were positively correlated to each other and those situated in the opposite side are negatively correlated (Figure 7).

Discussion

The IITA cassava breeding program continually evaluates many clones in several target locations over years aiming to identify clones with high yield productivity and stability and to assess adaptability across a wide range of diverse environmental conditions. This evaluation necessitates the establishment of multi-environment trials annually to determine the clones' yield performance across various agro-ecological zones in Nigeria. The release of new cultivars arises when the clones possess specific characteristics that prove their desirable performance for a given geographical region, emphasizing the importance of MET in plant breeding programs (Oliveira et al., 2020).

Studying the patterns of MET data for decision making cannot be adequately investigated using conventional statistical methods due to some limitations as pointed out by Bakare et al. (2022). Therefore, factor analytic structures fitted in the linear mixed model framework as used in this study are flexible and robust for modeling complex genetic variance structure and more parsimonious for MET analyses than unstructured models (Smith

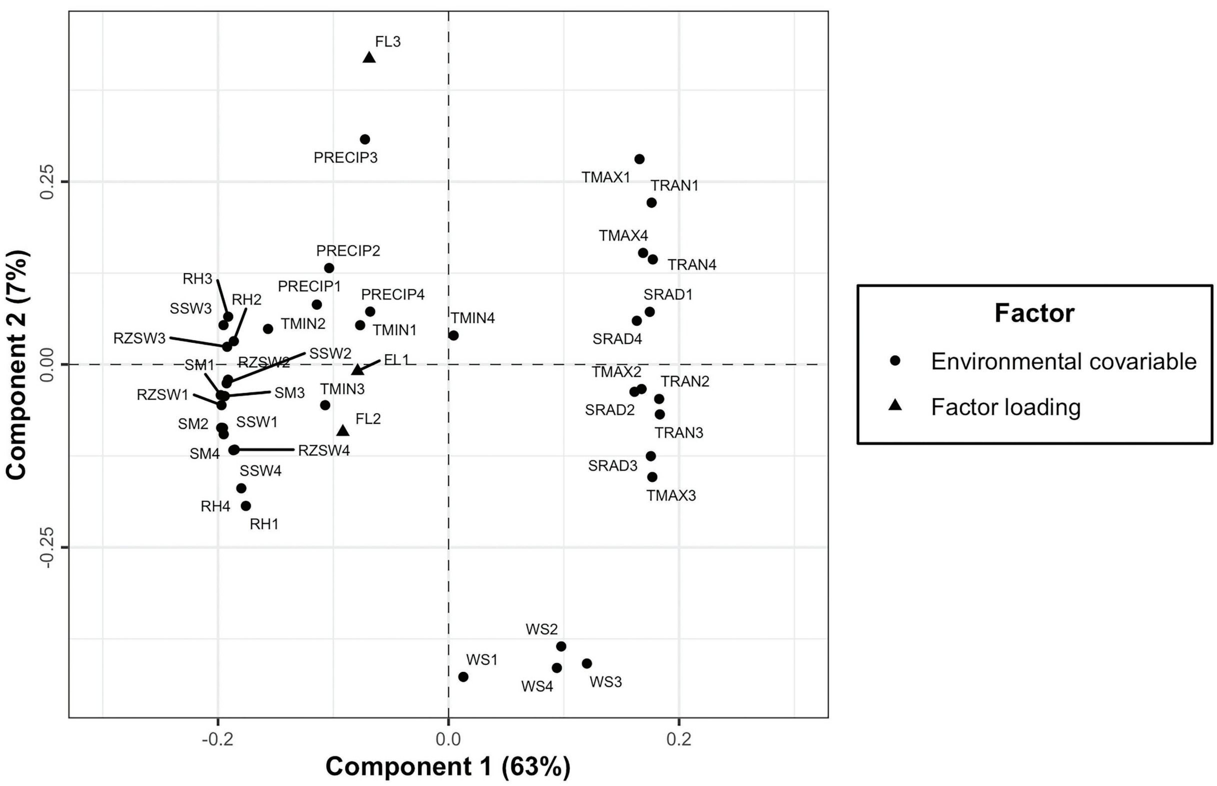


FIGURE 7
A plot of X and Y loadings revealing the association of factor loadings resulting from FA3 model to environmental covariables across four developmental phases of cassava.

et al., 2001a). Linear mixed models show great flexibility in handling unbalanced data that occur in METs due to unforeseen circumstances. The analyses of MET data have been broadly implemented using FA structures to understand the stability and adaptability of genotypes across testing environments (Li et al., 2017; Dias et al., 2018), and also to delineate mega-environments in plant breeding (Smith et al., 2015; Monteverde et al., 2018; Smith and Cullis, 2018).

Our study is the first to implement the FA model and to identify the factors influencing GEI in cassava. Furthermore, to our awareness, this is the first study that explored the extent of association between environmental covariables and factor loadings to examine the potential factors influencing GEI for fresh root yield in cassava. In this study, the Pearson's correlation between the environmental covariables and the factor loadings was used to describe the likely factors affecting GEI, as proposed by (Sae-Lim et al., 2014). This information is helpful to ascertain the effect each covariable has on genotypic performance across environments, toward identifying the most likely covariables affecting GE in a given set of environments. In general, relative humidity and temperature were the environmental covariates that explained the most genetic variability of fresh root yield across the environments. This information can support the breeders in recommending cassava clones for particular environments based on environmental covariables observed there historically. This ability will also facilitate the optimization of the number of testing environments for late stages of the breeding program, prioritizing environments with diverse environmental conditions.

The PLSR approach was found to be effective in clustering the testing environments from the southern region separately from that of the northern region of Nigeria based on factor loadings and environmental covariables incorporated into the model. The *X* and *Y* loadings biplot (Figure 7) showed that the GEI in the southern part of Nigeria was driven mostly by weather conditions such as minimum temperature, relative humidity, precipitation, surface soil wetness, root zone soil wetness, and soil moisture across the developmental phases of cassava. However, differential genotypic sensitivity across the environments in the north of Nigeria was mostly determined by wind speed, maximum temperature, temperature range and soil radiation. This study was limited to the environments where cassava breeders operate in Nigeria. The findings from PLSR could be used to restructure Nigerian breeding programs and adjust evaluation locations accordingly. However, future studies should explore how the environmental covariates could be used to forecast the performance of cassava clones in locations that were not within those evaluated in the previous MET.

The iClasses and ward.D2 hierarchical cluster were two approaches used to group the environments using the factor loadings. The former identified 4 clusters of environments based on the positive or negative signs of the loadings. Meanwhile the latter classified the environments into 3 cluster groups in terms of minimizing the change in variance. The two approaches showed a degree of similarity in terms of clustering environments from the same geographical regions together.

The use of latent regression plots to study yield stability and adaptability of genotypes across testing environments was recommended (Smith et al., 2015). In their approach, the predicted breeding values of genotypes are regressed on the factor loadings of the FA model. This study used FA structures and latent regression plots to identify cassava clones with high overall performance (TMS13F1376P0018 and IITA-TMS-IBA000070) and stability (TMS14F1306P0020 and TMS13F1365P0029) with their respective predicted genotypic scores for the first three factors (Supplementary Table 7).

Conclusion

This study demonstrated that the factor analytic model was the most parsimonious variance model to dissect and account for complex patterns of GEI by separating genetic effects into common and specific variance components. The delineation of testing environments or locations into clusters through a factor analytic model was an efficient way to optimize the resources by using one location per cluster group. The use of partial least squares regression proved to be an effective tool for identifying relevant environmental covariables affecting differential genotypic sensitivity in the context of multi-environment trials where a number of external environmental covariables are incorporated in modeling. Among the environmental covariables explored in this study, minimum temperature, precipitation, relative humidity, surface soil wetness, root zone soil wetness, and soil moisture were identified as the strongest influence on genotypic responses across the testing environments in the southern region of Nigeria. This was in contrast to maximum temperature, wind speed, and temperature range (difference between maximum and minimum temperature), and solar radiation affecting GEI in the northern region of Nigeria.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/mab658/variance_covariance_model_GxE.

Author contributions

MB, IR, J-LJ, CE, and PK: design and study conceptualization. IR, MB, SK, CA, and EP: study methodology and implementation. MB, SK, MW, J-LJ, and IR: formal data curation, analysis, and manuscript drafting. MB, SK, CA, MW, J-LJ, PK, and IR: manuscript reviewing and editing. IR, J-LJ, and PK: supervision, coordination and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

The research reported in this manuscript was generously funded by the Bill & Melinda Gates Foundation (https://www.gatesfoundation.org) and the United Kingdom's Foreign, Commonwealth & Development Office (FCDO https://www.gov.uk/government/organisations/foreign-commonwealth-development-office) through the "Next Generation Cassava Breeding project" (https://www.nextgencassava.org; grant no. INV-007637) managed by Cornell University.

Acknowledgments

The authors thank the UK's Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation (Grant INV-007637, http://www.gatesfoundation.org) for their financial support. We also extend our deep appreciation to the dedicated staff of the International Institute of Tropical Agriculture (IITA) of the cassava breeding program that assisted with the fieldwork, screen house evaluation, data collection and laboratory analyses. We thank reviewers BC and H-PP for their valuable suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.978248/full#supplementary-material

References

Aastveit, A. H., and Martens, H. (1986). ANOVA interactions interpreted by partial least squares regression. *Biometrics* 42, 829–844.

Akinwale, M., Akinyele, B. O., Odiyi, A. C., and Dixon, A. G. O. (2011). Genotype X environment interaction and yield performance of 43 improved cassava (*Manihot esculenta* Crantz) genotypes at three agro-climatic zones in Nigeria. *British Biotechnol. J.* 1, 68–84. doi: 10.9734/bbj/2011/475

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A., Resende, M. F. Jr., and Muñoz, P. R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. *plant Genome* 9:2016-01. doi: 10.3835/plantgenome2016.01.0009

Bakare, M. A., Kayondo, S. I., Aghogho, C. I., Wolfe, M. D., Parkes, E. Y., Kulakow, P., et al. (2022). Exploring genotype by environment interaction on cassava yield and yield related traits using classical statistical methods. *PLoS One* 17:e0268189. doi: 10.1371/journal.pone.0268189

Browning, B. L., and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98, 116–126. doi: 10.1016/j.ajhg.2015.11.020

Burgueño, J., Crossa, J., Cornelius, P. L., and Yang, R. C. (2008). Using factor analytic models for joining environments and genotypes without crossover genotype x environment interaction. *Crop Sci.* 48, 1291–1305. doi: 10.2135/cropsci2007.11.0632

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2017). ASReml-R reference manual version 4. ASReml-R Reference Manual, 176.

Crossa, J. (2012). From genotype × environment interaction to gene × environment interaction. *Curr. Genom.* 13, 225–244. doi: 10.2174/138920212800543066

Cullis, B. R., Jefferson, P., Thompson, R., and Smith, A. B. (2014). Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a Pinus radiata breeding programme. *Theor. Appl. Genet.* 127, 2193–2210. doi: 10.1007/s00122-014-2373-0

Cullis, B. R., Smith, A. B., Beeck, C. P., and Cowling, W. A. (2010). Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. *Genome* 53, 1002–1016. doi: 10.1139/G10-080

Dellaporta, S. L., Wood, J., and Hicks, J. B. (1983). A plant DNA minipreparation: version II. *Plant Mol. Biol. Report.* 1, 19–21. doi: 10.1007/BF02712670

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Parentoni, S. N., Guimarães, P. E. D. O., Carneiro, N. P., et al. (2018). Estimating genotype × environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. *Crop Sci.* 58, 72–83. doi: 10.2135/cropsci2016.07.0566

Dixon, A., and Ssemakula, G. (2007). Genotype X environment interaction, stability and agronomic performance of carotenoid-rich cassava clones. *Sci. Res. Essays* 2, 390–399. doi: 10.5897/SRE.900052

Egesi, C. N., Ilona, P., Ogbe, F. O., Akoroda, M., and Dixon, A. (2007). Genetic variation and genotype X environment interaction for yield and other agronomic traits in cassava in Nigeria. *Agron. J.* 99, 1137–1142. doi: 10.2134/agronj2006.0291

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Falconer, D. S. (1996). *Introduction to Quantitative Genetics*. Delhi, India: Pearson Education India.

Finlay, K. W., and Wilkinson, G. N. (1963). The analysis of adaptation in a plant-breeding programme. *Aust. J. Agric. Res.* 14, 742–754. doi: 10.1071/AR9630742

Gauch, H. G. Jr. (2016). *Model Selection and Validation for Yield Trials with Interaction*. Washington. DC: International Biometric Society Stable.

Gauch, H. G., and Zobel, R. W. (1997). Identifying mega-environments and targeting genotypes. *Crop Sci.* 37, 311–326. doi: 10.2135/cropsci1997.0011183X003700020002x

Gezan, S., de Oliveira, A., and Murray, D. (2021). *ASRgenomics: An R Package with Complementary Genomic Functions*. Hemel Hempstead: VSN International.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 51, 1440. doi: 10.2307/2533274

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9:e90346. doi: 10.1371/journal.pone.0090346

Hamblin, M. T., and Rabbi, I. Y. (2014). The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in cassava (*Manihot esculenta*). *Crop Sci.* 54, 2603–2608. doi: 10.2135/cropsci2014.02.0160

Jiwuba, L., Danquah, A., Asante, I., Blay, E., Onyeka, J., Danquah, E., et al. (2020). Genotype by environment interaction on resistance to cassava green mite associated traits and effects on yield performance of cassava genotypes in Nigeria. *Front. Plant Sci.* 11:572200. doi: 10.3389/fpls.2020572200

Kelly, A. M., Smith, A. B., Eccleston, J. A., and Cullis, B. R. (2007). The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Sci.* 47, 1063–1070. doi: 10.2135/cropsci2006.08.0540

Li, Y., Suontama, M., Burdon, R. D., and Dungey, H. S. (2017). Genotype by environment interactions in forest tree breeding: review of methodology and perspectives on research and application. *Tree Genet. Genomes* 13, 1–18. doi: 10.1007/s11295-017-1144-x

Liland, K. H., Mevik, B.-H., Wehrens, R., and Hiemstra, P. (2021). Pls: Partial Least Squares and Principal Component Regression. R package version 2.8-0.

Martini, J. W., Schrauf, M. F., Garcia-Baccino, C. A., Pimentel, E. C., Munilla, S., Rogberg-Muñoz, A., et al. (2018). The effect of the H− 1 scaling factors τ and ω on the structure of H in the single-step procedure. *Genet. Sel. Evol.* 50, 16–19. doi: 10.1186/s12711-018-0386-x

Meyer, K. (2009). Factor-analytic models for genotype × environment type problems and structured covariance matrices. *Genet. Sel. Evol.* 41, 1–11. doi: 10.1186/1297-9686-41-21

Misztal, I., Aguilar, I., Legarra, A., Tsuruta, S., Johnson, D., and Lawlor, T. (2010). *A unified approach to utilize phenotypic, full pedigree, andgenomic information for genetic evaluation.*

Monteverde, E., Rosas, J. E., Blanco, P., Pérez de Vida, F., Bonnecarrère, V., Quero, G., et al. (2018). Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of rice. *Crop Sci.* 58, 1519–1530. doi: 10.2135/cropsci2017.09.0564

Mrode, R. A. (2014). *Linear Models for the Prediction of animal Breeding Values*, (*3rd Edn.*). CABI, Oxfordshire, UK.

Nduwumuremyi, A., Melis, R., Shanahan, P., and Theodore, A. (2017). Interaction of genotype and environment effects on important traits of cassava (*Manihot esculenta* Crantz). *Crop J.* 5, 373–386. doi: 10.1016/j.cj.2017.02.004

Oakey, H., Verbyla, A. P., Cullis, B. R., Wei, X., and Pitchford, W. S. (2007). Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theor. Appl. Genet.* 114, 1319–1332. doi: 10.1007/s00122-007-0515-3

Oliveira, I. C. M., Guilhen, J. H. S., Ribeiro, P. C. D. O., Gezan, S. A., Schaffert, R. E., Simeone, M. L. F., et al. (2020). Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates. *Field Crop Res.* 257:107929. doi: 10.1016/j.fcr.2020.107929

Piepho, H.-P. (1997). Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms. *Biometrics* 53, 761–766.

Piepho, H. P. (1998a). Empirical best linear unbiased prediction in cultivar trials using factor analytic variance-covariance structures. *Agron. J.* 91, 154–160. doi: 10.2134/agronj1999.00021962009100010024x

Piepho, H. P. (1998b). Methods for comparing the yield stability of cropping systems—A review. *J. Agron. Crop Sci.* 180, 193–213. doi: 10.1111/j.1439-037X.1998.tb00526.x

Prochnik, S., Marri, P. R., Desany, B., Rabinowicz, P. D., Kodira, C., Mohiuddin, M., et al. (2012). The cassava genome: current Progress, future directions. *Trop. Plant Biol.* 5, 88–94. doi: 10.1007/s12042-011-9088-z

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sae-Lim, P., Komen, H., Kause, A., and Mulder, H. A. (2014). Identifying environmental variables explaining genotype-by-environment interaction for body weight of rainbow trout (Onchorynchus mykiss): reaction norm and factor analytic models. *Genet. Sel. Evol.* 46, 1–11. doi: 10.1186/1297-9686-46-16

Sayre, R., Beeching, J. R., Cahoon, E. B., Egesi, C., Fauquet, C., Fellman, J., et al. (2011). The BioCassava plus program: biofortification of cassava for sub-Saharan Africa. *Annu. Rev. Plant Biol.* 62, 251–272. doi: 10.1146/annurev-arplant-042110-103751

Smith, A. B., and Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* 214, 1–19. doi: 10.1007/s10681-018-2220-5

Smith, A., Cullis, B., and Gilmour, A. (2001a). The analysis of crop variety evaluation data in Australia. *Aust. N.Z. J. Stat.* 43, 129–145. doi: 10.1111/1467-842X.00163

Smith, A., Cullis, B., and Thompson, R. (2001b). Analyzing variety by environment data using mulitplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147. doi: 10.1111/j.0006-341X.2001.01138.x

Smith, A. B., Cullis, B. R., and Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *J. Agric. Sci.* 143, 449–462. doi: 10.1017/S0021859605005587

Smith, A. B., Ganesalingam, A., Kuchel, H., and Cullis, B. R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theor. Appl. Gene.* 128, 55–72. doi: 10.1007/s00122-014-2412-x

Smith, A., Norman, A., Kuchel, H., and Cullis, B. (2021). Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. *Front. Plant Sci.* 12:737462. doi: 10.3389/fpls.2021.737462

Talbot, M., and Wheelwright, A. (1989). The analysis of genotype× environment interactions by partial least squares regression. *Biuletyn Oceny Odmian* 21, 19–25.

Tier, B., Meyer, K., and Ferdosi, M. (2015). Which genomic relationship matrix? *Proceedings of the 21st Conference of the Association for the Advancement of Animal Breeding and Genetics* 21: 461–464. 28–30.

Tumuhimbise, R., Melis, R., Shanahan, P., and Kawuki, R. (2014). Genotype × environment interaction effects on early fresh storage root yield and related traits in cassava. *Crop J.* 2, 329–337. doi: 10.1016/j.cj.2014.04.008

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Vargas, M., Crossa, J., Sayre, K., Reynolds, M., Ramírez, M. E., and Talbot, M. (1998). Interpreting genotype × environment interaction in wheat by partial least squares regression. *Crop Sci.* 38, 679–689. doi: 10.2135/cropsci1998.0011183X003800030010x

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Houston, TX: Springer-Verlag.

Yan, W., Hunt, L. A., Sheng, Q., and Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci.* 40, 597–605. doi: 10.2135/cropsci2000.403597x

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 October 2022
DOI 10.3389/fpls.2022.1016170

OPEN ACCESS

EDITED BY
Mulatu Geleta,
Swedish University of Agricultural Sciences, Sweden

REVIEWED BY
John Saviour Yaw Eleblu,
University of Ghana, Ghana
Abel Gari,
International Livestock Research Institute, Ethiopia

*CORRESPONDENCE
Ismail Y. Rabbi
i.rabbi@cgiar.org

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 10 August 2022
ACCEPTED 15 September 2022
PUBLISHED 12 October 2022

CITATION
Ige AD, Olasanmi B, Bauchet GJ, Kayondo IS, Mbanjo EGN, Uwugiaren R, Motomura-Wages S, Norton J, Egesi C, Parkes EY, Kulakow P, Ceballos H, Dieng I and Rabbi IY (2022) Validation of KASP-SNP markers in cassava germplasm for marker-assisted selection of increased carotenoid content and dry matter content.
Front. Plant Sci. 13:1016170.
doi: 10.3389/fpls.2022.1016170

Validation of KASP-SNP markers in cassava germplasm for marker-assisted selection of increased carotenoid content and dry matter content

Adenike D. Ige[1,2], Bunmi Olasanmi[3], Guillaume J. Bauchet[4], Ismail S. Kayondo[1], Edwige Gaby Nkouaya Mbanjo[1], Ruth Uwugiaren[1,5], Sharon Motomura-Wages[6], Joanna Norton[6], Chiedozie Egesi[1,7], Elizabeth Y. Parkes[1], Peter Kulakow[1], Hernán Ceballos[8], Ibnou Dieng[1] and Ismail Y. Rabbi[1]*

[1]International Institute of Tropical Agriculture (IITA), Ibadan, Oyo State, Nigeria, [2]Pan African University Life and Earth Sciences Institute (including Health and Agriculture), University of Ibadan, Ibadan, Nigeria, [3]Department of Crop and Horticultural Sciences, University of Ibadan, Ibadan, Nigeria, [4]Boyce Thompson Institute, Ithaca, NY, United States, [5]Molecular Plant Sciences program, Washington State University, Pullman, WA, United States, [6]College of Tropical Agriculture and Human Resources, University of Hawaii at Manoa, Hilo, HI, United States, [7]Cornell University, Ithaca, NY, United States, [8]The Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT), Cali, Colombia

Provitamin A biofortification and increased dry matter content are important breeding targets in cassava improvement programs worldwide. Biofortified varieties contribute to the alleviation of provitamin A deficiency, a leading cause of preventable blindness common among pre-school children and pregnant women in developing countries particularly Africa. Dry matter content is a major component of dry yield and thus underlies overall variety performance and acceptability by growers, processors, and consumers. Single nucleotide polymorphism (SNP) markers linked to these traits have recently been discovered through several genome-wide association studies but have not been deployed for routine marker-assisted selection (MAS). This is due to the lack of useful information on markers' performances in diverse genetic backgrounds. To overcome this bottleneck, technical and biological validation of the loci associated with increased carotenoid content and dry matter content were carried out using populations independent of the marker discovery population. In the present study, seven previously identified markers for these traits were converted to a robust set of uniplex allele-specific polymerase chain reaction (PCR) assays and validated in two independent pre-breeding and breeding populations. These assays were efficient in discriminating marker genotypic classes and had an average call rate greater than 98%. A high correlation was observed between the predicted and observed carotenoid content as inferred by root yellowness intensity in the breeding (r = 0.92) and pre-breeding (r = 0.95) populations. On the other hand, dry matter content-markers had moderately low predictive accuracy in both populations (r< 0.40) due to the more quantitative nature of the trait. This work confirmed the markers' effectiveness in multiple backgrounds, therefore, further strengthening their value in cassava biofortification to ensure nutritional security as well as dry matter content productivity. Our study provides a framework to guide future marker validation, thus leading to the more routine use of markers in MAS in cassava improvement programs.

KEYWORDS

cassava (*Manihot esculenta* Crantz), provitamin A content, dry matter content, allele-specific PCR, predictive accuracy, marker-assisted selection

1 Introduction

Cassava (*Manihot esculenta* Crantz) is a principal starchy root crop for both the rural and urban populations in the tropics, particularly in sub-Saharan Africa. The continent accounts for more than half of the total world's production of 303 million tonnes (FAOSTAT, 2020). Due to its ability to grow with few agricultural inputs in marginal environments characterized by poor soils and water stress, the crop takes on the crucial role of being a key food security crop in sub-Saharan Africa (Burns et al., 2010). In Africa, cassava roots are usually consumed fresh after short boiling and are also processed into various fermented products such as gari and fufu or unfermented products such as flour and starch. Besides its role as food, cassava is increasingly relied upon globally as an industrial raw material for the production of paper, textiles, plywood, glue, biofuel, animal feed and beverages (Balagopalan, 2002).

Among the major staple sources of carbohydrates, cassava has one of the longest breeding cycles, ranging from five to eight years (Ceballos et al., 2004; Ceballos et al., 2012). This is due to its long growth cycle of 12 - 18 months; clonal propagation, which results in low multiplication rates of planting propagules; its high levels of heterozygosity; and difficulty in making crosses due to poor and asynchronous flowering as well as low seed set per cross (Jennings and Iglesias, 2002; Ceballos et al., 2012). These challenges notwithstanding, breeding programs around the world have developed improved varieties that address various production constraints, including biotic and abiotic stresses, improved yield and dry matter content (Kawano, 2003; Okechukwu and Dixon, 2008), as well as enhanced micronutrient content, particularly of provitamin A carotenoid (Ilona et al., 2017; Andersson et al., 2017). However, as the demand for cassava for food, feed, and industrial raw materials continues to grow due to an increase in the population (Anyanwu et al., 2015; Parmar et al., 2017), breeding programs need to adopt modern breeding technologies and tools such as marker-assisted selection or genomic selection to increase the rate of genetic gain to meet the demands in an ecologically sustainable manner (Ceballos et al., 2015).

Marker Assisted Selection (MAS) is one of the most important applications of molecular marker technology in plant breeding (Collard and Mackill, 2008). It facilitates the indirect selection of new plants based on the presence of a favorable allele at a marker that is closely linked to a trait of interest (Collard and Mackill, 2008). In cassava, MAS can be used at the early stages of the breeding scheme to select individuals with favorable alleles for storage-root traits that would otherwise only be phenotypically evaluated at maturity. This has several advantages, namely: 1) reduction in the time it takes to decide to advance a clone to the next stage of testing; 2) reduction in the number of clones to be advanced to larger plot trials, thereby saving scarce phenotyping resources, and 3) in some cases, the cost of marker assay is lower than those that are usually expended on the actual trait phenotyping. A good example is carotenoid quantification using the spectrophotometry method and High-Performance Liquid Chromatography (HPLC) which can be many-fold more expensive than a SNP assay (Semagn et al., 2014; Andersson et al., 2017). Therefore, the adoption of MAS can increase the efficiency of selection, leading to a more rapid rate of genetic gain, and fewer cycles of phenotypic evaluation, thus, reducing the time for varietal development (Collard and Mackill, 2008).

The prerequisite for the application of MAS is the identification of major genes or genomic regions associated with a trait of interest. Over the last 15 years, quantitative trait

Abbreviations: BLAST, Basic Local Alignment Search Tool; BLUP, Best Linear Unbiased Prediction; CET, Clonal Evaluation Trial; GWAS, Genome-Wide Association Study; HPLC, High-Performance Liquid Chromatography; IITA, International Institute of Tropical Agriculture; KASP, Kompetitive Allele-Specific PCR; MAE, Mean Absolute Error; MAS, Marker-Assisted Selection; PCR, Polymerase Chain Reaction; *PSY2*, Phytoene Synthase 2; QTL, Quantitative Trait Loci; REML, Restricted Maximum Likelihood; RMSE, Root Mean Square Error; SD, Standard Deviation; SN, Seedling Nursery; SNP, Single Nucleotide Polymorphism.

loci (QTL) mapping studies of different traits in cassava have been published (Fregene et al., 2001; Akano et al., 2002; Balyejusa et al., 2007; Okogbenin et al., 2012; Morillo et al., 2013; Rabbi et al., 2014). Most of these studies used segregating populations developed from either selfed or bi-parental crosses between parents with contrasting trait levels (Rabbi et al., 2014). More recently, association or linkage disequilibrium mapping using a genome-wide association study (GWAS) has become an approach for unraveling the molecular genetic basis underlying the natural phenotypic variation (Davey et al., 2011). The advantages of GWAS over QTL mapping are the higher mapping resolution and the identification of a broader set of alleles in large and diverse germplasm (Yu and Buckler, 2006). Several GWAS have been conducted on key cassava traits, including cassava mosaic disease resistance (Wolfe et al., 2016; Rabbi et al., 2020), carotenoids content (Esuma et al., 2016; Rabbi et al., 2017; Ikeogu et al., 2019; Rabbi et al., 2020), and dry matter content (Rabbi et al., 2020) in diverse cassava populations to discover significant loci. Despite this progress, the output from discovery research has not been translated into assays that breeders can easily use to support selection decisions (Chagné et al., 2019). To overcome this bottleneck and bridge the gap between discovery and routine usage, new trait-linked markers must be technically and biologically validated, preferably using independent populations (Platten et al., 2019; Ige et al., 2021). This process informs the breeder whether the expected allelic phenotypic effects are reproducible in different genetic backgrounds from the one in which the marker-trait association was originally identified (Li et al., 2013).

Cassava is very efficient in carbohydrate production, but its starchy roots lack essential micronutrients, including provitamin A carotenoid (Sayre et al., 2011; Ceballos et al., 2017). Vitamin A deficiency is a public health problem in more than half of all countries, especially in Africa and South-East Asia (WHO, 2022). This deficiency often leads to several severe health and economic consequences, including increased incidence of night blindness; suppressed immunity, leading to an increased mortality rate, especially among pregnant women and young children as well as reduced productivity (Sayre et al., 2011; WHO, 2022). Dry matter content is a crucial yield component and is a key determinant of variety acceptance by growers, processors, and consumers (Sánchez et al., 2014; Bechoff et al., 2018). Varieties with low dry matter content (less than 30%) are often less preferred than those with moderate to high dry matter content. Like carotenoid content, dry matter content can only be assessed on mature storage roots at the end of the growing season. Marker-assisted selection is expected to provide breeders with the ability, for example, to screen either for genotypes with high levels of these traits or eliminate those with undesirable levels at the early stages of testing, thereby allocating their limited field plots to high-value genotypes. The objective of the study was, therefore, to convert and validate SNP markers associated with increased provitamin A carotenoid biofortification and dry matter content; two important traits under active improvement in many breeding programs in the world.

2 Materials and methods

2.1 Retrieving significant loci linked to increased carotenoid and dry matter contents

The marker discovery, development, and validation workflow used in the present study is presented in Figure 1. The SNP markers linked to increased carotenoid and dry matter contents validated in the present study (Table 1) were derived from (Udoh et al., 2017; Rabbi et al., 2020). Sequencing of four carotenoid pathway candidate genes in 167 cassava accessions from the International Institute of Tropical Agriculture (IITA), Nigeria, uncovered two important SNPs on phytoene synthase 2 (*PSY2*) (Udoh et al., 2017). The most significant SNP on *PSY2* (position 572) is a causal mutation resulting in a non-synonymous amino acid substitution (Welsch et al., 2010). This marker was converted to a Kompetitive allele-specific PCR (KASP) assay and renamed as per its chromosomal position on the version 6.1 reference genome to S1_24155522. Additional markers associated with the study traits were obtained from a recent GWAS that used a large panel of 5130 diverse clones developed at IITA in Nigeria (Rabbi et al., 2020). The population was genotyped at more than 100K genome-wide SNP markers *via* genotyping-by-sequencing. For carotenoid content, a major locus on chromosome 1 tagged by three markers (S1_24159583, S1_24636113, and S1_30543962) as well as five new genomic regions associated with this trait on chromosomes 5, 8, 15, and 16 were identified. Of these, three (S1_30543962, S5_3387558, and S8_25598183) were selected for KASP conversion and validation in the present study. The markers associated with dry matter content were S1_24197219, S6_20589894, and S12_5524524.

2.2 Development of kompetitive allele-specific PCR assays

Fifty nucleotide bases flanking the target SNP on each side were obtained from the cassava (*Manihot esculenta*) reference genome (version 6.1) available at https://phytozome-next.jgi.doe.gov/info/Mesculenta_v6_1. Then, a nucleotide-nucleotide Basic Local Alignment Search Tool (BLAST) was used to check for locus-specificity of the assays to minimize the possibility of cross-amplification of the marker in non-target regions of the genome. Primers were designed using a proprietary Kraken^{TM} software system from LGC Biosearch Technologies, UK, with the default parameters.

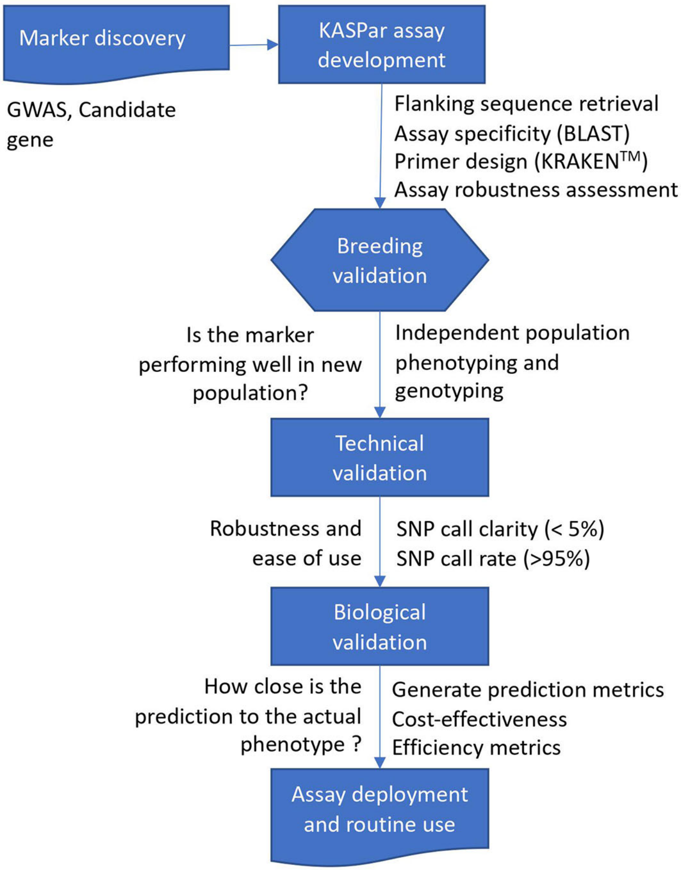

Marker discovery
GWAS, Candidate gene
KASPar assay development
Flanking sequence retrieval
Assay specificity (BLAST)
Primer design (KRAKEN™)
Assay robustness assessment
Breeding validation
Is the marker performing well in new population?
Independent population phenotyping and genotyping
Technical validation
Robustness and ease of use
SNP call clarity (< 5%)
SNP call rate (>95%)
Biological validation
How close is the prediction to the actual phenotype ?
Generate prediction metrics
Cost-effectiveness
Efficiency metrics
Assay deployment and routine use

FIGURE 1
Schematic overview of marker discovery, assay development, and validation of trait-linked markers for molecular breeding.

Assay technical validation was carried out using a panel of 188 genetically diverse cassava genotypes that are known to segregate at the SNP assays. A no-template control was included in the SNP genotyping. The robustness of the assays was assessed under four DNA concentrations (Dilution 1 = 10X, Dilution 2 = 100X, Dilution 3 = 24X, Dilution 4 = 240X) using metrics such as ease of scoring the three expected genotype classes, tightness, and distinctiveness of the genotypic classes on cluster plots, percentage call rate, and percentage clarity.

2.3 Validation of kompetitive allele-specific PCR assays in independent populations

The KASP assays' performances were assessed in two independent populations from IITA, Nigeria. These populations, consisting of breeding and pre-breeding germplasm, were different from the panel used for GWAS marker discovery.

2.3.1 Description of the study populations

The breeding population is part of IITA's regular recurrent selection pipeline and was derived from controlled crosses among elite genotypes carried out in 2017. Yield, multiple stress tolerance, and dry matter content are the major traits for improvement in this population. The cohort was evaluated initially at the seedling nursery (SN) stage consisting of 22,420 progenies from 563 families (mean family size of 40, ranging from 1 to 220) in 2018 in Ibadan, Nigeria (7°24′ N, 3°54′ E; 200 m above sea level). The SN trial was planted at a spacing of 1 m × 0.25 m and harvested 12 months after planting; a selection of 1599 genotypes based on disease resistance, plant vigor, plant architecture, and root yield was advanced to clonal evaluation trial (CET) at Ikenne, Nigeria (6°52′ N 3°42′ E; 61 m above sea level).

The pre-breeding population was developed using a polycross hybridization between twenty-three (23) IITA and nineteen (19) CIAT (International Center for Tropical Agriculture) parental clones. To ensure safe germplasm exchange between Africa and Latin America, the hybridization

TABLE 1 Description of the seven SNP markers, their flanking sequences and KASP primers.

Traits	SNP name*	Chr	Position (bp)	Flanking sequences	Favorable allele	Unfavorable allele	Forward primer (Allele X and Allele Y)	Common Reverse Primer
Carotenoid content	S1_24155522	1	24155522	GACAGATGAGCTTGTTGATGGACCTAATGCTTCACACATAACGCCAACAG[A/C]TTTAGATAGGTGGGAAGCAAGGTTGGAAGATATGTTTCGAGGTCGTCCCT	A	C	Allele X: ATGCTTCACACATAACGCCAACAGA Allele Y: TGCTTCACACATAACGCCAACAGC	CCAACCTTGCTTCCCACCTATCTAA
	S1_30543962	1	30543962	GGAGGTTTTTTTATGTGGCATTCTCAGCAGCTGCAGGAATCTCATTGTTCTTTACAATTCCAAGGCTCTTTCTTGCAATTAAAGGTGGGGAAGGTGCCCC[A/G]GACCTCTGGGGAACTGCTGGAAATGCTGCCATTAATATTGGTGGTAAATGCTTTAACCTTTCTCTGTCATATGAAGAAAATGAGTTAATTGATGTATAAT	G	A	Allele X: TCCAGCAGTTCCCCAGAGGTCT Allele Y: CAGCAGTTCCCCAGAGGTCC	CTTTACAATTCCAAGGCTCTTTCTTGCAA
	S5_3387558	5	3387558	GTTACACTTAGACCCTTGTCATTAAACATTACTGAGGCTGCAGTTGAAGTGTAAACAACTCTTTTCACTGTCTTTGATTCCAAGCATGTCCTTAATATCC[C/T]TAGCAATCCATCCACGGCTATTTTGGTCACACTTTCTTCAGGTTCTTTTCCATAATGATCCATTGGGTGAGCCACATGGAAGACTCCAATACAACCTTCA	T	C	Allele X: TGATTCCAAGCATGTCCTTAATATCCC Allele Y: TTGATTCCAAGCATGTCCTTAATATCCT	CCAAAATAGCCGTGGATGGATTGCTA
	S8_25598183	8	25598183	TAAATTCTGACTGTCTTGGCATGACTGTCCAGGTAGTCCCCGAAAATGAGAATGCTGCTCTCTACTCCACTCATTCATTCAAGATTTTGTTCAAGGAAGG[G/T]GGTTGTGGAACCTTCATTCCGCTCTTTTTCAACTTGCTCTCTTCAGTAAGGCAATACAATCAGCAACAAACCTCTGGAATGGGGCCCCAGATGAACCCTT	T	G	Allele X: CGGAATGAAGGTTCCACAACCC Allele Y: GCGGAATGAAGGTTCCACAACCA	CTCTCTACTCCACTCATTCATTCAAGATT
Dry matter content	S1_24197219	1	24197219	GATGTAGGCATGTTACATATAAGGGCTACATACACATTAGCAGCTAAAATGAGACCCGGATACCGAGCAATGCCATCAATTGAGAGATGAACTCAGGGTG[C/T]CCTGGCCATGCAGCTCCAGTAACCAAATTTTCATGAGTGTAGCAACGATGTATTGGATCAGGTTCTAGCCATGTTGCCCCAGCCAAGACCACGTTAATCT	C	T	Allele X: AATTGAGAGATGAACTCAGGGTGC Allele Y: TCAATTGAGAGATGAACTCAGGGTGT	TCATGAAAATTTGGTTACTGGAGCTGCAT
	S6_20589894	6	20589894	ATTGATGATTTTTTATTCATGATATGTAGCTATCAAAGTTACTCAGCAATGTCCTTGTTTTAGCCATGCTAGCAGCATGTTTTGTTGCGACAACAGTTGG[A/G]AGTTGTAT	G	A	Allele X: CATGTTTTGTTGCGACAACAGTTGGA Allele Y: ATGTTTTGTTGCGACAACAGTTGGG	CTGCCCAATGATATTCTGCATACAAGATA

(Continued)

TABLE 1 Continued

Traits	SNP name*	Chr	Position (bp)	Flanking sequences	Favorable allele	Unfavorable allele	Forward primer (Allele X and Allele Y)	Common Reverse Primer
				GAATATTGTTTTATCTTGTA TGCAGAATATCATTGGGCAGGAAGCAG GGAAAAGCGTGATTGAGGAATATTTAC GTCGTAGGGGTCACTCAG				
	S12_5524524	12	5524524	TGAATTATTTTAACTCTTTGATTGCTTC GCCAGTGCCTGGTCTCCAGAATGTGTG TGTTGCTTTGGTTTGTAGTTCCAAAGG TGAGCTGTGGCAATTTTA[T/C]TGCAGC CCCACTGGCATTAGACGCAGT AAATTATATCAGGACGAAGTAAGTTCA TCCTTCAAAGGAAATGATAATGGTCAA TTTGTGGGGAGCAAAGGTT	C	T	Allele X: TCTAATGCCAGTGGGGCTGCAA Allele Y: TCTAATGCCAGTGGGGCTGCAG	GTTCCAAAGGTGAGCTGTGGCAATT

*SNP marker position in base pairs (bp) is based on the cassava reference genome v6.1 (Bredeson et al., 2016).

was carried out in Hilo, Hawaii (19°38'24.57"N 155°4'57.76"W; 204m above sea level), which has a mild tropical climate that is suitable for cassava survival as well as prolific flowering. The objective of developing the population was to enhance provitamin A biofortification by introgressing a new source of novel alleles for Africa and to develop germplasm incorporating resistance to cassava mosaic disease, high content of provitamin A and starch, and tolerance to acid soils and drought for Latin America. Like the breeding population, a SN evaluation trial was established in Ibadan for 5,608 genotypes planted at a spacing of 1 m x 0.25 m. The mean family size was 16, ranging from 1 to 165 clones. The trial was harvested 10 months after planting and approximately 14% of the genotypes (790) were advanced to CET at Ikenne, Nigeria (6°52′ N 3°42′ E; 61 m above sea level) based on vigor alone.

2.3.2 Field experiment and phenotyping of cassava storage roots for carotenoid and dry matter contents

Genotypes at the first CET were used for the validation study. A CET was preferred because of the large size (typically several hundred) and diversity of most of the traits. The trials were laid out in an augmented design to accommodate a large number of entries. Each genotype was planted at a spacing of 1 m between rows and 0.5 m within rows. For the breeding population, the experiment comprised of 58 to 60 plots per 30 sub-blocks with five checks (IITA-TMS-IBA00070, IITA-TMS-IBA30572, TMEB419, IITA-TMS-IBA982101, IITA-TMS-IBA980581) randomly assigned to each sub-block. This trial was planted in June 2018 and harvested in June 2019. The pre-breeding population trial carried out between October 2018 and October 2019 consisted of 900 plots (50 plots per 18 sub-blocks) with four checks (TMEB419, IITA-TMS-IBA30572, IITA-TMS-IBA070593, and IITA-TMS-IBA000070) in each block. All field management practices were performed according to the technical recommendations and standard agricultural practices for cassava (Abass et al., 2014; Atser et al., 2017).

Direct estimation of total carotenoid content using laboratory extraction followed by spectrophotometry and HPLC is not only expensive but also has low throughput for routine germplasm screening, particularly at the early stages of breeding selection. Due to a large number of genotypes in this study, we used two color-based methods to assess the variation among the cassava genotypes for carotenoid content. Utilization of color intensity as a proxy for the carotenoids content in cassava is justified because of the well-established linear relationship between root yellowness and total carotenoids content (Pearson's coefficient, r, ranges from 0.81 to 0.84) (Iglesias et al., 1997; Chávez et al., 2005; Marín Colorado et al., 2009; Sánchez et al., 2014; Esuma et al., 2016) as well as with total beta-carotene (Udoh et al., 2017). Moreover, 80 to 90% of total carotenoid content in cassava is provitamin A compared to other crops, making color-based assessment a good proxy for estimating not only total carotenoids content but also total β-carotene content (Wong et al., 2004; Ceballos et al., 2017; Jaramillo et al., 2018). In maize, kernel color is not correlated with the primary carotenoid of interest, that is, β-carotene, which has the highest pro-vitamin A activity due to the presence of other carotenoids such as β-cryptoxanthin, zeaxanthin, and lutein (Wong et al., 2004).

The first method is a standard visual assessment of the yellowness of root parenchyma using a color chart with a scale ranging from 1 (white root) to 7 (orange root) (Supplementary Figure 1). The second method is a surface color measurement using a CR-410 chromameter (Konica Minolta). The chromameter's three-dimensional color space defined by l*, a*, and b* coordinates provides a more objective and precise assessment of surface color and its intensity. The Commission Internationale de l'Éclairage (CIELAB) l* coordinate value represents sample lightness ranging from 0 (black) to 100 (diffuse white). The a* values represent either red (positive coordinate values) or green (negative coordinate values). Of importance in our study is the b* coordinate, whose positive values measure the degree of yellowness and therefore provide an indirect estimate of carotenoid content.

For the chromameter color measurements, eight roots per plot were peeled, washed, grated, and thoroughly mixed. A subsample was transferred into a transparent sampling bag (Whirl-Pak$^{\text{TM}}$) and scanned at four independent positions. The CR-410 chromameter was calibrated each day using a white ceramic and illuminant D65 was used as a source of light.

Root dry matter content was assessed using the oven-drying method. Eight fully developed roots were randomly selected from each plot, peeled, washed, grated, and thoroughly mixed. For each sample, 100 g was weighed and oven-dried for 72 h at 80°C. The dry samples were then weighed, and the dry matter content was expressed as the percentage of dry weight relative to fresh weight.

2.3.3 Genotyping

Young leaves were sampled three months after planting from the evaluation plots. Three 6mm diameter leaf discs were obtained from each genotype into 96-well plates on ice, and freeze-dried for at least 72 hours. The samples were shipped to a genotyping service provider (Intertek, Sweden) for automated DNA extraction and SNP genotyping using four markers linked to increased carotenoid content and three markers linked to increased dry matter content (Table 1) using the KASP assay. Two blank controls were included in each plate during genotyping.

The KASP assay protocol is provided in the KASP manual (LGC Genomics, 2013). In brief, genotyping was carried out using the high-throughput PCR SNPline workflow using 1 μL reaction volume in 1536-well PCR plates. The KASP genotyping reaction mix is comprised of three components: (i) sample DNA (10 ng); (ii) marker assay mix consisting of target-specific primers; and (iii) KASP-TF^{TM} Master Mix containing two universal fluorescence resonant energy transfer cassettes (FAM and HEX), passive reference dye (ROX^{TM}), Taq polymerase, free nucleotides, and $MgCl_2$ in an optimized buffer solution. The SNP assay mix is specific to each marker and consists of two kompetitive allele-specific forward primers and one common reverse primer (Table 1). After PCR, the plates were fluorescently read, and allele calls were made using KRAKEN^{TM} software.

2.3.4 Data analysis

2.3.4.1 Phenotypic data analysis

A linear mixed model was implemented using restricted maximum likelihood (REML) to estimate the best linear unbiased predictions (BLUPs) for each genotype in the CETs of breeding and pre-breeding populations. The model was fitted using the *asreml* package (Butler, 2020) in R software version 4.0.3 (R Development Core Team, 2020). The mathematical model used for the incomplete block design analysis is represented as follows:

$$Y_{ijk} = \mu + G_i + R_k + B_{jk} + e_{ijk}$$

where Y_{ijk} is the vector of phenotype data of the i^{th} genotype of the j^{th} block nested into the k^{th} replication, μ is the overall mean, G_i is the effect of the i^{th} genotype, R_k is the effect of the k^{th} replication, B_{jk} is the effect of the j^{th} block nested into the k^{th} replication, and e_{ijk} is the residual, modeled as a sum of measurement error and a spatially dependent random process. A first-order auto-regressive process in both row and column directions was used for the spatial trend (Gilmour et al). All effects except μ were assumed to be random.

Broad-sense heritability was calculated as:

$$H^2 = \sigma_g^2 / \left(\sigma_g^2 + \sigma_e^2\right)$$

where H^2 is the broad-sense heritability; σ_g^2 and σ_e^2 are the variance components for the genotype effect and the residual error, respectively.

Pairwise correlation analysis of the traits using the BLUP estimates was determined using the *corr.test* function in the psych package (R Development Core Team, 2020).

2.3.4.2 Technical and biological validation of kompetitive allele-specific PCR markers

Technical performance metrics used to assess the robustness of markers include SNP call rate and call clarity. Call rate is the proportion of samples with non-missing genotype calls. Call clarity is defined by the ease of assigning samples to a genotype class based on their position on a fluorescence cluster Cartesian plot. The tighter and more distinct the cluster, the easier and more consistent it is to call the respective genotype class, namely homozygous for either allele 1 or 2 or heterozygous in the case of biallelic SNPs and a diploid genome.

Biological validation of the converted markers was assessed using three complementary approaches. First, the allele substitution effect was visualized using boxplots, and the difference in carotenoid and dry matter content BLUP values among the genotypic classes at each marker locus was assessed using a pairwise t-test. Second, the predictive ability of the SNP markers was estimated using a multiple linear regression model. As shown in the linear model below, marker alleles and the observed phenotypes were considered as the independent and response variables, respectively.

$$Y = \mu + \mu_1 + \mu_2 + \ldots + \mu_n + e$$

where: Y = phenotypic observations of traits, μ = overall mean of the population, μ_1 , μ_2 , μ_n = marker effects, e = residual value.

Bootstrap resampling was carried out to obtain robust estimates of model parameters, specifically the magnitude and confidence intervals of the allele-substitution effects for the markers associated with the two traits (Davison and Hinkley, 1997). The *reg_intervals* function in the tidymodels R package (Kuhn and Wickham, 2020) was used to generate 1000 bootstrap resamples and fit the multiple linear regression model on each one.

Finally, a 5-fold cross-validation analysis repeated 10 times was carried out to obtain marker performance metrics including predictive accuracy (R^2), root mean square error (RMSE, the square root of the mean squared difference between observed and predicted trait values), and mean absolute error (MAE, the average absolute difference between the predictions made by the model and the actual observations). To achieve this, the breeding and pre-breeding population data were partitioned into training and testing sets in a 3:1 ratio with a stratification based on the target traits (chromameter b* value or dry matter content). The regression model developed in the training set was used to predict the trait values in the hold-out testing set. All model training and cross-validation analyses were implemented in the R caret package (Kuhn, 2008).

3 Results

3.1 Phenotypic variation for root yellowness and dry matter content

Out of the evaluated clones, 81% of the breeding population and 52% of the pre-breeding population had white storage roots, while the remaining showed a range of yellow color (visual score of between 2 and 5), suggesting varying levels of carotenoid content (Figure 2). The average visual score of root yellowness was 1.30 (standard deviation (sd) = 0.72) in the breeding population and 1.74 (sd = 0.95) in the pre-breeding population. The chromameter b* values showed a bi-modal distribution in the two populations (Figure 2). The first peak (b*values from 11 to 22) is associated with clones that produced

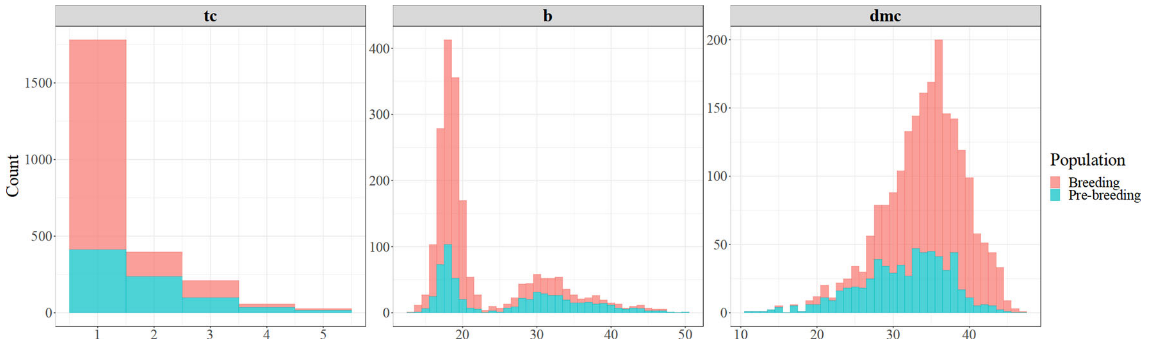


FIGURE 2
Frequency distribution of cassava genotypes for root yellowness intensity (color-chart [tc] and chromameter [b*]) and dry matter content (dmc %) in the breeding and pre-breeding populations.

white roots, while the second peak (b* values from 22 to 50) is associated with the variations among clones with yellow roots. The average chromameter measures of yellow color intensity were 21.0 (sd = 6.12) and 26.2 (sd = 8.82) for breeding and pre-breeding populations, respectively. The dry matter content of the clones evaluated in the two populations was normally distributed (Figure 2), ranging from 11.2 to 47.4, with averages of 31.5 (sd = 5.92) in the pre-breeding population and 35.1 (sd = 4.80) in the breeding population.

The broad-sense heritability of the visual assessment from the color chart and chromameter values were 0.87 and 0.88, respectively, for the breeding population, and 0.81 and 0.93, respectively for the pre-breeding population (Table 2). The heritability estimate for dry matter content in the pre-breeding population (0.61) was lower than that of the breeding population (0.70) (Table 2).

The two measures of root yellowness intensity; visual assessment and chromameter b* value were significantly and positively correlated (~0.90) in the two populations suggesting that visual scoring is a good proxy for yellow-color intensity. Significant negative correlations ranging from -0.27 to -0.20 were observed between root yellowness and dry matter content in the two populations. However, a lower magnitude of correlation coefficient was observed between visual assessment and dry matter content (-0.20) as well as between chromameter b* value and dry matter content (-0.23) in the pre-breeding population.

3.2 Technical validation of kompetitive allele-specific PCR assays

3.2.1 SNP call rate, call clarity, and genotypic frequencies

All markers were successfully converted to allele-specific KASP assays. The call rate and clarity were high for a wide range of DNA dilution levels tested during marker development, indicating that the assays are robust and suitable for routine use (Supplementary Figure 2). The overall call rate was above 98% for all the markers in the two populations genotyped (mean = 99%, sd = 0.53) (Supplementary Table 1). As expected, three distinct clusters were observed for all the SNPs except for marker S5_3387558 where the frequency of cluster TT was very low (Supplementary Figure 3).

Allelic and genotypic frequencies of the markers are presented in Supplementary Figures 4, 5, respectively. The favorable alleles across all the carotenoid-linked markers were more common in the pre-breeding population (ranging from 11 to 34%) compared to the breeding population (ranging from 3 to 11%) (Supplementary Figure 4). The favorable allele A at marker S1_24155522 had a frequency of 34% and 11% in the pre-breeding and breeding populations, respectively (Supplementary Figure 4). More than 15% of the individuals were homozygous for allele A at this marker in the pre-breeding population (Supplementary Figure 5).

TABLE 2 Broad-sense heritability calculated on a mean plot basis for root visual assessment, chromameter value, and dry matter content in the two cassava populations.

Traits	Breeding population			Pre-breeding population		
	σ_g^2	σ_e^2	H^2	σ_g^2	σ_e^2	H^2
Visual assessment	0.500	0.072	0.87	0.637	0.145	0.81
Chromameter b* value	34.641	4.572	0.88	65.676	4.879	0.93
Dry matter content	16.231	6.831	0.70	21.806	14.039	0.61

σ_g^2 is the clonal genotypic variance, σ_e^2 is the residual variance, and H^2 is the broad-sense heritability.

The percentage was much lower in the breeding population with only 2.3% of the individuals fixed for the same allele. In the two populations, between 0.4 to 7.3% of the individuals were fixed for the favorable alleles at the three remaining markers suggesting an opportunity to use these markers to increase their frequencies in the population (Supplementary Figure 5). For dry matter content, the favorable alleles at the linked SNPs occurred at intermediate to high frequencies ranging from 28 to 76% in both populations (Supplementary Figure 4). The percentage of individuals that were fixed for the favorable alleles was higher in the breeding than in the pre-breeding population for this trait (Supplementary Figure 5). About 27 to 53% of the individuals in the pre-breeding population were fixed for the unfavorable alleles (Supplementary Figure 5).

3.2.2 Biological validation

3.2.2.1 Allelic substitution effects on carotenoid and dry matter contents

Significant pairwise differences between genotypic classes at all the markers associated with carotenoid content were observed (Figure 3). Most of the markers displayed an additive mode of action with individuals carrying two copies of the favorable alleles having a higher intensity of root yellowness (b*) than those with only one copy while those that are fixed for non-favorable alleles had white roots. For instance, the mean b* values for genotype classes AA, CA, and CC for marker S1_24155522 were 38.53 ± 2.85, 31.64 ± 3.89, and 18.37 ± 2.48, respectively in the pre-breeding population (Figure 3B).

The genotype classes at the dry matter content-linked markers were not as differentiated as those for carotenoid content (Figure 4). Nonetheless, significant differences were observed among the genotypes at marker S6_20589894 in the two populations. In the pre-breeding population, there was no significant difference among CC, CT, and TT at marker S12_5524524 (Figure 4).

3.2.2.2 Marker-trait regression, confidence intervals, and models' predictive performances

The estimates of marker-trait regression parameters from bootstrap resampling analysis for the two traits are presented in Figures 5 and 6. The regression model with all the four markers for carotenoid variation produced R^2 values of 0.85 in the breeding population and 0.91 in the pre-breeding population. However, in a subset of the breeding and pre-breeding populations consisting of only genotypes with yellow roots, the R^2 values decreased to 0.46 and 0.53, respectively. SNP S1_24155522 had the strongest effect on variation in root yellowness. The effect size of having a single copy of a favorable allele (A) on the increase in root yellowness intensity (chromameter b* value) was 10.8 and 12.1 in the breeding and pre-breeding populations, respectively. Having two copies of the same

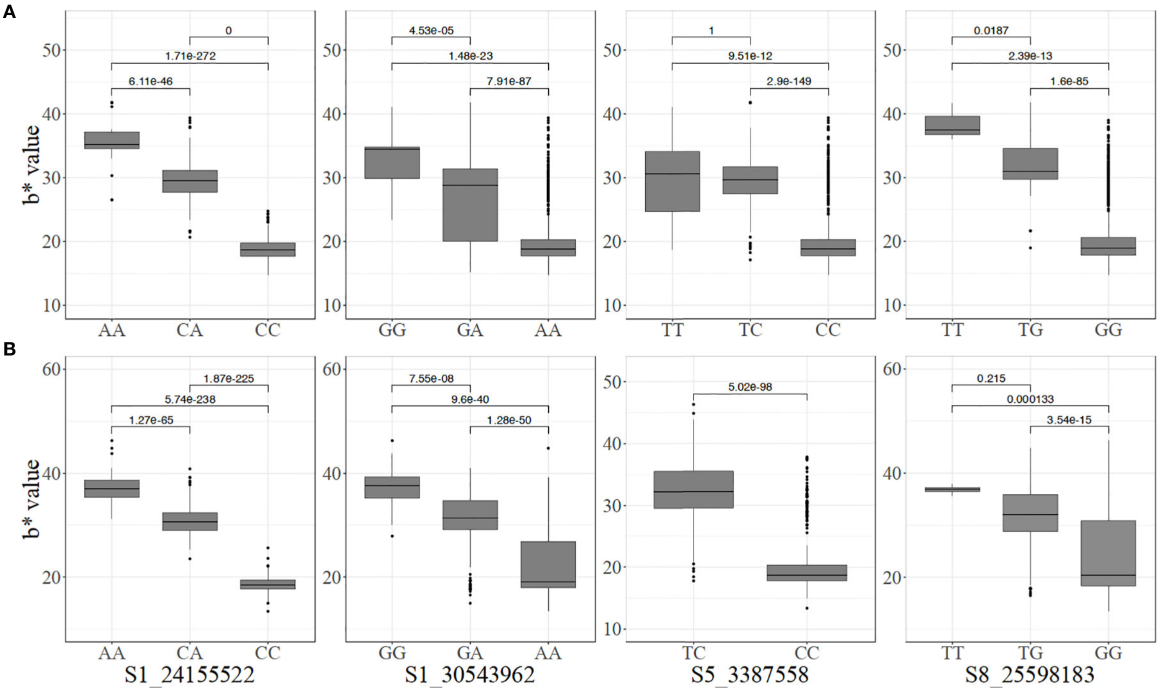


FIGURE 3
Allelic substitution effects of the markers associated with increased carotenoid content in the **(A)** breeding, and **(B)** pre-breeding populations (For marker S5_3387558, the mean and standard deviation cannot be estimated because one genotype had TT).

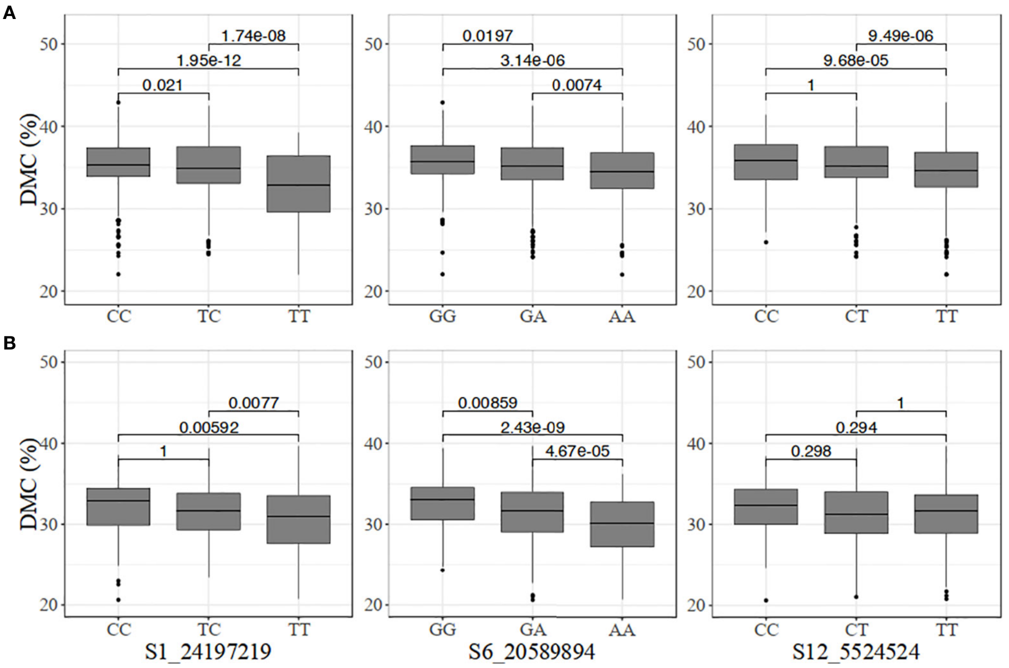


FIGURE 4
Allelic substitution effects of the markers associated with increased dry matter content (DMC) in the **(A)** breeding, and **(B)** pre-breeding populations.

allele resulted in an even larger effect size of 15.5 and 17.8, respectively, in the two populations. The confidence intervals of these marker genotypes were narrow, indicating higher precision of the marker prediction. After controlling for the major locus (S1_24155522) effect in the two populations, the other three markers had a low to moderate effect on the trait (Supplementary Figure 6). The effect sizes of the minor SNPs were more significant in the breeding compared to the pre-breeding population, particularly for markers S5_3387758 and S8_25598183.

The regression model with all three markers for dry matter content produced low R^2 values of 0.06 in the breeding and 0.09 in the pre-breeding population. Having two copies of favorable alleles across all SNPs was associated with an increase in dry matter content percentage from between 1.01 and 2.50 percentage units in the breeding population. A similar direction of effects was observed in the pre-breeding population except for marker S12_5524524 which did not contribute to the multiple regression model. A notable observation is a reversal in the effects of markers S1_24197219 and S6_20589894 across the two populations, suggesting a QTL by genetic background interaction.

The predictive accuracy of the carotenoid markers from the cross-validation regression analysis ranged from 0.84 to 0.91 with a mean of 0.87. In the pre-breeding population, the value was higher and approximately 0.90 in the training and testing sets (Table 3, Supplementary Figure 7). However, low predictive accuracy values were obtained for dry matter content-linked markers in the breeding population (0.07 for the training set and 0.05 for the testing set) and pre-breeding population (0.08 for the training set and 0.07 for the testing set) (Table 3, Supplementary Figure 7). In the breeding population, RMSE and MAE values for carotenoid markers were 1.88 and 1.43, respectively, in the training set, and 2.03 and 1.52, respectively, in the testing set (Table 3). The values of RMSE and MAE were 2.31 and 1.71, respectively, in the training set, and 2.35 and 1.68 in the testing set of the pre-breeding population. These values were higher for dry matter content-markers in both populations compared to those of carotenoid content-markers. The use of RMSE and MAE is very common in model evaluation, and they are good measures of prediction accuracy.

4 Discussion

The present study focused on the development and validation of markers for carotenoids and dry matter contents, two traits that are of primary importance to cassava breeding programs worldwide (Sánchez et al., 2006; Okechukwu and Dixon, 2008; Bouis et al., 2011; Saltzman et al., 2013; Talsma et al., 2013). Similar to our observations, several studies that used diverse cassava germplasm, particularly from Africa have reported that dry matter content and carotenoid content parameters such as total

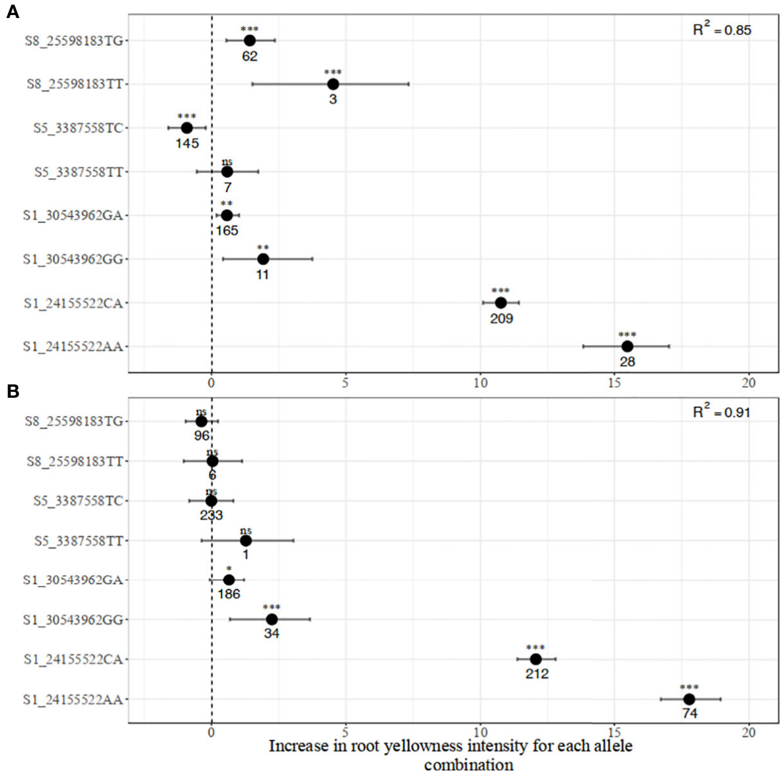


FIGURE 5
Distribution of marker allelic effects associated with increased carotenoid content in **(A)** breeding, and **(B)** pre-breeding populations.

carotenoid content, root yellowness intensity, and visual assessment of storage roots are negatively correlated with r-values ranging from 0.1 to 0.6 (Marín Colorado et al., 2009; Akinwale et al., 2010; Njoku et al., 2015; Esuma et al., 2016; Rabbi et al., 2017). On the contrary, these traits are independent in Latin American cassava populations (Ceballos et al., 2013; Sánchez et al., 2014). Although the selection of genotypes based on high intensities of root yellowness at the early stage of the breeding scheme saves time and costs associated with carotenoid quantification, it would indirectly select for lower dry matter content (Sánchez et al., 2014).

As part of the breeders' toolbox for MAS, markers validated can be used to select for the study traits simultaneously and are expected to address the challenges associated with vitamin A deficiency and higher demand for varieties with higher dry matter content. Vitamin A deficiency is a widespread nutritional public health problem in sub-Saharan Africa, with women and children being the most affected (Gegios et al., 2010; Stephenson et al., 2010). Breeding of clones with enhanced carotenoid levels is one of the most cost-effective and sustainable approaches to helping the communities burdened by vitamin A deficiency (Pfeiffer and McClafferty, 2007; Bouis et al., 2011; Talsma et al., 2013). While we have explored the performance of the markers in the IITA pre-breeding and breeding populations, these assays should have wide application in other breeding programs where the QTLs are present and are linked to the same SNP alleles. More importantly, these markers can be used for rapid mobilization of the favorable alleles in new populations developed using parents that are known to carry the associated trait alleles.

Trait discovery in cassava has been an active area of research with the advent of genome-wide SNP markers from genotyping-by-sequencing (Wolfe et al., 2016; Esuma, 2016; Rabbi et al., 2017; Udoh et al., 2017; Ikeogu et al., 2019; Rabbi et al., 2020). However, these trait discoveries have not been translated into deployable assays, obscuring their utility in MAS. Here, we have provided a framework for translating the outputs from genetic mapping to a set of easy-to-use, robust, and predictive allele-specific uniplex assays. The framework includes both technical and biological validation of the assays in a range of diverse germplasm to ascertain the relevance of the markers for predicting the trait values in independent populations. The KASP SNP platform was chosen due to its amenability for genotyping of any combination of individual samples and marker assays, and ease of automation to achieve high-

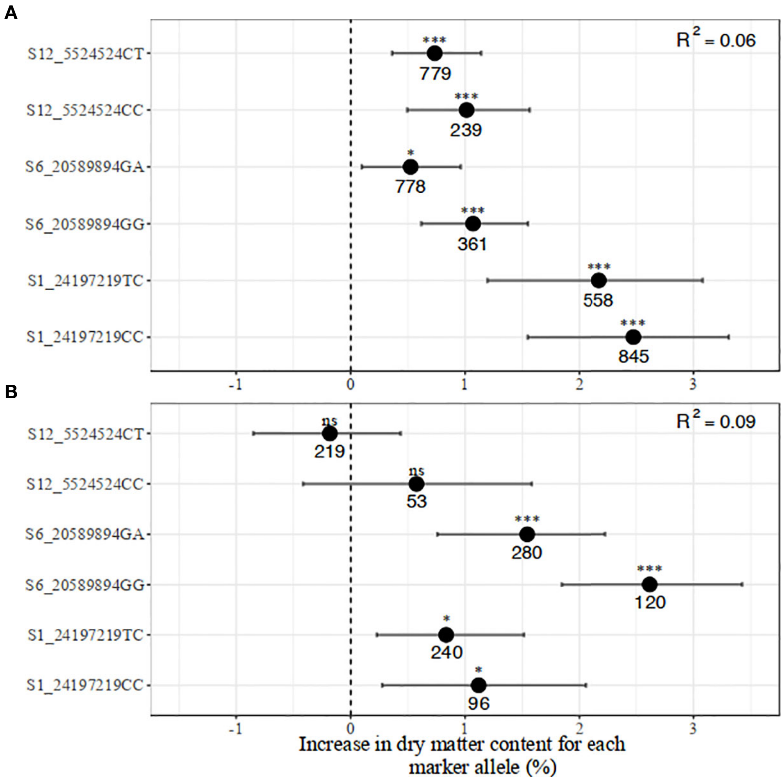


FIGURE 6
Distribution of the marker allelic effects associated with increased dry matter content in **(A)** breeding, and **(B)** pre-breeding populations.

throughput population screening (Semagn et al., 2014; Ogbonna et al., 2020; Ige et al., 2021). The designed SNP assays were found to work under a wide range of DNA concentrations. Even though the tightness of the cluster plots differed between the standard and low DNA concentrations, they were sufficiently distinct to allow for a high genotype call rate and call clarity. This suggests that the assays are expected to work under diverse DNA concentrations and most likely from different sample preparation methods, including fresh, frozen, lyophilized, or oven-dried (Semagn et al., 2014).

The best way to measure the predictive ability of a model is to test it on a dataset that is independent of the data used to train the model (Wani et al., 2018). The *k*-fold cross-validation, where the original dataset is randomly partitioned into equally sized *k*-subsets (a single subset is retained as the validation data for testing the model, and the remaining *k* - 1 subsets are used as training data), is one of the most commonly used cross-validation methods (Refaeilzadeh et al., 2009; Mathew et al., 2015). It is routinely used to assess genomic prediction accuracies (Okeke et al., 2017; de Andrade et al., 2019;

TABLE 3 Prediction performance metrics of the markers associated with increased carotenoid and dry matter contents in the training and testing sets of the breeding and pre-breeding populations.

Traits	Populations		N	R^2	RMSE	MAE
Chromameter b* value (Carotenoid content)	Breeding	Training set	1030	0.84	1.88	1.43
		Testing set	345	0.84	2.03	1.52
	Pre-breeding	Training set	396	0.91	2.31	1.71
		Testing set	133	0.90	2.35	1.68
Dry matter content (%)	Breeding	Training set	1102	0.07	3.13	2.48
		Testing set	368	0.05	3.20	2.48
	Pre-breeding	Training set	402	0.08	3.70	3.00
		Testing set	136	0.07	3.07	2.53

N, Number of observations; R^2, Prediction accuracy; RMSE, Root mean square error; MAE, Mean absolute error.

Phumichai et al., 2022). To our knowledge, this is the first study to use this metric for marker validation in cassava. In the present study, the performance of the regression model in an independent data set, that is, the testing set in terms of predictive accuracy for chromameter b* values were 0.84 in the breeding population and 0.90 in the pre-breeding population. These values are quite similar to those obtained in the training sets, suggesting that the models developed are stable and reliable. The low values of RMSE and MAE recorded in the breeding population compared to that of the pre-breeding population indicated that the markers are more accurate in predicting the carotenoid content in the breeding population. Both measures of cross-validation accuracy for this trait suggest that the designed assays can be deployed for routine use in breeding pipelines with carotenoid biofortification as a breeding goal. On the other hand, the predictive accuracy of the dry matter content markers (mean = 0.07) across populations was lower than the values obtained for carotenoid content markers. This could be due to the quantitative nature of dry matter content (Kawano et al., 1987). In the discovery population (Rabbi et al., 2020), also reported low predictive ability ($R^2 <$ 0.11) of these markers.

Moreover, for both traits, we used a bootstrapping regression approach to provide robust estimates of allele substitution effects and their confidence intervals (Fox and Weisberg, 2018). The multiple regression analysis of carotenoid content markers revealed that marker S1_24155522 was the main driver in carotenoid accumulation while the other markers played additional but minor roles. This result is consistent with earlier observations that the *PSY2* gene, which hosts marker S1_24155522 is a key rate-limiting step in the carotenoid pathway in cassava (Welsch et al., 2010; Rabbi et al., 2020). In a candidate gene-based association study, Udoh et al. (2017) reported that total carotenoid content and β-carotene were significantly associated with this marker, which occurs at position 572 of the *PSY2* gene. Indeed, the previously identified SNPs from other candidate genes such as *lcyE, lcyB*, and *crtRB* were hardly significantly associated with the trait (Udoh et al., 2017). On the other hand, markers S1_24197219 and S6_20589894 had small but significant effects on dry matter content in both populations, while marker S12_5524524 showed an effect in the pre-breeding population. Marker S6_20589894 was reported to occur close to the gene Manes.06G103600 (Bidirectional sugar transporter Sweet4-Related) which mediates fructose transport across the tonoplast of roots (Rabbi et al., 2020).

While we have assessed the performance of selected markers across the two diverse populations, we acknowledge that these markers may be tagging only a subset of major loci underlying the studied traits, particularly dry matter content. Ongoing and future GWAS and biparental QTL mapping studies will likely uncover additional QTLs. Such markers can be validated using the framework provided in this study and incorporated into the breeders' toolset, thus increasing the accuracy of predicting these traits. Moreover, other traits that are of importance for which major associations have recently been reported but not converted to marker assays include cassava green mite (Rabbi et al., 2020), cassava brown streak disease (Kayondo et al., 2018) and root mealiness (Uchendu et al., 2021). A major caveat of our study is the use of single-marker assays to tag each major locus for the two traits. The top SNPs at these loci are expected to be tightly linked to the causal allele based on the large GWAS population used in the discovery, with more than 5000 individuals genotyped at more than 100K genome-wide positions. However, factors such as independent emergence or evolution of favorable alleles at specific genes and nearby SNP can result in non-perfect association, hence resulting in false-positive and false-negative. This and other limitations of single marker analysis can be addressed by a haplotype-based approach through, for example, amplicon sequencing (AmpSeq) of targeted genomic regions (Yang et al., 2016). Further work is required to establish the viability of Amplicon Sequencing as a platform for haplotype-based MAS in cassava.

Data availability statement

The data presented in this study can be found in an online repository. The names of the repository and accession number (s) can be found at: https://cassavabase.org/ftp/manuscripts/Ige_et_al_2022/Ige_et_al_2022.csv.

Author contributions

ADI, BO, IYR, and PK designed the study. SM-W and JN developed the pre-breeding population. ADI and RU performed the experiment. ADI, GB, ID, and IYR analyzed the data. ADI drafted the manuscript. BO, GB, ISK, EGNM, CE, EP, PK, HC, and IYR revised the manuscript. All authors have read, edited, and approved the current version of the manuscript.

Funding

This work was supported by the UK's Foreign, Commonwealth, and Development Office (FCDO) and the Bill & Melinda Gates Foundation (NextGen Cassava: Grant INV-007637) as well as the Roots, Tubers, and Bananas Program of the CGIAR. The first author was supported by the African Union to pursue a Ph.D. in Plant Breeding at Pan African University Life and Earth Sciences Institute (including Health and Agriculture), University of Ibadan, Ibadan, Nigeria.

Acknowledgments

The authors gratefully acknowledge the technical support provided by the staff of the Cassava Breeding Program at the International Institute of Tropical Agriculture, Ibadan, Nigeria during the study. We also thank the Excellence-in-Breeding Platform of the CGIAR for supporting the conversion of the trait-linked markers to allele-specific PCR assays.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1016170/full#supplementary-material

References

Abass, A. B., Towo, E., Mukuka, I., Okechukwu, R. U., Ranaivoson, R., Tarawali, G., et al. (2014). *Growing cassava: a training manual from production to postharvest* (Ibadan, Nigeria: International Institute of Tropical Agriculture).

Akano, O., Dixon, O., Mba, C., Barrera, E., and Fregene, M. (2002). Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *Theor. Appl. Genet.* 105, 521–525. doi: 10.1007/s00122-002-0891-7

Akinwale, M. G., Aladesanwa, R. D., Akinyele, B. O., Dixon, A. G. O., and Odiyi, A. C. (2010). Inheritance of carotene in cassava (Manihot esculenta crantz). *Int. J. Genet. Mol. Biol.* 2, 198–201. doi: 10.5897/IJGMB.9000034

Andersson, M. S., Saltzman, A., Virk, P. S., and Pfeiffer, W. H. (2017). Progress update: crop development of biofortified staple food crops under HarvestPlus. *Afr J. Food Agric. Nutr. Dev.* 17, 11905–11935. doi: 10.18697/ajfand.78.HarvestPlus05

Anyanwu, C. N., Ibeto, C. N., Ezeoha, S. L., and Ogbuagu, N. J. (2015). Sustainability of cassava (Manihot esculenta crantz) as industrial feedstock, energy and food crop in Nigeria. *Renew Energy* 81, 745–752. doi: 10.1016/j.renene.2015.03.075

Atser, G., Dixon, A., Ekeleme, F., Chikoye, D., Dashiell, K. E., Ayankanmi, T. G., et al. (2017). *The ABC of weed management in cassava production in Nigeria: a training manual* (Ibadan, Nigeria: International Institute of Tropical Agriculture).

Balagopalan, C. (2002). *Cassava utilization in food, feed and industry*. CABI

Balyejusa, K. E., Rönnberg-Wästljung, A.-C., Egwang, T., Gullberg, U., Fregene, M., and Westerbergh, A. (2007). Quantitative trait loci controlling cyanogenic glucoside and dry matter content in cassava (Manihot esculenta crantz) roots. *Hereditas* 144, 129–136. doi: 10.1111/j.2007.0018-0661.01975.x

Bechoff, A., Tomlins, K., Fliedel, G., Lopez-lavalle, L. A. B., Westby, A., Hershey, C., et al. (2018). Cassava traits and end-user preference: Relating traits to consumer liking, sensory perception, and genetics. *Crit. Rev. Food Sci. Nutr.* 58, 547–567. doi: 10.1080/10408398.2016.1202888

Bouis, H. E., Hotz, C., McClafferty, B., Meenakshi, J. V., and Pfeiffer, W. H. (2011). Biofortification: A new tool to reduce micronutrient malnutrition. *Food Nutr. Bull.* 32, S31–S40. doi: 10.1177/15648265110321S105

Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., et al. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* 34, 562–570. doi: 10.1038/nbt.3535

Burns, A., Gleadow, R., Cliff, J., Zacarias, A., and Cavagnaro, T. (2010). Cassava: The drought, war and famine crop in a changing world. *Sustainability* 2, 3572–3607. doi: 10.3390/su2113572

Butler, D. (2020). "Asreml: fits the linear mixed model," in *R package version 4.1*, (UK: VSNi) vol. 0.143.

Ceballos, H., Davrieux, F., Talsma, E. F., Belalcazar, J., Chavarriaga, P., and Andersson, M. S. (2017). *Carotenoids in cassava roots* (UK: IntechOpen). doi: 10.5772/intechopen.68279

Ceballos, H., Iglesias, C. A., Perez, J. C., and Dixon, A. G. (2004). Cassava breeding: opportunities and challenges. *Plant Mol. Biol.* 56, 503–516. doi: 10.1007/s11103-004-5010-5

Ceballos, H., Kawuki, R. S., Gracen, V. E., Yencho, G. C., and Hershey, C. H. (2015). Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. *Theor. Appl. Genet.* 128, 1647–1667. doi: 10.1007/s00122-015-2555-4

Ceballos, H., Kulakow, P., and Hershey, C. (2012). Cassava breeding: Current status, bottlenecks and the potential of biotechnology tools. *Trop. Plant Biol.* 5, 73–87. doi: 10.1007/s12042-012-9094-9

Ceballos, H., Morante, N., Sánchez, T., Ortiz, D., Aragón, I., Chávez, A. L., et al. (2013). Rapid cycling recurrent selection for increased carotenoids content in cassava roots. *Crop Sci.* 53, 2342–2351. doi: 10.2135/cropsci2013.02.0123

Chagné, D., Vanderzande, S., Kirk, C., Profitt, N., Weskett, R., Gardiner, S. E., et al. (2019). Validation of SNP markers for fruit quality and disease resistance loci in apple (Malus $\times$ domestica borkh.) using the OpenArray® platform. *Hortic. Res.* 6, 30. doi: 10.1038/s41438-018-0114-2

Chávez, A. L., Sánchez, T., Jaramillo, G., Jm, B., Echeverry, J., Bolaños, E. A., et al. (2005). Variation of quality traits in cassava roots evaluated in landraces and improved clones. *Euphytica* 143, 125–133. doi: 10.1007/s10681-005-3057-2

Collard, B. C. Y., and Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R Soc. Lond B Biol. Sci.* 363, 557–572. doi: 10.1098/rstb.2007.2170

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap methods and their application* (UK: Cambridge university press).

de Andrade, L. R. B., Sousa, M. B., Oliveira, E. J., de Resende, M. D. V., and Azevedo, C. F. (2019). Cassava yield traits predicted by genomic selection methods. *PloS One* 14, e0224920. doi: 10.1371/journal.pone.0224920

Esuma, W. (2016). *Genetic analysis and genome-wide association mapping of carotenoid and dry matter content in cassava* (South Africa: PhD Thesis. University of the Free State). 66, 627–635

Esuma, W., Herselman, L., Labuschagne, M. T., Ramu, P., Lu, F., Baguma, Y., et al. (2016). Genome-wide association mapping of provitamin a carotenoid content in cassava. *Euphytica* 212, 97–110. doi: 10.1007/s10681-016-1772-5

Esuma, W., Kawuki, R. S., Herselman, L., and Labuschagne, M. T. (2016). Diallel analysis of provitamin a carotenoid and dry matter content in cassava (Manihot esculenta crantz). *Breed Sci.* 15159 (66), 627–635. doi: 10.1270/jsbbs.15159

FAOSTAT (2020). Available at: http://www.fao.org/faostat/en/#data/QC (Accessed January 11, 2021).

Fox, J., and Weisberg, S. (2018). *An r companion to applied regression* (USA: Sage publications).

Fregene, M., Okogbenin, E., Mba, C., Angel, F., Suarez, M. C., Janneth, G., et al. (2001). Genome mapping in cassava improvement: Challenges, achievements and opportunities. *Euphytica* 120, 159–165. doi: 10.1023/A:1017565317940

Gegios, A., Amthor, R., Maziya-Dixon, B., Egesi, C., Mallowa, S., Nungo, R., et al. (2010). Children consuming cassava as a staple food are at risk for inadequate zinc, iron, and vitamin a intake. *Plant Foods Hum. Nutr.* 65, 64–70. doi: 10.1007/s11130-010-0157-5

Gilmour, A. R., Cullis, B. R., and Verbyla, A. P. (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *J. Agric. Biol. Environ. Stat.* 1997, 269–293. doi: 10.2307/1400446

Ige, A. D., Olasanmi, B., Mbanjo, E. G. N., Kayondo, I. S., Parkes, E. Y., Kulakow, P., et al. (2021). Conversion and validation of uniplex SNP markers for selection of resistance to cassava mosaic disease in cassava breeding programs. *Agronomy* 11, 420. doi: 10.3390/agronomy11030420

Iglesias, C., Mayer, J., Chavez, L., and Calle, F. (1997). Genetic potential and stability of carotene content in cassava roots. *Euphytica* 94, 367–373. doi: 10.1023/A:1002962108315

Ikeogu, U. N., Akdemir, D., Wolfe, M. D., Okeke, U. G., Chinedozi, A., Jannink, J.-L., et al. (2019). Genetic correlation, genome-wide association and genomic prediction of portable NIRS predicted carotenoids in cassava roots. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01570

Ilona, P., Bouis, H. E., Palenberg, M., Moursi, M., and Oparinde, A. (2017). Vitamin a cassava in Nigeria: crop development and delivery. *Afr J. Food Agric. Nutr. Dev.* 17, 12000–12025. doi: 10.4314/ajfand.v17i2

Jaramillo, A. M., Londoño, L. F., Orozco, J. C., Patiño, G., Belalcazar, J., Davrieux, F., et al. (2018). A comparison study of five different methods to measure carotenoids in biofortified yellow cassava (Manihot esculenta). *PloS One* 13, e0209702. doi: 10.1371/journal.pone.0209702

Jennings, D. L., and Iglesias, C. (2002). *Cassava: Biology, production and utilization* (Wallingford UK: CABI) 18, 149–166. doi: 10.1079/9780851995243.0149

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

Kawano, K., Fukuda, W., and Cenpukdee, U. (1987). Genetic and environmental effects on dry matter content of cassava Root1. *Crop Sci.* 27, 69–74. doi: 10.2135/cropsci1987.0011183X002700010018x

Kayondo, S. I., Del Carpio, D. P., Lozano, R., Ozimati, A., Wolfe, M., Baguma, Y., et al. (2018). Genome-wide association mapping and genomic prediction for CBSD resistance in manihot esculenta. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-19696-1

Kuhn, M. (2008). Building predictive models in r using the caret package. *J. Stat. Softw* 28, 1–26. doi: 10.18637/jss.v028.i05

Kuhn, M., and Wickham, H. (2020). "Tidymodels: Easily install and load the'Tidymodels' packages," in *R package version 0.1. 0*. Available at: https://tidymodels.tidymodels.org/.

LGC Genomics (2013). *KASP genotyping chemistry user guide and manual* (Teddingt UK: LGC Ltd).

Li, L., Tacke, E., Hofferbert, H.-R., Lübeck, J., Strahwald, J., Draffehn, A. M., et al. (2013). Validation of candidate gene markers for marker-assisted selection of potato cultivars with improved tuber quality. *Theor. Appl. Genet.* 126, 1039–1052. doi: 10.1007/s00122-012-2035-z

Marín Colorado, J. A., Ramírez, H., and Fregene, M. (2009). Genetic mapping and QTL analysis for carotenes in a s1 population of cassava. *Acta Agronómica* 58, 15–21.

Mathew, B., Léon, J., and Sillanpää, M. J. (2015). Integrated nested Laplace approximation inference and cross-validation to tune variance components in estimation of breeding value. *Mol. Breed* 35, 99. doi: 10.1007/s11032-015-0248-y

Morillo, A. C., Morillo, C. Y., and Ceballos, L. H. (2013). Identification of QTLs for carotene content in the genome of cassava (Manihot esculenta crantz) and S1 population validation. *Acta Agronómica* 62, 196–206.

Njoku, D., Gracen, V., Offei, S., Asante, I., Egesi, C., Kulakow, P., et al. (2015). Parent-offspring regression analysis for total carotenoids and some agronomic traits in cassava. *Euphytica* 206, 657–666. doi: 10.1007/s10681-015-1482-4

Ogbonna, A. C., De Andrade, L. R. B., Rabbi, I. Y., Mueller, L. A., De Oliveira, E. J., and Bauchet, G. J. (2020). Genetic architecture and gene mapping of cyanide in cassava (Manihot esculenta crantz.). *BioRxiv*. doi: 10.1101/2020.06.19.159160

Okechukwu, R. U., and 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. *J. Crop Improv* 22, 181–208. doi: 10.1080/15427520802212506

Okeke, U. G., Akdemir, D., Rabbi, I., Kulakow, P., and Jannink, J.-L. (2017). Accuracies of univariate and multivariate genomic prediction models in African cassava. *Genet. Sel Evol.* 49, 88. doi: 10.1186/s12711-017-0361-y

Okogbenin, E., Egesi, C. N., Olasanmi, B., Ogundapo, O., Kahya, S., Hurtado, P., et al. (2012). Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. *Crop Sci.* 52, 2576–2586. doi: 10.2135/cropsci2011.11.0586

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

Pfeiffer, W. H., and McClafferty, B. (2007). HarvestPlus: breeding crops for better nutrition. *Crop Sci.* 47, S–88. doi: 10.2135/cropsci2007.09.0020IPBS

Phumichai, C., Aiemnaka, P., Nathaisong, P., Hunsawattanakul, S., Fungfoo, P., Rojanaridpiched, C., et al. (2022). Genome-wide association mapping and genomic prediction of yield-related traits and starch pasting properties in cassava. *Theor. Appl. Genet.* 135, 145–171. doi: 10.1007/s00122-021-03956-2

Platten, J. D., Cobb, J. N., and Zantua, R. E. (2019). Criteria for evaluating molecular markers: Comprehensive quality metrics to improve marker-assisted selection. *PloS One* 14, e0210529. doi: 10.1371/journal.pone.0210529

Rabbi, I., Hamblin, M., Gedil, M., Kulakow, P., Ferguson, M., Ikpan, A. S., et al. (2014). Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. *Crop Sci.* 54, 1384–1396. doi: 10.2135/cropsci2013.07.0482

Rabbi, I. Y., Kayondo, S. I., Bauchet, G., Yusuf, M., Aghogho, C. I., Ogunpaimo, K., et al. (2020). Genome-wide association analysis reveals new insights into the genetic architecture of defensive, agro-morphological and quality-related traits in cassava. *Plant Mol. Biol* 109, 195–213. doi: 10.1101/2020.04.25.061440

Rabbi, I. Y., Udoh, L. I., Wolfe, M., Parkes, E. Y., Gedil, M. A., Dixon, A., et al. (2017). Genome-wide association mapping of correlated traits in cassava: Dry matter and total carotenoid content. *The Plant Genome* 10. doi: 10.3835/plantgenome2016.09.0094

R Development Core Team (2020). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing).

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. *Encycl Database Syst.* 5, 532–538. doi: 10.1007/978-0-387-39940-9_565

Saltzman, A., Birol, E., Bouis, H. E., Boy, E., Moura, F. F. D., Islam, Y., et al. (2013). Biofortification: Progress toward a more nourishing future. *Glob Food Secur* 2, 9–17. doi: 10.1016/j.gfs.2012.12.003

Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., et al. (2014). Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and hunter color techniques. *Food Chem.* 151, 444–451. doi: 10.1016/j.foodchem.2013.11.081

Sánchez, T., Chávez, A., Ceballos, H., Rodriguez-Amaya, D., Nestel, P., and Ishitani, M. (2006). Reduction or delay of post-harvest physiological deterioration in cassava roots with higher carotenoid content. *J. Sci. Food Agric.* 86, 634–639. doi: 10.1002/jsfa.2371

Sayre, R., Beeching, J. R., Cahoon, E. B., Egesi, C., Fauquet, C., Fellman, J., et al. (2011). The BioCassava plus program: biofortification of cassava for sub-Saharan Africa. *Annu. Rev. Plant Biol.* 62, 251–272. doi: 10.1146/annurev-arplant-042110-103751

Semagn, K., Babu, R., Hearne, S., and Olsen, M. (2014). Single nucleotide polymorphism genotyping using kompetitive allele specific PCR (KASP): overview of the technology and its application in crop improvement. *Mol. Breed* 33, 1–14. doi: 10.1007/s11032-013-9917-x

Stephenson, K., Amthor, R., Mallowa, S., Nungo, R., Maziya-Dixon, B., Gichuki, S., et al. (2010). Consuming cassava as a staple food places children 2-5 years old at risk for inadequate protein intake, an observational study in Kenya and Nigeria. *Nutr. J.* 9, 9. doi: 10.1186/1475-2891-9-9

Talsma, E. F., Melse-Boonstra, A., de Kok, B. P. H., Mbera, G. N. K., Mwangi, A. M., and Brouwer, I. D. (2013). Biofortified cassava with pro-vitamin a is sensory and culturally acceptable for consumption by primary school children in Kenya. *PloS One* 8, 1–8. doi: 10.1371/journal.pone.0073433

Uchendu, K., Njoku, D. N., Paterne, A., Rabbi, I. Y., Dzidzienyo, D., Tongoona, P., et al. (2021). Genome-wide association study of root mealiness and other texture-associated traits in cassava. *Front. Plant Sci.* 12, 770434. doi: 10.3389/fpls.2021.770434

Udoh, L. I., Gedil, M., Parkes, E. Y., Kulakow, P., Adesoye, A., Nwuba, C., et al. (2017). Candidate gene sequencing and validation of SNP markers linked to carotenoid content in cassava (Manihot esculenta crantz). *Mol. Breed* 37, 123. doi: 10.1007/s11032-017-0718-5

Wani, F. J., Rizvi, S. E. H., Sharma, M. K., and Bhat, M. I. J. (2018). A study on cross validation for model selection and estimation. *Int. J. Agric. Sci.* 14, 165–172. doi: 10.15740/HAS/IJAS/14.1/165-172

Welsch, R., Arango, J., Bär, C., Salazar, B., Al-Babili, S., Beltrán, J., et al. (2010). Provitamin a accumulation in cassava (manihot esculenta) roots driven by a single nucleotide polymorphism in a phytoene synthase gene. *Plant Cell* 22, 3348–3356. doi: 10.1105/tpc.110.077560

WHO (2022) *Nutr landsc inf syst*. Available at: https://www.who.int/data/nutrition/nlis/info/vitamin-a-deficiency#:~:text=Deficiency%20of%20vitamin%

20A%20is,outcomes%20of%20pregnancy%20and%20lactation (Accessed January 11, 2021).

Wolfe, M. D., Rabbi, I. Y., Egesi, C., and Hamblin, M. (2016). Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *The Plant Genome*. 9. doi: 10.3835/plantgenome2015.11.0118

Wong, J. C., Lambert, R. J., Wurtzel, E. T., and Rocheford, T. R. (2004). QTL and candidate genes phytoene synthase and ζ-carotene desaturase associated with the accumulation of carotenoids in maize. *Theor. Appl. Genet.* 108, 349–359. doi: 10.1007/s00122-003-1436-4

Yang, S., Fresnedo-Ramírez, J., Wang, M., Cote, L., Schweitzer, P., Barba, P., et al. (2016). A next-generation marker genotyping platform (AmpSeq) in heterozygous crops: A case study for marker-assisted selection in grapevine. *Hortic. Res.* 3, 16002. doi: 10.1038/hortres.2016.2

Yu, J., and Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* 17, 155–160. doi: 10.1016/j.copbio.2006.02.003

TYPE Original Research
PUBLISHED 17 October 2022
DOI 10.3389/fpls.2022.974795

Genetic variability and genotype by environment interaction of two major cassava processed products in multi-environments

Cynthia Idhigu Aghogho[1,2], Saviour J. Y. Eleblu[1], Moshood A. Bakare[3], Ismail Siraj Kayondo[2], Isaac Asante[1], Elizabeth Y. Parkes[2], Peter Kulakow[2], Samuel Kwame Offei[1] and Ismail Rabbi[2]*

[1]West Africa Centre for Crop Improvement (WACCI), College of Basic and Applied Sciences University of Ghana, Legon, Ghana, [2]International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria, [3]Plant Breeding and Genetics Section, School of Integrative Plant Science, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, United States

OPEN ACCESS

EDITED BY
Madhav Bhatta,
Bayer Crop Science, United States

REVIEWED BY
Rodrigo R. Amadeu,
Bayer Crop Science, United States
Hossein Sabouri,
Gonbad Kavous University, Iran

*CORRESPONDENCE
Ismail Rabbi
I.Rabbi@cgiar.org

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 21 June 2022
ACCEPTED 15 August 2022
PUBLISHED 17 October 2022

CITATION
Aghogho CI, Eleblu SJY, Bakare MA, Kayondo IS, Asante I, Parkes EY, Kulakow P, Offei SK and Rabbi I (2022) Genetic variability and genotype by environment interaction of two major cassava processed products in multi-environments. *Front. Plant Sci.* 13:974795. doi: 10.3389/fpls.2022.974795

Conversion of cassava (*Manihot esculenta*) roots to processed products such as gari and fufu before consumption is a common practice worldwide by cassava end-user for detoxification, prolonged shelf life or profitability. Fresh root and processed product yield are supposed to be equivalent for each genotype, however, that is not the case. Developing genotypes with high product conversion rate is an important breeding goal in cassava as it drives the adoption rates of new varieties. The objective of this study was to quantify the contribution of genetic and genotype-by-environment interaction (GEI) patterns on cassava root conversion rate to gari and fufu. Sixty-seven advanced breeding genotypes from the International Institute of Tropical Agriculture (IITA) were evaluated across eight environments in Nigeria. Root conversion rate means across trials ranges from 14.72 to 22.76% for gari% and 16.96–24.24% for fufu%. Heritability estimates range from 0.17 to 0.74 for trial bases and 0.71 overall environment for gari% and 0.03–0.65 for trial bases and 0.72 overall environment for fufu% which implies that genetic improvement can be made on these traits. Root conversion rate for both gari and fufu% showed a negative but insignificant correlation with fresh root yield and significant positive correlation to Dry Matter content. For all fitted models, environment and interaction had explained more of the phenotypic variation observed among genotypes for both product conversion rates showing the presence of a strong GEI. Wrickle ecovalence (Wi) stability analysis and Geometric Adaptability index (GAI) identified G40 (TMS14F1285P0006) as part of top 5 genotypes for gari% but no overlapping genotype was identified by both stability analysis for fufu%. This genotypic performance across environments suggests that it is possible to have genotype with dual-purpose for high gari and fufu conversion rate.

KEYWORDS

genetic variability, heritability, genotype by environment interaction, gari and fufu conversion rate, cassava processed products breeding

Introduction

Cassava (*Manihot esculenta*) is an affordable carbohydrate source for about 800 million people in Africa, Asia, and Latin America (Montagnac et al., 2009; Dusunceli, 2019). This clonally propagated starchy root crop is considered a food security or insurance crop for smallholder farmers due to its year-round availability, and its ability to grow in marginal environments characterized by water scarcity and poor soils. More than 90% of cassava produced in Africa is used for human food compared with 50% in Asia and 43% in South America, while the remaining 10% is used for animal feed production (Nweke, 2004).

Cassava roots are consumed either fresh or processed into various products. Processing products differ depending on the consumption preference and processing method used. Sweet genotypes with low cyanide content goes through minimal processing such as boiling, roasting or frying while non-sweet genotypes goes through a more rigorous processing which include grating and fermentation before consumption (Lancaster et al., 1982; Nweke, 1994). Various processing techniques are used to add value, extend the shelf life of the products as well as detoxify the roots by removing cyanogenic glucosides (Westby, 2002; Cardoso et al., 2005). Without processing, commercialization of cassava roots for urban markets will be difficult to achieve (Coulibaly et al., 2014). By far, processed products account for the largest proportion of cassava food consumption in Africa absorbing more than 90% of the produced roots (Nweke, 2004).

The processed products derived from cassava roots include flour, starch and various forms of fermented products. In Africa, cassava roots are converted into a diverse set of products, the most important of which are gari and fufu, as well as tapioca, lafun, and attieke (Ezemenari et al., 1998). The processing method which may or may not include fermentation or starch gelatinization (Sanchez et al., 2010) and cell-wall disintegration (Eggleston and Asiedu, 1994) results in products with different attributes related to sensory, pasting and functional properties (Sanni et al., 2003; Onitilo et al., 2007). The end products may be categorized as flour, thick paste, or semolina-like particles (Awoyale et al., 2021).

Also known as cassava semolina or "farinha de mandioca," gari is the most common processed product of cassava in West Africa due to its long shelf-life and its ready-to-consume characteristic (Oluwafemi and Udeh, 2016). Cassava roots are rendered into gari by peeling, washing, grating, squeezing out water, and roasting on a dry hot surface. The grated mash can either be directly processed or fermented to produce a product with varying degrees of sourness. The resulting product is dry crispy, fine to coarse granular flour. As a result of its pre-gelatinization property, gari can be eaten in the uncooked form, or soaked in cold water like cereal or added to hot water to produce a thick dough called eba and consumed with vegetable sauce (Sanni et al., 1998).

Fufu is the second most common product after gari in Africa (Sanni et al., 1998). It is produced from retted roots after steeping in water for several days to allow for microbial fermentation and tissue disintegration. The raw mash is sieved to remove insoluble fiber (vascular bundles) and cooked directly into doughy meals or dried and milled into flour for longer shelf life (Akingbala et al., 1991). There are different variations to these processing methods depending on region or desired taste (Okpokiri et al., 1985).

Product conversion rate, defined as the percentage of final product relative to a unit of starting fresh roots, is an important factor in determining variety acceptability by growers and processors. Although this has increased productivity on a fresh root weight basis, processing traits related to conversion rates have not been adequately addressed by breeders (Wossen et al., 2017). Processing the same quantity of roots from different varieties may result in different quantities of derived products (on a dry weight basis). This has an important efficiency and economic implication given the fixed cost of product processing such as labor, time, energy and other resources such as transportation. Varieties with high conversion rate are more preferable when everything is held constant (Wossen et al., 2017). Additionally, overall product yield in tons/ha, defined as fresh root yield (t/ha) times the conversion rate of the product is an important overall productivity metric that can vary among varieties. There is also a trade-off between fresh root yield and conversion rate as it influences production and processing efficiency. For example, a variety with high fresh yield but low conversion rate may be less preferable than a moderately yielding variety with high conversion.

Cassava breeding cycle through phenotypic recurrent selection takes up to 8 years before the release of a new variety (Ceballos et al., 2020). Cassava breeding scheme from botanical seed production to varietal release comprises of six selection stages which include progeny testing (F_1) clonal evaluation stage, preliminary yield stage, advance yield stage, uniform yield stage (regional trial) and one multiplication stages which include farm level testing (Ceballos et al., 2012). Historically cassava breeding has focused on yield improvement, increases nutritional content and pest and disease resistance in developing new varieties which have largely been addressed (Hahn, 1989; Okechukwu and Dixon, 2008). Currently, due to increased cassava cultivation and commercialization worldwide and in Africa, there is increased incentive to breed for varieties that are not only high yielding but also high product conversion rate. The lack of prioritization of these traits in breeding programs can often lead to low adoption of modern varieties (Nweke, 2004; Wossen et al., 2017).

Trait improvement is highly dependent on the availability of information related to genetic component and trait behavior across environments. Genetic variability, heritability, and stability of expression of root conversion rate traits as well as their correlation in cassava is limited, thereby hindering the ability of breeders to improve them through recurrent selection schemes. Heritability, defined as the proportion of phenotypic variation that is attributable to genetic variance, is important in crop improvement as it influences traits evaluation and selection accuracy (Kempthorne, 1970). Breeding selection is relatively easy for traits with large genetic variance and heritability. The influence of the genotype-by-environment interactions (GEI) in the expression of a trait in different environments also need to be considered by breeders in order to identify superior genotypes and of the location that best represents the target environment. Previous studies on conversion rate of processed products have focused on either the effect of different root storage method before processing, different processing methods or genotype harvesting age (Oghenechavwuko et al., 2013; Adegbola et al., 2014; Oyeyinka et al., 2019). Studies associated with genetic variability are limited, descriptive in nature and assessed a limited number of genotypes in one or few environments (Etudaiye et al., 2009; Bassey, 2018). These studies have been unable to estimate the genetic contribution and genotype by environment interaction (GEI) effect as well as relationship between cassava processed product traits and key agronomic variables.

In the present study, we carried out multi-location, multi-year phenotyping trials using advanced breeding lines to: (1) evaluate the genetic variation and heritability of 12 traits related to product conversion rate and overall product yield for two major processed products (gari and fufu); (2) monitor the effect of different environments and to estimate genotype by environment (GEI) interactions; and (3) understand the relationship between processing traits and key agronomic variables such as yield and yield components. We used a collection of advanced breeding clones developed by the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria.

Materials and methods

Plant materials

A total of 62 advanced genotypes and 5 checks from the uniform yield trials (UYT) in the IITA cassava breeding program were used in this study. These genotypes, derived from a second generation of a genomic selection-based population improvement pipeline (Wolfe et al., 2017) were selected based on their performance on fresh root yield, dry matter content, harvest rate, root number. All accessions are resistant to the cassava mosaic disease which is known to negatively influence productivity in cassava (Thresh et al., 1997). The genotypes had white root pulp with moderate to high dry matter percentage. The genotypes were randomly divided into two sets of trials (UYT36setA and UYT36setB). Each set has 31 clonal lines and 5 standard checks in common, making a unique set of clones of 67 genotypes (**Supplementary Table 1**). The 5 standard checks used for the study include TMS30572 and TMEB419 (most adopted and popular varieties in Nigeria), TMS-IBA000070, a recently released variety, TMS-IBA980581 and TMS-IBA982101 both of which are high yielding and disease resistant varieties.

Experimental sites and design

Field experiments were conducted across four unique locations in 3 years, the locations used varied from 1 year to another as did the number of trials making a total of eight unique environment with year and location combined (**Table 1**). These locations represent two major agro-ecological zones in Nigeria (**Table 1**) which was selected to represent the major regions where gari and fufu products are produced and consumed (Ezedinma et al., 2007). The agroecological variability of the locations include the humid forest characterized by high precipitation and derived savanna with moderate precipitation (Iloeje, 1965). The selected locations represent the major The trials were established in an alpha-lattice design with two replications. Plot dimension was 6 m × 7 m consisting of 42 planted at a spacing of 1 m × 0.8 m^2 plants between and within rows, respectively. The row and column numbers for each genotype within-trial sets were recorded for spatial trend analysis.

Trial harvesting and yield trait phenotyping

The trials were harvested at maturity, 12 months after planting. To reduce the border effect on genotype agronomic performance, only the net plot consisting of 20 plants were harvested for phenotyping. During harvest, plot-level data on root number (RTNO), root weight (RTWT), and root size (RTSZ) were recorded. Harvest index was recorded as the ratio of root biomass relative to total biomass. Fresh root yield (FYLD) expressed as tons per hectare and was calculated from RTWT adjusted by plant spacing of 0.8 m^2. Dry root yield

TABLE 1 Summary of trial locations, agro-ecological zones, and seasons.

Location	Agro-ecological zone	Year	Latitude	Longitude
Ago-Owu	Derived Savanna	2017, 2018, 2019	7°20′ N	4°16′ E
Ibadan	Derived Savanna	2018, 2019	7°49′ N	3°90′ E
Ikenne	Humid forest	2017, 2019	6°84′ N	3°69′ E
Ubiaja	Humid forest	2019	6°67′N	6°34′ E

(DYLD) was derived as a product of Dry matter (DM) content and FYLD. RTSZ was recorded categorically as 3 (small), 5 (average), and 7 (large) as recommended by Fukuda et al. (2010). All data was captured using FieldBook app (Rife and Poland, 2014) and traits recorded using established ontologies from www.cassavabase.org.

Product processing

Marketable roots from each plot were selected for dry matter estimation and processing into gari and fufu products. To access the root dry matter content, 6–8 roots were randomly sampled, peeled, and grated after removing proximal and distal ends to reduce fibrous material. After thorough mixing, 100 g samples of the root grates were oven-dried at 95°C for 48 h until constant weight and the dry matter was expressed as a percentage of fresh weight. Care was taken to ensure rotted or damaged roots are not included in the sampling.

Each product was processed from 20 kg of roots but 10 kg was used when sufficient quantity was not available. The roots from each plot were packed in separate pre-labeled bags and transported to centralized facilities to ensure processing is done the same day and reduce post-harvest physiological deterioration.

Gari production

Gari processing was carried out as described in Ukhun (1989). Peeled roots were washed, and grated into fine particles. The grated mash was transferred into woven polypropylene sacks and allowed to undergo spontaneous fermentation for 48 h. Water was pressed out of each sample using a hydraulic method to eliminate about 60 percent of the remaining water. The semi-dried cakes were then sieved and toasted in a hot stainless-steel frying tray to form gelatinized, dry and crispy granules. The temperature of the copper tray before frying ranges between 143.67°C and 148.87°C and the final temperature of the fried product is 88.01°C–90.93°C. Finally, the gari product was allowed to cool to room temperature and stored in barcoded nylon bags after recording the product weight in kg. Conversion rate was calculated as a percentage of starting fresh root as follows:

$$Gari\,\% = \frac{Final\ gari\ weight}{Starting\ root\ weight}\ X\ 100$$

Dried fufu processing

Fufu processing followed the method of Achi and Akomas (2006). This method produces either odorless fufu paste or dried to produce a flour-like product. In this study, we generated dry fufu product for estimating conversion rates. The dry fufu is suitable for long term storage.

After peeling, roots were cut into small chunks and soaked in individual plastic buckets using 40 l of water for 4 days to undergo lactic acid fermentation until the roots are softened. After softening, the starchy pulp was separated from insoluble fiber using a 0.3 cm pore size sieve over a clean bucket. The pulp filtrate was washed twice with 20 l of clean water and allowed to sediment for 4 h until the water clears. The water was carefully decanted and the product transferred to a cotton bag followed by straining of the remaining water. Finally, the product was spread on clean flat stainless-steel trays and oven-dried at 60°C for 48 h to a constant weight. The resulting odorless fufu was allowed to cool to room temperature and stored in barcoded nylon bags after recording the product weight in kg.

$$Fufu\,\% = \frac{dried\ fufu\ weight}{Starting\ root\ weight}\ X\ 100$$

Processing losses in terms of peel weight and dried insoluble fiber removed from fufu mash were also recorded. Conversion rate is a function of % moisture content and peel waste, which can be up to 35% of root proportion (Omosuli et al., 2017). For fufu, processing losses also includes insoluble fiber that is removed after sieving and placed in a 60°C oven for a period of 48 h before weighing in kg. Peel loss and fiber content were converted to percentage of initial weight of root used for processing into product as described below:

$$Peel\ loss\,\% = \frac{Peel\ weight}{Starting\ root\ weight}\ X\ 100$$

$$Fibre\ content\,\% = \frac{Dried\ fibre\ weight}{Starting\ root\ weight}\ X\ 100$$

Statistical analyses

Descriptive statistics per environment and trait were generated and visualized in R (R Core, 2020). The distribution of observed traits using BLUPs was visualized with violin plot with boxplot superimposed and stacked plots across trials using the ggplot2 package (Wickham, 2016) in R (R Core, 2020).

In order to estimate variance components of traits, a single trait linear mixed model was fitted using the ASReml-R package version 4.1 (Butler et al., 2017), considering the genotype, replication, environment, and genotype by environment as random effects while sets, rows and columns of each trial as fixed effects of accounting for trial design-related variables.

We fitted a model as shown in Eq. [1]:

$$y = X\beta + \mathrm{Zu} + \mathrm{e}\ [1]$$

With $u \sim N\left(0, \mathbf{I}\,\sigma_u^2\right)$ and $\mathbf{e} \sim N\left(0, \mathbf{I}\,\sigma_e^2\right)$, where $\mathbf{y}$ is the response vector of a trait for a given location, β is the vector of fixed effects with the design matrix $\mathbf{X}$ (relating observations to fixed effects which include grand mean, row number nested within set and column number nested within set); **u** is the vector of random genetic effects with the design matrix **Z** (relating trait values to genotype, environment, replication nested within environment and GEI) and e is the residual. Test of significance of variance components was done using *z*-test as done in ASReml-R package version 4.1 (Butler et al., 2017).

The BLUP represents an estimate of each individual's total genetic value across environments for the genotype effect. For the sake of correlation analysis, it is important to estimate a single value that encapsulates all the information available on the individual; we estimated the de-regressed BLUPs (D-RBLUPS) by dividing by their reliability $deregressed\ BLUP = \frac{BLUP}{\left(1-\frac{PEV}{\sigma_i^2}\right)}$ (Garrick et al., 2009) where PEV is the predicted error variance of the BLUP and σ_i^2 is the clonal variance component.

Correlation analysis of the traits using the D-RBLUPS estimates was determined using the *corrr* package and visualize using *ggcorrplot* in core R version 4.1.1 (R Core, 2020).

Broad-sense heritability was estimated using two methods. First, the standard method (H2_standard) based on error plot variance across all environments was derived from variance components estimated as $H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \left(\frac{\sigma_{ge}^2}{e}\right) + \left(\frac{\sigma_\varepsilon^2}{er}\right)}$ where σ_g^2 refers to the variance of genotype, σ_{ge}^2 is GEI variance, σ_ε^2 is the environmental variance, *e* is the number of environments, *r* is the number of replicates of genotypes per environment, and other terms were described above. The second broad-sense heritability (H_Cullis) proposed by Cullis et al. (1996) was estimated using genotype standard error calculated as; $H^2_{cullis} = 1 - \frac{\overline{v}_\triangle^{BLUP}}{2\sigma_g^2}$ where σ_g^2 refers to genetic variance, $\overline{v}_\triangle^{BLUP}$ to the average standard error of the genotypic BLUPs.

To carry out GEI, we generated table of genotype means by fitting a mixed model with genotype as fixed effect as shown in Eq. [2]:

$$y = X\beta + Zu + e \quad [2]$$

With $u \sim N\left(0, \mathbf{I}\,\sigma_u^2\right)$ and $\mathbf{e} \sim N\left(0, \mathbf{I}\,\sigma_e^2\right)$, where **y** is the response vector of a trait for a given location, β is the vector of fixed effects with the design matrix **X** (relating observations to fixed effects which include grand mean and genotype); **u** is the vector of random genetic effects with the design matrix **Z** (relating trait values to environment, row number nested within set and column number nested within set) and e is the residual.

The resulting table of Best Linear Unbiased Estimates (BLUEs) was used to model the GEI using three approaches shown in **Table 2**; Malosetti et al. (2013) using the statgenGxE package version 1.0 (van Rossum et al., 2021) in R version 4.1.1 (R Core, 2020).

The first approach was suggested by Finlay and Wilkinson (1963) (FW), a regression analysis which has been widely used to describe stability and GEI in various cultivars (Mulusew et al., 2014; Swanckaert et al., 2020). The Finlay-Wilkinson regression model estimates the heterogeneity of slopes and sensitivity of a genotype by regressing mean phenotypic performance of individual genotypes on an environmental index using within-line ordinary least squares (OLS) regression (Lian and de los Campos, 2016). However, FW linear regression is not sufficient to fully explain the genotype phenotypic stability.

Before fitting the FW model (**Table 2**), trait values were scaled to mean of zero and a standard deviation of 1, following the equation below as: adjusted phenotype mean scaled $= \frac{[y_{ij} - mean(Y)]}{sd(Y)}$ where y_{ij} is the adjusted phenotypic mean value of ith genotype in jth environment and sd (Y) is the standard deviation of the overall mean of the adjusted phenotypic response of all clones in all environments. The scaling allowed the comparison of MSE and sensitivity values across traits that are originally on different scales and units measurement (Falcon et al., 2020).

The second and third approaches are Fixed-effect linear-bilinear models Additive Main-effects and Multiplicative Interaction (AMMI) and Genotype Main Effect plus Genotype-Environment Interaction (GGE). Both approaches depend on analysis of variance (ANOVA) for estimating genotype and environment main effect, principal component analysis (PCA) for decomposing GEI structure into Interactive Principal Component Axes (IPCAs) and biplot for graphical presentations. The AMMI model can be further used to delineate the testing environments into mega environments using principal component axes scores. AMMI gives a suitable approach in separating genotypic effect from genotype by environment effect with cultivar ranking in mega-environment (Hagos and Abay, 2013) while GGE is suitable for grouping sites and cultivars without cultivar rank change (Yan and Hunt, 2001). Despite their different approaches, both models complement each other in order to strengthen decision making thereby permitting increased reliability in the selection of superior cultivars and test environments.

Stability of genotype performance across environments

We assessed stability of genotypes for traits observed using both static and dynamic stability measures. Static stability was measured using Wrickle ecovalence (Wi) as proposed by Wricke (1962) and described $W_i^2 = \sum\left(X_{ij} - \overline{X}i. - \overline{X}.j + \overline{X}..\right)^2$ where X_{ij} is the observed trait respond (average across replication), $\overline{X}i.$ correspond to the mean yield of genotype *i*, $\overline{X}.j$ is the mean yield of the environment *j* and $\overline{X}$ is the grand means.

According to the ranking of genotypes by Wi, stable genotype has lower Wi preferably close to 0. These are genotypes that have smaller deviation from the environmental mean. Likewise, genotypes with high Wi indicates instability in genotype performance across the environments and a large contribution of the genotype to the GEI.

TABLE 2 Description of models and references.

Model	Formular	Variables	References
FW	$y_{ij} = \mu + G_i + E_j + b_i E_j + e_{ij}$	y_{ij} is the mean yield of genotype i in environment j μ is the grand mean G_i is the genotypic effect; E_j is the environment effect; $b_i E_j$ is a sensitivity parameters; e_{ij} is the residual.	Malosetti et al., 2013
AMMI	$y_{ij} = \mu + g_i + e_j + \sum_{k=1}^{K} \lambda_k a_{ik} \gamma_{ij} + \varepsilon_{ij}$	y_{ij} is the mean yield of genotype i in environment j; μ is s the grand mean. g_i is the genotype fixed effect of jth environment. The GEI component is decomposed into K multiplicative terms (k = 1, 2, . . ., K), each multiplicative term is a product of kth eigenvalue (k); genotypic score (ik); and environmental scores (jk); and ij is the residual	Gauch, 1992
GGE	$y_{ij} = \mu + e_j + \sum_{k=1}^{K} \lambda_k a_{ik} \gamma_{ij} + \varepsilon_{ij}$	Terms are similar to AMMI model but without g_i which is the genotype fixed effect of jth environment.	Yan et al., 2000

Geometric adaptability index (GAI) is a measure of the adaptability of a genotype and is classified as a dynamic concept of stability (Mohammadi and Amri, 2008) and described as: GAI $= \sqrt[E]{\overline{X}1 + \overline{X}2 + \ldots + \overline{X}l}$. Where $\overline{X}1$, $\overline{X}2$, and $\overline{X}l$ are the mean yields of the first, second and ith genotypes across environments and E is the number of environments. According to the ranking of genotypes by GAI, genotypes with the high GAI (low ranks) are desirable (Mohammadi and Amri, 2008; Pourdad, 2011), Wi and GAI was estimated using metan package version 1.15 (Olivoto and Lúcio, 2020) in R version 4.1.1 (R Core, 2020).

Results

Analysis of variation for cassava gari, fufu, and related yield traits evaluated in multiple environments

Across environments and genotypes, a considerable range of phenotypic values was observed among genotypes and trials as shown in the distribution of estimated BLUPs (**Figure 1**). We observed differences between trials for FYLD which ranged from low performing environment (UB18, 20.18 t/ha) to high performing environment (IB19, 39.22 t/ha) with an average of 33.03 t/ha. In the study population we also observed an average of 36.94% for DM with genotypes ranging from 24.67 to 43.26 across environments. There was also variation in DM among trials used in this study ranging from lowest average DM content in AG19 (31.99) and highest IB19 (32.66). An average of 19.23% was observed for gari% with a variation between genotypes ranging from 11.82 to 25.27% across the eight environments used in this study. Between trials UB20 (20.24%) had the lowest average gari% while IB19 (21.29%) had the highest average gari%. Fufu% also appears to have an average phenotypic variation of 19.68% across the eight environments with genotypes ranging from 13.72 to 25.34%. We also observed trial variation for fufu% with UB20 (19.22%) having the lowest average and AG20 (21.25%) having the highest average fufu%. The results from phenotypic variability for fufu% is not far from what we observed in gari%. In the study population the average peel loss% was 20.80% but can range from as low as 10.75% to as much as 31.97% for some genotypes. Among the trials we had the lowest average peel loss% recorded in IK20 (15.93%) and highest was observed in IB19 (25.58%).

Relationship between all studied traits is represented using a correlation matrix (**Figure 2**). There was a positive significant ($P < 0.05$) correlation between DM and gari% ($r = 0.80$) and between DM and fufu% ($r = 0.82$) as well as between gari and fufu% ($r = 0.84$, $P < 0.05$). Interestingly, there appears to be no relationship between yield and the conversion rates as well as DM. We could see from the correlation matrix that DYLD, FYLD, gari yield, and fufu yield (yield traits) were highly correlated among each other. The yield traits correlation is not surprising because all yield traits were derivatives of RTWT. We observed a negative correlation between Fiber content% and fufu% ($r = –0.35$, $P < 0.05$). The fiber content% of a genotype is highly dependent on the ability of the genotype to soften during the fermentation stage in processing. Genotypes with high softening ability will produce less fiber which will be desirable for production of high fufu%. The correlation matrix also showed a negative correlation of peel loss% and gari% ($r = –0.30$, $P < 0.05$) and FYLD ($r = –0.43$, $P < 0.05$).

To further understand the relationship between both conversion rates (gari% and fufu%) and processing losses (peel loss% and fiber content%) can be visualized in **Figure 3**, we observed a 1:1 ratio in average values among genotypes for gari% (19.23%), fufu% (19.68%), and peel loss% (20.87%). This ratio means that the overall performance of a genotype is dependent on peel loss%. We recorded fiber content% which is related to fufu processing to be about 4.5% across trials. Among both processing losses, peel loss% has the highest contribution compared to fiber content%.

Because of the observed variation among trials, heritability estimates were computed using the mean of the genotypes within each trial and square of the standard error of the genetic estimates (**Figure 4** and **Supplementary Table 2**). We computed both H2_standard and H2_Cullis estimates

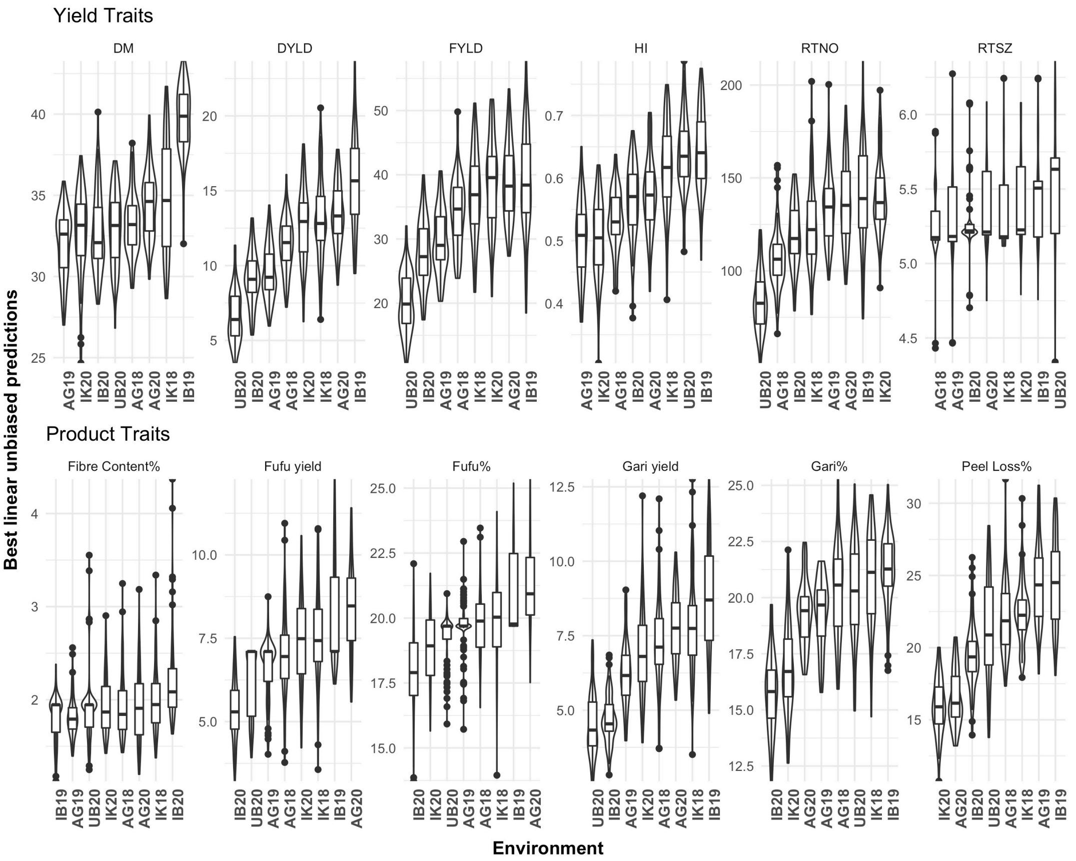


FIGURE 1
Phenotypic variation of 12 yield and product traits of cassava across eight environments. IB19, Ibadan 2019; IK18, Ikenne 2018; AG19, AgoOwo 2019; UB20, Ubiaja 2019; IK20, Ikenne 2020; Env, Environment; DM, Dry matter; DYLD, Dry root yield; FYLD, Fresh root yield; HI, Harvest rate; RTNO, Root number; RTSZ, Root size.

were comparable for all 12 traits, ranging from 0.31 to 0.75 and 0.6–0.8, respectively. Gari% and fufu% had a higher H2_standard (>0.70) compared to H2_Cullis (>0.40). We observed a within trial heritability estimated using H2_Cullis method (Figure 4) for gari% ranging from 0.18 to 0.74 with standard deviation (error bars) of 0.19 and fufu% ranges from 0.03 to 0.65 with standard deviation of 0.21. Some traits such as FYLD, DYLD gari and fufu yield had heritability estimates higher than 0.60 using the H2_standard method and higher than 0.40 using the H2_Cullis method. The lower H2_Cullis heritability estimates observed for these traits is due to inter-trials variations. The most relevant processing traits (peel loss%) had a heritability estimate of 0.54 using the H2_Cullis method and 0.81 using the H2_standard. The moderate to high levels of heritability observed for gari and fufu% indicates a higher proportion of genetic to total phenotypic variability which is suitable for genetic improvement of these traits through recurrent selection.

Results from the linear mixed model fitted to explain the contribution of genotype, environment, GEI, replication within environment and residual to phenotypic variation is presented in Figure 5 and sorted according to the descending H2_standard heritability estimates. We observed that environment had a significant effect on all traits (P < 0.01) ranging from 6.10 in RTSZ to 54.33% in peel loss% and explained the largest percentage of variation for most traits except for RTSZ (6.10%), fiber content% (16.87%) and HI (16.87%). Gari and fufu% had 45.48 and 36.14% of phenotypic variation explained by the environment, respectively, which was higher than what was explained by genotype effect. About 8.59 and 9.37% of phenotypic variation was explained for gari and fufu%, respectively, by the genotype term. The genotypic effect was also found to be significant (*P* < 0.001) for all traits and explained between 7.09% (DYLD) and 12.76% (DM). The significant genotypic terms suggest that these traits are influenced by genes and not only the environment. Change in relative performance of genotypes across environments is explained by the GEI term. There was a significant (P < 0.001) contribution of genotype x Env term to phenotypic variation observed among genotypes for most of

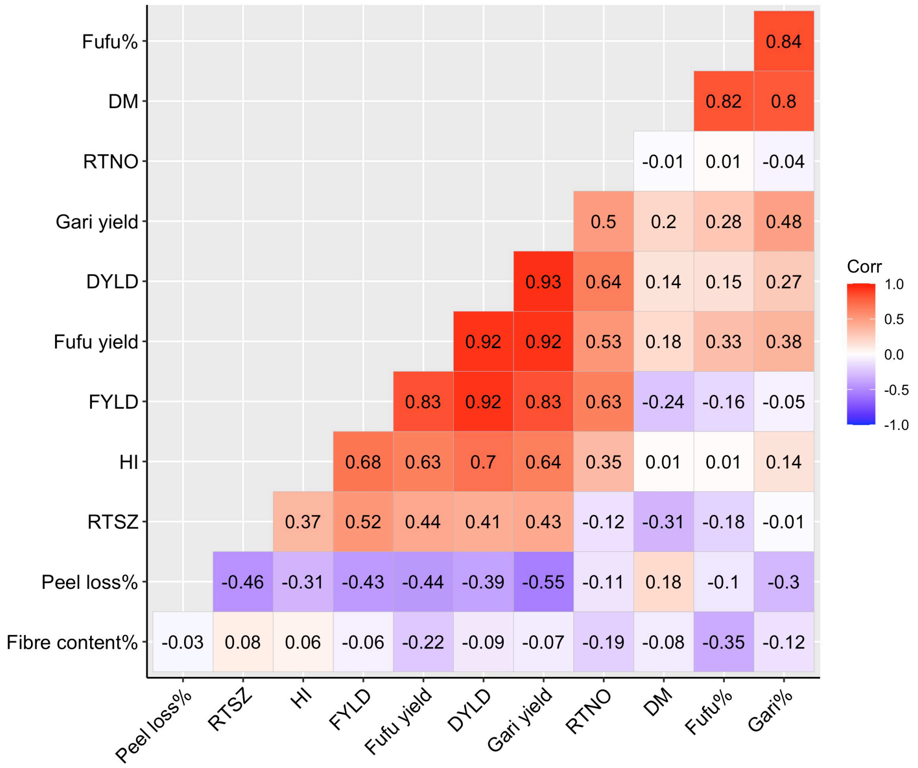


FIGURE 2
Correlation plot for processed products traits, and yield and root morphological traits using deregressed BLUPS form 8 environments. DM, Drymatter; RTNO, Root number; DYLD, Dry root yield; FYLD, Fresh root yield; HI, Harvest index; RTSZ, Root size.

trait in our study population except for fiber content% and explained between 6.03% (peel loss%) to 16.71% (DM) of phenotypic variation. Gari and fufu% phenotypic variance explained by GEI as 12.39 and 8.65%, respectively, which is close to what was explained by the genotype term. The replication nested within the environment which explains the experimental design effect captured the least percentage variation (0.50–11.48%) and was insignificant (Supplementary Table 3). The second-largest source of variation after the environment team accounting for up to 69.27% for RTSZ (Figure 6) is the residual term. This means there are some unexplained variations that could not be explained by the other terms in the model.

Genotype by environment interaction

Finlay-Wilkinson regression

The genotypic and environmental main effects of the Finlay-Wilkinson (FW) model were highly significant ($P < 0.001$) for all observed traits (Supplementary Table 4). However, the interaction effect was not significant for all traits except DYLD and RTSZ. According to the partitioning of the total sums of squares (TSS) (Figure 6), the environment term had the largest contribution for most traits ranging from 42.34% (HI) to 61.81% (gari yield). Of all the terms in the FW model, the interaction term has the least contribution to the TSS while the error term contributed almost twice the percentage contributed by genotype term.

Significant differences in regression slope (sensitivity) among genotypes on the environmental mean was found for all traits except dry matter content (Supplementary Table 5). In other words, there was variation in genotypic response for all traits but not dry matter with respect to changes in environment mean. Genotype and trait sensitivity to GEI was explained using the variance of the slopes and the variance of the mean square deviation, respectively, which was extracted from the FW regression analysis.

The genotypic sensitivity values were ranked from the most stable (low sensitivity values) to the least stable for each trait (high sensitivity values) for each trait (Supplementary Table 6). Using the FW we identified G38, G32, G24, G3, and G17 as top five stable genotypes for gari% and a different sets of genotypes as stable for fufu% which include G19, G35, G32, G39, and G5 except for G32.

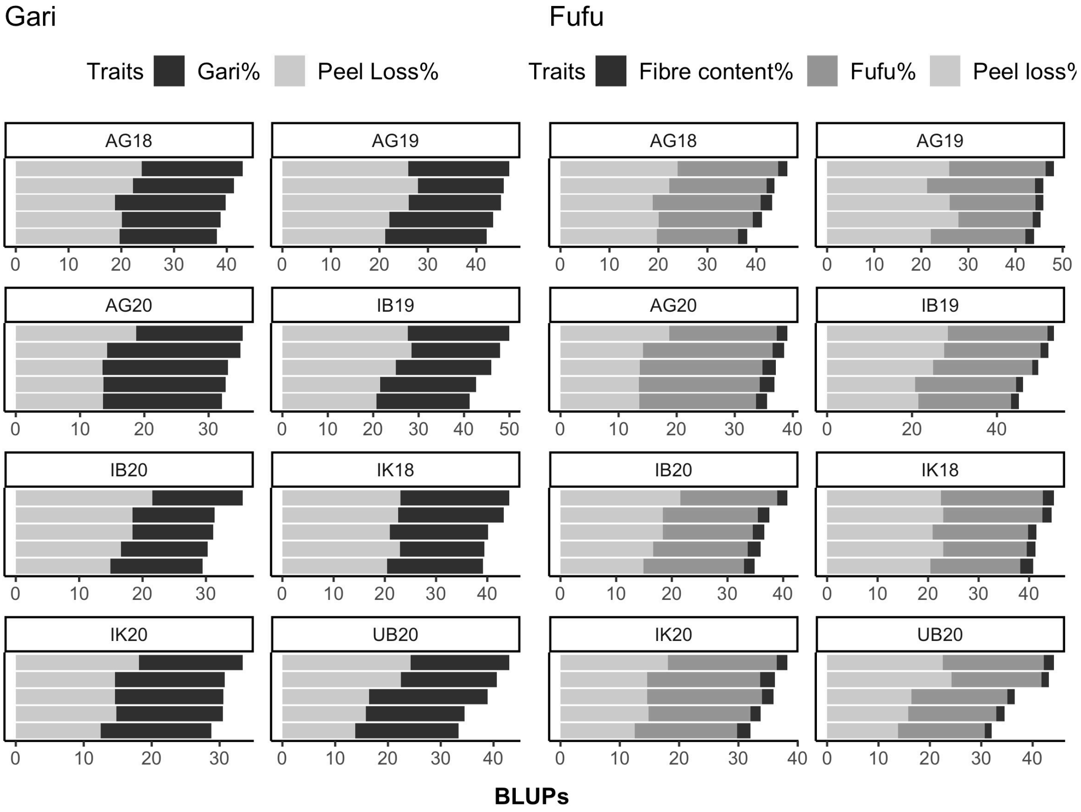


FIGURE 3
Relationship between gari and fufu % and processing loss traits for eight environments. IB19, Ibadan 2019; IK18, Ikenne 2018; AG19, AgoOwo 2019; UB20, Ubiaja 2019; IK20, Ikenne 2020.

Traits-specific environmental stability can be approximated using slope of the regression. Traits with narrow tolerance to distribution (higher slope) are more sensitive to the effect of environmental stress. The slope variance observed among traits varied from 0.04 (gari%) to 23.14 (fiber content%) with their corresponding slope median values varied from 1.01 and 0.95, respectively (Supplementary Table 4). The higher variance observed for fiber content% from FW regression analysis makes it a rather difficult trait to phenotype because of the large variation in some environments in comparison with others. Fufu% had a slope variance of 0.16 and median of 1.06. Peel loss had the lowest median MS deviation of all traits (median = 0.21) and the variance of MSE (variance = 0.03). Root size and Fiber content had the highest MS deviation. Among all traits observed, genotypes used as checks did not rank first as stable genotypes.

Additive main effect and multiplicative interaction

The Additive main effect and multiplicative interaction (*AMMI*) analysis revealed significant variation in the main effects of genotype, environment and their interactions (GEI) ($P < 0.001$) for all observed traits (Supplementary Table 4). The partition of total sum of squares (TSS) (Figure 6) showed that the environment main effect accounted for the highest amount of variation varying from 5.72% (RTSZ) to 54.85% (gari%). Traits with high TSS explained by environment indicate the existence of a group of environments sharing the same genotype(s) as best performing with large differences among environmental means (Yan and Rajcan, 2002). In the study population, the genotype contribution to TSS varied between 11.39% (DYLD) and 23.55% (HI). It is interesting to find out that DM (17.44%) had an almost 1:1 ratio of TSS explained by genotypes for gari% (13.75%), peel loss% (16.03%), and fufu% (14.97%). The ratio observed between traits points out the presence of genetic control for gari and fufu% that can be exploited for traits improvement through recurrent selection in multiple environments. We further decomposed the variation due to GEI for gari and fufu% using the first and second IPCAs and found out that both IPCAs accounted for 20.27 and 27.65% the TSS. For all traits measured in this study, the first and second IPCAs accounted for between 20.27% (gari%) and 49.96% (fiber content%) of the TSS due to GEI. Residual term explained between 10.01% (peel loss%) and 24.43% (RTSZ) of the TSS. We

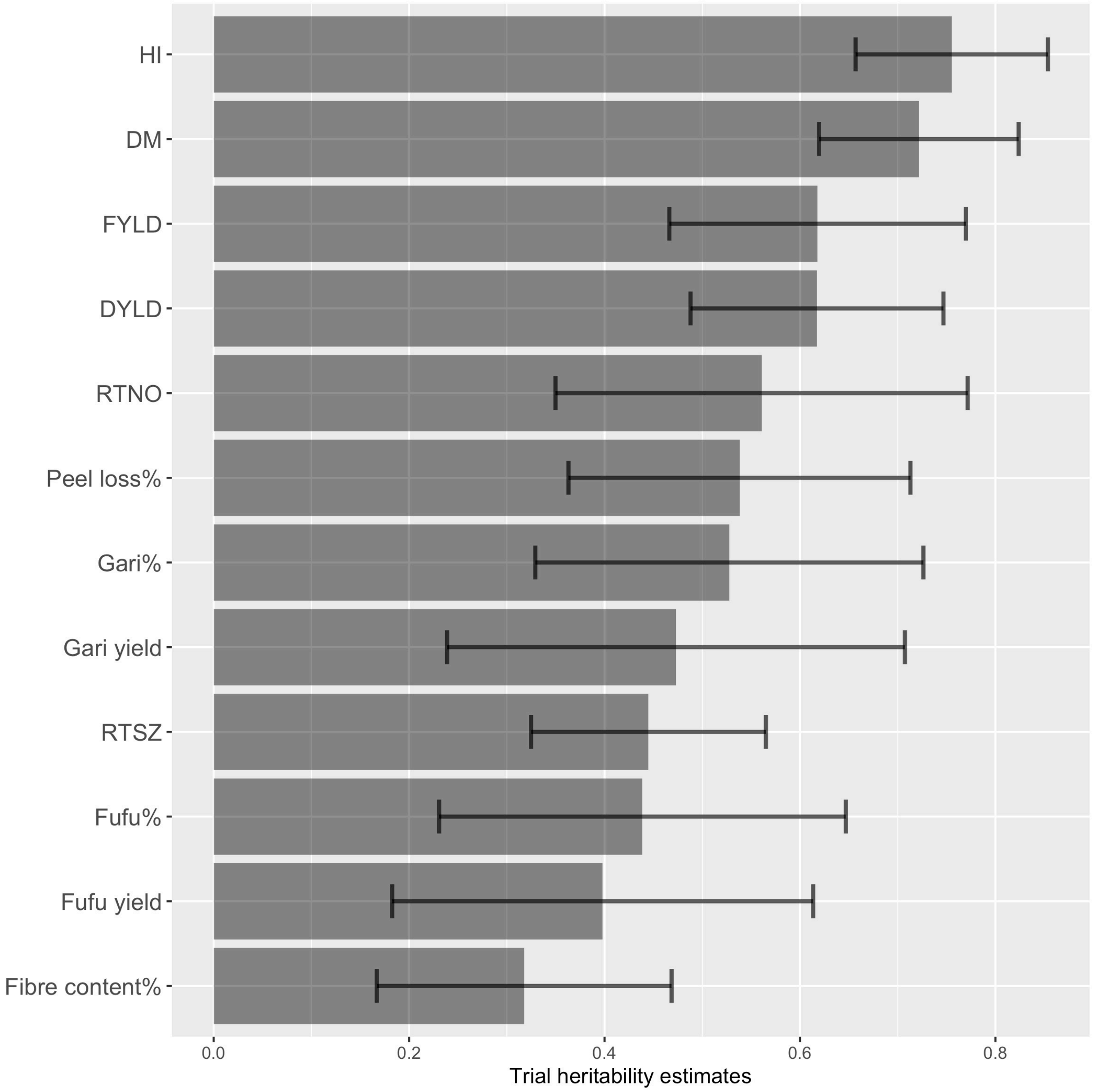


FIGURE 4
Broad-sense heritability estimated for 12 processed product and yield traits based on H2_Cullis method with error bars (standard deviation). DM, Drymatter, RTNO, Root number, DYLD, Dry root yield; FYLD, Fresh root yield; HI, Harvest index; RTSZ, Root size.

observed an equivalent proportion of TSS explained by residual and genotype terms for all traits except for HI which had a lower % explained by the residual term.

Genotype and genotype by environment

The GGE analysis of variance for 67 genotypes revealed a significant main effect of environment and combined genotype and GEI effect ($P < 0.001$) for the observed traits (**Supplementary Table 4**). To discern the contribution of different terms fitted in the model we partition the environment and interaction (first and second IPCAs) sum of squares as percentage of the total sum of squares. After partitioning the TSS we observed that between 5.38% (RTSZ) and 55.06% (fufu yield) of TSS was explained by the environment term and 31.93% (DYLD) to 66.68% (fiber content%) attributed to the interaction term. For gari and fufu%, TSS explained by the environment term was 53.10% which was larger than what

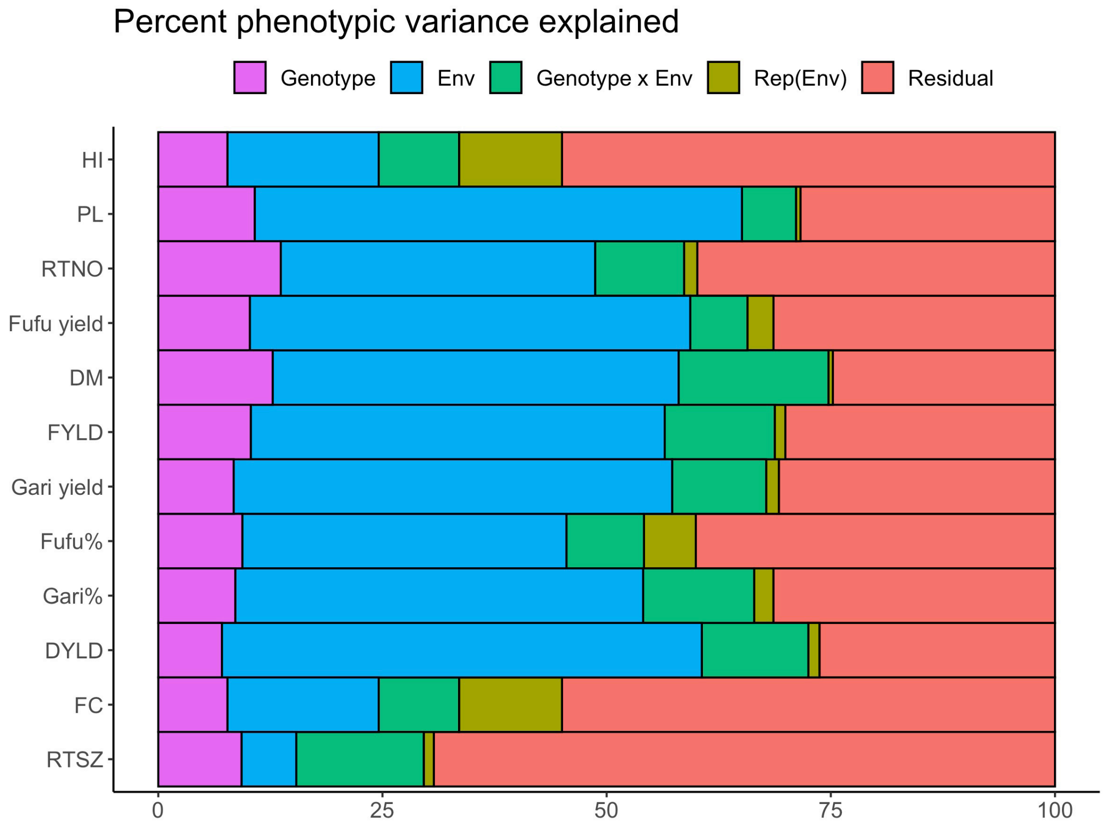


FIGURE 5
Percent of phenotypic variance explained by each fixed model analysis of variance model for 12 processed product and yield related traits. Env, Environment; rep, Replication; DM, Drymatter; RTNO, Root number; DYLD, Dry root yield; FYLD, Fresh root yield; HI, Harvest index, RTSZ, Root size.

the interaction term explained (Figure 6). However, for fufu% both the environment (42.92%), and interaction (41.46%) terms explained a 1:1 ratio of TSS.

Comparing the statistical models used in dissecting the GEI effects of traits observed in this study, we observed that most of the phenotypic variation seen in most traits were explained by the environment term. We also noticed that the contribution of interaction varies greatly among all statistical methods, GGE and FW had the highest and lowest contribution of the Interaction terms to traits phenotypic variation, respectively. This may be due to the removal of the genotype term when fitting GGE model and the sensitivity parameter in the FW model. Of all the variance terms measured, the environment had the highest contribution to traits expression, followed by interaction and residual before genotype. However, a single conclusion that can be drawn from the output indicates that the genotypes present different behavior for different traits in all environments used for this study and the environment was the primary source of variation. The significant interaction terms can affect the attainment of genetic advance from phenotypic selection due to differential response of genotypes under the target test environments.

GGE biplot which allows the visualization of genotype, environment and interaction based on symmetric scaling was used to understand the type of GEI in this study for observed traits (Figure 7, Supplementary Figures 1–10). The GGE biplot explained about 57.81 and 61.05%, of the total G + GE interactions (PC1 + PC2) for gari% and fufu%, respectively. We recognized a crossover type of GEI for gari and fufu %, which was indicated by the biplot principal component scores (PC1 and PC2) having both negative and positive values. Additionally, the polygon vertices of the biplots are markers for highly projected genotypes indicating performance in environments in the polygon vector. G45, G44, and G55 had the highest gari% above the environmental means in IK18, AG18, UB20, IK20, and AG20 while G56 and G21 were projected as the best performing genotypes in terms of fufu% in IB19, AG18, UB20, IK20, and AG20. Though the projected best performing genotypes for gari and fufu% in all environments were different, the environment seems to be correlated as indicated by the distance of the environment from origin and angle with other environments. The environmental correlation implies that one environment can represent another in screening for gari and fufu% (Solonechnyi et al., 2015; Temesgen et al., 2015).

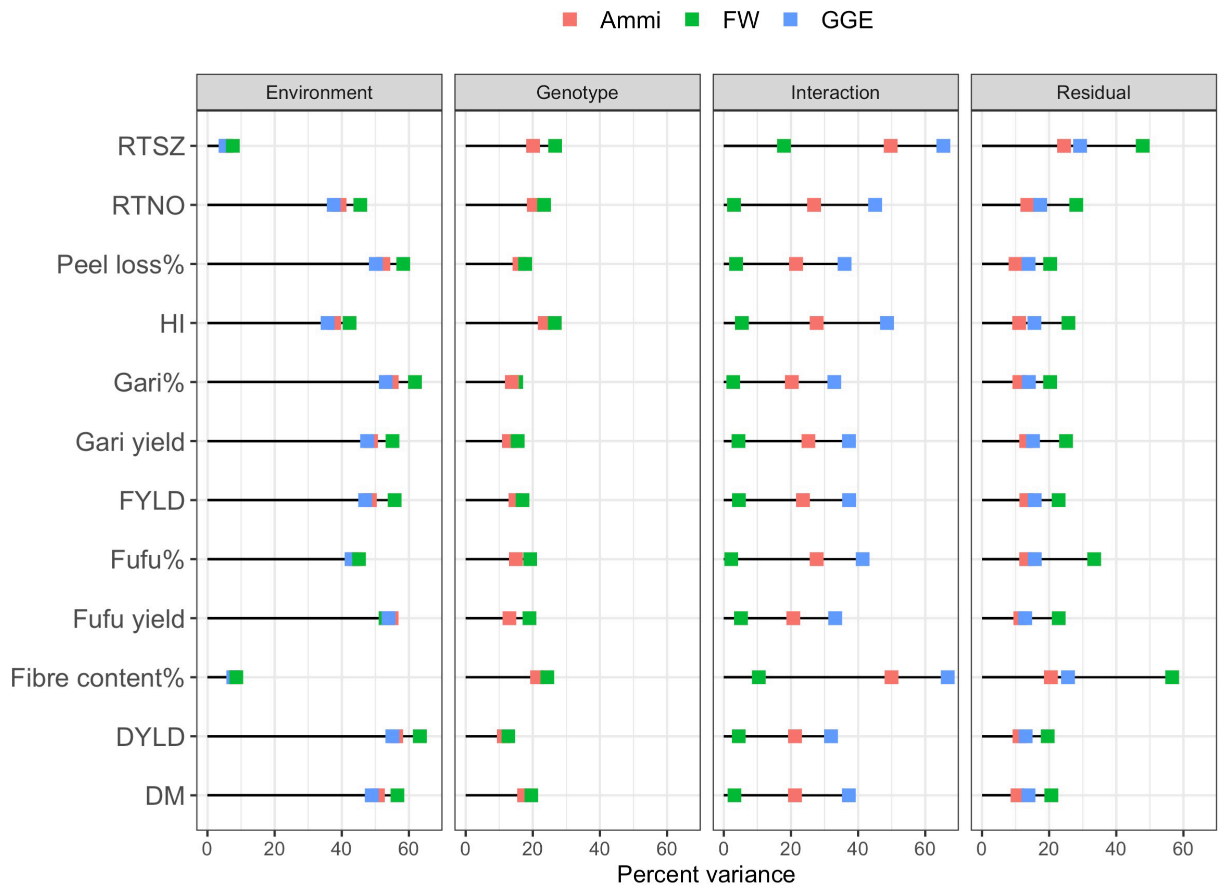


FIGURE 6

Partitioning of total sum of square (TSS) based on total sum of squares captured by each factor from fitting additive main effect and multiplicative interaction (AMMI) (represented by the orange square), Finlay Wilkinson (FW) (represented by green square), and genotype and genotype by environment (GGE) (represented by blue square) models on 67 elite cassava genotypes evaluated in 8 environments. DM, Drymatter; RTNO, Root number; DYLD, Dry root yield; FYLD, Fresh root yield; HI, Harvest index; RTSZ, Root size.

Stability of genotype performance across environments

Wrickle's ecovalence

According to Wrickle (1962) genotypes with low ecovalence index have smaller fluctuation in performance across environments and are desirable because they are more stable. These genotypes, G13, TMS980581, G40, G47, and G12 had the lowest Wi less than 0.48 for gari while G50, G52, G14 G8, and G18 had the lowest Wi less than 0.17 for fufu% (**Supplementary Table 7**). These genotype had limited differential response to the different environments used in this study. Among the genotypes that showed low Wi for gari%, G40, G47, and G12 while G50, G14, and G18 for fufu% were above the population mean and can be recommended for wide adaptation (Seife and Tena, 2020). Unstable genotypes for gari% include G3, G39, G31, G6, and G21 while for fufu% include TMS982101, G26, G56, G21, and G31.

Geometric adaptability index

GAI is used to evaluate genotypes stability based on the geometric mean of genotypes across environments; thus, genotypes with high GAI and low GAI rank are desired (Mohammadi and Amri, 2008; Pourdad, 2011). Results from the GAI (**Supplementary Table 7**), top 5 genotypes with the lowest ranks for gari% includes G56, G50, G40, G23, and G55 while top 5 genotypes for the lowest ranks for fufu% include G56, G24, G21,TMEB419, and G23. These genotypes had relatively high gari and fufu% above the population mean of 20.10 and 20.93%, respectively. Genotypes with the highest ranks for gari% G45, G44, G49, G35, and G37 while G52, G8, G29, G27, and G30 had the highest rank for fufu%.

Discussion

Accessing genetic variability and heritability, GEI pattern and relationship between processing traits and key agronomic

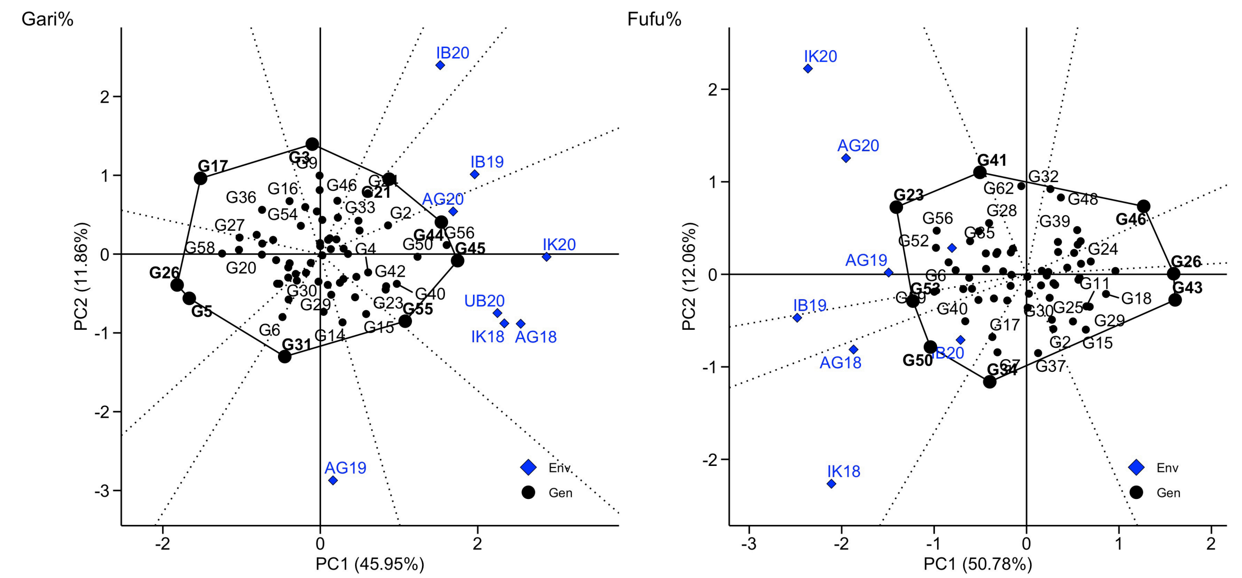


FIGURE 7
Vector views of PC2 are plotted against PC1 for environment relations, winning genotypes, and environment correlations.

variables is an important step in trait genetic improvement through recurrent selection schemes. In this study we assessed 12 traits including processed product and key agronomic variable performance of 62 breeding lines compared with five checks, in multi-environment field trials. The environments used in this study include major areas where cassava is converted into processed products and falls under two major agro-ecological zones in Nigeria. The study population revealed large phenotypic variation among genotypes in regards to all traits and between trials (**Figure 1**). The range of genetic variability observed among genotypes in this study for gari% was in range with what (Ibe and Ezedinma, 1981; Achinewhu et al., 1998; Amoah et al., 2010) reported for five genotypes. However, the range of genetic variability measured for gari% (11.82–25.27%) in this study was lower than what was reported for the two traits harvested at different age which ranged from 22.00 to 52.00% (Adegbola et al., 2014). The range of genetic variability observed among genotypes in this study for fufu % was similar with what (Awoyale et al., 2020) reported for ten improved varieties to assess their suitability for fufu production. Therefore, genetic improvement can be done through recurrent selection for conversion rate of processed products.

Selection efficiency can be improved with a proper understanding of the relationships between traits. The correlation analysis done between processing traits and key agronomic variables reveal a strong positive correlations between gari and fufu% with dry matter content which connotes that dry matter content could be used as a proxy selection parameter to evaluate genotypes which agrees with the results of Laya et al. (2018). The findings in this study supports previous suggestions from Apea-Bah et al. (2009) and Teeken et al. (2021) that an increase in dry matter content would increase conversion rates of processed products. The correlation observed between gari and fufu% with dry matter content is not surprising because the final stage of root conversion into gari% (Frying) and fufu% (oven drying) aim to remove moisture for maximum increase in shelf life of products.

Contrary to expectations, we did not find any relationship between FYLD and gari and fufu% meaning that genotypes with high FYLD do not translate to high root conversion rate for both processed products in the population used for this study. This finding is in line with what Ibe and Ezedinma (1981) reported for 12 cassava cultivars. However, FYLD is still an essential trait for cassava breeding as it is the first attraction to farmers to a variety before processed products' potential. Another interesting relationship found in this study was the 1:1 ratio between gari% and fufu% with peel loss% (**Figure 4**) which was similar to what Hahn (1989) reported for 12 varieties. Furthermore, RTSZ may not be directly related to gari and fufu% but it is a strong determinant for peel loss% making root size an essential component for selection when breeding for high Gari% and Fufu%.

Both methods of heritability estimates (H2_Standard and H2_Cullis) for traits revealed high to moderate estimates which emphasizes genetic contribution to these traits in the population used in this study. However, we observed low heritability estimates from the H2_Cullis methods for some trials and should not be misunderstood as a consequence of no genetics contributing to the expression of the trait but may be due to the effect of either environment or processing. There is room for improvement on achieving increased heritability estimates for these traits by further optimizing processing methods or high throughput phenotyping methods.

According to the linear mixed model, FW, AMMI and GGE analysis of variance, a significant genotype effect was observed for the traits measured; this indicated that genotypes were significantly different, hence genetic improvement could be achieved through hybridization and recurrent selection. Furthermore, we observe a change in genotypic performance in different trials which was confirmed by the largest percentage of total sums of squares and the significant effects of environment in all models for gari and fufu%. Also, a significant effect of GEI was found for these traits, thereby complicating the breeders' quest for developing a stable variety, because neither genotype nor environment effect can effectively capture all variation observed. This will require testing of genotypes in diverse environments before selection can be made. The GGE biplot enables a visual comparison of the locations and genotypes, and their interrelationships and performance potential of genotypes. The biplot (**Figure 7**) explained 57.81 and 61.05% of the total G + GE interactions of gari and fufu%, respectively, indicating that there is more environment and interaction contribution to performance of a genotype.

Genotypes that are above population means and stable must be considered as selection candidates simultaneously to exploit the beneficial effects of GEI and to have a more accurate selection for traits improvement. Based on genotypic performance across environment and traits, different genotypes emerged as stable performers for different traits as shown by the computed Wi and GAI (**Supplementary Table 7**). The genotype ranking from both stability analyses suggests that there are some generic relationships between traits which supports the correlation analysis done earlier in this study. However, Wi and GAI ranked genotypes differently as their top 5, this is not surprising as previous studies suggest that Wi and GAI are negatively correlated in measuring stability (Mohammadi and Amri, 2008; Mohammadi and Nader Mahmoodi, 2008). Notwithstanding, Wi and GAI ranked G40 as part of top5 genotypes for gari% but there were no overlapping genotypes for fufu% in the study. For plant breeding purposes, the dynamic stability measure is preferred for genotype selection and trait improvement because it assumes that all genotypes responds differently to change in environmental conditions (Changizi et al., 2014). Therefore, genotypes selected as top 5 using GAI for gari% (G56, G50, G40, G23, and G55) and fufu% (G56, G24, G21, TMEB419, and G23) are recommended.

Conclusion

There is a measurable degree of genetic variation among genotypes for the root conversion rate for gari and fufu%, making it possible to make progress through conventional selection and advancement of clonal genotypes. However, there is a need for further optimization of the data collection process and introducing high throughput data collection methods. Dry Matter content and Peel loss% had the highest correlation and could be used as a selection proxy for gari and fufu conversion rate. However, the rate of the genetic gain obtained per year might not be the same for as recorded for Dry Matter content and needs further investigation. It would be interesting to know if dry matter and root conversion rate of gari and fufu is controlled by the same genomic region/s in the cassava genome and the influence of dry matter content on quality of processed products. Apart from genetic variation, environment and interaction had a huge role to play in gari and fufu% in this study. Environmental variance is typically the most prominent most significant component of variance in populations in natural conditions. This genotypic performance suggests that a genotype with dual-purpose for high percent gari and fufu conversion rate can be bred for using one or two of the correlated environments in addition to a contrasting environment for evaluation. We have identified genotypes that performed better than the checks used in this study for gari and fufu%.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author/s.

Author contributions

CA, IR, and, PK: design and study conceptualization. CA, IR, and IK: study methodology, implementation, and manuscript drafting. CA, IR, IK, and MB: formal data curation and analysis. IR, IK, SE, MB, and EP: manuscript reviewing and editing. IR, SE, IA, SO, and EP: supervision, coordination, and fund acquisition. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Deutscher AkademischerAustauschdienst (DAAD) to pursue a Ph.D in Plant Breeding at the West Africa Centre for Crop Improvement (WACCI), University of Ghana, Legon, Accra, Ghana. We acknowledged the support from Next Generation Cassava Breeding Project (Grant INV-007637) through funds received from the Bill and Melinda Gates Foundation

(BMGF) and the United Kingdom's Foreign, Commonwealth & Development Office (FCDO).

Acknowledgments

We extend our appreciation to the dedicated staff of the International Institute of Tropical Agriculture (IITA) of the cassava breeding program that assisted with the fieldwork and processing center. We express our sincere gratitude to reviewers for their valuable time and suggestions to improve the manuscript. We Hope the published version honors their effort.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.974795/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Vector views of PC2 are plotted against PC1 fufu_yield.

SUPPLEMENTARY FIGURE 2
Vector views of PC2 are plotted against PC1 Gari_yield.

SUPPLEMENTARY FIGURE 3
Vector views of PC2 are plotted against PC1 Harvest_index.

SUPPLEMENTARY FIGURE 4
Vector views of PC2 are plotted against PC1 Root size.

SUPPLEMENTARY FIGURE 5
Vector views of PC2 are plotted against PC1 Root number.

SUPPLEMENTARY FIGURE 6
Vector views of PC2 are plotted against PC1 Dry root yield.

SUPPLEMENTARY FIGURE 7
Vector views of PC2 are plotted against PC1 Fresh root yield.

SUPPLEMENTARY FIGURE 8
Vector views of PC2 are plotted against PC1 Fiber content.

SUPPLEMENTARY FIGURE 9
Vector views of PC2 are plotted against PC1 Peel loss.

SUPPLEMENTARY FIGURE 10
Vector views of PC2 are plotted against PC1 Dry matter content.

References

Achi, O. K., and Akomas, N. S. (2006). Comparative assessment of fermentation techniques in the processing of fufu, a traditional fermented cassava product. *Pak. J. Nutr.* 5, 224–229. doi: 10.3923/pjn.2006.224.229

Achinewhu, S. C., Barber, I., and Ljeoma, I. O. (1998). *Plant Foods Hum Nutr.* 52: 133–40. doi: 10.1023/a:1008029101710

Adegbola, A. G., Dauda, M., and Rahmatallah, A. (2014). Yield, suitability and sensory evaluation of gari produced from two cassava varieties at different age. *Int. J. Nat. Sci. Curr. Future Res. Trends* 7, 1–10.

Akingbala, J. O., Oguntimein, G. B., and Abass, A. B. (1991). Effect of processing methods on quality and acceptability of fufu from low cyanide cassava. *J. Sci. Food Agric.* 57, 151–154. doi: 10.1002/jsfa.2740570118

Amoah, R. S., Sam-Amoah, L. K., Boahen, C. A., and Duah, F. (2010). Estimation of the material losses and gari recovery rate during the processing of varieties and ages of cassava into gari. *Asian J. Agric. Res.* 4, 71–79. doi: 10.3923/ajar.2010.71.79

Apea-Bah, F. B., Oduro, I., Ellis, W. O., and Safo-Kantanka, O. (2009). Principal components analysis and age at harvest effect on quality of gari from four elite cassava varieties in Ghana. *Afr. J. Biotechnol.* 8, 1943–1949.

Awoyale, W., Alamu, E. O., Chijioke, U., Tran, T., Takam Tchuente, H. N., Ndjouenkeu, R., et al. (2021). A review of cassava semolina (gari and eba) end-user preferences and implications for varietal trait evaluation. *Int. J. Food Sci. Technol.* 56, 1206–1222. doi: 10.1111/ijfs.14867

Awoyale, W., Asiedu, R., Kawalawu, W. K., Abass, A., Maziya-Dixon, B., Kromah, A., et al. (2020). Assessment of the suitability of different cassava varieties for gari and fufu flour production in Liberia. *Asian Food Sci. J.* 14, 36–52. doi: 10.9734/afsj/2020/v14i230128

Bassey, E. E. (2018). Evaluation of nine elite cassava (Manihot esculenta Crantz) genotypes for tuber and gari yields and gari quality in four locations in Akwa Ibom State, Nigeria. *Am. Res. J. Agric.* 4, 1–10. doi: 10.21694/2378-9018.18003

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2017). *ASReml-R reference manual version 4*. Hemel Hempstead: VSN International Ltd.

Cardoso, A. P., Mirione, E., Ernesto, M., Massaza, F., Cliff, J., Rezaul Haque, M., et al. (2005). Processing of cassava roots to remove cyanogens. *J. Food Compos. Anal.* 18, 451–460. doi: 10.1016/j.jfca.2004.04.002

Ceballos, H., Kulakow, P., and Hershey, C. (2012). Cassava breeding: Current status, bottlenecks and the potential of biotechnology tools. *Trop. Plant Biol.* 5, 73–87. doi: 10.1007/s12042-012-9094-9

Ceballos, H., Rojanaridpiched, C., Phumichai, C., Becerra, L. A., Kittipadakul, P., Iglesias, C., et al. (2020). Excellence in cassava breeding: Perspectives for the future. *Crop Breed. Genet. Genom.* 2:e200008.

Changizi, M., Choukan, R., Heravan, E. M., Bihamta, M. R., and Darvish, F. (2014). Evaluation of genotype× environment interaction and stability of corn hybrids and relationship among univariate parametric methods. *Can. J. Plant Sci.* 94, 1255–1267. doi: 10.4141/cjps2013-386

Coulibaly, O. N., Arinloye, A. D., Faye, M., and Abdoulaye, T. (2014). *Regional cassava value chains analysis in West Africa: Case study of Nigeria*. Dakar: West and Central African Council for Agricultural Research and Development. doi: 10.13140/2.1.2510.2403

Cullis, B. R., Thomson, F. M., Fisher, J. A., Gilmour, A. R., and Thompson, R. (1996). The analysis of the NSW wheat variety database. I. Modelling trial error variance. *Theor. Appl. Genet.* 92, 21–27. doi: 10.1007/BF00222947

Dusunceli, F. (2019). *Protecting cassava, a neglected crop, from pests and diseases.* 2.

Eggleston, G., and Asiedu, R. (1994). Effect of boiling on the texture of cassava clones: A comparison of compressive strength, intercellular adhesion and physiochemical composition of the tuberous roots. *Trop. Sci.* 34, 259–273.

Etudaiye, H. A., Nwabueze, T. U., and Sanni, L. O. (2009). Quality of fufu processed from cassava mosaic disease (CMD) resistant varieties. *Afr. J. Food Sci.* 3, 061–067.

Ezedinma, C., Ojioko, I. A., Okechukwu, R. U., Lemchi, J., Umar, A. M., Sanni, L. O., et al. (2007). *The cassava food commodity market and trade network in Nigeria*. Ibadan: IITA.

Ezemenari, K., Nweke, F., and Strauss, J. (1998). *Consumption patterns and expenditure elasticities of demand for food staples in rural Africa—focus on cassava growing areas*. COSCA Working Paper No 24. Ibadan: International Institute of Tropical Agriculture.

Falcon, C. M., Kaeppler, S. M., Spalding, E. P., Miller, N. D., Haase, N., AlKhalifah, N., et al. (2020). Relative utility of agronomic, phenological, and morphological traits for assessing genotype-by-environment interaction in maize inbreds. *Crop Sci.* 60, 62–81. doi: 10.1002/csc2.20035

Finlay, K., and Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Aust. J. Agric. Res.* 14:742. doi: 10.1071/AR9630742

Fukuda, W. M. G., Guevara, C. L., Kawuki, R., and Ferguson, M. E. (2010). *Selected morphological and agronomic descriptors for the characterization of cassava*. Ibadan: IITA, 28.

Garrick, D. J., Taylor, J. F., and Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evolution* 41, 1–8. doi: 10.1186/1297-9686-41-55

Gauch, H. G. Jr. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Amsterdam: Elsevier Science Publishers.

Hagos, H. G., and Abay, F. (2013). Ammi and gge biplot analysis of bread wheat genotypes in the northern part of Ethiopia. *J. Plant Breed. Gen.* 1, 12–18.

Hahn, S. K. (1989). An overview of african traditional cassava processing and utilization. *Outlook Agric.* 18, 110–118. doi: 10.1177/003072708901800303

Ibe, D. G., and Ezedinma, F. O. C. (1981). *Gari yield from cassava: Is it a fonction of root yield? in tropical root crops: Research strategies for the 1980s: Proceedings of the first triennial root crops symposium of the international society for tropical root crops-Africa branch, 8-12 Sept. 1980*. Ottawa, ON: IDRC.

Iloeje, N. P. (1965). *A new geography of West Africa*. Ikeja Lagos: Longman.

Kempthorne, O. (1970). *An introduction to genetic statistics*. Ames, IA: Iowa State University Press, xvii+545.

Lancaster, P. A., Ingram, J. S., Lim, M. Y., and Coursey, D. G. (1982). Traditional cassava-based foods: Survey of processing techniques. *Econ. Bot.* 36, 12–45. doi: 10.1007/BF02858697

Laya, A., Koubala, B. B., Kouninki, H., and Nchiwan Nukenine, E. (2018). Effect of harvest period on the proximate composition and functional and sensory properties of gari produced from local and improved cassava (*Manihot esculenta*) Varieties. *Int. J. Food Sci.* 2018, 1–15. doi: 10.1155/2018/6241035

Lian, L., and de los Campos, G. (2016). FW: An R Package for Finlay–Wilkinson Regression that Incorporates Genomic/Pedigree Information and Covariance Structures Between Environments. *G3 Genes Genomes Genet.* 6, 589–597. doi: 10.1534/g3.115.026328

Malosetti, M., Ribaut, J.-M., and van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Front. Physiol.* 4:44. doi: 10.3389/fphys.2013.00044

Mohammadi, R., and Amri, A. (2008). Comparison of parametric and non-parametric methods for selecting stable and adapted durum wheat genotypes in variable environments. *Euphytica* 159, 419–432. doi: 10.1007/s10681-007-9600-6

Mohammadi, R., and Nader Mahmoodi, K. (2008). Stability analysis of grain yield in barley (*Hordeum vulgare* L.). *Int. J. Plant Breed.* 2, 74–78.

Montagnac, J. A., Davis, C. R., and Tanumihardjo, S. A. (2009). Nutritional value of cassava for use as a staple food and recent advances for improvement. *Compr. Rev. Food Sci. Food Saf.* 8, 181–194. doi: 10.1111/j.1541-4337.2009.00077.x

Mulusew, F. D. J. B., Tadele, T., and Amsalu, A. (2014). Comparison of biometrical methods to describe yield stability in field pea (*Pisum sativum* L.) under south eastern Ethiopian conditions. *Afr. J. Agric. Res.* 9, 2574–2583. doi: 10.5897/AJAR09.602

Nweke, F. I. (1994). *Processing potential for cassava production growth in Africa*. COSCA. doi: 10.17660/ActaHortic.1994.380.49

Nweke, F. I. (2004). *New challenges in the cassava transformation in Nigeria and Ghana*. EPTD Discussion Paper No. 118. Washington DC: International Food Policy Research Institute.

Oghenechavwuko, U. E., Saka, G. O., Adekunbi, T. K., and Taiwo, A. C. (2013). Effect of processing on the physico-chemical properties and yield of gari from dried chips. *J. Food Process. Technol.* 4, 01–06.

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

Okpokiri, A. O., Ijioma, B. C., Alozie, S. O., and Ejiofor, M. A. N. (1985). Production of improved cassava fufu. *Niger. Food J.* 2:3.

Olivoto, T., and Lúcio, A. D. (2020). metan: An R package for multi-environment trial analysis. *Methods Ecol. Evol.* 11, 783–789. doi: 10.1111/2041-210X.13384

Oluwafemi, G. I., and Udeh, C. C. (2016). Effect of fermentation periods on the physicochemical and sensory properties of gari. *J. Environ. Sci.* 10, 37–42.

Omosuli, S., Ikujenlola, A., and Abisuwa, A. (2017). Quality assessment of stored fresh cassava roots and 'fufu' flour produced from stored roots. *J. Food Sci. Nutr. Ther.* 3, 009–013. doi: 10.17352/jfsnt.000008

Onitilo, M. O., Sanni, L. O., Oyewole, O. B., and Maziya-Dixon, B. (2007). Physicochemical and functional properties of sour starches from different cassava varieties. *Int. J. Food Properties* 10, 607–620. doi: 10.1080/10942910601048994

Oyeyinka, S. A., Adeloye, A. A., Smith, S. A., Adesina, B. O., and Akinwande, F. F. (2019). Physicochemical properties of flour and starch from two cassava varieties. *Agrosearch* 19, 28–45.

Pourdad, S. S. (2011). Repeatability and relationships among parametric and non-parametric yield stability measures in safflower (*Carthamus tinctorius* L.) genotypes. *Crop Breed. J.* 1, 109–118.

R Core (2020). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Rife, T. W., and Poland, J. A. (2014). Field book: An open-source application for field data collection on android. *Crop Sci.* 54, 1624–1627. doi: 10.2135/cropsci2013.08.0579

Sanchez, T., Dufour, D., Moreno, I. X., and Ceballos, H. (2010). Comparison of pasting and gel stabilities of waxy and normal starches from potato, maize, and rice with those of a novel waxy cassava starch under thermal, chemical, and mechanical stress. *J. Agric. Food Chem.* 58, 5093–5099. doi: 10.1021/jf1001606

Sanni, L. O., Akingbala, J. O., Oguntunde, A. O., Bainbridge, Z. A., Graffham, A. J., and Westby, A. (1998). Processing of fufu from cassava in Nigeria: Problems and prospects for development. *Sci. Technol. Dev.* 16, 58–71.

Sanni, L. O., Oyewole, O. B., Adebowale, A. R. A., and Adebayo, K. (2003). Current trends in the utilization of roots and tubers for sustainable development. *Food Based Approaches Healthy Nutr.* 11, 23–28.

Seife, A., and Tena, E. (2020). Genotype x environment interaction and yield stability analysis of sugarcane (*Saccharum officinarum* L.) genotypes. *Int. J. Adv. Res. Biol. Sci.* 7, 14–26.

Solonechnyi, P., Vasko, N., Naumov, A., Solonechnaya, O., Vazhenina, O., Bondareva, O., et al. (2015). GGE biplot analysis of genotype by environment interaction of spring barley varieties. *Zemdirbyste Agric.* 102, 431–436. doi: 10.13080/z-a.2015.102.055

Swanckaert, J., Akansake, D., Adofo, K., Acheremu, K., De Boeck, B., Eyzaguirre, R., et al. (2020). Variance component estimations and mega-environments for sweetpotato breeding in West Africa. *Crop Sci.* 60, 50–61. doi: 10.1002/csc2.20034

Temesgen, M., Alamerew, S., Eticha, F., and Mehari, M. (2015). Genotype X environment interaction and yield stability of bread wheat genotypes in South East Ethiopia. *World J. Agric. Sci.* 11, 121–127.

Teeken, B., Agbona, A., Bello, A., Olaosebikan, O., Alamu, E., Adesokan, M., et al. (2021). Understanding cassava varietal preferences through pairwise ranking of gari?eba and fufu prepared by local farmer-processors. *Int. J. Food Sci. Technol.* 56, 1258–1277.

Thresh, J. M., Otim-Nape, G. W., Legg, J. P., and Fargette, D. (1997). African cassava mosaic virus disease: The magnitude of the problem. *Afr. J. Root Tuber Crops* 2, 13–19.

Ukhun, M. (1989). The hydrocyanic acid (HCN) content of garri flour made from cassava (*Manihot* spp.) and the influence of length of fermentation and location of source. *Food Chem.* 33, 107–113. doi: 10.1016/0308-8146(89)90113-1 van Rossum, B.-J., van Eeuwijk, F., Boer, M., and Malosetti, M. (2021). *Package 'statgenGxE.' CRAN*. Available online at: https://CRAN.R-project.org/package=statgenGxE (accessed April, 23 2022).

Westby, A. (2002). "Cassava utilization, storage and small-scale processing," in *Cassava: Biology, production and utilization*, eds R. J. Hillocks and J. M. Thresh (Wallingford: CABI), 281–300. doi: 10.1079/9780851995243.0281

Wickham, H. (2016). *Getting Started with ggplot2," in ggplot2*. Berlin: Springer, 11–31. doi: 10.1007/978-3-319-24277-4_2

Wolfe, M. D., Del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., et al. (2017). Prospects for genomic selection in cassava breeding. *Plant Genome* 10, 1–19. doi: 10.3835/plantgenome2017.03.0015

Wossen, T., Girma, G., Abdoulaye, T., Rabbi, I., Olanrewaju, A., Kulakow, P., et al. (2017). *The cassava monitoring survey in Nigeria*, Vol. 80. Ibadan: International Institute of Tropical Agriculture (IITA).

Wricke, G. (1962). Uber eine Methode zur Erfassung der okologischen Streubreite in Feldverzuchen. *Z. Pflanzenzuech* 47, 92–96.

Yan, W., and Hunt, L. A. (2001). Interpretation of genotype ϰ environment interaction for winter wheat yield in Ontario. *Crop Science* 41:7. doi: 10.2135/cropsci2001.41119x

Yan, W., and Rajcan, I. (2002). Biplot analysis of test sites and trait relations of soybean in Ontario. *Crop Sci.* 42, 11–20. doi: 10.2135/cropsci2002.1100

Yan, W., Hunt, L. A., Sheng, Q., and Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci.* 40, 597–605. doi: 10.2135/cropsci2000.403597x

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 November 2022
DOI 10.3389/fpls.2022.990250

Check for updates

OPEN ACCESS

EDITED BY
Valerio Hoyos-Villegas,
McGill University, Canada

REVIEWED BY
Dinakaran Elango,
Iowa State University, United States
Xu Wang,
University of Florida, United States

*CORRESPONDENCE
Edwige Gaby Nkouaya Mbanjo
e.mbanjo@cgiar.org

SPECIALTY SECTION
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

RECEIVED 09 July 2022
ACCEPTED 14 September 2022
PUBLISHED 08 November 2022

CITATION
Mbanjo EGN, Hershberger J, Peteti P, Agbona A, Ikpan A, Ogunpaimo K, Kayondo SI, Abioye RS, Nafiu K, Alamu EO, Adesokan M, Maziya-Dixon B, Parkes E, Kulakow P, Gore MA, Egesi C and Rabbi IY (2022) Predicting starch content in cassava fresh roots using near-infrared spectroscopy. *Front. Plant Sci.* 13:990250. doi: 10.3389/fpls.2022.990250

Predicting starch content in cassava fresh roots using near-infrared spectroscopy

Edwige Gaby Nkouaya Mbanjo[1*], Jenna Hershberger[2], Prasad Peteti[1], Afolabi Agbona[1,3], Andrew Ikpan[1], Kayode Ogunpaimo[1], Siraj Ismail Kayondo[1], Racheal Smart Abioye[1], Kehinde Nafiu[1], Emmanuel Oladeji Alamu[1], Michael Adesokan[1], Busie Maziya-Dixon[1], Elizabeth Parkes[1], Peter Kulakow[1], Michael A. Gore[4], Chiedozie Egesi[1,4,5] and Ismail Yusuf Rabbi[1]

[1]International Institute of Tropical Agriculture (IITA), Ibadan, Oyo State, Nigeria, [2]Department of Plant and Environmental Sciences, Pee Dee Research and Education Center, Clemson University, Florence, SC, United States, [3]Molecular & Environmental Plant Sciences, Texas A&M University, College Station, TX, United States, [4]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States, [5]National Root Crops Research Institute (NRCRI), Umuahia, Nigeria

The cassava starch market is promising in sub-Saharan Africa and increasing rapidly due to the numerous uses of starch in food industries. More accurate, high-throughput, and cost-effective phenotyping approaches could hasten the development of cassava varieties with high starch content to meet the growing market demand. This study investigated the effectiveness of a pocket-sized SCiO™ molecular sensor (SCiO) (740–1070 nm) to predict starch content in freshly ground cassava roots. A set of 344 unique genotypes from 11 field trials were evaluated. The predictive ability of individual trials was compared using partial least squares regression (PLSR). The 11 trials were aggregated to capture more variability, and the performance of the combined data was evaluated using two additional algorithms, random forest (RF) and support vector machine (SVM). The effect of pretreatment on model performance was examined. The predictive ability of SCiO was compared to that of two commercially available near-infrared (NIR) spectrometers, the portable ASD QualitySpec® Trek (QST) (350–2500 nm) and the benchtop FOSS XDS Rapid Content™ Analyzer (BT) (400–2490 nm). The heritability of NIR spectra was investigated, and important spectral wavelengths were identified. Model performance varied across trials and was related to the amount of genetic diversity captured in the trial. Regardless of the chemometric approach, a satisfactory and consistent estimate of starch content was obtained across pretreatments with the SCiO (correlation between the predicted and the observed test set, (R^2_{P}): 0.84–0.90; ratio of performance deviation (RPD): 2.49–3.11, ratio of performance to interquartile distance (RPIQ): 3.24–4.08, concordance correlation coefficient (CCC): 0.91–0.94). While PLSR and SVM showed comparable prediction abilities, the RF model yielded the lowest performance. The heritability of the 331 NIRS spectra varied across trials and spectral regions but was highest ($H^2 > 0.5$) between 871–1070 nm in most trials. Important wavelengths corresponding to absorption bands associated with starch and water were identified from 815 to 980 nm. Despite its limited spectral range, SCiO provided satisfactory prediction, as did BT, whereas QST showed less optimal calibration models. The SCiO spectrometer may be a cost-effective solution for phenotyping the starch content of fresh roots in resource-limited cassava breeding programs.

KEYWORDS

cassava, starch, NIRS, spectrophotometers, SCiO, spectra, heritability

Introduction

The global starch market is experiencing increased demand, with an estimated value of US$ 51.5 billion in 2021 and a projected value of US$70.5 billion by 2027[1]. Starch is a polysaccharide that plants produce as a carbohydrate reserve. Approximately 54% of the starches produced globally are used for food. In comparison, 46% are used in non-food products such as textiles, pharmaceuticals, pulp and paper, adhesives for packing industries, and cosmetics manufacturing (Omojola, 2013; Desta and Tigabu, 2018; Raji, 2020). Cassava starch, with its excellent characteristics and favorable physicochemical and functional properties, could be an alternative source of starch in a market traditionally dominated by cereal and potato starches (Oladunmoye et al., 2014; Spencer and Ezedinma, 2017; Chisenga et al., 2019).

Cassava (*Manihot esculanta* Crantz) is a climate-resilient crop owing to its tolerance to drought, poor soils, and wide adaptability to various climate and cropping systems. It is also a poverty alleviating crop, primarily grown for human consumption. Cassava is gradually evolving into an industrial crop (Chisenga et al., 2019). The significantly increased demand for starch and starch-based products combined with the inability of traditional exporters to meet market demand provides new opportunities for the crop in sub-Saharan Africa. Cassava has the potential to contribute to income, social progress and development, and economic growth in countries that produce it (Dada, 2016). Therefore, cassava production in sub-Saharan Africa should increase to meet rising market demand. In the face of resource depletion, land scarcity, urbanization, and rapid population growth, increasing starch production by expanding cassava cultivation land areas is not a sustainable solution. An alternative solution to close the demand gap is developing high starch content cassava varieties.

Breeding efforts for cassava varieties with high starch content could be accelerated by developing high-throughput phenotyping tools that can rapidly and precisely assess numerous genotypes at an early stage. Phenotyping remains one of the major limitations hindering the power of genetic analysis of key traits and accurate selection of superior genotypes at all stages of the breeding process (Cobb et al., 2013; Reynolds et al., 2020). Several high-throughput, non-invasive phenotyping technologies, such as image analysis (Baek et al., 2020), satellite imaging, and remote sensing with unoccupied aerial vehicles (Chawade et al., 2019), have recently been developed, opening up new opportunities in breeding. In cassava, spectroscopy-based approaches that use near-infrared (NIR) regions of the electromagnetic spectrum have shown promise for the rapid estimation of key traits (Sánchez et al., 2014; Abincha et al., 2021; Hershberger et al., 2022). Near-infrared technology could replace the laborious and time-consuming approach currently used for root starch content quantification.

NIR spectroscopy studies the spectral properties of an object when exposed to electromagnetic radiation. Light from the NIR region may be absorbed, reflected, or transmitted. The resulting spectrum is associated with molecular vibrational excitation caused by overtones and a combination of a specific set of chemicals bound from within a molecule (Ozaki et al., 2020; Beć et al., 2021). The NIR region is further classified into three sub-categories: region I (800–1200 nm), also known as the Herschel region, region II (1200 –1800 nm), and region III (1800–2500 nm). Technological advancement has fostered the development of miniaturized NIR devices with limited spectral ranges but offering significant advantages in terms of price and portability over traditional spectrometers with full spectral ranges. However, these advantages may come at the expense of accuracy and robustness. As a result, such devices must be assessed for analytical performance and model reliability (Ozaki et al., 2020; Beć et al., 2021).

Our study investigates the potential of a miniaturized SCiO™ spectrometer as an alternative phenotyping method for determining starch content in fresh cassava roots. Cassava clones (hereafter referred to as genotypes) from 11 trials were harvested and their starch was extracted and quantified. Using three chemometric modeling

[1] https://www.businesswire.com/news/home/20220408005379/en/Global-Industrial-Starches-Market-2022-to-2027—Growth-Trends-COVID-19-Impact-and-Forecasts—ResearchAndMarkets.com

approaches, random forest (RF), support vector machine (SVM), and partial least squares regression (PLSR), the relationship between reference values and NIR spectra collected with the SCiO™ molecular sensor was established. The heritability of individual wavelengths was investigated to determine the degree to which variation for a wavelength is due to genetic variation among genotypes. The most effective wavelengths in this experiment for predicting starch content were identified using variable importance analysis. The SCiO™ sensor's performance was compared to two different NIRS instruments: the portable ASD QualitySpec® Trek and the benchtop FOSS XDS Rapid Content™ Analyzer. It was established that SCiO™ could be a rapid analytic tool for measuring starch content, allowing breeders to screen large populations at an early stage.

Materials and methods

Plant material

The set of genotypes used in this study was composed of genotypes from preliminary yield trials (PYTs), advanced yield trials (AYTs), uniform yield trials, regional nationally coordinated research program (NCRP) trials, and genomic selection (GS) cycles (Table 1). These trials were established across three locations in Nigeria (Ikenne, Ibadan, and Ago-Owu) in the 2019 and 2020 rainy seasons. Mature roots were harvested 12 months after planting (MAP). In total, 344 unique genotypes from 11 field trials were evaluated. These included an early-stage evaluation (PYT Trial A) with 174 unique genotypes planted in one environment; three late-stage evaluation trials (UYT and AYT Trials J, H, I) with between 36 and 40 genotypes planted in two replicates in a single environment; a pre-release trial (NCRP Trials E, F) with 18 unique genotypes planted in three replicates across two environments; and trials of two germplasm collections maintained by the IITA cassava breeding program. The first collection comprised a popular landrace and improved varieties widely cultivated in Nigeria with 33 unique genotypes and was planted in two replicates across three environments (Trial C, D, G). The second collection, which comprised 52 unique genotypes, was planted in replicated PYTs across two environments (Trial B and K). This second collection (also considered a "core collection") was selected from a large pool

TABLE 1 Cassava breeding field trial metadata and summary statistics for root starch content.

Cassava base trial name	Abbreviated trial name	Date of planting	Date of harvest	Trial type[a]	Trial design[b]	Location	Min[b]	Max[c]	SD[d]	CV[e]	Plots used	Unique genotypes
19.GS.C4B.PYT.500.IK	Trial A	4 Aug.2019	27.Oct. 2020	PYT	RCBD	Ikenne	18.7	41.6	3.45	0.11	261	174
19.CASS.PYT.52.IK	Trial B	25 June 2019	27 Oct.2020	PYT	RCBD	Ikenne	4.1	38.8	8.07	0.37	97	50
19.CMSSurveyVarieties.AYT33.IK	Trial C	10 May 2019	23 Apr. 2020	AYT	Alpha-Lattice	Ikenne	19.9	37.8	4.21	0.15	65	32
19.CMSSurveyVarieties.AYT.33.IB	Trial D	29 Apr 2019	20 Apr. 2020	AYT	Alpha-Lattice	Ibadan	10.1	30.3	5.01	0.25	52	31
19NCRPAG	Trial E	2 July 2019	27 July 2020	NCRP	Alpha-Lattice	Ago-Owu	18.6	35.3	3.6	0.14	36	18
19NCRPIK	Trial F	4 Aug 2019	28 July 2020	NCRP	Alpha-Lattice	Ikenne	20.2	37	3.77	0.13	36	18
20.CMSSurveyVarieties.AYT.33.IB	Trial G	24 Apr 2020	24 Apr. 2021	AYT	Alpha-Lattice	Ibadan	14.3	31.35	4.34	0.18	65	33
20.GS.C2.UYT.36.SetA.IB	Trial H	15 May 2020	15 May 2021	UYT	Alpha-Lattice	Ibadan	14.5	31.8	3.57	0.14	71	36
20.GS.C2.UYT.36.SetB.IB	Trial I	15 May 2020	18 May 2021	UYT	RCBD	Ibadan	17.55	31.65	2.62	0.1	72	36
20.GS.C4B.AYT.40.IB	Trial J	10 June 2020	30 July 2021	AYT	RCBD	Ibadan	13.65	27.85	2.94	0.13	80	40
20.CASS.PYT.52.IK	Trial K	12 July 2020	17 March 2021	PYT	RCBD	Ikenne	10.1	34.6	6.59	0.3	88	50
							4.10	41.6	6.23	0.24	921	518/344

a, minimum; b, maximum, c, standard deviation; d, coefficient of variation.

[a]AYT, Advanced yield trial; PYT, Preliminary yield trial; UYT, Uniform yield trial; NCRP, National coordinate research program.

[b]RCBD, randomized complete block design.

[c]Min, Minimum value; Max, Maximum value.

[d]Standard deviation.

[e]Coefficient of variation.

of historically important breeding lines from IITA (Okechukwu and Dixon, 2008) and contains substantial variation for important agronomic traits, including fresh root yield and starch content.

Reference data measurement

Six healthy storage roots of varying sizes were randomly selected from each plot to ensure representativeness. Selected roots were free of defects such as decomposition, disease, and bruises. These roots were harvested, placed in labeled sampling bags, and immediately taken to the laboratory for starch extraction with the protocol adapted from Matsumoto et al. (2021). Briefly, roots were washed and peeled, and the proximal and distal ends of each root were removed. The top, middle, and bottom sides of each selected root were shredded with a hand grater (3-mm hole diameter). All shredded roots from each plot were mixed together. The starch of individual genotypes was extracted using a wet-milling approach with 3 L of water. One hundred grams of the mixed, shredded tissue was milled with 200 mL of water for one minute with two-second breaks. The slurry was filtered through a sieve with a mesh size of 180 µm. This filtering process was repeated until the residue turned pale white, at which point the remaining water was added to the precipitate. The precipitate was left at room temperature for three hours to allow the starch granules to settle. The supernatant was slowly decanted, and the sediment (starch) was air-dried for 72 hours at room temperature before being oven-dried for 24 hours at 40°C. The RSC, expressed as a percentage of fresh root yield, was calculated by weighing the dry sediments. The amount of starch was determined using the following equation:

$$RSC\,(\,\%\,) = \frac{DSM}{FM} \times 100$$

Where dry starch mass (DSM) is the weight of starch extracted from a known weight of the root matter and fresh root mass (FM) is the known weight of the root matter.

Spectra acquisition

NIR spectra were acquired primarily using a pocket-sized SCiO^{TM} (SCiO) molecular sensor (Consumer physics, Tel Aviv, Israel) that collected spectral information from 740 to 1070 nm with a resolution of 1 nm. The SCiO sensor was synced with a tablet *via* Bluetooth, enabling communication between the two devices for digital data transfer from the SCiO sensor to the SCiO cloud *via* the SCiO smartphone application. The sensor was calibrated before sample capture using a built-in reference standard in the SCiO case. The thoroughly mixed, shredded cassava roots were placed in quartz cell glasses. The SCiO optical shade was connected to the sensor and placed on the top of the cell quartz with the optical head facing down. The light source illuminated the samples and the reflected lighted captured by the detector was uploaded to the online SCiO cloud database. Each genotype was measured in three technical replicates (i.e., three independent tissue samples), and each sample was scanned three times in different positions by rotating the quartz cell glass. The spectra were downloaded as comma-separated value files from the SCiO cloud database. The various repeated scans per sample were averaged and the averaged spectrum was used for further analyses.

Reference data analysis

Descriptive statistics for each trait [minimum and maximum values, coefficient of variation (CV), and standard deviation (SD)] were obtained using the R package pastecs (Grosjean et al., 2018). Boxplots were used to visualize starch variation in each trial. Significant differences (P< 0.05) between the trials were estimated using the Kruskal-Wallis rank test.

Spectra data analysis

The raw spectra were used to classify cassava genotypes into homogeneous groups using principal component analyses (PCA). This analysis was performed using the R package factorMineR (Lê et al., 2008). The PCA plot was visually inspected to identify extreme values, and the two genotypes that deviated from most data were removed. Model development and validation were performed using the R package waves version 0.2.4 (Hershberger et al., 2021). Twelve combinations of mathematical pretreatments, standard normal variate (SNV), first derivative (D1), second derivative (D2), standard normal variate and first derivative (SNV1D), standard normal variate and second derivative (SNV2D), Savitzky-Golay filter (SG), standard normal variate and Savitzky-Golay filter (SNVSG), gap-segment derivative (window size = 11) (SGD1), Savitzky-Golay filter first derivative (window size = 5) (SG.D1W5), Savitzky-Golay filter and first derivative (window size = 11) (SG.D1W11), Savitzky-Golay filter and second derivative (window size = 5) (SG.D2W5), and Savitzky-Golay filter and second derivative (window size = 11) (SG.D2W11), were implemented within the waves R package version 0.2.4 (Hershberger et al., 2021) to minimize the effect of uncontrolled covariates (scattering effects, particle size, variation in the light path, etc.), remove noise from NIR spectra, correct non-linear trends and additive and/or multiplicative effects in the spectrum, and enable a thorough search for optimum prediction. The Mahalanobis distance of each spectrum was calculated, and outliers were removed based on Mahalanobis distance > 3. Individual trials were modeled using PLSR. When data from all 11 trials were combined, two other modeling approaches, RF and SVM with a radial kernel, were evaluated. The genotypes were divided into two sets for internal cross-validation, one for calibration (training set) and one for validation (test set). The calibration set was chosen randomly and accounted for 70% of the total genotypes, while the test set accounted for 30% of the total genotypes. Five-fold cross-validation was used to identify the model with the best prediction ability. This process was repeated 50 times (niter = 50). Several statistical parameters, including the squared Pearson's correlation between predicted and observed values in the test set (R^2_{P}), the coefficient of determination extracted from the PLSR model (R^2_{CV}), and the root mean squared error of prediction as calculated using predicted and observed values from the test set (RMSE_{P}) were used to assess the model's goodness of fit. Other parameters included the root mean square error of cross validation extracted from the PLSR model (RMSE_{CV}), the ratio of performance deviation (RPD), standard error of prediction (SEP), ratio of performance to interquartile distance (RPIQ), and Lin's concordance correlation (CCC).

Four additional cross-validation schemes that mimic scenarios commonly encountered by plant breeders (CV2, CV1, CV0, and CV00) were applied across the 11 trials tested. Each trial was treated as an independent environment, as described by Jarquín et al. (2017). For CV2 (tested genotypes in tested environments), 30% of the genotypes from a given trial made up the test set. All remaining genotypes and all genotypes from other trials were combined to form the training set. The test sets for CV1 (untested genotypes in tested environments) were the same as for CV2; however, genotypes present in the test set were entirely removed from the training set. Each trial underwent 50 iterations of training, each with a different random sample of genotypes in the test set. For CV0 (tested genotypes in untested environments), an entire trial was included as the test set. All other trials, regardless of whether they contained genotypes represented in the test set trial, constituted the training set. CV00 (untested genotypes in untested environments) followed the same procedure as CV0, but all test set genotypes were removed from the training set prior to model training. For CV0 and CV00, only a single iteration was performed (Hershberger et al., 2022).

Variable importance and heritability estimate

RF and PLSR models were used to assess the significance of each wavelength in predicting root starch content by calculating variable importance for each wavelength. The possibility of heritable variation along the spectra was investigated. The heritability of root starch content was also evaluated for each trial. Variance components were estimated for both scenarios using a mixed linear model and the R package lme4 (Bates et al., 2015). The trial design was used to define the model. The following model was used for the randomized complete block design (RCBD) trials:

$$Y_{ij} = \mu + G_i + b_j + e_{ij} \begin{cases} eij \sim N(0,\ \sigma^2) \\ Gi \sim N(0,\ \sigma_G^2) \\ bj \sim N(0,\ \sigma_b^2) \end{cases}$$

Where Y_{ij} represents the reflectance data of the wavelength derived from genotype i with block j;

μ represents the overall mean; G_i is the random effect of genotype i, b_j is the effect of block j, and e_{ij} is the error associated with the observation. All random effects were assumed to have a normal distribution. The following model was used for the alpha lattice trials:

$$Y_{ijk} = \mu + G_i + \text{Rep}_j + b_{k(j)} + e_{ijk} \begin{cases} eijk \sim N(0,\ \sigma^2) \\ Gi \sim N(0,\ \sigma_G^2) \\ bk \sim N(0,\ \sigma_b^2) \\ Repj \sim N(0,\ \sigma_b^2) \end{cases}$$

Where Y_{ijk} denotes the reflectance value of each of the wavelengths derived from genotype i in replicate j and block k. Rep_j is the effect of the replicate j; $b_{k(j)}$ is the effect of block k nested within replicate j, and e_{ijk} is the error associated with the observation of genotype i in block k within replicate j. All random effects were assumed to have a normal distribution. Variance components estimated above were used to calculate heritability. Broad-sense heritability (H^2) was calculated for root starch content and each wavelength as follows:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{\text{nRep}}} \times 100$$

Where $\sigma_g{}^2$ is the genotypic variance; $\sigma_e{}^2$ is the residual variance, and nRep is the mean number of repetitions for one genotype in the trial. The estimated heritabilities of the entire measured NIR spectrum were plotted using the ggplot function from the ggplot^2 package (Wickham, 2016) in R (R Core Team, 2021).

Instrument comparison

Root spectra were also captured using two additional devices to enable instrument comparison in the five trials from the 2021 harvest season (Trials G, H, I, J, and K) (Table 1). These spectrometers include the full range (350 to 2500 nm) portable instrument ASD QualitySpec® Trek (QST; Malvern, Panalytical, Cambridge, UK) with a spectral interval of 1 nm and the benchtop FOSS XDS Rapid Content™ Analyzer NIR spectrometer (BT; FOSS, Hillerød, Denmark) with a spectral range from 400 to 2490 nm and a spectral interval of 10 nm. For the QST, a reference reading was taken when starting a scanning session. Each genotype was measured three times, with each spectrum representing the average of 50 scans. BT spectra were collected in reflectance mode. Three separate samples per genotype were placed in cell quartz glasses and measured three times each. For this spectrometer, each spectrum represents an average of 60 scans. Data from all five trials were combined. Based on the raw spectra from each spectrometer, PCA was used to classify cassava genotypes into homogeneous groups visually. This analysis was performed using the R package factorMineR (Lê et al., 2008). The performance of the three devices to predict root starch content was carried out using the same sample sets. Three approaches were used: (**1**) the initial full spectral range of the three devices; (**2**) comparison of the three devices in the overlapping regions (740 −1070 nm); and (**3**) the spectral data from the QST and BT were trimmed at the beginning (< 600 nm) and the end of the spectra (> 1900 nm) to remove potential noise. The selected range was determined after graphical visualization of the raw spectra. Background noise was evident with QST. The BT spectra were trimmed to the same range as the QST spectra for consistency and ease of comparison.

Results

Reference data exploration

Root starch content ranged from 4.1 to 41.6% among the 344 unique genotypes in this study. Furthermore, we observed root starch content varied within and between trials, over time, and across environments (Figure 1). The Kruskal-Wallis rank test revealed significant (*P*< 0.05) differences in starch content between trials. Trial B (coefficient of variation = 0.37) had the most genotype variation, followed by Trial K (coefficient of variation = 0.3). Trial I displayed the lowest level of variability (coefficient of variation = 0.1). Table 1 shows descriptive statistics for root starch content and the number of genotypes used for calibrating each trial. Supplementary Table 1 shows the number of common genotypes between trials.

Principal component analysis of the raw spectral data

A total of 301 averaged scans were recorded using the SCiO device across trials (Supplementary Table 2). Supplementary Figure 1 depicts the averaged raw spectra recorded on ground fresh cassava roots. PCA revealed variation between genotypes and subtle differences between trials (Figure 2). The overlap between trials could be attributed to common genotypes present and their close relatedness. The overlap may also be due to overlap in the mean and range of root starch content across trials (Supplementary Figure 2). The first PC accounted for 97.0% of the variation in the NIR spectra, while PC2 accounted for 2.9%. Overall, this exploratory PCA revealed the potential of spectral information in characterizing genotypes.

Analysis of SCiO spectra data using partial least squares regression

Assessment of prediction accuracy between trials

Several metrics were used to evaluate model prediction, including R^2_P, RPD, RPIQ, RMSE, and SEP. These metrics, which indicate model fitness for each trial, are reported in Supplementary Table 3. The prediction of root starch content differed between trials (Figure 3). A high-quality model should have higher R^2_P and R^2_{CV} values, lower $RMSE_P$ and $RMSE_{CV}$, and SEP and bias close to zero. Standard guidelines for the

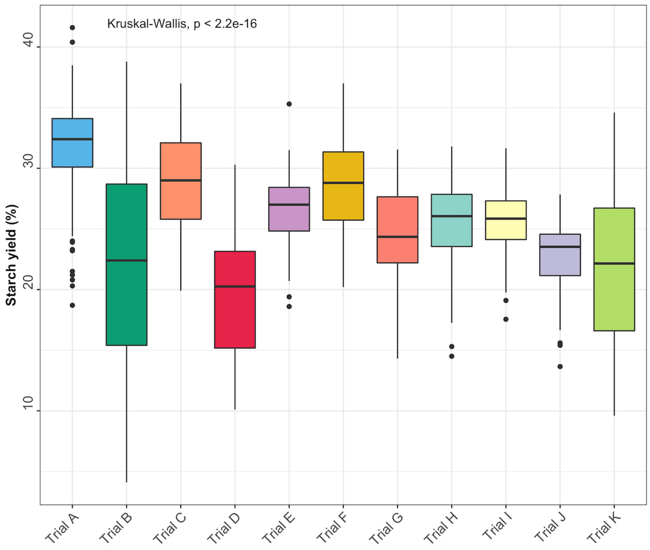


FIGURE 1
Root starch content distributions for the 11 evaluated cassava trials.

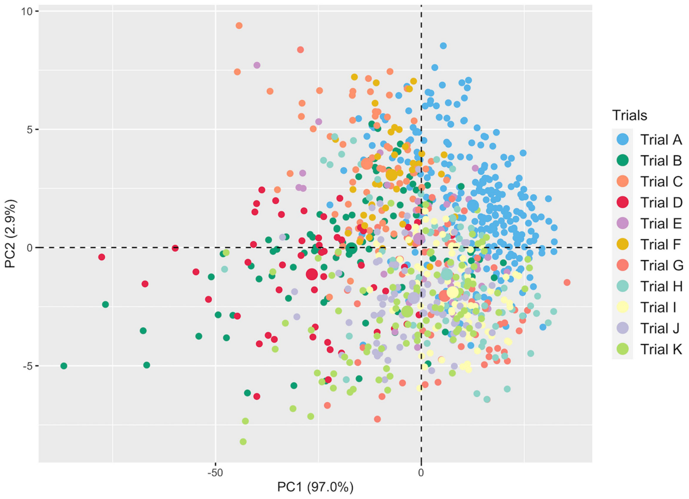


FIGURE 2
Principal component analysis of NIR spectral data from fresh cassava root scans captured with the SCiO.

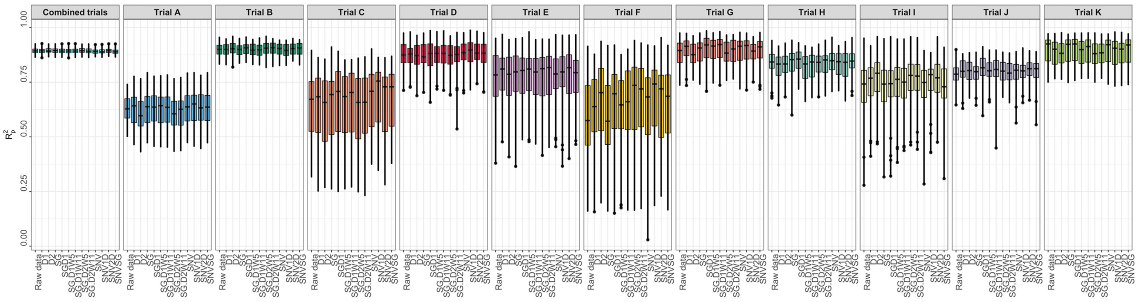


FIGURE 3
Individual trial performance using partial least squares regression. Pearson's correlation between predicted and observed values in the test set ($R^2{}_p$); no spectral pretreatment (raw data); standard normal variate (SNV); standard normal variate and first derivative (SNV1D); standard normal variate and second derivative (SNV2D); first derivative (D1); second derivative (D2); Savitzky-Golay with window size = 11 (SG); standard normal variate and Savitzky-Golay (SNVSG), gap segment derivative with window size = 11 (SGD1), Savitzky-Golay with window size = 5 and first derivative (SG.D1W5); Savitzky-Golay with window size = 11 and first derivative (SG.D1W11); Savitzky-Golay with window size = 5 and second derivative (SG.D2W5); and Savitzky-Golay with window size = 11 and second derivative (SG.D2W11).

interpretation of $R^2{}_P$ (Williams and Norris, 2001; Lebot et al., 2009; Polinar et al., 2019) and RPD (Williams and Norris, 2001; Williams, 2014; Polinar et al., 2019) were applied. Based on the $R^2{}_P$ values for each pretreatment and trial and the R^2 interpretation guidelines suggested by Williams and Norris (2001), the trained models could be used for: (a) rough screening (Trial A; $R^2{}_P$ =0.61.-0.64); (b) screening and other approximate calibration (Trials E, I, J; $R^2{}_p$ = 0.71-0.81); (c) most applications but with caution (Trials B, D, G, K, and Combined; $R^2{}_p$ = 0.86-0.90); (d) rough screening to screening and other approximate calibration (Trials C and F, $R^2{}_p$ = 0.59-0.68); and (e) screening and other approximate calibration to use for most applications but with caution (Trial H; $R^2{}_p$ =0.82-0.85). Based on the RPD values, the predicted models could be used for screening (Trials B, K, and Combined; RPD ≥ 3) and very rough screening (Trials D and H; RPD = 1.593 – 2.306) in some trials, but not in others (Trials A, C, E, F, J; RPD = 1.595 - 2.306). A combination of factors, including the small sample range between the reference data and the number of samples used, could have hampered efficient model prediction in trials E and F. Trials E and F are both NCRP trials; the final testing stages before varieties can be commercialized. These trials include superior genotypes with high yield and starch content and little variation because they are all high performers. Variation in the environment is also an important factor that could have influenced model prediction. Trials E and F, which had similar genotypes, were tested at two different phenotyping sites (Ikenne and Ago-Owu). Similar findings apply to Trials C and D, which were also evaluated in two different agroecological zones (Ikenne and Ibadan). Trials B, K, and the combined trials showed good predictive ability (RPIQ ≥ 4.0) consistent with their RPD values (Williams, 2014). The similarity of the $RMSE_P$ and $RMSE_{CV}$ for most trials confirmed the fair and robust fitting of the validation samples. Overall, Trials B, D, G, K, and Combined best predicted root starch content. An effect of spectral pretreatment on model prediction was observed in some cases. (Figure 3). The spectral pretreatments with the best performing models from Trial B were D2, SGD2W5, and SGD2W11. The model based on SGD1 and SGD1W11 pretreated spectra performed best in Trial G. The optimal model from Trial D was pretreated with SNV1D. The best performing pretreatment in the Combined trial was SG. The variability of genotypes within each trial had a greater impact on model prediction performance (r_s = 0.78, $P< 0.05$) than the number of genotypes (r_s = 0.18, $P< 0.05$) (Supplementary Figures 3, 4). In terms of prediction, trials with the highest coefficient of variation performed better. Trial A had the most genotypes (261) but a lower R^2_P value than Trials B (97), K (88), and D (52), which had smaller sample sizes but high coefficients of variation. As a result, model prediction can be linked to the level of sample variability.

Comparison of different prediction models using aggregated data

Spectral data from all 11 trials were combined and the model prediction was assessed using three chemometric modeling approaches: RF, PLSR, and SVM. Using PLSR, high prediction accuracies were obtained regardless of the pretreatment applied (R^2_P = 0.89, RPD > 3.0, RPIQ > 3.9, SEP ≤ 2.07%) (Figure 4; Supplementary Table 4). SVM performance was also consistently satisfactory across the 11 pretreatments. This was supported by a high RPD value (> 3), RPIQ (>3.9), low SEP (≤ 2.07%), and low bias (0.01–0.06) (Figure 4; Supplementary Table 4). Regarding statistical performance parameters, SVM and PLSR models yielded comparable predictions. When the RF model was applied, models based on SNV1D (R^2_P = 0.89, RPD = 3.03, RPIQ = 3.97, SEP = 2.07%) and SNV2D (R^2_P = 0.89, RPD = 3.04, RPIQ = 4.01, SEP = 2.04%) were deemed reasonable for root starch content prediction, while other pretreatments showed only a fair RPD value (2.5 ≤ RPD ≤ 2.9). The lowest predictability for the RF model was obtained when no spectral pretreatment was applied (R^2_P = 0.84, RPD = 2.49, RPIQ = 3.24, SEP = 2.52%) (Figure 4; Supplementary Table 4). Four cross-validation schemes (CV1, CV2, CV0, and CV00) were used to evaluate the ability of each model to correctly predict root starch content across a range of realistic scenarios. Supplementary Tables 5, 6 show the performance statistics of the models developed using PLSR and SVM. The SVM and PLSR models performed nearly identically across all CV schemes. (Figure 5; Supplementary Figure 5). The overall mean model performance based on R^2_P ranged from 0.76 to 0.79, which is lower than within-trial R^2_P (0.89 to 0.90). The performance of other metrics was also lower (Supplementary Table 7), but the difference in performance between groups was negligible. Model prediction was slightly improved when the tested set of genotypes was represented in the training set. Likewise, the SEP decreased in the schemes in which the test set environment was represented in the training set (CV1,

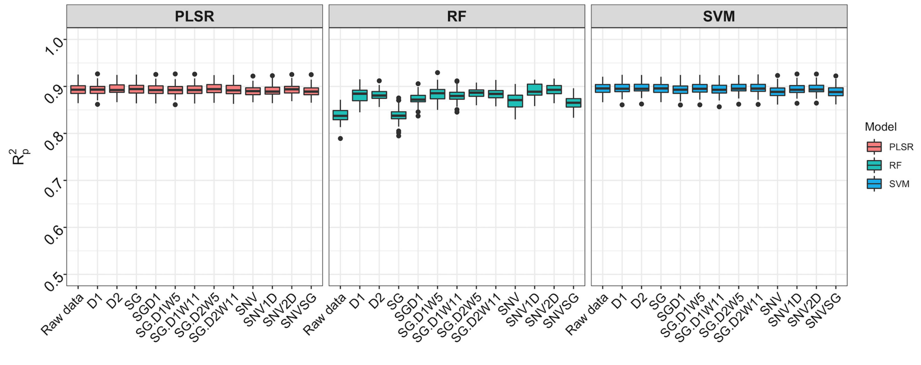


FIGURE 4
Comparison of three chemometric modeling approaches using SCiO spectral data and all accessions combined. Pearson's correlation between predicted and observed values in the test set is represented by the y-axis (R^2_P). The various pretreatment approaches and the model without spectral pretreatment (raw data) are depicted on the x-axis (standard normal variate (SNV), standard normal variate and first derivative (SNV1D), standard normal variate and second derivative (SNV2D), first derivative (D1), second derivative (D2), Savitzky-Golay with window size = 11 (SG), standard normal variate and Savitzky-Golay (SNVSG), gap segment derivative with window size = 11 (SGD1), Savitzky-Golay with window size = 5 and first derivative (SG.D1W5), Savitzky-Golay with window size = 11 and first derivative (SG.D1W11), Savitzky-Golay with window size = 5 and second derivative (SG.D2W5), and Savitzky-Golay with window size = 11 and second derivative (SG.D2W11)). Three modeling approaches were evaluated: random forest (RF), partial least square regression (PLSR), and support vector machine (SVM) (SVM). SVM and PLSR both produced consistent and comparable predictions.

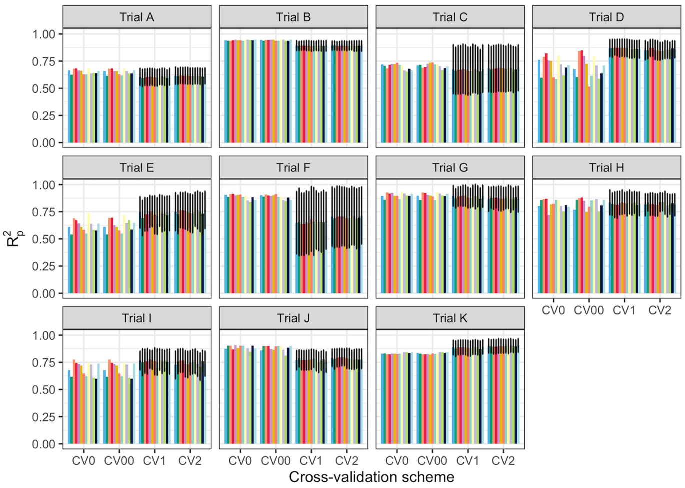


FIGURE 5
Prediction of cassava root starch content using four cross-validation schemes and the partial least squares regression algorithm. The x-axis displays the four cross-validation (CV) schemes. The y-axis shows the squared Pearson's correlation between predicted and observed values in the test set (R^2_p) for 50 iterations of the waves prediction pipeline with no spectral pretreatment. The colors represent the various pretreatments. CV0 indicates leave-one-trial-out CV and CV00 indicates that there was no overlap between genotypes and environments in the training and test sets. CV1 indicates overlap in the environment but not genotypes between the training and test sets. CV2 indicates an overlap of both genotypes and environments in the training and test sets. However, genotypes with multiple replicates within a trial were sorted together in all cases. Error bars show the standard deviation for schemes with subsampling (CV1 and CV2). As no subsampling occurred in either the CV0 or CV00 schemes, the standard deviation was not calculated and, hence, no error bars are displayed.

CV2). The genetic similarity of the genotypes may have contributed to the comparable model performance observed across scenarios.

Wavelengths of importance and heritability

The variable importance analysis identified the relative contribution of wavelengths in predicting root starch content. The most informative wavelengths for both PLSR and RF models were between 815 and 980 nm, corresponding to a) the third overtone of C-H and C-H_2 stretching related to the presence of carbohydrates and b) the second overtone for O-H bands, the most prominent signal for water (Bantadjan et al., 2020a; Bantadjan et al., 2020b; Farhadi et al., 2020) (Table 2; Supplementary Figure 6).

The extent to which NIR spectral variation is due to genetic variation among genotypes was examined by computing the heritability of NIR reflectance values for each trial. The heritability of NIR spectra varied between trials and across spectral regions (Figure 6). Trials K and B had higher heritabilities across all wavelengths ($H^2 \geq 0.79$), whereas Trial H had the lowest range of heritability ($H^2 < 0.4$) (Supplementary Table 8). This finding implies that most of the variation in NIR spectral patterns is due to the genetic variation among genotypes. Spectra from 871 to 1070 nm, a range that contains spectral bands strongly related to root starch content (Bantadjan et al., 2020a; Bantadjan et al., 2020b; Farhadi et al., 2020), showed higher heritability ($H^2 > 0.5$) (Figure 6). The heritability of root starch content was also computed for each trial. The heritability of root starch content on a mean entry basis ranged from moderate (H^2 = 0.53; Trial J) to high (H^2 = 0.88; Trial B), except for Trial I, which had a lower heritability (H^2 = 0.29). The heritability of root starch content also varied between years and locations (Supplementary Table 9). Except for Trial H (H^2: 0.28-0.32), heritability estimates based on spectral data were slightly higher than root starch content heritability estimates, supporting the possibility of using spectral information *via* indirect selection to improve traits in cassava breeding.

Instrument comparison

Reflectance values varied across wavelengths, but the patterns of reflectance recorded by the QST and BT devices were quite similar. The SCiO patterns, on the other hand, appear to be different, possibly due to distinct optical parameters and operational characteristics of this miniaturized device. Furthermore, the spectral pattern differences could be explained by the proprietary algorithm used to remove noises from the raw signals captured by the SCiO sensor before storing the raw spectral data in the cloud (Figure 7). PCA of raw spectra from the different devices revealed further similarities (Supplementary Figure 7). For BT, PC1 explained 83.7% of the

TABLE 2 Top wavelengths identified through variable importance analysis for predicting root starch content with partial least squares regression (PLSR) and random forest (RF) models using data captured with SCiO^{TM} for all combined cassava breeding field trials at IITA.

		Wavelength (nm)											
Model	PLSR		878	879	880	911	912	*959*	*960*		*973*	*974*	*975*
	RF	815				913		963	964	965	979	980	

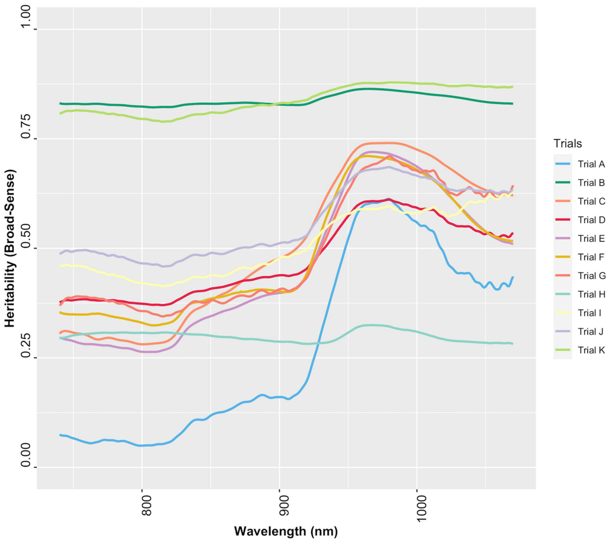


FIGURE 6
Broad-sense heritability of each wavelength of cassava root NIR spectra collected with the SCiO.

variability in the raw spectral data, while PC2 explained 9.9%. Similarly, for QST, PC1 accounted for 85% of the variability, while PC2 accounted for 9%. In contrast to BT and QST, PC1 and PC2 captured 97.1 and 2.7% of the variability of the SCiO, respectively. Differences between genotypes and subtle differences between trials were observed regardless of the instrument used (Supplementary Figure 7).

When the entire spectral range (SCiO: 740–1070 nm; QST: 350–1070 nm; BT: 400–2490 nm) was used, adequate and consistent prediction was achieved across pretreatments using the SCiO spectrometer with PLSR (R^2_P = 0.89-0.90, RPD = 3.10–3.19, RPIQ = 3.74–3.85, SEP = 1.47–1.52%) and SVM (R^2_P = 0.89–0.90, RPD = 2.99–3.31, RPIQ = 3.61–4.00, SEP = 1.47–1.52%) models. In general, the RF model statistics were lower (R^2_P = 0.77–0.87, RPD = 2.08–2.82, RPIQ = 2.53–3.37, SEP = 1.67–2.22%) (Supplementary Figure 8). Model statistics for untrimmed spectral data derived from the BT varied greatly depending on pretreatments and the chemometric model used. The optimal BT PLSR model was obtained when the SGD1W11 pretreatment was applied (R^2_P = 0.87, RPD = 2.99, RPIQ = 3.60, SEP = 1.66%), while the highest statistical indicators from RF were obtained when spectra data was processed by SNV2D (R^2_P = 0.89, RPD = 3.06, RPIQ = 3.70, SEP = 1.54%) and SNV1D (R^2_P = 0.89, RPD = 3.03, RPIQ = 3.64, SEP = 1.56%). The best pretreatment approach for the SVM model was SG (R^2_P = 0.89, RPD = 3.12, RPIQ = 3.77, SEP = 1.52%), followed by SGD1 (R^2_P = 0.88, RPD = 3.11, RPIQ = 3.75, SEP = 1.56%). Less optimal calibration models were observed with the QST spectrometer (R^2_P =0.10-0.83, RPD = 0.99-2.57, RPIQ =1.20-3.10, SEP = 1.89-4.68%) (Supplementary Table 10).

When the overlapping region between the three devices (740–1070 nm) was used, model prediction with the BT and QST spectrometers improved considerably (Figure 8). Model statistics revealed that depending on the pretreatment applied, certain models were suitable for predicting root starch content and, in some cases, were slightly superior to the model developed with the SCiO (Supplementary Table 11). The optimal models for the BT were obtained using the SVM (R^2_P = 0.91; RPD = 3.39; RPIQ = 4.19, SEP = 1.39) and PLSR (R^2_P = 0.91; RPD = 3.37; RPIQ = 4.17, SEP = 1.38) algorithm. Here as well, the QST produced the models with the poorest performance (R^2_P = 0.40–0.85; RPD = 1.06–2.70; RPIQ = 1.28–3.26, and SEP = 1.80–4.40%) (Figure 8 and Supplementary Table 11).

After trimming the spectra to remove noise (Supplementary Figure 9), model calibration obtained with the BT spectrometer

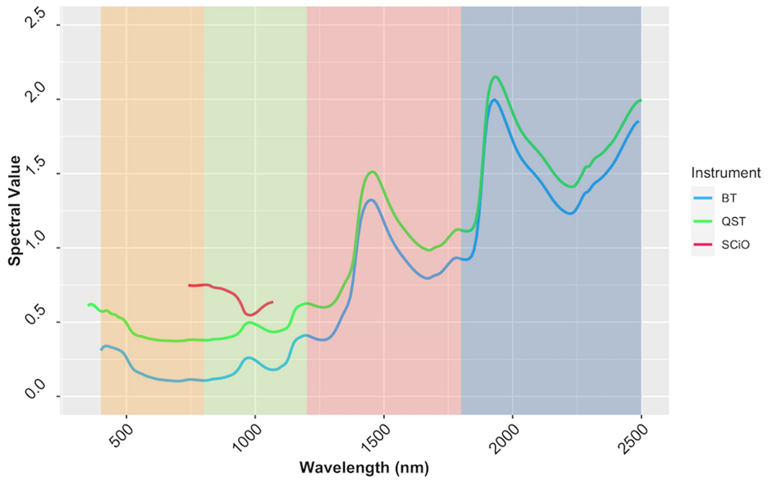


FIGURE 7
The average spectrum of the cassava accessions obtained using ASD QualitySpec® Trek (QST), Benchtop FOSS XDS Rapid Content™ Analyzer NIR spectrometer (BT) and pocket-sized SCiO™ (SCiO). The various NIRS regions are highlighted on the background, yellow (visible; 400–800 nm), green (region1; 800–1200 nm), pink (region2; 1200–1800 nm), and blue (Region 3; 1800–2500 nm).

was slightly superior (R^2_P = 0.82–0.91; RPD = 2.37–3.52; RPIQ = 2.86–4.35; SEP = 1.32–1.98%) to the model developed with the SCiO (R^2_P = 0.77–0.90; RPD = 2.08–3.31; RPIQ = 2.53–4.00; SEP = 1.43–2.22%) depending on the pretreatment and the chemometric model used (Supplementary Figure 10, Supplementary Table 12). It is critical to use the appropriate pretreatments to achieve more accurate predictions. Regarding root starch content prediction, although the effect of pretreatment on model prediction was more pronounced when using the benchtop device, both the BT and SCiO outperformed QST. Even though the SCiO sensor only captures information in the second and third overtones, its limited spectral range did not affect root starch content prediction in this study. Supplementary Tables 13–15 show the average prediction value from multiple predictions from a random sampling of the test set, while Supplementary Figure 11 shows the correlation between the obtained predicted values and the reference values. However, newly selected samples from the next harvesting season would accurately correlate the observed laboratory values with the root starch content obtained from the three devices.

Discussion

Trial selection, sample coverage, and model prediction

Breeding programs devoted to developing cassava varieties with high root starch content for industry necessitate robust, fast, and low-cost methods for screening breeding populations, particularly at the early stages of selection when many entries are evaluated. Laboratory-based quantification of root starch content is tedious and time-consuming. The potential of NIRS technology for quantifying root starch content was investigated. The importance of training set composition, including consideration of the trial type and phenotypic variation within a trial, was demonstrated in developing a robust model. The current study found that some trials with more genotypes (e.g., Trial A) had lower prediction accuracy than trials with fewer genotypes and a wider range of root starch content (e.g., Trials B and K), highlighting the importance of capturing a diverse range of phenotypes (Cafferky et al., 2020; Zerihum et al., 2020). Environment factors may also impact trait prediction. This is evidenced by Trials C and D, which were carried out in two distinct agroecological zones. Trial C was conducted in Ikenne (a rainforest) and Trial D was conducted in Ibadan (a derived savanna). The effects of edaphic and climatic conditions on cassava root content and their physiochemical properties have been previously reported (Benesi et al 2004; Gu et al., 2013). For a robust model, selecting a set of genotypes representative of the breeding pool from different selection stages, locations, growing seasons, and years is preferable to maximize the number of genotypes. Routine model updates capturing new variations are advised to prevent bias (Lebot, 2012).

Assessment of model prediction

Various pretreatments were used to correct spectral data. A recent study by Hershberger et al. (2022) evaluated the ability of the SCiO to predict cassava root dry matter content. Consistent

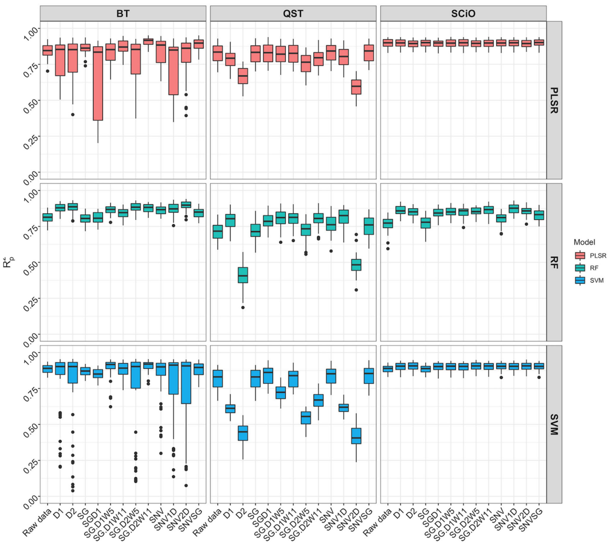


FIGURE 8
Comparison of model prediction using Partial least squares regression (PLSR), support vector machine (SVM) and random forest (RF) algorithms between ASD QualitySpec® Trek (QST), the Benchtop FOSS XDS Rapid Content^{TM} Analyzer NIRS spectrometer (BT) and the pocket-size SCiO^{TM} (SCiO) using the overlapping wavelengths (740 1070 nm) between the three devices. The Y-axis shows the squared Pearson's correlation between predicted and observed values in the test set (R^2_{P}). The X-axis indicates the model without spectral pretreatment (raw data) and the different pretreatment approaches used [standard normal variate (SNV), standard normal variate and first derivative (SNV1D), standard normal variate and second derivative (SNV2D), first derivative (D1), second derivative (D2), Savitzky-Golay with window size = 11 (SG), standard normal variate and Savitzky-Golay (SNVSG), gap segment derivative with window size = 11 (SGD1), Savitzky-Golay with window size = 5 and first derivative (SG.D1W5), Savitzky-Golay with window size = 11 and first derivative (SG.D1W11), Savitzky-Golay with window size = 5 and second derivative (SG.D2W5), and Savitzky-Golay with window size = 11 and second derivative (SG.D2W11)].

results were obtained across the same 12 combinations of pretreatments with PLSR and SVM, but the effect of spectral pretreatment was evident with the RF model. Previous research has highlighted the effect of pretreatment on prediction accuracy for NIRS (Agussabti et al., 2020; Cafferky et al., 2020). Because there is no one-size-fits-all pretreatment, care should be taken to avoid model bias when selecting a spectral pretreatment method.

A thorough evaluation is critical to ensure that models are appropriate for their intended uses. R^2_{p} and R^2_{cv} are commonly used to assess model fit and predictive strength, but these metrics should not be used as stand-alone indicators of model performance. RPD and RPIQ are additional statistical parameters used in the current study to evaluate model prediction accuracy. RPD is inappropriate when the assumption of a normal distribution is violated, and its interpretation varies from study to study (Lebot, 2012; Williams, 2014; Zerihum et al., 2020; Zhao et al., 2021). As a result, RPIQ was also considered when evaluating model fit (Bellon-Maurel et al., 2010). Other meaningful metrics to measure model fit that were examined include RMSE, which gives the standard residual error, model SEP, bias, and CCC.

Algorithm choice is also critical for model development. We found that SVM models performed similarly to those trained with PLSR, a more traditional NIRS modeling approach. This is consistent with the findings of Mendez et al. (2019) and Hershberger et al. (2022). They reported a marginal improvement in predictive ability for SVM over PLS, contradicting Ludwig et al. (2019) and Wang et al. (2021) who found SVM superior to PLSR. The observed variation in algorithm performance between studies could be attributed to differences in the trait investigated, data distribution, and sample variability (Frizzarin et al., 2021). Consistent with previous studies, SVM and PLSR outperformed the RF algorithm in this study (Mendez et al., 2019; Abincha et al., 2021; Hershberger et al., 2022). PLSR may remain the go-to model for trait prediction with NIRS due to its sensitivity and computational efficiency.

A decrease in model prediction accuracy was observed when tested with additional cross-validation schemes. While it is important to adequately account for environmental and genotype variability to ensure broad-based calibration, it is equally important to minimize sample bias through an adequate calibration set and genotype representativeness (Au et al., 2020; Hershberger et al., 2022). Within-trial cross-validation should be interpreted cautiously because it can produce overly optimistic statistics and may not reflect the conditions observed in practice. Thus, the four additional cross-validation schemes tested may provide a more realistic assessment of the ability of the SCiO to predict unknown samples (Li et al., 2018; Patel et al., 2020).

The wavelength of importance and heritability

Important wavelengths for predicting cassava root starch content were identified between 815 and 980 nm through variable importance analysis. This interval has previously been linked to spectral bands associated with starch and water absorption (Bantadjan et al., 2020a; Wang et al., 2021). In this interval, the third overtone associated with C-H, C-H_2 stretching was reported at 900, 910, 914, 915, and 930 nm (Bantadjan et al., 2020a). A signal from water O-H bonds was captured between 970 and 975 nm (Bantadjan et al., 2020a; Bantadjan et al., 2020b; Farhadi et al., 2020). The peak at 980 nm is likely related to carbohydrates and water in the root samples (Wang et al., 2021). Given that variable importance is used to identify wavelengths that may correspond to the most relevant information for predicting phenotypes, the preferential targeting of these identified wavelengths of importance could simplify the modeling process. Fitting fewer wavelengths would also require less computing time (Li et al., 2020; Wang et al., 2021).

Although broad-sense heritability estimates varied across the NIR spectrum, highly heritable regions were identified. This indicates that NIR spectral bands are influenced mainly by genetic effects (Hein and Chaix, 2014). Heritable NIR signatures, especially those also predictive of root starch content, could be used to identify desirable cassava genotypes (Hein and Chaix, 2014). Highly heritable spectral regions may also aid in deciphering root starch content genetics (Fujimoto et al., 2015; Razar et al., 2021). Such findings highlight the utility of spectral data in conjunction with, for example, genomics-assisted breeding approaches.

Instrument comparison

Miniaturized NIR spectrometers have the potential to offer more cost-effective and appropriate high-throughput phenotyping procedures for plant breeding programs. Their effectiveness, however, is still under debate. Despite its limited spectral range, more accurate predictions were obtained using the pocket-sized SCiO compared to the QST, regardless of the pretreatment method applied. This contradicts the hypothesis that spectrometers with broader spectral ranges can provide superior predictions. When models trained with the overlapping region of the three devices (740 −1070 nm) were compared, the SCiO still had an advantage, as evidenced by its higher predictive ability. The overlapping region may contain the most influential bands for predicting root starch content. Bittante et al. (2021) made a similar observation, emphasizing the importance of capturing the most informative portion of the spectrum. Rukundo et al. (2021) reported that the limited spectral range of the smartphone NIR spectrometer used in their study did not affect model performance. The improved prediction obtained after spectral trimming could be attributed to an increase in signal-to-noise ratio, emphasizing the negative effect of the discarded spectral regions. The poor performance of the QST in all scenarios could be attributed to complex information captured, making their extraction more difficult. Differences in device technology and operational characteristics cannot be ruled out as potential contributors to model prediction disparities between instruments (Stocco et al., 2019; Ozaki et al., 2020; Beć et al., 2021). The number of reports in the literature on using a miniaturized SCiO sensor for trait prediction is growing. Some studies have pointed out the strong performance of spectral data from the SCiO in trait prediction (Li et al., 2018; Riu et al., 2020; McVey et al., 2021; Hershberger et al., 2022), while others have found models developed using SCiO data to be unreliable (Berzaghi et al., 2021). In other cases, the analytical performance of the SCiO sensor was comparable to that of widely used benchtop devices, the go-to instruments in NIR spectroscopy (Li et al., 2018; Wiedemair et al., 2019).

The routine use of near-infrared spectroscopy for trait prediction in cassava breeding

Recent studies have reported the value of NIRS for predicting key cassava traits such as dry matter, carotenoids, cyanogenic glucosides, and starch content in fresh cassava roots (Sánchez et al., 2014; Ikeogu et al., 2019; Bantadjan et al., 2020a; Bantadjan et al., 2020b; Abincha et al., 2021; Hershberger et al., 2022). NIR sensors, particularly miniaturized devices, will be helpful in cassava breeding programs where thousands of samples are processed, and data turnaround is critical. A significant amount of time spent on starch extraction will be saved. Another anticipated benefit of routinely implementing NIRS technology in cassava breeding is lower selection costs and a lower risk of advancing lines with inadequate starch content. Aside from analytical performance, the cost of technology is an essential factor that influences its adoption and use. The SCiO sensor is much cheaper than the QST and the BT and could appeal to breeding programs with limited resources. However, one potential barrier to the routine use of this device by programs with limited budgets is access to cloud-based data storage. This necessitates a license and the need to operate *via* an internet connection, which is impractical in remote breeding sites. One crucial point to emphasize is that it is misleading to believe that the ability of SCiO to predict traits such as dry matter content and starch content implies that it applies to all traits. The situation may be different for other traits. As a result, the device's ability to predict other traits should be assessed on a case-by-case basis.

Conclusion

The ability of the pocket-size SCiO^{TM} spectrometer to predict starch content was investigated, and its performance compared to that of the Benchtop FOSS XDS Rapid Content ^{TM}

Analyzer and ASD QualitySpec Trek$^{®}$. The relevance of spectral information was also evaluated. The SCiO sensor successfully predicted starch content in fresh, shredded cassava roots despite its limited spectral range. After removing noise at the beginning and end of the spectrum, model calibration using the BT spectrometer slightly outperformed the SCiO sensor. With the QST, suboptimal calibration was achieved. The SCiO could be an economically viable solution for breeding programs with limited resources looking for a quick analytical tool to predict cassava root starch content. We demonstrated that spectral information could also characterize accessions. The heritability of the spectra highlighted the possibility of using spectral information for quantitative genetic analyses and improvement. Capturing new variations and continual prediction model updates will help ensure adequate predictive performance and avoid incorrect decisions caused by a miscalibrated model.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

EM designed the study, analyzed the data, wrote the first manuscript, reviewed, and edited it. JH provided critical assistance in data analysis, as well as critically reviewed and edited the manuscript. PP coordinated the experiment and curated the data. AI, KO, and KN collected spectra data. RA collected root starch content data. EA and MA participated in designing some of the experiments and provided critical edits. The manuscript was reviewed and edited by AA, SK, BM-D, EP, PK, and CE. MG provided software assistance and critical edits. IR conceptualized, and coordinated the experiments, as well as edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The authors thank the UK's Foreign, Commonwealth & Development Office (FCDO) and the Bill and Melinda Gates Foundation (Grant INV-007637, http://www.gatesfoundation.org) for their financial support. 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.). 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 USAID under Cooperative Agreement No. 7200AA-19LE-00005 (M.A.G.).

Acknowledgments

We acknowledge the IITA Nigeria Cassava Breeding team at Yam Barn and the Food Science team for the logistic and support provided during harvesting and data collection; the RTBfoods project (https://rtbfoods.cirad.fr):" Breeding RTB products for end-user preferences (RTBfoods)" led by the French Agricultural Research Centre for International Development (CIRAD), Montpellier, France, specifically Dominique Dufour for initial discussions that led to the initiation of this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.990250/full#supplementary-material

References

Abincha, W., Ikeogu, U. N., Kawuki, R., Egesi, C., Rabbi, I., Parkes, E., et al. (2021). Portable spectroscopy calibration with inexpensive and simple sampling reference alternatives for dry matter and total carotenoid contents in cassava roots. *Appl. Sci.* 11, 1714. doi: 10.3390/app11041714

Agussabti, R., Satriyo, P., and Munawar, A. A. (2020). Data analysis on near infrared spectroscopy as a part of technology adoption for cocoa farmer in aceh province, Indonesia. *Data Br.* 29, 105251. doi: 10.1016/j.dib.2020.105251

Au, J., Youngentob, K. N., Foley, W. J., Moore, B. D., and Fearn, T. (2020). Sample selection, calibration and validation of models developed from a large dataset of near infrared spectra of tree leaves. *J. Near. Infrared. Spectrosc.* 28, 186–203. doi: 10.1177/0967033520902536

Baek, J., Lee, E., Kim, N., Kim, S. L., Choi, I., Ji, H., et al. (2020). High throughput phenotyping for various traits on soybean seeds using image analysis. *Sensors* 20, 248. doi: 10.3390/s20010248

Bantadjan, Y., Rittiron, R., Malithong, K., and Narongwongwattana, S. (2020a). Establishment of an accurate starch content analysis system for fresh cassava roots using short-wavelength near infrared spectroscopy. *ACS Omega.* 5, 15468–15475. doi: 10.1021/acsomega.0c01598

Bantadjan, Y., Rittiron, R., Malithong, K., and Narongwongwattana, S. (2020b). Rapid starch evaluation in fresh cassava root using a developed portable visible and near-infrared spectrometer. *ACS Omega.* 5, 11210–11216. doi: 10.1021/acsomega.0c01346

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Beć, K. B., Grabska, J., and Huck, C. W. (2021). Principles and applications of miniaturized near-infrared (NIR) spectrometers. *Chem. - A. Eur. J.* 27, 1514–1532. doi: 10.1002/chem.202002838

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J. M., and McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends Anal. Chem.* 29, 1073–1081. doi: 10.1016/j.trac.2010.05.006

Benesi, I. R., Labuschagne, M. T., Dixon, A. G. O., and Mahungu, N. M. (2004) Genotype x environment interaction effects on native cassava starch quality and potential for starch use in the commercial sector. *Afr. Crop Sci. J.* 12, 205–216. doi: 10.4314/acsj.v12i3.27880

Berzaghi, P., Cherney, J. H., and Casler, M. D. (2021). Prediction performance of portable near infrared reflectance instruments using preprocessed dried, ground forage samples. *Comput. Electron. Agric.* 182, 106013. doi: 10.1016/j.compag.2021.106013

Bittante, G., Savoia, S., Cecchinato, A., Pegolo, S., and Albera, A. (2021). Phenotypic and genetic variation of ultraviolet–visible-infrared spectral wavelengths of bovine meat. *Sci. Rep.* 11, 13946. doi: 10.1038/s41598-021-93457-5

Cafferky, J., Sweeney, T., Allen, P., Sahar, A., Downey, G., Cromie, A. R., et al. (2020). Investigating the use of visible and near infrared spectroscopy to predict sensory and texture attributes of beef *M. longissimus thoracis* et. *lumborum. Meat. Sci.* 159, 107915. doi: 10.1016/j.meatsci.2019.107915

Chawade, A., Van Ham, J., Blomquist, H., Bagge, O., Alexandersson, E., and Ortiz, R. (2019). High-throughput field-phenotyping tools for plant breeding and precision agriculture. *Agronomy* 9, 258. doi: 10.3390/agronomy9050258

Chisenga, S. M., Workneh, T. S., Bultosa, G., and Laing, M. (2019). Characterization of physicochemical properties of starches from improved cassava varieties grown in Zambia. *AIMS. Agric. Food* 4, 939–966. doi: 10.3934/agrfood.2019.4.939

Cobb, J. N., DeClerck, G., Greenberg, A., Clark, R., and McCouch, S. (2013). Next-generation phenotyping: Requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. *Theor. Appl. Genet.* 126, 867–887. doi: 10.1007/s00122-013-2066-0

Dada, A. D. (2016). Taking local industry to global market: The case for Nigerian cassava processing companies. *J. Econ. Sustain. Dev.* 7, 2222–1700.

Desta, T. A., and Tigabu, Y. T. (2018) *Starch production, consumption, challenges and investment potentials in Ethiopia: The case of potato starch.* Available at: https://www.agrobig.org/documents/Potato_Starch_Production_Consumption_Challenges_and_Investment_potentials_2016 (Accessed June 20 2022).

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Farhadi, R., Afkari-Sayyah, A. H., Jamshidi, B., and Gorji, A. M. (2020). Prediction of internal compositions change in potato during storage using visible/near-infrared (Vis/NIR) spectroscopy. *Int. J. Food Eng* 16, 395–422. doi: 10.1515/ijfe-2019-0110

Frizzarin, M., Gormley, I. C., Berry, D. P., Murphy, T. B., Casa, A., Lynch, A., et al. (2021). Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *J. Dairy. Sci.* 104, 7438–7447. doi: 10.3168/jds.2020-19576

Fujimoto, T., Chiyoda, K., Yamaguchi, K., and Isoda, K. (2015). Heritability estimates for wood stiffness and its related near-infrared spectral bands in sugi (*Cryptomeria japonica*) clones. *J. For. Res.* 20, 206–212. doi: 10.1007/s10310-014-0464-z

Grosjean, P., Frederic, I., and Etienne, M. (2018) *Package "pastecs": Package for analysis of space-time ecological series.* Available at: https://cran.r-project.org/web/packages/pastecs/pastecs.pdf (Accessed June 24, 2022).

Gu, B., Yao, Q., Li, K., and Chen, S. (2013). Change in physicochemical traits of cassava roots and starches associated with genotypes and environmental factors. *Starch/Staerke* 65, 253–263. doi: 10.1002/star.201200028

Hein, P. R. G., and Chaix, G. (2014). NIR spectral heritability: A promising tool for wood breeders? *J. Near. Infrared. Spectrosc.* 22, 141–147. doi: 10.1255/jnirs.1108

Hershberger, J., Mbanjo, E. G. N., Peteti, P., Ikpan, A., Ogunpaimo, K., Nafiu, K., et al. (2022). Low-cost, handheld near-infrared spectroscopy for root dry matter content prediction in cassava. *Plant Phenome. J.* 5, e20040. doi: 10.1002/ppj2.20040

Hershberger, J., Morales, N., Simoes, C. C., Ellerbrock, B., Bauchet, G., Mueller, L. A., et al. (2021). Making waves in breedbase: An integrated spectral data storage and analysis pipeline for plant breeding programs. *Plant Phenome. J.* 4, e20012. doi: 10.1002/ppj2.20012

Ikeogu, U. N., Akdemir, D., Wolfe, M. D., Okeke, U. G., Chinedozi, A., Jannink, J. L., et al. (2019). Genetic correlation, genome-wide association and genomic prediction of portable nirs predicted carotenoids in cassava roots. *Front. Plant Sci.* 10, 1570. doi: 10.3389/fpls.2019.01570

Jarquín, D., Lemes da Silva, C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., et al. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype × environment interactions in Kansas wheat. *Plant Genome* 10, 1–15. doi: 10.3835/plantgenome2016.12.0130

Lebot, V. (2012). Near infrared spectroscopy for quality evaluation of root crops: Practical constraints, preliminary studies and future prospects. *J. Root. Crop* 38, 3–14.

Lebot, V., Champagne, A., Malapa, R., and Shiley, D. (2009). NIR determination of major constituents in tropical root and tuber crop flours. *J. Agric. Food Chem.* 57, 10539–10547. doi: 10.1021/jf902675n

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An r package for multivariate analysis. *J. Stat. Software* 25, 1–18. doi: 10.18637/jss.v025.i01

Li, L., Lin, D., Wang, J., Yang, L., and Wang, Y. (2020). Multivariate analysis models based on full spectra range and effective wavelengths using different transformation techniques for rapid estimation of leaf nitrogen concentration in winter wheat. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00755

Li, M., Qian, Z., Shi, B., Medlicott, J., and East, A. (2018). Evaluating the performance of a consumer scale SCiO^{TM} molecular sensor to predict quality of horticultural products. *Postharv. Biol. Technol.* 145, 183–192. doi: 10.1016/j.postharvbio.2018.07.009

Ludwig, B., Murugan, R., Parama, V. R. R., and Vohland, M. (2019). Accuracy of estimating soil properties with mid-infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. *Soil Sci. Soc Am. J.* 83, 1542–1552. doi: 10.2136/sssaj2018.11.0413

Matsumoto, R., Asfaw, A., De Koeyer, D., Muranaka, S., Yoshihashi, T., Ishikawa, H., et al. (2021). Variation in tuber dry matter content and starch pasting properties of white guinea yam (*Dioscorea rotundata*) genotypes grown in three agroecologies of Nigeria. *Agronomy* 11(1944):1–15. doi: 10.3390/agronomy11101944

McVey, C., Gordon, U., Haughey, S. A., and Elliott, C. T. (2021). Assessment of the analytical performance of three near-infrared spectroscopy instruments (Benchtop, handheld and portable) through the investigation of coriander seed authenticity. *Foods* 10, 956. doi: 10.3390/foods10050956

Mendez, K. M., Reinke, S. N., and Broadhurst, D. I. (2019). A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* 15, 150. doi: 10.1007/s11306-019-1612-4

Okechukwu, R. U., and 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. *J. Crop Improv.* 22, 181–208. doi: 10.1080/15427520802212506

Oladunmoye, O. O., Aworh, O. C., Maziya-Dixon, B., Erukainure, O. L., and Elemo, G. N. (2014). Chemical and functional properties of cassava starch, durum wheat semolina flour, and their blends. *Food Sci. Nutr.* 2, 132–138. doi: 10.1002/fsn3.83

Omojola, M. (2013). Tacca starch: A review of its production, physiocochemical properties, modification and industrial uses. *Afr. J. Food. Agric. Nutr. Dev.* 13, 7972–7985. doi: 10.18697/ajfand.59.12930

Ozaki, Y., Huck, C., Tsuchikawa, S., and Engelsen, S. B. (2020). *Near-infrared spectroscopy: Theory, instrumentation, and applications* (Singapore: Springer), 593p.

Patel, N., Toledo-Alvarado, H., Cecchinata, A., and Bittante, G. (2020). Predicting the content of 20 minerals in beef by different portable near-infrared (NIR) spectrometers. *Foods* 9, 1389. doi: 10.3390/foods9101389

Polinar, Y. Q., Yaptenco, K. F., Peralta, E. K., and Agravante, J. U. (2019). Near-infrared spectroscopy for non-destructive prediction of maturity and eating quality of 'carabao' mango (*Mangifera indica l.*) fruit. *Agric. Eng. Int. CIGR. J.* 21, 209–219.

Raji, A. O. (2020). "Utilization of starch in food and allied industries in Africa: Challenges and prospects," in *Innovation in the food sector through the valorization of food and agro-food by-products* (London, UK: IntechOpen), 24p.

Razar, R. M., Makaju, S., and Missaoui, A. M. (2021). QTL mapping of biomass and forage quality traits measured using near-infrared reflectance spectroscopy (NIRS) in switchgrass. *Euphytica* 217, 51. doi: 10.1007/s10681-021-02788-x

R Core Team (2021)R: A language and environment for statistical. computing. In: *R foundation for statistical computing* (Vienna, Austria). Available at: https://www.R-project.org/ (Accessed March 15, 2022).

Reynolds, M., Chapman, S., Crespo-Herrera, L., Molero, G., Mondal, S., Pequeno, D. N. L., et al. (2020). Breeder friendly phenotyping. *Plant Sci.* 295, 110396. doi: 10.1016/j.plantsci.2019.110396

Riu, J., Gorla, G., Chakif, D., Boqué, R., and Giussani, B. (2020). Rapid analysis of milk using low-cost pocket-size NIR spectrometers and multivariate analysis. *Foods* 9, 1090. doi: 10.3390/foods9081090

Rukundo, I. R., Danao, M. G. C., MacDonald, J. C., Wehling, R. L., and Weller, C. L. (2021). Performance of two handheld NIR spectrometers to quantify crude protein of composite animal forage and feedstuff. *AIMS. Agric. Food* 6, 462–477. doi: 10.3934/agrfood.2021027

Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., et al. (2014). Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and hunter color techniques. *Food Chem.* 151, 444–451. doi: 10.1016/j.foodchem.2013.11.081

Spencer, D. S. C., and Ezedinma, C. (2017). "Cassava cultivation in sub-Saharan africa," in *Achieving sustainable cultivation of cassava*. Ed. C. H. Hershey(Burleigh Dodds: Burleigh Dodds Science Publishing Limited), 1–26. doi: 10.19103/AS.2016.0014.06

Stocco, G., Cipolat-Gotet, C., Ferragina, A., Berzaghi, P., and Bittante, G. (2019). Accuracy and biases in predicting the chemical and physical traits of many types of cheeses using different visible and near-infrared spectroscopic techniques and spectrum intervals. *J. Dairy. Sci.* 102, 9622–9638. doi: 10.3168/jds.2019-16770

Wang, F., Wang, C., Song, S., Xie, S., and Kang, F. (2021). Study on starch content detection and visualization of potato based on hyperspectral imaging. *Food Sci. Nutr.* 9, 4421–4431. doi: 10.1002/fsn3.2415

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

Wiedemair, V., Langore, D., Garsleitner, R., Dillinger, K., and Huck, C. (2019). Investigations into the performance of a novel pocket-sized near-infrared spectrometer for cheese analysis. *Molecules* 24, 428. doi: 10.3390/molecules24030428

Williams, P. (2014). The RPD statistic: A tutorial note. *NIR. News* 25, 22–26. doi: 10.1255/nirn.1419

Williams, P., and Norris, K. (2001). "Implementation of near-infrared technology," in *Near-infrared technology in the agricultural and food industries*. Eds. P. Williams and K. Norris (American Association of Cereal Chemists), 145–169.

Zerihum, M., Fox, G., Nega, A., Seyoum, A., Minuye, M., Jordan, D., et al. (2020). Near-infrared reflectance spectroscopy (NIRS) for tannin, starch and amylase determination in sorghum breeding programs. *Int. J. Food Nutr. Sci.* 7, 455–450. doi: 10.15436/2377-0619.20.2716

Zhao, D., Arshad, M., Li, N., and Triantafilis, J. (2021). Predicting soil physical and chemical properties using vis-NIR in Australian cotton areas. *Catena* 196, 104938. doi: 10.1016/j.catena.2020.104938

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/fpls.2022.1017275

Validation of KASP markers associated with cassava mosaic disease resistance, storage root dry matter and provitamin A carotenoid contents in Ugandan cassava germplasm

Williams Esuma[1]*, Oscar Eyoo[1,2], Francisca Gwandu[2], Settumba Mukasa[2], Titus Alicai[1], Alfred Ozimati[1,2], Ephraim Nuwamanya[1], Ismail Rabbi[3] and Robert Kawuki[1]

[1]National Crops Resources Research Institute, Kampala, Uganda, [2]College of Natural Sciences, Department of Plant Sciences, Microbiology and Biotechnology, Makerere University, Kampala, Uganda, [3]International Institute of Tropical Agriculture (IITA), Oyo, Nigeria

OPEN ACCESS

EDITED BY
Mathieu Rouard,
Alliance Bioversity International and CIAT, France

REVIEWED BY
Xiaofei Zhang,
International Center for Tropical Agriculture (CIAT), Colombia
Salej Sood,
Indian Council of Agricultural Research (ICAR), India
Mohini Prabha Singh,
Punjab Agricultural University, India

*CORRESPONDENCE
Williams Esuma
esumawilliams@yahoo.co.uk

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 11 August 2022
ACCEPTED 08 November 2022
PUBLISHED 23 November 2022

CITATION
Esuma W, Eyoo O, Gwandu F, Mukasa S, Alicai T, Ozimati A, Nuwamanya E, Rabbi I and Kawuki R (2022) Validation of KASP markers associated with cassava mosaic disease resistance, storage root dry matter and provitamin A carotenoid contents in Ugandan cassava germplasm. *Front. Plant Sci.* 13:1017275. doi: 10.3389/fpls.2022.1017275

Introduction: The intrinsic high heterozygosity of cassava makes conventional breeding ineffective for rapid genetic improvement. However, recent advances in next generation sequencing technologies have enabled the use of high-density markers for genome-wide association studies, aimed at identifying single nucleotide polymorphisms (SNPs) linked to major traits such as cassava mosaic disease (CMD) resistance, dry matter content (DMC) and total carotenoids content (TCC). A number of these trait-linked SNPs have been converted to Kompetitive allele-specific polymerase chain reaction (KASP) markers for downstream application of marker assisted selection.

Methods: We assayed 13 KASP markers to evaluate their effectiveness in selecting for CMD, DMC and TCC in 1,677 diverse cassava genotypes representing two independent breeding populations in Uganda.

Results: Five KASP markers had significant co-segregation with phenotypes; CMD resistance (2), DMC (1) and TCC (2), with each marker accounting for at least 30% of the phenotypic variation. Markers located within the chromosomal regions for which strong marker-trait association loci have been characterised (chromosome 12 markers for CMD, chromosome 1 markers for DMC and TCC) had consistently superior ability to discriminate the respective phenotypes.

Discussion: The results indicate varying discriminatory abilities of the KASP markers assayed and the need for their context-based use for MAS, with

PSY2_572 particularly effective in selecting for high TCC. Availing the effective KASP markers on cost-effective genotyping platforms could facilitate practical implementation of marker-assisted cassava breeding for accelerated genetic gains for CMD, DMC and provitamin A carotenoids.

KEYWORDS

allele-specific PCR assay, genetic gain, *Manihot esculenta*, marker-assisted breeding (MAB), vitamin A deficiency

Introduction

Cassava (*Manihot esculenta* Crantz), a climate-resilient crop grown on approximately 18 million hectares in Africa, offers great potential to end extreme hunger, achieve food security, improve nutrition and eradicate poverty if ideal varieties are deployed (Kolawole et al., 2010). Cassava's prominence in Africa's subsistence farming systems is attributed primarily to the crop's competitive advantage to produce reasonable yields under adverse environments where other crops would fail (Ogola and Mathews, 2011) or where resource-poor farmers cannot afford modern inputs required for intensive farming. The crop is mainly cultivated by smallholder farmers for its starchy roots, which are consumed when boiled or processed into products such as flour for preparing meals (Iragaba et al., 2020).

However, the inherent heterozygous nature, long breeding cycles and high sensitivity to environmental variations make conventional cassava breeding a difficult and expensive task (Ceballos et al., 2015). Integrating breeding innovations such as marker-assisted selection (MAS) into cassava improvement programs is expected to increase the efficiency and speed of variety development (Ferguson et al., 2011). Indeed, marker-assisted introgression of cassava mosaic disease (CMD) resistance into Latin American cassava germplasm prior to its introduction to Africa was a pioneer success story of classical MAS (Okogbenin et al., 2012).

Application of MAS in cassava breeding programs would help in several ways, including: (i) elimination of genotypes with unfavorable alleles at the early stage of the selection scheme, thus allowing for field testing of a reduced number of genotypes, (ii) rapid introgression of resistant genes into existing cassava clones, in places where the disease has recently spread to, for example in Southeast Asian countries (Uke et al., 2022), (iii) early selection of traits that are best measured at the late crop developmental stages, (iv) selection for resistance under environments where the disease pressure is low; and (v) discrimination of genotypes homozygosity and/or heterozygosity. Despite the promise from MAS, its use in cassava breeding is limited and lags considerably behind progress in other major crops such as maize, rice and wheat. An important technical impediment in deploying markers in cassava breeding is the failure to translate genomic knowledge into user-friendly assays that are robust enough to support selection decisions (Ferguson et al., 2011).

The advent of next generation sequencing technologies has renewed the hope for MAS in plant breeding by enabling the identification and use of trait-linked single nucleotide polymorphism (SNP) markers in selection. With sequencing and annotation of the cassava genome nearly complete (Prochnik et al., 2012), it is now possible to use genome-wide markers (Elshire et al., 2011) in genome-wide association studies (GWAS) for identification of trait-linked genomic regions and precisely anchor SNPs at such loci (Oliveira et al., 2012). Recent studies have used genome-wide mapping to identify SNPs associated with provitamin A carotenoids content (Esuma et al., 2016), cassava brown streak diseases resistance (Kayondo et al., 2018), CMD resistance (Rabbi et al., 2014), green mite resistance (Ezenwaka et al., 2018) and dry matter content (Rabbi et al., 2022). To facilitate their use for MAS in cassava breeding, some of the trait-linked SNPs identified by Rabbi et al. (2022) have been converted into Kompetitive allele-specific polymerase chain reaction (KASP) assays (Ige et al., 2021). KASP markers have three major advantages over other molecular marker techniques: low cost, high throughput, and high specificity and sensitivity. These attributes have already made KASP markers a popular choice for MAS in crops such as rice (Sandhu et al., 2022), wheat (Grewal et al., 2022) and soybean (Rosso et al., 2021).

In the case of cassava, the use of KASP markers in selection decisions is yet to be fully integrated into breeding programs. Effective deployment of markers for routine MAS requires an assessment of their predictive ability through technical and biological validations in independent populations (Chagné et al., 2019). Technical validation would provide information on the marker call rate and clarity of genotype classes, while biological validation pinpoints the ability of markers to predict a phenotype. This study was undertaken to evaluate the effectiveness of selected KASP markers CMD resistance, dry matter content (DMC) and total carotenoids content (TCC) in Uganda's cassava breeding population.

Materials and methods

Germplasm and phenotyping

Two diverse cassava populations from Uganda's breeding program at National Crops Resources Research Institute (NaCRRI) were used as independent genetic resource for the KASP marker validation. One population was constituted by 653 genotypes segregating for DMC and TCC in the provitamin A cassava breeding pipeline (pVAC population) while the second trial had 1,024 genotypes (white-fleshed population) segregating for CMD resistance. The two populations were phenotyped concurrently at Namulonge, central Uganda, during 2018/2019 cropping season. Namulonge is a known hotspot for CMD due to the high prevalence of cassava geminiviruses and super abundance of whitefly vector population in the area. Each trial was laid down in an augmented design using 10 plants per plot and two checks replicated within each block, all planted at spacing of 1 x 1 m. Furthermore, rows of CMD infected plants of cultivar BAO (highly susceptible to CMD) were planted as source of inoculum within and around the trial for white-fleshed population to increase disease pressure. Plants were allowed to grow in field under natural conditions for 12 months, with weeding done regularly when needed.

CMD severity was scored on plant basis at three, six and nine months after planting using the 1-5 scale (IITA, 1990), where 1 = no observable symptoms; 2 = mild chlorotic pattern on entire leaflets, mild distortion on the leaves; 3 = pronounced mosaic pattern on the entire leaf, narrowing and distortion of the lower one third of the leaflets; 4 = severe mosaic pattern, distortion of two thirds of leaflets and general reduction of leaf size and stunting plants; 5 = very severe mosaic pattern, distortion of four fifths or more of leaflets, twisted and severe reduction of leaf size in most leaves and severe stunting of plants. At 12 months, plants were uprooted for assessment of TCC and DMC in roots. TCC was assessed by visually scoring the intensity of pigmentation of the root parenchyma on a qualitative scale of 1-8 (Chávez et al., 2005). We used the visual color scale as a high throughput measurement of TCC, aware that previous reports have indicated strong positive correlation between carotenoid content assessed visually and quantitatively (Esuma et al., 2016). To estimate DMC, approximately 200 g of fresh root samples were dried in oven to a constant weight at 105 °C for 24 hours. DMC was then computed as:

$$\text{DMC}(\%) = \frac{\text{DSW}}{\text{FSW}} \times 100$$

where DSW = dry sample weight and FSW = fresh sample weight.

KASP marker genotyping

Thirteen KASP markers associated with CMD (3), DMC (4) and TCC (6) were selected from genomic resource under the Next Generation Cassava Breeding project (https://www.nextgencassava.org/). The markers were a product of GWAS using a West African cassava germplasm (Rabbi et al., 2022) from which SNPs with significant marker-trait association were converted to KASP markers at Intertek Laboratory, Australia, as a central repository for coordinated genotyping service. Specific details of the conversion and validation of trait-linked SNPs into uniplex KASP genotyping assays is provided by Ige et al. (2021). Table 1 presents some statistical and genomic profiles of markers tested.

Tissues were collected from young newly expanded leaves of plants growing under natural field conditions. Four leaf discs of

TABLE 1 Summary statistics of 13 KASP markers validated for CMD, DMC and TCC.

Trait	Marker	Intertek ID	Chr	Minor allele	Major allele	β	SE	p-value
CMD	S12_7926132	snpME0021	12	G	T*	0.89	0.02	p≈0
CMD	S12_7926163	snpME0022	12	A	G*	0.89	0.02	p≈0
CMD	S14_4626854	snpME0025	14	A*	G	-0.23	0.03	$1.0\text{x}10^{-14}$
DMC	S1_24197219	snpME0027	1	T	C*	0.77	0.04	p≈0
DMC	S6_20589894	snpME0038	6	G*	A	0.78	0.09	$1.7\text{x}10^{-16}$
DMC	S12_5524524	snpME0040	12	C*	T	0.69	0.01	$8.0\text{x}10^{-12}$
DMC	S15_1012346	snpME0029	15	C	T*	-0.84	0.01	$4.0\text{x}10^{-17}$
TCC	PSY2_572	snpME0001	1	A*	C	0.37	0.01	$1.3\text{x}10^{-219}$
TCC	S1_24636113	snpME0043	1	G*	A	0.57	0.02	$1.3\text{x}10^{-270}$
TCC	S1_30543962	snpME0047	1	G*	A	0.40	0.02	$2.4\text{x}10^{-76}$
TCC	S5_3387558	snpME0053	5	T*	C	0.20	0.02	$2.0\text{x}10^{-16}$
TCC	S8_25598183	snpME0056	8	T*	G	0.18	0.03	$8.0\text{x}10^{-12}$
TCC	S15_7659426	snpME0049	15	G	T*	-0.06	0.01	$4.0\text{x}10^{-17}$

CMD, cassava mosaic disease; DMC, dry matter content; TCC, total carotenoid content; β, SNP effect from associated GWAS; SE, standard error; p-value, probability value for marker-trait association; *Favorable allele. Marker information extracted from Rabbi et al. (2022) and Ige et al. (2021).

6 mm diameter were punched into wells of sample collection plates and desiccated using silica gel for 48 hours. Plates containing dry leaf tissues were shipped to Intertek, Australia. The KASP marker details are available through the Excellence in Breeding repository here: https://excellenceinbreeding.org/module3/kasp; individual marker IDs are further provided in Table 1. Details of the procedure for preparation and running of KASP reactions are provided in the KASP manual (available online: https://www.biosearchtech.com/). Briefly, the genotyping used the high throughput PCR SNPline workflow using 1 µL reaction volume in 96-well plates for PCR. The KASP genotyping reaction mix comprised three components: (i) sample DNA (~10 ng); (ii) marker assay mix consisting of target-specific primers; and (iii) KASP-TFTM Master Mix containing two universal FRET (fluorescence resonant energy transfer) cassettes (FAM and HEX), passive reference dye (ROX^{TM}), Taq polymerase, free nucleotides, and $MgCl_2$ in an optimized buffer solution. The SNP assay mix was specific to each marker and consisted of two Kompetitive allele-specific forward primers and one common reverse primer. Finally, the PCR products were fluorescently read, and allele calls made using KRAKENTM software.

Data analysis

Phenotypic data analysis

Because some genotypes infected with CMD tend to recover during the plant growing stage, we used CMDs scored at nine months as the optimal data for subsequent analyses. Phenotypic data for each trial were considered independent and fitted separately into linear mixed models with the *lme4* package for R statistical software (Bates et al., 2015) to allow for extraction of best linear unbiased predictions (BLUPs) of the genotype effects for CMD, TCC and DMC. In each case, we fitted the following linear mixed model:

$$y = X\beta + Z_{g^c} + Z_{\text{block}^b} + \epsilon$$

Where *y* was the vector of raw phenotype, β was a fixed effect of grand mean with the corresponding incidence matrix *X*, vector *c* and corresponding incidence matrix $Z_g$was the random effect for genotypes (*g*) such that $c \sim N(0, I\sigma_e^2)$, Z_{block^B} represented the random effect for blocks and ϵ was the residual such that $e \sim N(0, I\sigma_e^2)$. In the model, checks were considered as fixed effects while accessions and blocks were considered random effects. Variance components were extracted from the models for estimation plot-based heritability (H^2) as:

$$H^2 = \frac{\sigma_c^2}{(\sigma_c^2 + \sigma_e^2)}$$

where σ_c^2 was the genotype variance and σ_e^2 was the model residual variance. BLUPs for each genotype were extracted using the *ranef* function in *lme4* package.

Marker segregation and marker effects

Boxplots drawn with the *ggpubr* package in R were used to visualize segregation of marker genotypes for each phenotype, and statistical differences among the genotypes were compared using the Kruskal-Wallis test. Marker effects were further evaluated by regressing the marker genotypes onto respective phenotypes for estimation of phenotypic variance accounted to by each marker. In this case, linear regression was performed using the *lm* function in R such that marker genotypes and the corresponding phenotypes were treated as independent and response variables, respectively.

Estimation of biological metrics for CMD markers

The KASP markers for CMD resistance targeted the CMD2 locus (Rabbi et al., 2014; Wolfe et al., 2016), which has been classified as dominant monogenic trait (Okogbenin et al., 2012) and is expected to segregate in Mendelian fashion. Therefore, we used confusion matrix to estimate some performance statistics to determine the ability of each CMD KASP marker in predicting the response of genotypes for CMD resistance or susceptibility. The performance statistics included: a) accuracy (ACC), which is the proportion of correctly predicted genotypes, either as resistant or susceptible; b) false positive rate (FPR), which is the proportion of the genotypes predicted to be resistant but were diseased (also referred to as type I error); and c) the false-negative rate (FNR) which is the proportion of genotypes predicted to be susceptible but were resistant (type II error). These statistics were computed as:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{FN + TP}$$

Where FP = false positive, TN = true negative, FN = false negative, TP = true positive (Ige et al., 2021).

Results

Phenotypic variation and broad-sense heritability for CMD, DMC and TCC

All the three traits phenotyped had considerable variation, with CMDs showing typical bimodal distribution that had clear-cut separation between symptomless plants and those with varying severity levels (Figure 1). The white-fleshed population evaluated for CMD had 73.4% of the genotypes scoring 1 or 2 for

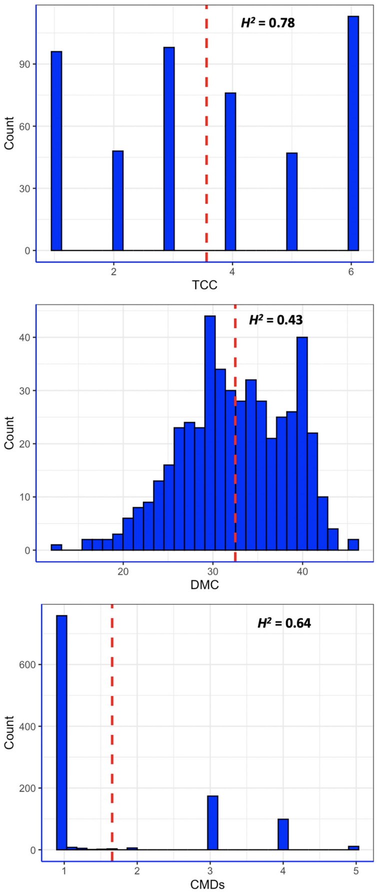


FIGURE 1
Distribution of phenotypic values of three traits measured in this study. CMDs = severity of cassava mosaic disease; DMC = dry matter content; TCC = total carotenoids content; H^2 = broad-sense heritability; red dashed line is the mean value for each trait.

CMD severity and were subsequently categorized as resistant (Lim et al., 2022), while the remaining genotypes were susceptible, with CMDs varying between 3 and 5. The mean CMDs score in the population was 1.67. In contrast, DMC and TCC exhibited continuous distribution pattern, akin to quantitative traits. DMC varied between 12.3% and 45.4%, with an average of 31.5%. Root pigmentation, a qualitative measure of TCC, varied from 1 (white) to 6 (deep yellow), with a mean of 4.0. In fact, 85.1% of the pVAC population had TCC score varying between 2 and 6. Broad-sense heritability of CMD, DMC and TCC were 0.64, 0.43 and 0.78, respectively (Figure 1).

Allele profiles and frequencies of markers tested

Based on fluorescence profiles of the KASP assays, the allele calls of each marker clustered into three distinct groups: homozygous genotypes with a HEX-type allele, homozygous genotypes with a FAM-type allele and heterozygotes (Figure 2). The allelic states corresponded to homozygous genotypes for minor alleles, homozygous genotypes for major alleles and heterozygotes having both alleles. Meanwhile, there was an overall high call rate (>97%) for the 13 KASP markers assayed, except for marker S15_1012346 (DMC) which had a comparably low call rate of 84.5% (Table 2). In the case of CMD, all the three markers (S12_7926132, S12_7926163 and S14_4626854) had a consistent pattern of allele distribution, with heterozygotes having the highest frequency (>56%) followed by homozygous genotypes for major alleles. The frequency of the homozygous state of minor alleles for CMD markers varied between 15% (marker S14_4626854) and 20% (markers S12_7926132 and S12_7926163), while the frequency of major alleles ranged from 20% to 29%. Meanwhile, distribution of allele frequencies for DMC and TCC markers did not follow any specific pattern. For example, the homozygous state of minor and major alleles of marker S15_1012346 (DMC) had equal frequency (16.7%) while the heterozygotes were 51.1%. On the contrary, homozygous state of the minor allele for marker S8_25598183 (TCC) had frequency of 70.8%, while the homozygous state of its major allele frequency was 1.1%.

Marker effects on traits

Marked differences were observed in the allele substitution effects across all the 13 markers and within markers for each trait. CMD markers S12_7926132 and S12_7926163, which are only 31 bp apart, exhibited segregation patterns typical of dominant markers. For marker S12_7926132, genotypes with at least one copy of the favorable (resistance) allele (TG or TT) significantly co-segregated with low CMD severity compared to the unfavorable allele for which the associated genotypes had mean CMD severity >3 (Figure 3). Similarly, genotypes with a copy of the favorable allele G (GA or GG) for marker S12_7926163 showed significant co-segregation with CMD resistant clones. For both markers, the heterozygotes had different performance when compared to that of genotypes homozygous for favorable allele. For marker S14_4626854, genotypes with the favorable allele A (AA or AG) had significantly low scores of CMDs compared to those carrying GG. However, in this marker, the heterozygotes had similar performance to that of genotypes with homozygous favorable allele (AA).

Three markers tested for TCC (PSY2_572, S1_24636113 and S1_30543962) also had segregation patterns likened to dominant gene effect. For marker PSY2_572, genotypes with at least one copy of the favorable allele (AA or CA) co-segregated with high levels of TCC (Figure 4). Similarly, markers S1_24636113 and S1_30543962 had significantly higher levels of TCC in genotypes carrying favorable alleles in the form of GA/GG and TC/TT,

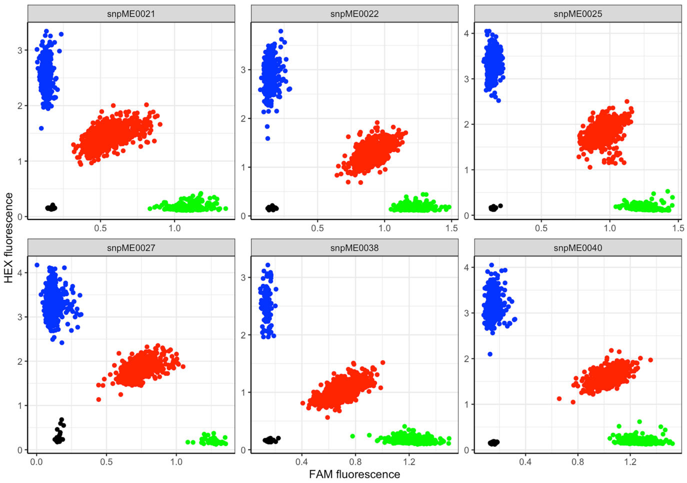


FIGURE 2
Scatter plot for selected KASP assays showing clustering of genotypes on the *Y*- and *X*-axes. Genotypes colored blue have a HEX-type allele; genotypes colored green have a FAM-type allele; genotypes colored red are heterozygotes; black dots represent non-template controls.

respectively. Evidently, these markers exhibited additive allelic effects, with mean TCC levels consistently higher in genotypes homozygous for favorable alleles than those in heterozygotes. The other three markers evaluated for TCC (S5_3387558, S8_25598183 and S15_7659426) did not show segregation patterns consistent with expression of the trait in genotypes assayed. For example, there were non-significant differences between marker-phenotype co-segregation of the two homozygous allelic states for each of the three markers.

For DMC, two markers (S1_24197219 and S6_20589894) exhibited dominant segregation patterns (Figure 5). In the case of marker S1_24197219, genotypes CC and TC co-segregated with high DMC. A similar observation was noted for marker S6_20589894, for which genotypes GA and GG were significantly associated with high values of DMC. Although not pronounced, these two markers manifested additivity, with genotypes carrying homozygous favorable alleles having the highest levels of DMC. Meanwhile, segregation patterns in the other two DMC markers (S12_5524524 and S15_1012346) were inconsistent with the phenotypic distribution, where the two homozygous allelic states of each marker had nonsignificant co-segregation with the trait.

An inverse relationship was noted between TCC and DMC (Figure 6) for genotypes assayed. When the allelic further profiles

TABLE 2 Frequency of 13 KASP marker genotypes segregating for CMD, DMC and TCC in Uganda's cassava breeding population.

Trait	Marker	N	Hom1	Het	Hom2	%Hom1	%Het	%Hom2	% Null
CMD	S12_7926132	1,024	GG	TG	TT	19.9	59.3	20.2	0.6
CMD	S12_7926163	1,024	AA	GA	GG	19.9	59.5	20.5	0.1
CMD	S14_4626854	1,024	AA	AG	GG	14.9	56.1	28.7	0.3
DMC	S1_24197219	653	TT	TC	CC	46.5	28.3	7.0	0.2
DMC	S6_20589894	653	GG	GA	AA	30.2	52.5	16.7	0.6
DMC	S12_5524524	653	CC	CT	TT	8.1	49.3	42.1	0.5
DMC	S15_1012346	653	CC	CT	TT	16.7	51.1	16.7	15.5
TCC	PSY2_572	653	AA	CA	CC	61.1	29.9	8.3	0.8
TCC	S1_24636113	653	GG	GA	AA	55.3	32.5	9.8	2.5
TCC	S1_30543962	653	GG	GA	AA	19.9	47.3	30.5	2.3
TCC	S5_3387558	653	TT	TC	CC	0.6	75.8	23.1	0.5
TCC	S8_25598183	653	TT	TG	GG	70.8	27.6	1.1	0.6
TCC	S15_7659426	653	GG	GT	TT	49.3	44.1	6.4	0.2

Hom1, homozygous for minor allele; Het, heterozygote; Hom2, homozygous for major allele; Null, uncallable genotypes.

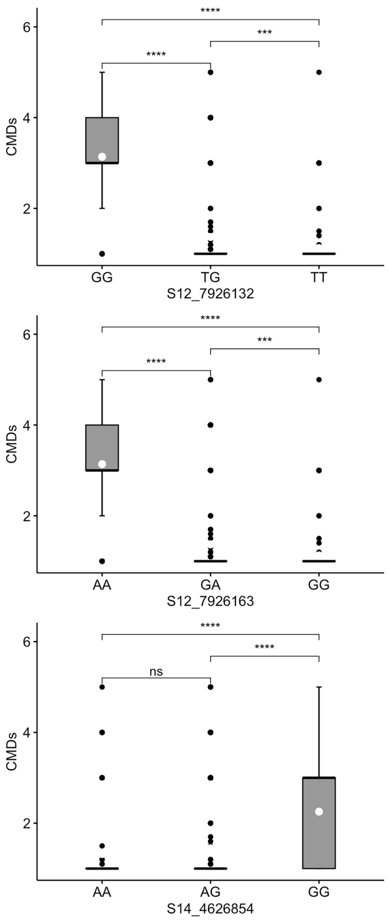


FIGURE 3
Box plots showing co-segregation of three KASP markers with cassava mosaic disease severity (CMDs). ****, *** significant at p ≤ 0.0001, p ≤ 0.001 and p ≤ 0.01, respectively; ^{ns} = non-significant.

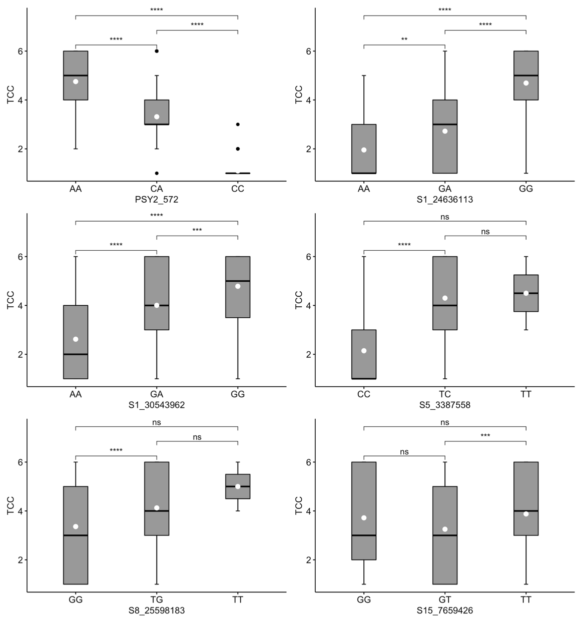


FIGURE 4
Box plots showing co-segregation of six KASP markers with total carotenoids content (TCC). ****, ***, ** significant at p ≤ 0.0001, p ≤ 0.001 and p ≤ 0.01, respectively; ^{ns} = non-significant.

of best performing markers for TCC (PSY2_572) and DMC (S1_24197219) were further examined, all genotypes homozygous for the favorable allele for TCC (AA) did not have the favorable allele for DMC (Figure 6) and had generally low levels of DMC. Similarly, genotypes homozygous for the favorable allele for DMC (CC) did not have the favorable allele for TCC and had low total carotenoids content, with exception of one genotype scoring CA/CC. However, there were 116 genotypes combining heterozygous states of the two markers (CA/TC).

When we regressed the marker genotypes onto the respective phenotypes, all markers showed significant association with traits, except two markers (S8_25598183 and S15_7659426) for TCC (Table 3). However, the proportion of phenotypic variance attributable to the markers was generally low, with the highest value recorded for CMD markers S12_7926132 and S12_7926163 (R^2 = 0.45). Only one marker (S1_24197219) accounted for up to 30% of variation in DMC, with the remaining three explaining ≤5% of the phenotypic variance. Meanwhile, markers PSY2_572 and S1_24636113 accounted for the 42% and 30%, respectively. All other TCC markers had low R^2 values, with S1_30543962 accounting for 15% of the phenotypic variation and S15_7659426 explaining none of the observable variation in the trait.

Principal component analysis further revealed similarities between markers located within the same chromosomal positions. For example, CMD markers S12_7926132 and S12_7926163 had identical contribution to the total variation in the principal components; the same observation was true for TCC markers PSY2_572, S1_24636113 (Supplementary Figure 1). Such markers would be considered redundant to each other.

Predictive ability of CMD markers based on biological metrics

We used biological metrics to further investigate the predictive ability of CMD markers, given qualitative nature of the trait. In this case, markers S12_7926132 and S12_7926163 had comparable predictive abilities. For example, both markers had high prediction accuracy (86.7%) and relatively low false positive rate (~38%) (Table 4). In contrast, marker S14_4626854 had somewhat low prediction accuracy (74.5%) and a relatively high false positive rate (44.3%).

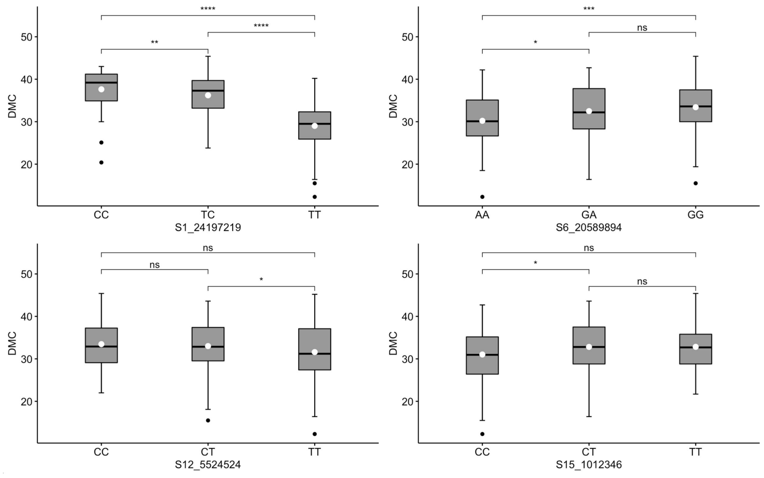


FIGURE 5
Box plots showing co-segregation of six KASP markers with dry matter content (DMC). ****, ***, **, * significant at $p \leq 0.0001$, $p \leq 0.001$, $p \leq 0.01$ and $p \leq 0.05$, respectively; ^{ns} = non-significant.

Discussion

The global agricultural production is currently facing unprecedented challenges imposed by rapid human population growth, limited arable land and adverse effects of climate change. These challenges call for greater efforts to optimize and deploy appropriate tools, technologies and methods that can accelerate genetic gains from breeding programs for rapid delivery of high-capacity varieties to farmers. Advances in sequencing technologies can now allow for cost-effective use of genome-wide markers (Elshire et al., 2011) for identification of and use of trait-linked DNA polymorphisms. In the case of cassava, various research investments, including those through the Next Generation Cassava Breeding project (https://www.nextgencassava.org/) have yielded important genomic resources, such as well annotated genome sequence (Prochnik et al., 2012) and collation SNP markers for various agronomic and quality traits (Rabbi et al., 2022). Thus, this study tested the effectiveness of

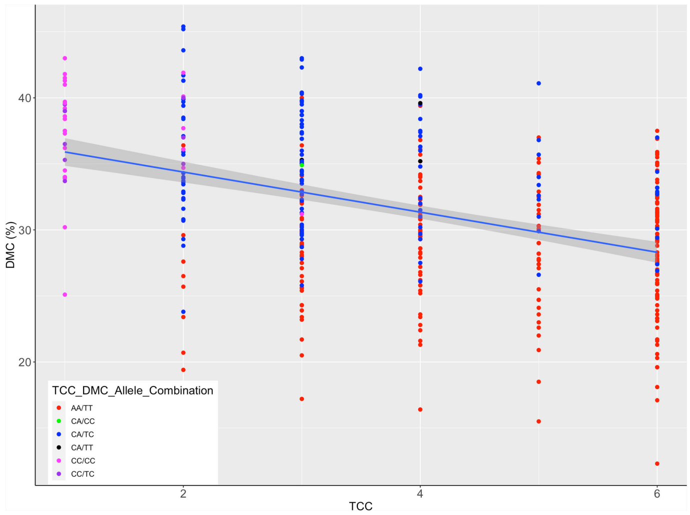


FIGURE 6
Scatter plot of total carotenoids content (TCC) and dry matter content (DMC) scaled by alleles for markers PSY2_572 and S1_24197219; the legend shows TCC/DMC allelic combinations of the two markers in genotypes assayed; the allelic states correspond to homozygous favorable alleles, heterozygotes and the wild type, as indicated in Figures 4, 5.

TABLE 3 Regression coefficients for 13 KASP markers tested for CMD, DMC and TCC in Ugandan cassava germplasm.

Trait	Marker	df	ms	R^2
CMD	S12_7926132	2	322.8***	0.45
CMD	S12_7926163	2	322.6***	0.45
CMD	S14_4626854	2	96.6***	0.13
DMC	S1_24197219	2	1897.6***	0.30
DMC	S6_20589894	2	299.6***	0.05
DMC	S12_5524524	2	241.0***	0.04
DMC	S15_1012346	2	261.7***	0.04
TCC	PSY2_572	2	301.5***	0.42
TCC	S1_24636113	2	210.2***	0.30
TCC	S1_30543962	2	56.9***	0.15
TCC	S5_3387558	2	91.5***	0.06
TCC	S8_25598183	2	3.9	0.01
TCC	S15_7659426	2	2.7	0.00

CMD, cassava mosaic disease; DMC, dry matter content; TCC, total carotenoid content; df, degree of freedom; ms, mean square; R^2, coefficient of determination and; ***, significance at $p \leq 0.001$.

selected trait-linked KASP markers identified from West African gene pool for MAS in Uganda's cassava breeding population as an independent validation set. By deploying trait-linked markers for MAS, cassava breeders could efficiently select genotypes with desired trait combinations at seedling stage so that inferior individuals for a specific trait are quickly discarded. With a reduced number of clones in subsequent selection stages, breeders can then shift to index selection where multiple traits are improved simultaneously, thereby increasing the speed of variety development and reducing cost of breeding operations (Ceballos et al., 2015).

The cassava improvement program in Uganda implements a demand-led breeding, in which a stage-gate product development is inspired by needs of two main market segments: food (boiled roots and flour) and industry. Dry matter content, provitamin A carotenoids content and virus disease resistance are must-have traits for cassava product profile in Uganda (Iragaba et al., 2020; Esuma et al., 2021). The two cassava breeding populations used in this study have three important attributes that would warrant the use of MAS for rapid improvement of these traits. Firstly, the wide phenotypic variability for CMD, DMC and TCC justifies the use of markers that can efficiently pinpoint desired genotypes in a large segregating population for forward breeding. In fact, DMC and TCC exhibited substantial degree of quantitative variation for which phenotypic selection alone can be slow and expensive. Secondly, the low broad-sense heritability of DMC implies the need for robust field phenotyping before accurate decisions can be taken on a genotype's genetic merit for the trait. Thirdly, all the three traits can only be optimally estimated on physiologically mature (≥ months old) plants.

The robust and high call rates for the marker genotypes demonstrated their usefulness in screening cassava germplasm of broad genetic background. However, the wide variability in the predictive ability of the markers suggests the need for their case-by-case deployment in cassava breeding. For example, the consistent and significant co-segregation of allelic states of markers S12_7926132 and S12_7926163 with CMD severity indicates their reliability for MAS. In fact, markers S12_7926132 and S12_7926163 may be tightly linked to the functional gene in conferring CMD resistance in cassava given their close proximity to CMD2 resistance locus previously identified on chromosome 12 (Okogbenin et al., 2012; Wolfe

TABLE 4 Predictive ability of three KASP markers for CMD in Uganda's cassava breeding population.

Marker	Prediction	True phenotype		Accuracy (%)	FPR (%)	FNR (%)
		R	S			
S12_7926132	R	699	114	86.7	38.4	2.9
	S	21	183			
S12_7926163	R	703	115	86.7	38.7	2.9
	S	21	182			
S14_4626854	R	594	132	74.5	44.3	17.7
	S	128	166			

R, Resistant; S, Susceptible; FPR, false positive rate; FNR, false negative rate.

et al., 2016; Rabbi et al., 2022). Given, the short distance separating markers S12_7926132 and S12_7926163, using either of them would be sufficient for implementing MAS for CMD2 resistance.

The allelic segregation of markers S12_7926132 and S12_7926163 points to two important aspects. Foremost, the consistent dominant allelic effects exhibited in the marker segregation underpin CMD2 is a qualitative trait under additive genetic effect. Despite the classical qualitative segregation exhibited for CMDs, there was a considerable variation in severity levels for susceptible clones, which could be attributed to differences in plot-based viral loads or other genetic factors controlling plant fitness. Secondly, the large proportion of variation in CMDs unexplained by the marker effects indicates the need for continuous improvement for effective operationalization of MAS platforms in cassava. Through whole genome sequencing and genetic variant analysis, (Lim et al., 2022) fine-mapped the CMD2 locus to a 190 kilobase and identified additional nonsynonymous SNP in *DNA polymerase δ subunit 1 (MePOLD1)* as the functional gene on chromosome 12 responsible for CMD2 resistance. That study generated eight novel KASP markers that, when incorporate into the existing genomic resource, could reinforce prospects of MAS for CMD2 resistance in cassava breeding.

TCC markers PSY2_572, S1_24636113 and S1_30543962 segregated in typical dominant fashion, with additive effects exhibited for allele substitution. These markers are located in a close proximity of *Manes.01G124200.1*, which is a *phytoene synthase* (*PSY*) gene known to increase accumulation of provitamin A carotenoids in cassava roots (Esuma et al., 2016; Rabbi et al., 2017). A detailed characterization of the *PSY* locus by Welsch et al. (2010) indicated that a SNP in *PSY2-Y-2* gene co-segregated with high carotenoids content in cassava roots and the polymorphism resulted in a single amino acid change in a highly conserved region of the protein which increased catalytic activity in *Escherichia coli*. Indeed, *PSY* has also been reported to encode the expression of threonine, a major substance in the carotenoid biosynthetic pathway, in other plant species such as maize varieties with yellow endosperm (Shumskaya et al., 2012; Shumskaya and Wurtzel, 2013) and golden rice (Paine et al., 2005). Thus, it is likely that markers PSY2_572, S1_24636113 and S1_30543962 are in strong linkage disequilibrium with the functional locus controlling synthesis and/or accumulation of provitamin A carotenoids content in cassava roots, and their use for MAS for TCC would be effective. However, marker S1_30543962 accounted for a low proportion of variation in TCC, indicating its ineffectiveness for MAS when used in isolation.

The continuous variability in DMC was typical of a quantitative trait be controlled by many genes with small effects. Nonetheless, markers S1_24197219 and S6_20589894 showed strong co-segregation with DMC. Rabbi et al. (2017) identified a genomic region on chromosome 1 associated with DMC in cassava. The authors annotated two genes *UDP-glucose pyrophosphorylase* and *sucrose synthase* within the DMC association signal. Both genes are known to be essential in the synthesis of sucrose and polysaccharide (Zabotina et al., 2021). Thus, marker S1_24197219, which accounted for the highest proportion of variation in DMC (30%) in our study, could be tightly linked to the functional genes for the trait on chromosome 1 and would be effective for MAS. Based on the segregation pattern and low R^2 values for DMC, other markers would be ineffective for implementing MAS, especially in the genetic material evaluated in this study.

The apparent negative correlation between TCC and DMC is a manifestation of the current challenges baffling breeding efforts aimed at delivering cassava varieties with desired end-user traits. In the case of market segment for boiled cassava, there is a high preference for roots with elevated levels of DMC (Iragaba et al., 2020). The KASP marker-based selection tested in this study shows the prospect for rapid and efficient identification of genotypes combining favorable alleles for TCC and DMC at early selection stages and could aid the implementation a rapid cycle recurrent selection scheme in cassava genetic improvement (Ceballos et al., 2015).

Taken together, data presented in this study highlight some prospects for MAS in cassava, especially for CMD, provitamin A carotenoids and DMC. Our data highlighted five markers with sufficient discriminatory ability for MAS: S12_7926132 and S12_7926163 for CMD, PSY2_572 and S1_24636113 for TCC, and S1_24197219 for DMC. The markers should be the focus for cassava breeding programs targeting immediate application of MAS for genetic improvement of these traits. As more genomic tools and resources get optimized for cassava, marker-assisted breeding will become a reality and greatly help increase genetic gains for important agronomic and quality traits that have been too complex to exploit through conventional breeding methods (Ceballos et al., 2012; Ceballos et al., 2015). In the meantime, efforts should be made to enhance the utility and deployment of these markers across the global cassava community to facilitate rapid delivery of varieties that can contribute to reducing poverty and ending hunger, as desired by the first two sustainable development goals.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://doi.org/10.6084/m9.figshare.21213605.v2.

Author contributions

WE conceptualized the study, guided field trials and wrote the original manuscript. OE phenotyped trial for CMD and reviewed manuscript. FG phenotyped trial for DMC and TCC. SM guided field trials and edited manuscript. TA guided field trials and edited manuscript. EN guided laboratory analyses for DMC/TCC and edited the manuscript. AO guided data analyses and reviewed manuscript. IR guided marker data analyses and edited manuscript. RK guided the study conceptualization and provided funds. All authors contributed to the article and approved the submitted version.

Funding

This work was supported through the Next Generation Cassava Breeding project (www.nextgencassava.org/) funded by the Bill and Melinda Gates Foundation through Cornell University (grant number OPP1048542).

Acknowledgments

We thank IR (IITA, Ibadan) for providing detailed technical information on the 13 KASP markers.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer XZ is currently organizing a Research Topic with one of the authors RK.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1017275/full#supplementary-material

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Ceballos, H., Kawuki, R. S., Gracen, V. E., Yencho, G. C., and Hershey, C. H. (2015). Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: A case study for cassava. *Theor. Appl. Genet.* 128, 1647–1667. doi: 10.1007/s00122-015-2555-4

Ceballos, H., Kulakow, P., and Hershey, C. (2012). Cassava breeding: Current status, bottlenecks and the potential of biotechnology tools. *Trop. Plant Biol.* 5, 73–87. doi: 10.1007/s12042-012-9094-9

Chagné, D., Vanderzande, S., Kirk, C., Profitt, N., Weskett, R., Gardiner, S. E., et al. (2019). Validation of SNP markers for fruit quality and disease resistance loci in apple (Malus × domestica borkh.) using the OpenArray® platform. *Horticul. Res.* 6, 30. doi: 10.1038/s41438-018-0114-2

Chávez, A. L., Sánchez, T., Jaramillo, G., Bedoya, J. M., Echeverry, J., Bolaños, E. A., et al. (2005). Variation of quality traits in cassava roots evaluated in landraces and improved clones. *Euphytica* 143, 125–133. doi: 10.1007/s10681-005-3057-2

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Esuma, W., Herselman, L., Labuschagne, M. T., Ramu, P., Lu, F., Baguma, Y., et al. (2016). Genome-wide association mapping of provitamin a carotenoid content in cassava. *Euphytica* 212, 97–110. doi: 10.1007/s10681-016-1772-5

Esuma, W., Ozimati, A., Kulakow, P., Gore, M. A., Wolfe, M. D., Nuwamanya, E., et al. (2021). Effectiveness of genomic selection for improving provitamin a carotenoid content and associated traits in cassava. *G3 Genes|Genomes|Genet.* 11, 1–10. doi: 10.1093/g3journal/jkab160

Ezenwaka, L., Del Carpio, D. P., Jannink, J., Rabbi, Y. R., Danquah, E., Asante, I. K., et al. (2018). Genome-wide association study of resistance to cassava green mite pest and related traits in cassava. *Crop Sci.* 28, 1907–1918. doi: 10.2135/cropsci2018.01.0024

Ferguson, M. E., Rabbi, I., Kim, D.-J., Gedil, M., Lopez-Lavalle, L. A. B., and Okogbenin, E. (2011). Molecular markers and their application to cassava breeding: Past, present and future. *Trop. Plant Biol.* 5, 95–109. doi: 10.1007/s12042-011-9087-0

Grewal, S., Coombes, B., Joynson, R., Hall, A., Fellers, J., Yang, C., et al. (2022). Chromosome-specific KASP markers for detecting amblyopyrum muticum segments in wheat introgression lines. *Plant Genome.* 15, e20193. doi: 10.1002/tpg2.20193

Ige, A. D., Olasanmi, B., Mbanjo, E. G. N., Kayondo, I. S., Parkes, E. Y., Kulakow, P., et al. (2021). Conversion and validation of uniplex SNP markers for selection of resistance to cassava mosaic disease in cassava breeding programs. *Agronomy* 11, 420. doi: 10.3390/agronomy11030420

IITA. (1990). Cassava in tropical Africa: A reference manual. Ibadan, Nigeria: International Institute of Tropical Agriculture. Available online: https://www.iita.org/wp-content/uploads/2016/06/Cassava_in_tropical_Africa_a_reference_manual_1990.pdf.

Iragaba, P., Hamba, S., Nuwamanya, E., Kanaabi, M., Nanyonjo, R. A., Mpamire, D., et al. (2020). Identification of cassava quality attributes preferred by Ugandan users along the food chain. *Int. J. Food Sci. Technol.* 56, 1184–1192. doi: 10.1111/ijfs.14878

Kayondo, S. I., Pino Del Carpio, D., Lozano, R., Ozimati, A., Wolfe, M., Baguma, Y., et al. (2018). Genome-wide association mapping and genomic prediction for CBSD resistance in manihot esculenta. *Sci. Rep.* 8, 1549. doi: 10.1038/s41598-018-19696-1

Kolawole, P. O., Agbetoye, L., and Ogunlowo, S. A. (2010). Sustaining world food security with improved cassava processing technology: The Nigeria experience. *Sustainability* 2, 3681–3694. doi: 10.3390/su2123681

Lim, Y.-W., Mansfeld, B. N., Schläpfer, P., Gilbert, K. B., Narayanan, N. N., Qi, W., et al. (2022). Mutations in DNA polymerase δ subunit 1 co-segregate with CMD2-type resistance to cassava mosaic geminiviruses. *Nat. Commun.* 13, 1–11. doi: 10.1038/s41467-022-31414-0

Ogola, J. B. O., and Mathews, C. (2011). Adaptation of cassava (Manihot esculenta) to the dry environments of Limpopo, south Africa: Growth, yield and yield components. *Afr. J. Agric. Res.* 6, 6082–6088. doi: 10.5897/AJAR11.764

Okogbenin, E., Egesi, C. N., Olasanmi, B., Ogundapo, O., Kahya, S., Hurtado, P., et al. (2012). Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. *Crop Sci.* 52, 2576–2586. doi: 10.2135/cropsci2011.11.0586

Oliveira, E. J., Resende, M. D. V., Silva Santos, V., Ferreira, C. F., Oliveira, G. A. F., Silva, M. S., et al. (2012). Genome-wide selection in cassava. *Euphytica* 187, 263–276. doi: 10.1007/s10681-012-0722-0

Paine, J. A., Shipton, C. A., Chaggar, S., Howells, R. M., Kennedy, M. J., Vernon, G., et al. (2005). Improving the nutritional value of golden rice through increased pro-vitamin a content. *Nat. Biotechnol.* 23, 482–487. doi: 10.1038/nbt1082

Prochnik, S., Marri, P. R., Desany, B., Rabinowicz, P. D., Kodira, C., Mohiuddin, M., et al. (2012). The cassava genome: Current progress, future directions. *Trop. Plant Biol.* 5, 88–94. doi: 10.1007/s12042-011-9088-z

Rabbi, I. Y., Hamblin, M. T., Kumar, P. L., Gedil, M. A., Ikpan, A. S., Jannink, J. L., et al. (2014). High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. *Virus Res.* 186, 87–96. doi: 10.1016/j.virusres.2013.12.028

Rabbi, I. Y., Kayondo, S. I., Bauchet, G., Yusuf, M., Aghogho, C. I., Ogunpaimo, K., et al. (2022). Genome-wide association analysis reveals new insights into the genetic architecture of defensive, agro-morphological and quality-related traits in cassava. *Plant Mol. Biol.* 109, 195–213. doi: 10.1007/s11103-020-01038-3

Rabbi, I. Y., Udoh, L. I., Wolfe, M., Parkes, E. Y., Gedil, M. A., Dixon, A., et al. (2017). Genome-wide association mapping of correlated traits in cassava: Dry matter and total carotenoid content. *Plant Genome.* 10, 1–14. doi: 10.3835/plantgenome2016.09.0094

Rosso, M. L., Chang, C., Song, Q., Escamilla, D., Gillenwater, J., and Zhang, B. (2021). Development of breeder-friendly KASP markers for low concentration of kunitz trypsin inhibitor in soybean seeds. *Int. J. Mol. Sci.* 22, 2675. doi: 10.3390/ijms22052675

Sandhu, N., Singh, J., Singh, G. P., Sethi, M., Singh, M., Pruthi, G., et al. (2022). Development and validation of a novel core set of KASP markers for the traits improving grain yield and adaptability of rice under direct-seeded cultivation conditions. *Genomics* 114, 110269. doi: 10.1016/j.ygeno.2022.110269

Shumskaya, M., Bradbury, L. M. T., Monaco, R. R., and Wurtzel, E. T. (2012). Plastid localization of the key carotenoid enzyme phytoene synthase is altered by isozyme, allelic variation, and activity. *Plant Cell.* 24, 3725–3741. doi: 10.1105/tpc.112.104174

Shumskaya, M., and Wurtzel, E. T. (2013). The carotenoid biosynthetic pathway: thinking in all dimensions. *Plant Sci.* 208, 58–63. doi: 10.1016/j.plantsci.2013.03.012

Uke, A., Tokunaga, H., Utsumi, Y., Vu, N. A., Nhan, P. T., Srean, P., et al. (2022). Cassava mosaic disease and its management in southeast Asia. *Plant Mol. Biol.* 109, 301–311. doi: 10.1007/s11103-021-01168-2

Welsch, R., Arango, J., Bar, C., Salazar, B., Al-Babili, S., Beltran, J., et al. (2010). Provitamin a accumulation in cassava (Manihot esculenta) roots driven by a single nucleotide polymorphism in a phytoene synthase gene. *Plant Cell.* 22, 3348–3356. doi: 10.1105/tpc.110.077560

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., et al. (2016). Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *Plant Genome.* 9, 1–13. doi: 10.3835/plantgenome2015.11.0118

Zabotina, O. A., Zhang, N., and Weerts, R. (2021). Polysaccharide biosynthesis: Glycosyltransferases and their complexes. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.625307

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/fpls.2022.1018156

Check for updates

OPEN ACCESS

EDITED BY
Bernardo Ordas,
(CSIC), Spain

REVIEWED BY
Rodrigo R. Amadeu,
St. Louis, United States
Suresh L. M.,
The International Maize and Wheat Improvement Center (CIMMYT), Kenya

*CORRESPONDENCE
Alfred A. Ozimati
ozimatiafred@gmail.com

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 12 August 2022
ACCEPTED 24 October 2022
PUBLISHED 23 November 2022

CITATION
Ozimati A, Esuma W, Manze F, Iragaba P, Kanaabi M, Ano C, Egesi C and Kawuki RS (2022) Utility of Ugandan genomic selection cassava breeding populations for prediction of cassava viral disease resistance and yield in West African clones. *Front. Plant Sci.* 13:1018156. doi: 10.3389/fpls.2022.1018156

Utility of Ugandan genomic selection cassava breeding populations for prediction of cassava viral disease resistance and yield in West African clones

Alfred A. Ozimati[1,2]*, Williams Esuma[1], Francis Manze[1], Paula Iragaba[1], Michael Kanaabi[1], Chukwuka Ugochukwu Ano[3], Chiedozie Egesi[3,4,5] and Robert S. Kawuki[1]

[1]National Crops Resources Research Institute, Kampala, Uganda, [2]Department of Plant Sciences, Microbiology and Biotechnology, College of Natural Sciences, Makerere University, Kampala, Uganda, [3]Plant Breeding and Genetics Section, College of Agricultare and Life Sciences, Cornell University, Ithaca, NY, United States, [4]National Root Crops Research Institute, Umudike, Nigeria, [5]International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria

Cassava (*Manihot esculenta* Crantz) is a staple crop for ~800 million people in sub-Saharan Africa. Its production and productivity are being heavily affected by the two viral diseases: cassava brown streak disease (CBSD) and cassava mosaic disease (CMD), impacting greatly on edible root yield. CBSD is currently endemic to central, eastern and southern Africa, if not contained could spread to West Africa the largest cassava producer and consumer in the continent. Genomic selection (GS) has been implemented in Ugandan cassava breeding for accelerated development of virus resistant and high yielding clones. This study leveraged available GS training data in Uganda for pre-emptive CBSD breeding in W. Africa alongside CMD and fresh root yield (FRW). First, we tracked genetic gain through the current three cycles of GS in Uganda. The mean genomic estimated breeding values (GEBVs), indicated general progress from initial cycle zero (C0) to cycle one (C1) and cycle two (C2) for CBSD traits and yield except for CMD. Secondly, we used foliar data of both CBSD and CMD, as well as harvest root necrosis and yield data to perform cross-validation predictions. Cross-validation prediction accuracies of five GS models were tested for each of the three GS cycles and West African (WA) germplasm as a test set. In all cases, cross-validation prediction accuracies were low to moderate, ranging from -0.16 to 0.68 for CBSD traits, -0.27 to 0.57 for CMD and -0.22 to 0.41 for fresh root weight (FRW). Overall, the highest prediction accuracies were recorded in C0 for all traits tested across models and the best performing model in cross-validation was G-BLUP. Lastly, we tested the predictive ability of the Ugandan training sets to predict CBSD in W. African clones. In general, the Ugandan training sets had low prediction accuracies for all traits across models in West African germplasm, varying from -0.18 to 0.1. Based on the findings of this study, the cassava breeding program in Uganda has made progress through application of GS for most target traits, but the utility of the training population for pre-emptive breeding in WA is limiting. In this case, efforts should be devoted to sharing Ugandan germplasm that possess resistance with the W. African breeding programs for hybridization to fully enable deployment of genomic selection as a pre-emptive CBSD breeding strategy in W. Africa.

KEYWORDS

cassava, genomic prediction, training population and viral diseases, models, validation set

Introduction

The raising energy demand globally in face of climate change is popularizing cassava as an alternative source of renewable fuel with full potential to replace fossil fuel in the developed countries (Kang et al., 2014). Besides, being a potential crop to generate renewable fuel at global level, cassava is a major of source of carbohydrate and staple food for over 800 million people in the world (Hammond et al., 2013). Because of the global importance of the cassava, its production has steadily increased world-wide in last the two decades from 162 MT in 1998 to 303 MT in 2018 (FAO, 2019), with the world's highest production of ~ 60 MT coming from Nigeria in W. Africa.

Despite the importance of cassava as food security, especially in sub-Saharan Africa, average yields still remain low (12 t/ha) compared with yield average of 20 t/ha recorded in Asia countries like Thailand (Nweke, 2004). A number of biotic and abiotic factors contribute to this yield gap in sub-Saharan Africa. The leading biotic stress being cassava brown steak disease (CBSD) and cassava mosaic disease (CMD) (Legg et al., 2014). While CMD is present in all cassava producing areas in Africa and Asia, CBSD is only endemic to Eastern and Southern and Central Africa and more recently the disease was reported in Angola, which is closer to West Africa especially Nigeria, the largest cassava producer and consumer in the world (Ano et al., 2021). In highly susceptible varieties, yield losses of up to 100% have been reported (Hillocks et al., 2002; Alicai et al., 2007). The recent epidemiological studies indicate that CBSD is fast spreading to West Africa (Patil et al., 2015), and thus posing an eminent threat to cassava production in the Western part of the continent.

Fortunately, cassava that lagged previously in terms of genomic resources relative to cereal crops like maize, rice and wheat and legumes such as common beans, ground nuts and soya beans, has received significant funding to develop complete reference genome assembly (Prochnik et al., 2012). With the availability of the genomic resources for cassava and low-cost genotyping technologies such as genotyping-by-sequencing (Elshire et al., 2011; Rabbi et al., 2015) and more recently the diversity array technology sequencing (DArTSeq) platform, cassava breeding is evolving from traditional phenotypic selection to selecting plants based on their genomic estimated breeding values (GEBVs). Genomic selection (GS), which uses high-density markers to cover the entire genome, was proposed by Meuwissen et al. (2001) as a new method for selection of individuals in a population based on the breeding values.

Genomic selection has been reported to offer some advantages over phenotypic selection breeding scheme: (i) genomic selection allows for more cycles of recombination per unit time than phenotypic selection, (ii) selection is solely based on estimates of marker effects without prior knowledge of the QTL and also captures variation due to loci with small effects (de Oliveira et al., 2012). Another argument put in favor of genomic selection is that genotyping cost will further decrease per sample; on the other hand, phenotyping costs do not exhibit the same downward trend, because they are dependent on human resources and agricultural inputs. The cost of these resources have historically been increasing (de Oliveira et al., 2012; Poland and Rife, 2012).

Through the Next Generation Cassava Breeding project, Uganda embraced genomic selection tool for cassava improvement in early 2010s. The breeding program has so far developed three recurrent genomic selection cycles, and some of the elite material has been channeled to the variety development pipeline. Our primary traits of focus include: CMD resistance, fresh root weight, end-user root quality attributes (Wolfe et al., 2017) and CBSD resistance. From historical data, Uganda has registered significant gains for CMD and CBSD resistance breeding efforts (Manze et al., 2021), and thus could offer CBSD resistant parents to west African cassava breeding programs such as Nigeria where CBSD is not yet a threat. However, because of the CBSD pandemic in Eastern, Southern and Central Africa, there is restriction on moving plant materials currently from Eastern, Central and Southern Africa to West Africa, where CBSD is non-existent (Ano et al., 2021). Nonetheless, illegal movement of plant materials due to porous borders, besides the whitefly supported transmission could pose a risk of CBSD reaching to West Africa (Legg et al., 2014).

We previously leveraged on the available genomic resources under Next Generation Cassava Breeding project (https://www.

nextgencassava.org/) to predict CBSD in West Africa using 35 clones shared from IITA, Ibadan (Ozimati et al., 2018). Generally, low predictive ability, ranging from 0.14 to 0.36 for CBSD foliar severity, and -0.29 to 0.11 for CBSD root necrosis (Ozimati et al., 2018) were recorded. Building on previous CBSD pre-emptive breeding study, which was limited by the sample size, we expanded on sample size of the West Africa test set used in the current study. Specifically, we assessed gains from genomic selection for virus disease resistance and fresh root weight in Ugandan GS training populations, and further evaluated effectiveness of the training sets in predicting CBSD, CMD resistance and fresh root yield in WA clones as pre-emptive CBSD breeding strategy.

Materials and Methods

Germplasm and field evaluation

The training population comprised three recurrent genomic selection cycles obtained from NaCRRI. These cycles were: cycle zero (C0), cycle one (C1) and cycle two (C2). Briefly, C0 population was derived from forty-nine diverse progenitors that were assembled from International Institute of Tropical Agriculture (IITA), International Center for Tropical Agriculture (CIAT) and NaCRRI. Germplasm from CIAT (Columbia) targeted improvement of quality and yield traits, while germplasm from the IITA (Tanzania), and NaCRRI (Uganda) breeding programs targeted improvement of CBSD resistance. Botanical seeds from crosses (full-sibs and half-sibs) of forty-nine progenitors were planted in a seedling nursery at Namulonge, and the sprouted seedlings were evaluated in an unreplicated seedling trial at Namulonge in 2012. A total of 466 C0 seedlings were selected visually as a training population for implementation of GS based on their CMD and CBSD resistance, and evaluated for two years (2013 and 2014) at Namulonge (central Uganda), Kasese (mid-western Uganda) and Ngetta (northern Uganda), using an alpha lattice design with two replications. Namulonge, Kasese and Ngetta were specifically chosen because of high viral disease pressure (cassava brown streak disease and cassava mosaic disease) and whitefly (vector) populations (Alicai et al., 2019). The C1 clones were derived from recurrent selection and recombination of the best a hundred C0 clones selected through GS. A total of 667 C1 seedlings were selected visually and evaluated in a clonal trial that was laid out using an augmented randomized block design at both Namulonge and Serere in 2016 and 2017. Similarly, the top hundred performers selected from C1 clonal trial were recombined to generate the C2 population. The C2 clonal trial comprised 421 clones and was also laid out using an augmented randomized block design in 2019 at Namulonge for one season. Selection of progenitors for constitution of C1 and C2 were based on CMD resistance, CBSD resistance, harvest index and fresh root yield.

All clones in the training set (C0, C1 and C2) were evaluated for CBSD and CMD severity, fresh root weight and harvest index. CBSD foliar severity was assessed at three (CBSD3S) and six CBSD6S) months after planting using a standard scale of 1-5; where 1 = no apparent symptoms, 2 = slight foliar chlorosis, but with no stem lesions, 3 = pronounced foliar chlorosis and mild stem lesions with no die back, 4 = severe foliar chlorosis and severe stem lesions with no die back, and 5 = defoliation, severe stem lesions and die back (Gondwe et al., 2003). Cassava mosaic disease was also assessed at three (CMD3S) and six (CMD6S) after planting using a scale of 1 to 5; where 1 = no visible disease symptoms, 2 = mild chlorotic pattern on entire leaflets or mild distortion at base of leaflets, rest of leaflets appearing green and healthy, 3 = strong mosaic pattern on entire leaf, and narrowing and distortion of lower one-third of leaflets, 4 = severe mosaic, distortion of two-thirds of leaflets and general reduction of leaf size, and 5 = severe mosaic, distortion of four-fifths or more of leaflets, twisted and misshapen leaves (IITA, 1990).

At twelve months after planting, clonal trials were harvested to allow evaluation of fresh root weight (FRW) and cassava brown streak root necrosis severity (CBSDRS). All the ten plants were harvested and partitioned into roots and above-ground biomass (leaves and stems). Fresh root weight (FRW) and above-ground biomass were separately weighed (kg plot^{-1}) using a hanging weighing scale of 200 kg capacity. On the other hand, CBSDRS was recorded on all harvested roots per plot using a scale of 1-5; where 1 = no observable necrosis, 2 = ≤ 5% of root necrotic, 3 = 6 to 25% of root necrotic, 4 = 26 to 50% of root necrotic with mild root constriction, and 5 = > 50% of root necrosis with severe root constriction (Gondwe et al., 2003).

The validation set comprised germplasm that was sourced from National Roots Crops Research Institute (NRCRI), Nigeria. A total of 5,000 botanical seeds were generated from bi-parental crosses involving forty-eight elite progenitors. The progenitors were selected per se based on their yielding ability and resistance to cassava mosaic disease (CMD). Accordingly, these seeds were shipped and planted in a seedling nursery at Namulonge. Out of the 5000 botanical seeds, 1980 successfully emerged, giving rise to 106 families. The 1980 seedlings were thus established in an unreplicated seedling trial during the second rains of 2018 (September/October). A total of 569 clones were selected from the seedling trial for further evaluation at the clonal stage during the 2019-2020 season, which was laid out using an augmented randomized block design at Namulonge. At the end of clonal evaluation, only 297 clones remained, as half of the clones were directly culled by CBSD. These 297 clones constituted the validation set for genomic prediction, and were assessed for CMD and CBSD severity fresh root yield as the 3rd trait evaluated, following the same procedure previously described for the training population.

Genotyping of the training and validation sets

Leaf samples were obtained from the clonal evaluation stage of the training (C0, C1, C2) and the validation (germplasm from Nigeria) populations and shipped to Intertek, Australia, for DNA extraction and genotyping. Both C0 and C1 clones were originally genotyped using genotyping by sequencing (GBS) platform with 46K single nucleotide polymorphism (SNP) chip at Genomic Diversity facility of Cornell University (Ozimati et al., 2018). However, because the National Cassava Breeding Program of Uganda recently opted for Diversity Arrays Technology (DArT) genotyping services for routine genomic selection work, SNP markers from GBS (for both C0 and C1) were later imputed with those from DArT platform, giving rise to 23K SNP markers for genomic selection. Genotyping of the C2 and validation population (germplasm from Nigeria) was therefore done by the DArT platform, Australia, using the same 23K SNP markers that had been used previously to genotype C0 and C1 populations. Missing markers of the genotyped individuals were filled in by imputation, using markers from the East Africa imputation reference panel using BEAGLE software version 5.0 (Browning and Browning, 2007). The markers were thereafter filtered, and those with minor allele frequency (MAF) greater than 0.01 (21,938 SNPs) were used for downstream analyses.

We used the 21,938 SNPs to assess the population structure of the training population from Uganda (C0, C1 and C2) and validation population from West Africa. The SNP genotypes were coded as -1, 0, or +1. Principal component analysis (PCA) was done on scaled SNP markers using the *prcomp* function in R. The first two principal components (PC) were used to visualize population structure.

Estimates of broad-sense heritability, genetic gain and accuracy of genomic prediction

To estimate heritability for each trait per cycle of GS population (C0, C1 and C2) and the WA clones, we fitted linear mixed models based on experimental design for each trial, followed by extraction of variance components using restricted maximum likelihood procedure (Spilke et al., 2005). The variance components were then used for estimation of broad-sense heritability per trait. Because C0 trial from Kasese in 2014 generally had low broad-sense heritability estimates across traits, the trial was not included for subsequent genomic prediction analyses.

For genomic prediction, we fitted a two-stage prediction model. At the first stage the raw phenotypes were merged across trials (training [C0, C1 and C2] and validation trial [WA]) into a single data set and fitted the linear mixed model using *lme4* package in R, accounting for the environmental differences as well as trial evaluation year as below:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}_{clone}c + \boldsymbol{Z}_{rep(loc/study\ \ year)}r + \epsilon$$

where, y represents raw phenotypic value; β represents fixed effect of the grand population mean, (C0, C1, C2 and WA), study year, and location, with $\boldsymbol{X}$ being the corresponding incidence matrix linking observations to those effects. c and r represent random effects of clones with $c \sim N(0, I\sigma_c^2)$, and replication nested in location-study year such that $r \sim N(0, I\sigma_r^2)$ with $\boldsymbol{Z}_{clone}$ and $\boldsymbol{Z}_{rep(loc/study\ \ year)}$ being corresponding incidence matrices for clones and replications nested in location-study year respectively. The residuals ϵ were distributed as: $\epsilon \sim N(0, I\sigma_\epsilon^2)$ with I representing the identity matrix. We extracted best linear unbiased predictors (BLUPs) for each clone using the *ranef* function available in *lme4* package (Bates et al., 2015), and these were preferred over fixed clone effects for the genomic prediction study due to imbalances in the dataset.

After extraction of BLUPs for each clone for each cycle, we fitted a G-BLUP model to estimate genomic estimated breeding values (GEBVs) that were used for assessment of gains from genomic selection using the three evaluated Uganda's GS cycles (C0, C1 and C2). A one-way analysis of variance was performed to test for significant differences among the means of the GEBVs for the three cycles for each trait using R (R Core Team, 2021). Mean GEBVs of three cycles were separated using Tukey's honestly significant difference. Gains from genomic selection were thereafter calculated as the difference between the mean performance of new cycle and mean performance of the previous cycle from which the new cycle was selected.

Furthermore, we carried out 5-fold cross-validation analyses for each training population (C0, C1 and C2) and WA clones. To do the cross-validation, the BLUPs that were extracted from the first stage analyses per trait were used as the response variable to fit a second stage prediction model for five genomic prediction models with different statistical assumptions. These models were: genomic best linear unbiased prediction (G-BLUP) (VanRaden, 2008; Endelman, 2011), Bayesian ridge regression BRR (Meuwissen et al., 2001), Bayesian least absolute shrinkage and selection operator (BL), Bayes A and Bayes B (Park and Casella, 2008). An excellent review of these models has already been provided by Heslot et al. (2012), and thus will be discussed briefly.

To implement G-BLUP, we fitted a model: $Y = 1\beta + Xg + \epsilon$, with g ~ N (0, Kσ^2_g) and ϵ ~ (0, Iσ^2_g), where Y represents the vector of BLUPs, β represents an overall population mean, X represents the design matrix linking observations to genomic values, g being vector of genomic estimated breeding values for each clone, and ϵ represents the vector of residuals. We assumed, g has a known covariance structure defined by the realized genomic relationship matrix K, while I representing identity matrix.

Additional, we implemented the four Bayesian models i.e. BRR, BL, Bayes A and Bayes B, following the same linear mixed model: $Y = 1\beta + Zg + \epsilon$, with g~N (0, Kσ^2_g) and ϵ ~ (0 Iσ^2_g), where Y represents the vector of BLUPs, β represents an overall population mean, X represents the design matrix linking observations to genomic values, g being vector of genomic estimated breeding values for each clone, and ϵ represents the vector of residuals. We assumed, g also has a known covariance structure defined by the realized genomic relationship matrix K and I representing identity matrix. Specifically, BRR assigns a Gaussian prior with common variance to each marker effect, and applies homogeneous shrinkage to all marker effects. BL employs a double-exponential prior distribution for marker effects, which places strong shrinkage to markers with little to no effect on the trait. Bayes A applies a scaled-t prior distribution to marker effects, and places slightly less shrinkage on markers with zero effect, thereby allowing more flexibility for marker effects. Lastly, Bayes B assumes that most of the markers have zero effect on the trait, and assumes that the markers with an effect on the trait will follow a scaled-t prior distribution as in the case of Bayes A, making it relatively more stringent when compared to Bayes A. All the four Bayesian models used in this study were fitted using the *BGLR* function available in the R package *BGLR* (Pérez and de los Campos, 2014). A Markov Chain Monte Carlo (MCMC) algorithm was applied with prior parameters defined following the procedure suggested by de los Campos et al. (2013). Computations were performed using a chain length of 10,000 iterations, with the first 1000 iterations discarded as burn-in (Pérez and de los Campos, 2014).

Briefly, during implementation of cross-validation within each population (C0, C1, C2 and validation), the clones were randomly split into five subsets (5-fold), where 4/5 of the subsets were used to train the model, while 1/5 was reserved for model validation and this was replicated 5 times. The accuracy of genomic prediction for each fold was then computed as Pearson correlation coefficient between the genomic estimated breeding values and BLUPs for each trait as a response variable.

Lastly, we carried out independent validation for the WA clones, the five evaluated genomic prediction models (G-BLUP, BRR, BL, Bayes A, and Bayes B) were trained using $C_{0,}$ C1 and C2 to predict disease severity and fresh root weight in the validation population (West African population) that comprised 297 clones. Similarly, the prediction accuracy for each model was assessed using Pearson's correlation between the GEBVs and the BLUP values per trait.

Results

Broad sense heritability for evaluated traits in the training and validation populations

Plot-based heritabilities were low to intermediate (Table 1). Estimates of plot-based broad-sense heritability for the training set were highest for disease traits, and these ranged from 0.04 to 0.99 for CBSD foliar severity, 0.2 to 0.86 for CBSD root necrosis severity and 0.00 to 0.99 for CMD severity. Differences in trait heritabilities for data collected at two time points (three and six months after planting) were not substantial, for both CBSD and CMD severity. Heritabilities for fresh root weight were generally modest, ranging from 0.00 to 0.99. Lowest heritabilities for disease traits were observed at Kasese in the mid-western Uganda, while highest heritability for both disease severity and fresh root weight was observed at Serere in Eastern Uganda. Namulonge (central Uganda) registered the lowest heritability for fresh root weight. Though heritability for fresh root weight in the validation population was 0.00, heritabilities for CBSD and CMD severity were moderately high (H^2 > 0.65).

TABLE 1 Plot based broad sense heritability estimates for disease severity and fresh root weight for training and validation populations evaluated at the different locations in Uganda between 2013 to 2019.

Population	Year	Location	CBSD3S	CBSD6S	CBSDRS	CMD3S	CMD6S	FRW
C0	2013	Kasese	0.31	0.30	0.45	0.64	0.45	0.40
C0	2013	Namulonge	0.33	0.37	0.60	0.42	0.74	0.40
C0	2013	Ngetta	–	0.52	0.68	0.75	–	0.47
C0	2014	Kasese	0.04	0.06	–	0.00	0.00	–
C0	2014	Namulonge	0.38	0.37	0.68	0.49	0.77	–
C1	2016	Namulonge	0.40	0.17	0.20	0.80	0.83	0.57
C1	2016	Serere	0.65	0.46	0.66	0.80	0.79	0.02
C1	2017	Namulonge	0.55	0.45	0.44	0.79	0.71	0.11
C1	2017	Serere	0.99	0.99	0.86	0.99	0.94	0.99
C2	2019	Namulonge	0.70	0.54	0.5	0.84	0.90	0.00
WA	2019	Namulonge	0.81	0.67	0.84	0.91	0.94	0.00

CBSD3S and CBSD6S, cassava brown streak disease foliar severity at 3 and 6 months after planting respectively; and CBSDRS, cassava brown streak disease root severity at 12 months after planting.

Gains from genomic selection in Uganda's breeding populations from 2013 to 2019

Using a boxplot, we summarized variations for genomic estimated breeding values across GS cycles for the six traits assessed, with overall genetic progress recorded for most traits except for CMD (Figure 1). Based on average genomic estimated breeding values per cycle (GEBVs), CBSD foliar severity at three months reduced from a mean GEBV of 0.016 for C0 to -0.008 in C1. CBSD foliar severity at six months and CBSD root necrosis severity exhibited a similar downward trend in disease severity when C0 clones where recombined and advanced to C1 using genomic selection (Table 2). With regard to CMD severity, mean GEBVs reduced from 0.006 to 0.004, and 0.013 to -0.002, for CMD3S and CMD6S, respectively, as clones were advanced from C0 to C1. Fresh root weight also increased from -0.017 in C0 to -0.004 to C1. From C1 to C2, all disease traits i.e. CBSD3, CBSD6S, CBSDRS, CMD3S and CMD6S further exhibited a downward trend in disease severity based on their mean GEBVs. Fresh root weight also continued to exhibit an upward trend when C1 clones were recombined and advanced to C2 of genomic selection. Highest response to selection was observed with fresh root weight, CBSD root necrosis resistance, fresh root weight, CBSD foliar severity, and lastly CMD severity.

Population structure between training and validation sets

Principal component analysis revealed a slight genetic differentiation between the Ugandan and West African cassava populations (Figure 2). Variations in genetic structure between the Ugandan and West African populations were moderate, as the first two principal components (PCs) explained approximately 53% of the variation, where the first and second PCs accounted for 35.5%, and 17.5%, respectively.

Cross-validation prediction accuracies within the training and validation populations

Cross-validation prediction accuracies were performed using five models (Bayes A, Bayes B, BRR, BL, and G BLUP) to assess prediction accuracy of genomic selection for CBSD resistance, CMD resistance and fresh root weight within the training and validation populations. We observed modest prediction accuracies for all evaluated traits and populations (Figure 3). Prediction accuracies in the training set ranged from -0.06 to 0.59 for CBSD3S, -0.16 to 0.68 for CBSD6S, -0.15 to 0.68 for CBSDRS, -0.21 to 0.57 for CMD3S, -0.27 to 0.59 for CMD6S, and -0.22 to 0.41for FRW. Of the three cycles in the training set, C0 registered the highest prediction accuracies for all traits, followed by C1 and lastly C2. Average prediction accuracies for C0 were: 0.37, 0.48, 0.48, 0.33, 0.40 and 0.26 for CBSD3S, CBSD6S, CBSDRS, CMD3S, CMD6S and FRW, respectively. Average prediction accuracies for C1 were: 0.32, 0.34, 0.12, 0.11, 0.08 and 0.08, for CBSD3S, CBSD6S, CBSDRS, CMD3S, CMD6S and FRW, respectively. Lastly, mean prediction accuracies for C2 were: 0.21, 0.30, 0.11, 0.16, 0.13 and 0.00 for CBSD3S, CBSD6S, CBSDRS, CMD3S, CMD6S and FRW, respectively. Across the three cycles and the evaluated

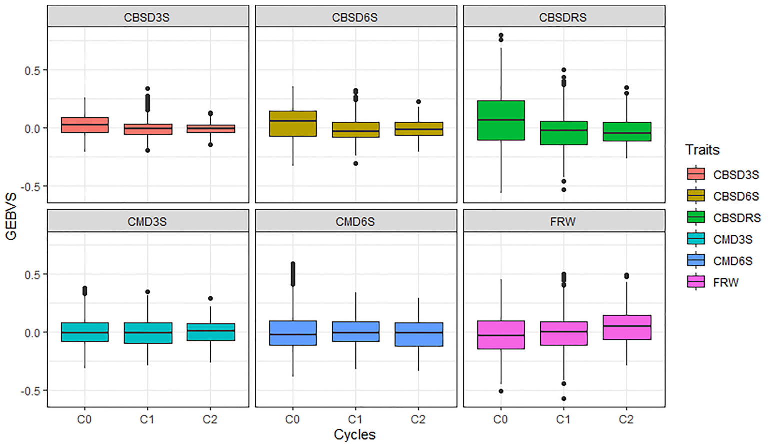


FIGURE 1
Performance of the three cycles (C0, C1 and C2) of Uganda's cassava genomic selection population for disease resistance. CBSD3S, cassava brown streak disease foliar severity scored at three months; CBSD6S, Cassava brown streak disease foliar severity scored at six months; CBSDRS, Cassava brown streak disease root severity scored at 12 months; CMD3S, Cassava mosaic disease severity scored at three months; CMD6S, Cassava mosaic disease severity scored at six months.

TABLE 2 Mean performance of genomic selection cycles and corresponding gains from selection for fresh root weight and virus disease resistance.

Cycle	FRW	CBSD3S	CBSD6S	CBSDRS	CMD3S	CMD6S
C_0	-0.017^a	0.016^a	0.034^a	0.063^a	0.006^a	0.013^a
C_1	-0.004^a	-0.008^b	-0.019^b	-0.031^b	-0.004^a	-0.002^a
C_2	0.044^b	-0.009^b	-0.013^b	-0.034^b	-0.001^a	-0.017^b
P-value	***	***	***	***	NS	*
Gains from selection						
C_1 - C_0	0.013	-0.024	-0.054	-0.094	-0.009	-0.015
C_2 - C_1	0.048	-0.001	0.006	-0.003	0.003	-0.014

CBSD3S, cassava brown streak disease foliar severity scored at three months; CBSD6S, Cassava brown streak disease foliar severity scored at six months; CBSDRS, Cassava brown streak disease root severity scored at 12 months; CMD3S, Cassava mosaic disease severity scored at three months; CMD6S, Cassava mosaic disease severity scored at six months; and FRW, fresh root weight. Letters indicate significant differences using Tukey's honestly significant difference (α = 0.05). * P < 0.05, *** P < 0.001, and NS = non-significant differences between average performance of selection cycles.

models, CBSD6S was predicted with the highest accuracy (0.37), followed by CBSD3S (0.29), CBSDRS (0.22) and lastly fresh root weight (0.11). We observed that GBLUP was slightly superior to all evaluated Bayesians models across the five traits and three populations in the training set. On the other hand, cross-validation predictions in the validation set (clones from West Africa) were relatively lower than those observed in the training population (clones from Uganda). Prediction accuracies ranged from -0.25 to 0.31 for CBSD3S, -0.09 to 0.38 for CBSD6S, -0.36 to 0.53 for CBSDRS, -0.17 to 0.29 for CMD3S, -0.19 to 0.37 for CMD6S, and -0.10 to 0.47 for FRW. On average, CBSDRS was predicted with the highest accuracy (0.29), followed by CMD6S (0.13) and lastly FRW (0.07).

Using Uganda's training population to predict traits in West African clones

Analyses were performed using C0, C1, and C2 to assess prediction accuracy of genomic selection for CBSD resistance, CMD resistance and FRW in the West African population (297 clones) that was part of the pre-breeding populations evaluated in Uganda for CBSD resistance. We observed extremely low prediction accuracies for all traits (Table 3). For example, prediction accuracies ranged from -0.07 to 0.10 for CBSD3S, -0.02 to 0.15 for CBSD6S, -0.18 to 0.05 for CBSDRS, 0.002 to 0.09 for CMD3S, -0.034 to 0.078 for CMD6S and lastly -0.076 to 0.086 for FRW. Average predictions were less than 0.1 for all evaluated

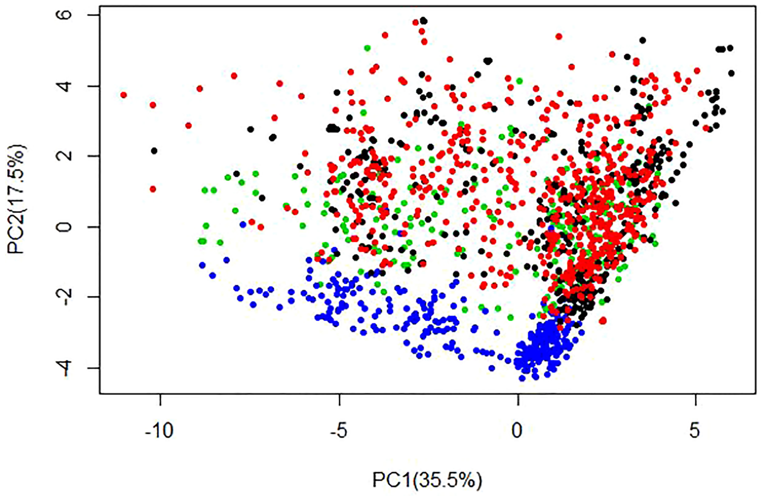


FIGURE 2
Population structure displayed by the first two principal components (PCs) generated for training set i.e. C0 (384 clones), C1 (638 clones), C2 (287 clones) and the validation population (279 clones from West Africa) using 21,938 SNP markers. The figure displays population structure from PC1 vs PC2, with associated variances for each PC represented in brackets. Black = C0, Red = C1, Green = C2 and Blue = WA clones.

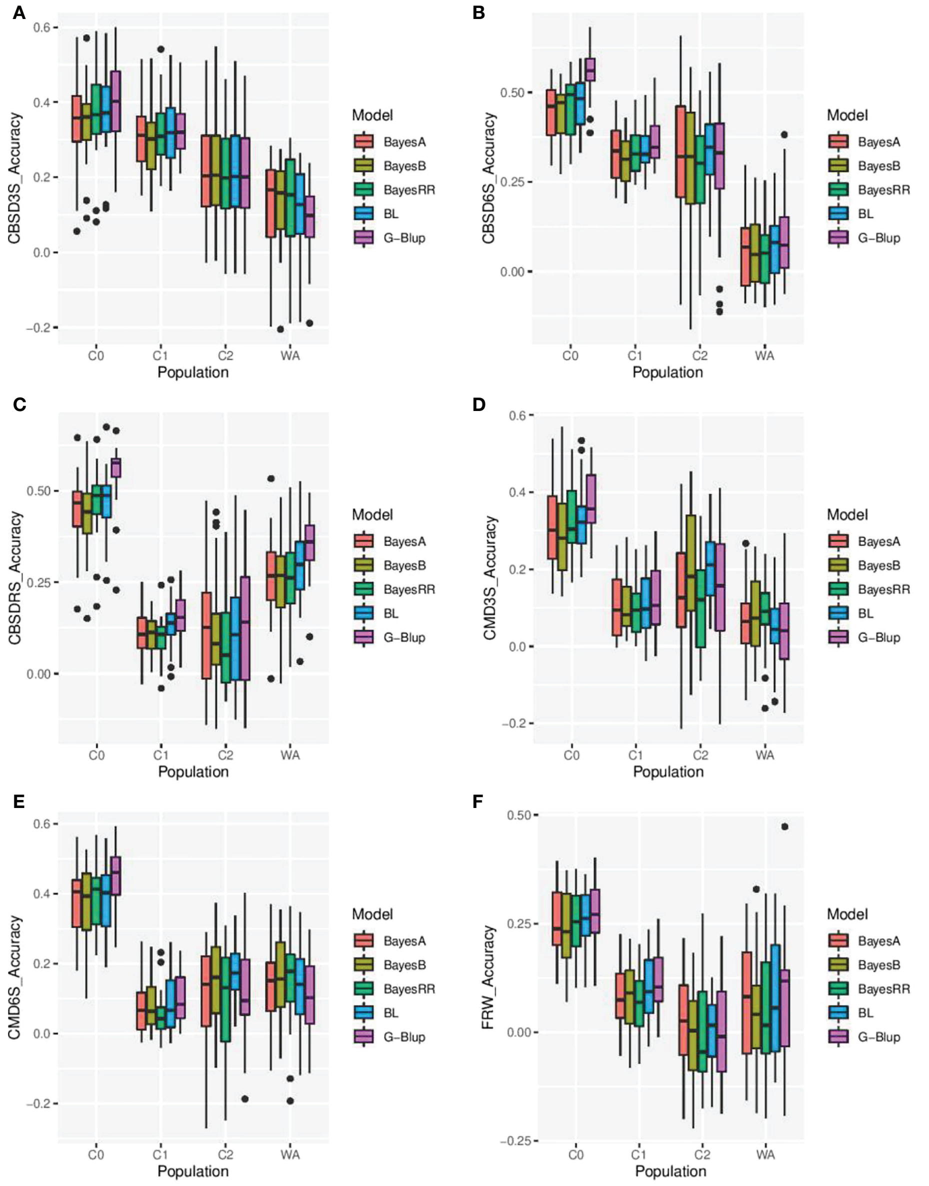


FIGURE 3
Cross validation prediction accuracies for cassava brown streak disease severity at three (CBSD3S), six (CBSD6S) and twelve months after planting (CBSDRS), cassava mosaic disease severity at three months (CMD3S) and six months after planting (CMD6S), and fresh root weight (FRW) using five genomic prediction models in the training (C0, C1, C2) and validation population (germplasm from West Africa). **(A–F)** Represent prediction accuracies for CBSD3S, CBSD6S, CBSDRS, CMD3S, CMD6S and FRW, respectively. C0, C1, C2 represent cycle zero, cycle one, cycle two of Uganda's cassava genomic selection population, while WA represent West African cassava germplasm from Nigeria. BL, Bayesian Least Absolute shrinkage and selection operator, BRR, Bayesian Ridge Regression, and G-Blup, Genomic Best Linear Unbiased Prediction.

traits. Though predictions were extremely low, C1 registered the highest prediction for CBSD6S (0.14) in WA population, and lowest prediction was observed when C0 was used to predict CBSDRS in the validation set. Since prediction accuracies were extremely low, it seemed unreasonable to assess how a combination of the three populations would affect prediction accuracies of GS in the WA population.

Discussion

The challenges of rapid human population growth and climate change invariably affect agricultural productivity, and thus the need for increased genetic gains (Hickey et al., 2017). Currently, there are concerted global efforts to combat CBSD, a disease that is endemic to East and Central Africa but posing a

TABLE 3 Independent validation prediction accuracies for cassava mosaic disease severity, cassava brown streak disease severity and fresh root weight using five genomic prediction models and three cycles of genomic selection.

Cycle	Model	CBSD3S	CBSD6S	CBSDRS	CMD3S	CMD6S	FRW
C0	BayesB	-0.057	-0.019	-0.085	0.007	0.032	0.007
C0	BayesA	-0.056	-0.011	-0.061	0.042	0.068	0.042
C0	BL	-0.057	-0.008	-0.084	0.074	0.078	0.074
C0	G-Blup	-0.039	0.019	-0.181	0.057	0.075	0.057
C0	BayesRR	-0.072	-0.007	-0.085	0.086	0.073	0.086
C1	BayesB	0.066	0.151	-0.086	0.036	0.044	-0.028
C1	BayesA	0.093	0.153	-0.151	0.002	0.014	-0.013
C1	BL	0.085	0.153	-0.053	0.042	0.043	-0.015
C1	G-Blup	0.088	0.107	-0.072	0.005	0.002	-0.022
C1	BayesRR	0.104	0.154	-0.089	0.037	0.066	0.012
C2	BayesB	0.028	0.124	0.056	0.032	0.021	-0.006
C2	BayesA	0.053	0.125	-0.016	0.052	-0.034	-0.017
C2	BL	0.052	0.141	0.012	0.092	-0.022	-0.076
C2	G-Blup	0.039	0.096	0.021	0.049	-0.010	-0.069
C2	BayesRR	0.017	0.131	-0.003	0.058	-0.010	-0.034

BL, Bayesian least absolute shrinkage and selection operator; G-BLUP, Genomic best linear unbiased prediction method; BRR, Bayesian Ridge Regression. CBSD3S, cassava brown streak disease foliar severity scored at three months; CBSD6S, Cassava brown streak disease foliar severity scored at six months; CBSDRS, Cassava brown streak disease root severity scored at 12 months; CMD3S, Cassava mosaic disease severity scored at three months; CMD6S, Cassava mosaic disease severity scored at six months; and FRW, fresh root weight.

significant threat to cassava production in West Africa, the world's largest producer and consumer of cassava (Legg et al., 2014). In this study, we leveraged genomic prediction approaches as a possible means to enable pre-emptive breeding for CBSD resistance in West Africa, using elite cassava populations from Uganda. Accordingly, three Uganda's populations segregating for CBSD severity comprised the training set, and these were used to predict CBSD resistance along with other equally important traits such as CMD resistance and fresh root weight in the WA population that was evaluated in Uganda, a hotspot for CBSD.

Broad sense heritability estimates for evaluated traits were low ($H^2< 0.2$) to high ($H^2 > 0.6$), and were well in range with heritability estimates in literature (Kayondo et al., 2018; Okul et al., 2018; Ozimati et al., 2018). These results underpin the general conclusion that the experimental sites were hotspots for CMD and CBSD i.e. the disease pressure was high enough to cause substantial variation in clone response to the virus diseases (Alicai et al., 2019). This finding further implies that Namulonge and Serere are suitable for screening of germplasm against CMD and CBSD, and could be used by breeding programs threatened by CBSD. The extremely low heritability estimates for CMD severity are attributable to low phenotypic variations for CMD in the evaluated Ugandan cassava populations. The low phenotypic variations for CMD resistance traits were attributable to the fact that breeding efforts targeting resistance to CMD have been ongoing since 1930s (Legg and Thresh, 2000), which is sufficient time for increasing the frequency of resistance alleles in the breeding populations (Hallauer et al., 1988), and thus we might have fixed CMD resistance alleles in our recently developed cassava germplasm. The low heritabilities of CBSD traits are also attributable to the low phenotypic variations in CBSD severity observed in C0, C1 and C2, which were also attributable to selection and recombination. These low phenotypic variations for CMD and CBSD resistance imply that the breeding program has attained a usable level of resistance to virus diseases in most of its elite material, and therefore, alleles for yield and end-user preferred traits need to be introgressed into disease resistance background to allow enhancement of yield traits.

We observed substantial gains for all evaluated traits in Uganda's GS cycles. This finding is agreement with findings from Sweeney et al. (2021) who reported increased gains from GS in the spring barley breeding program. The observed improvements in trait means based on their GEBVs is an indication that genomic selection successfully increased frequency of desirable alleles for target traits in the evaluated cassava populations. These findings further imply that even with low predictions accuracies of less than 0.40, genetic gains are possible with GS for low heritability traits. The low gains in CMD resistance could be due to low phenotypic variability in the evaluated traits i.e. clones exhibited a similar level of resistance both at three and six months after planting for the three evaluated cycles with a mean severity score of 1.4. With the observed downward trend in disease severity and a concurrent upward trend in fresh root weight, genomic selection is likely to fast-track variety replacement and/or increase variety turnover in cassava especially in this era of climate change and rapid population increase.

Having observed significant gains in traits using genomic selection, we evaluated the importance of our GS cassava populations in predicting cassava traits in West Africa, where CBSD is an eminent threat. Based on principal component analysis of SNP data, we observed a close relationship between the training set (Uganda's cassava populations) and validation set (West African cassava population), with a slight population structure and genetic differentiation between the two populations (Figure 2). This low genetic variability and lack of clear structure within these populations underpins the likelihood that the East and West African materials might have shared a common ancestry, a situation that could be attributed to germplasm exchange between east and west Africa in the 1930s, during the advent of cassava mosaic disease (CMD) pandemic (Jennings, 2002). The absence of clear population structure and low genetic variation between the evaluated populations also suggested the appropriateness of using Uganda's population (C0, C1, and C2) as a training population for genomic prediction of the West African populations and subsequent selection of individuals using GEBVs as a pre-emptive breeding strategy. Accordingly, analyses were performed to determine whether the close relationship between the populations would result into high prediction accuracies when Uganda's population was used to train models for prediction of disease resistance and fresh root weight in the WA population. Surprisingly, the west African population which was fairly genetically similar to Uganda's training population, was predicted with extremely low accuracy (ranging from -0.07 to 0.15) for all evaluated traits when C0, C1 and C2 were separately used as training populations, suggesting that there could be other factors that affected the prediction accuracy of GS other than the relationship between training and validation sets.

Several genomic prediction models have been developed to predict trait performance under different genetic architecture and the five GP models (Bayes A, Bayes B, BRR, BL and G-BLUP) chosen for this study also differed in assumptions about the genetic architecture of the evaluated traits. Results revealed that models performed similarly for the most part, but there also occasions where G-BLUP was slightly superior to Bayesian models used in this study. These results were in good agreement with earlier findings from Wolfe et al. (2016) and Kayondo et al. (2018). Superiority of G-BLUP could be that the true QTL effects for evaluated traits were relatively small and that the distribution of these effects could be less extreme. The superiority of G-BLUP could be also attributed to its ability to take advantage of the relationships among individuals at the causal loci for the traits under analysis (VanRaden, 2008), indicating that models that might estimate relationship information between training and test sets could be more valuable than those that estimate marker effects directly.

Average cross validation prediction accuracies across the three populations for CBSD and CMD resistance fresh root weight ranged between 0.26 to 0.48, and were comparable to findings by Wolfe et al. (2017); Wolfe et al. (2016); Kayondo et al. (2018) and Ozimati et al. (2018). The low cross validation prediction accuracies suggested that they could be attributed to the low phenotypic variations for the studied traits observed in the evaluated populations. These cross validation predictions within the three populations were encouraging and thus highlighting the utility of GS for improving CBSD and CMD resistance, and fresh root weight. On the other hand, cross validation prediction accuracies in the validation set (west African clones) were much lower than that was observed in the Ugandan training set, and this could be attributed to the fact that west African clones might be deficient in CBSD resistance alleles (Ano et al., 2021).

On the other hand, independent validation prediction accuracies of genomic selection were generally low, and they were lower than cross-validation prediction accuracies for CMD, CBSD traits and fresh root weight. Given that the training and validation populations were fairly genetically similar, the low prediction accuracies for independent validations could be attributed to genotype by environment interaction i.e. the training and validation populations were evaluated during different seasons. These low predictions could also be attributed to the fact that west African cassava populations were deficient in CBSD resistance alleles (Ano et al., 2021). No consistent superior performance was observed for any of the prediction models that were assessed, and this was in good agreement with Heslot et al. (2012) and Jannink et al. (2010). Although the models tested in this study assumed different distributions of marker effects (Meuwissen et al., 2001; Lorenz et al., 2011), their similarity in prediction accuracies could be interpreted as approximation to optimal genomic prediction models, where all the models capture the same or similar QTL effects across the genome (Su et al., 2014). In such a situation, choice of GS model would be less important than choice of training population.

The C0 training set yielded the lowest prediction accuracies, with negative average accuracies for all traits across all models. The disparity between the predictive ability of the C1, C2 and the C0 training sets might be because the C1 training set was able to capture more genetic signals for CBSD foliar and root symptom expression in the West African clones than C0 training set. This phenomenon was noted by Ozimati et al. (2018) who reported that optimized Ugandan training sets were able to capture more genetic signals and yielded higher prediction accuracies for CBSD resistance in IITA clones than random training sets. Another possible explanation might be that the quantitative trait loci (QTLs) responsible for CBSD resistance in the two populations were different. This might be due to recombination events that might have occurred in their genomes, resulting in the rearrangement of QTLs responsible for CBSD resistance in the three populations.

Conclusion

Based on the findings of this study, the breeding program in Uganda has made genetic progress through GS accelerated breeding cycles for most target traits, especially for CBSD root necrosis which is one of the must to have traits in a variety, demonstrating the worthwhile of GS for rapid population improvement and variety development. In general, low prediction accuracies were recorded from using Ugandan training set to predict traits in African clones, suggesting inadequacy of utilizing Ugandan training set, especially for CBSD pre-emptive breeding in WA. In this case, efforts should be devoted to sharing Uganda's germplasm that possess resistance with the W. African breeding programs for hybridization to fully enable deployment of genomic selection as a pre-emptive CBSD breeding strategy in W.A

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.nextgencassava.org.

Author contributions

OA Conceptualized the study, collected and analyzed data, and wrote the original manuscript. WE collected data and reviewed the manuscript. AU collected data and reviewed the manuscript. FM collected and analyzed the data, and reviewed the manuscript. PI collected data and reviewed the manuscript. MK collected data and reviewed the manuscript. CE acquisition of the funding and reviewed the manuscript. RK collected data and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Next Generation Cassava Breeding Project funded by Bill and Melinda Gates Foundation, funding granted through Cornell University (grant number OPP1048542).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alicai, T., Omongo, C. A., Maruthi, M. N., Hillocks, R. J., Baguma, Y., Kawuki, R., et al. (2007). Re-emergence of cassava brown streak disease in Uganda. *Plant Disease.* 91 (1), 24–29. doi: 10.1094/PD-91-0024

Alicai, T., Szyniszewska, A. M., Omongo, C. A., Abidrabo, P., Okao-okuja, G., Baguma, Y., et al. (2019). Expansion of the cassava brown streak pandemic in Uganda revealed by annual field survey data for 2004 to 2017. *Sci. Data* 6, 1–8. doi: 10.1038/s41597-019-0334-9

Ano, C. U., Ochwo-Ssemakula, M., Ibanda, A., Ozimati, A., Gibson, P., Onyeka, J., et al. (2021). Cassava brown streak disease response and association with agronomic traits in elite Nigerian cassava cultivars. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.720532

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67 (1), 1–47. doi: 10.18637/jss.v067.i01

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81 (5), 1084–1097. doi: 10.1086/521987

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193 (2), 327–345. doi: 10.1534/genetics.112.143313

de Oliveira, E. J., de Resende, M. D. V., da Silva Santos, V., Ferreira, C. F., Oliveira, G. A. F., da Silva, M. S., et al. (2012). Genome-wide selection in cassava. *Euphytica* 187, 263–276. doi: 10.1007/s10681-012-0722-0

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6, 1–10. doi: 10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrBLUP. *Plant Genome J.* 4, 250. doi: 10.3835/plantgenome2011.08.0024

Food and Agriculture Organization of the United Nations (FAOSTAT) (2019). *FAOSTAT Statistical Database.* (Rome: FAO).

Gondwe, F. M. T., Mahungu, N. M., Hillocks, R. J., Raya, M. D., Moyo, C. C., Soko, M. M., et al. (2003). "Economic losses experienced by small-scale farmers in Malawi due to cassava brown streak virus disease," in *Proceedings of an international workshop*. Eds. J. P. Legg and R. J. Hillocks(Mombasa, Kenya: Natural Resources International Limited).

Hallauer, A. R., Carena, M. J., and Filho, M. J.. (1988). *Quantitative Genetics in Maize Breeding* (First edit;J. Prohens, F. Nuez and M. J. Carena eds.). doi: 10.1007/978-1-4419-0766-0

Hammond, W., Jeerapong, L., Hoogeveen, J., Kienzle, J., Kumar, L., Chikelu, M., et al. (2013). *Save and grow: Cassava, a guide to sustainable production intensification*. (Rome: Food and Agriculture Organization of the United Nations)

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: A comparison of models.. *Crop Sci.* 52 (1), 146–160. doi: 10.2135/cropsci2011.06.0297

Hickey, L. T., Hafeez, A. N., Robinson, H., Jackson, S. A., Leal-bertioli, S. C. M., Tester, M., et al. (2017). Breeding crops to feed 10 billion. *Nat. Biotechnol* 37, 744–754. doi: 10.1038/s41587-019-0152-9

Hillocks, R. J., Thresh, J. M., Tomas, J., Botao, M., Macia, R., and Zavier, R. (2002). Cassava brown streak disease in northern Mozambique. *Int. J. Pest Management.* 48, 178–181. doi: 10.1080/09670870110087376

Jennings, D. L. (2002). Historical perspectives on breeding for resistance to cassava brown streak virus disease. In *Cassava Brown Streak Virus Disease: Past, Present and Future Cassava Brown Streak Virus Disease: Past, Present and Future* (Issue October).

IITA (1990). *Cassava in tropical Africa: A reference manual*, Vol. 3. 1–176 (United Kingdom: Chayce Publication Services). doi: 10.1002/ejoc.201200111

Jannink, J. L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings Func Genom. Proteom.* 9 (2), 166–177. doi: 10.1093/bfgp/elq001

Kang, M., Kanno, C. M., Reid, M. C., Zhang, X., Mauzerall, D. L., Celia, M. A., et al (2014). Direct measurements of methane emissions from abandoned oil and gas wells in Pennsylvania. *PNAS* 111 (51). 18173–18177. doi: 10.1073/pnas.1408315111

Kayondo, S. I., Del Carpio, D. P., Lozano, R., Ozimati, A., Wolfe, M., Baguma, Y., et al (2018). Genome-wide association mapping and genomic prediction for CBSD resistance in Manihot esculenta. *Scientific Rep.* 8 (1). 1–11. doi: 10.1038/s41598-018-19696-1

Legg, J., Somado, E. A., Barker, I., Beach, L., Ceballos, H., Cuellar, W., et al. (2014). A global alliance declaring war on cassava viruses in Africa. *Food Security* 6, 231–248. doi: 10.1007/s12571-014-0340-x

Legg, J. P., and Thresh, J. M. (2000). Cassava mosaic virus disease in East Africa: A dynamic disease in a changing environment. *Virus Res.* 71 (1–2), 135–149. doi: 10.1016/S0168-1702(00)00194-5

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al (2011). Genomic Selection in Plant Breeding. Knowledge and Prospects.. *Adv Agronomy* 110 (C). 77–123. doi: 10.1016/B978-0-12-385531-2.00002-5

Manze, F., Rubaihayo, P., Ozimati, A., Gibson, P., Esuma, W., Bua, A., et al (2021). Genetic Gains for Yield and Virus Disease Resistance of Cassava Varieties Developed Over the Last Eight Decades in Uganda. *Front. Plant Sci.* 12, 651992. doi: 10.3389/fpls.2021.651992

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense markers maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Nweke, F. (2004). New challenges in the cassava transformation in Nigeria and Ghana: Environment and production technology division international food policy research institute. *Food Policy* 67, 1–118.

Okul, V. A., Ochwo-Ssemakula, M., Kaweesi, T., Ozimati, A., Mrema, E., Mwale, E. S., et al (2018). Plot based heritability estimates and categorization of cassava genotype response to cassava brown streak disease. *Crop Protection* 108, 39–46. doi: 10.1016/j.cropro.2018.02.008

Ozimati, A., Kawuki, R., Esuma, W., Kayondo, I. S., Wolfe, M., Lozano, R., et al (2018). Training Population Optimization for Prediction of Cassava Brown Streak Disease Resistance in West African Clones. *Genes, Genomes Genetics*. doi: 10.1534/g3.118.200710

Park, T., and Casella, G. (2008). The Bayesian lasso. *J. Am. Stat. Assoc.* 103 (482), 681–686. doi: 10.1198/016214508000000337

Patil, B. L., Legg, J. P., Kanju, E., and Fauquet, C. M. (2015). Cassava brown streak disease: A threat to food security in Africa. *J. Gen. Virology.* 96, 956–968. doi: 10.1099/vir.0.000014

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198 (2), 483–495. doi: 10.1534/genetics.114.164442

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-Sequencing for plant breeding and genetics. *Plant Genome J.* 5 (3), 92. doi: 10.3835/plantgenome2012.05.0005

Prochnik, S., Marri, P. R., Desany, B., Rabinowicz, P. D., Kodira, C., Mohiuddin, M., et al. (2012). The cassava genome: Current progress, future directions. *Trop. Plant Biol.* 5, 88–94. doi: 10.1007/s12042-011-9088-z

Rabbi, I. Y., Kulakow, P. A., Manu-Aduening, J. A., Dankyi, A. A., Asibuo, J. Y., Parkes, E. Y., et al. (2015). Tracking crop varieties using genotyping-by-sequencing markers: A case study using cassava (Manihot esculenta crantz). *BMC Genet.* 16, 1–11. doi: 10.1186/s12863-015-0273-1

R Core Team. (2021). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). URL https://www.R-project.org/.

Spilke, J., Piepho, H., and Xiyuan, H. (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *J. Agricultural Biological Environ. Statistics.* 10 (3), 374–389. doi: 10.1198/108571105X58199

Su, G., Guldbrandtsen, B., Aamand, G. P., Strandén, I., and Lund, M. S. (2014). Genomic relationships based on X chromosome markers and accuracy of genomic predictions with and without X chromosome markers. *Genetics Selection Evol.* 46 (1), 47. doi: 10.1186/1297-9686-46-47

Sweeney, D. W., Rooney, T. E., and Sorrells, M. E. (2021). Gain from genomic selection for a selection index in two-row spring barley. *Plant Genome* 14, e20138. doi: 10.1002/tpg2.20138

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., et al (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). doi: /10.3835/plantgenome2015.11.0118

Wolfe, M. D., Del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., et al. (2017). Prospects for genomic selection in cassava breeding. *Plant Genome* 10 (3), 1–19. doi: 10.3835/plantgenome2017.03.0015

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 December 2022
DOI 10.3389/fpls.2022.1054909

Identification of cassava germplasms resistant to two-spotted spider mite in China: From greenhouse large-scale screening to field validation

Xiao Liang[1,2*†], Qing Chen[1,2*†], Ying Liu[1,2†], Chunling Wu[1,2], Kaimian Li[3], Mufeng Wu[1,2], Xiaowen Yao[1,2], Yang Qiao[1,2], Yao Zhang[1,2] and Yue Geng[1,2]

[1]Environment and Plant Protection Institute, Chinese Academy of Tropical Agricultural Sciences, Key Laboratory of Integrated Pest Management on Tropical Crops, Ministry of Agriculture and Rural Affairs, Haikou, Hainan, China, [2]Sanya Research Academy, Chinese Academy of Tropical Agriculture Science, Hainan Key Laboratory for Biosafety Monitoring and Molecular Breeding in Off-Season Reproduction Regions, Sanya, Hainan, China, [3]Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agriculture Sciences, Haikou, China

Introduction: Utilization of resistant germplasm is considered as an effective, economical and eco-friendly strategy for cassava pest management. *Tetranychus urticae*, known as the two-spotted spider mite (TSSM), is a devastating pest in Asian cassava planting countries as well as in China. However, the resistant levels of abundant cassava germplasms to TSSM remains largely unknown.

Methods: To fill this knowledge gap, we conducted screening of 202 cassava germplasm for resistance to TSSM in China based on the classification of mite damage phenotype, under both greenhouse and field conditions.

Results: The three rounds of large-scale greenhouse experiments had identified two highly resistant (HR) varieties (C1115 and MIANDIAN), five resistant (R) varieties (SC5, SC9, SC15, COLUMBIA-4D and LIMIN) and five highly susceptible (HS) varieties (KU50, BREAD, SC205, TMS60444 and BRA900), besides, these 'HR' and 'R' varieties would significantly repress the normal development and reproduction of TSSM. In addition, the 12 cassava varieties selected from the greenhouse screening were further subjected to consecutive five years of field validation at Danzhou, Wuming and Baoshan. The seven resistant varieties not only exhibited stable TSSM-resistance performance across the three field environments, but also possessed the same resistant levels as the greenhouse identification, while the resistant varieties SC5 was an exception, which was identified as moderate resistant in Baoshan, indicating the variety-environment interaction may affect its resistance. Furthermore, regional yield estimation suggested that the higher the resistance level was, the better capacity in reducing the yield losses.

OPEN ACCESS

EDITED BY
Robert Kawuki,
National Crops Resources Research Institute (NaCRRI), Uganda

REVIEWED BY
Evans N. Nyaboga,
University of Nairobi, Kenya
Abhay K. Pandey,
Tea Research Association, India

*CORRESPONDENCE
Xiao Liang
liangxiaozju@126.com
Qing Chen
chqingztq@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 03 October 2022
ACCEPTED 18 November 2022
PUBLISHED 07 December 2022

CITATION
Liang X, Chen Q, Liu Y, Wu C, Li K, Wu M, Yao X, Qiao Y, Zhang Y and Geng Y (2022) Identification of cassava germplasms resistant to two-spotted spider mite in China: From greenhouse large-scale screening to field validation.
Front. Plant Sci. 13:1054909.
doi: 10.3389/fpls.2022.1054909

Discussion: This study demonstrated that the TSSM-resistant varieties could be considered as ideal materials in mite control or in future breeding programme of mite-resistant cassava plant.

KEYWORDS

Manihot esculenta, *Tetranychus urticae*, resistance, mite pest, eco-friendly

Introduction

Cassava (*Manihot esculenta* Crantz), serving as food, animal feed as well as biomass energy (Wu et al., 2022), is widely cultivated in more than 100 countries (Parmar et al., 2017). In China, this crop is mainly used for the production of ethanol fuel, which accounts for approximately 70% of the consumption (Jiang et al., 2019). In the past couple of years, China is the world's largest importer of cassava, while most cassava products such as chips and flour were imported from Southeast Asian countries, *i.e.*, Thailand, Vietnam, Laos, and Cambodia (Tan et al., 2018). However, only a few provinces located in the south and southwest of China possess suitable growing conditions for cassava cultivation. Therefore, increasing the yield will be an important demand for increased ethanol production (Chen et al., 2016). Moreover, in order to ensure sufficient profit, farmers prefer a robust cultivar and easy-handling field management strategy.

Insect pest is of great threat to cassava yield, and the phytophagous mite is one of the most destructive pests (Chen et al., 2022a). The cassava green mite (CGM), *Mononychellus* spp., is one of the most widely distributed cassava pests in the world (Vásquez-Ordóñez and Parsa, 2014). Comparatively, the *Tetranychus* spp., known as red spider mite, were predominantly distributed in Asian countries. There were over 10 species found in cassava fields (Bellotti, 2008). In particular, the two-spotted spider mite (TSSM; *Tetranychus urticae;* Acari: Tetranychidae) can cause 50%–70% yield losses in China (Chen et al., 2019). To date, acaricide application is still the major strategy to control TSSM. However, the dense canopy of the cassava plant makes it difficult to target the acaricide. Moreover, the excessive use of acaricide may also largely reduce the natural enemies' population and bring about mite resistance problems. Up to now, the cases of TSSM resistance to several acaricides are continuously increasing, including acequinocyl (Leeuwen et al., 2008), spirodiclofen (Van Pottelberge et al., 2009), and cyenopyrafen (Khalighi et al., 2016).

Utilization of resistant germplasm is considered as an effective, economical, and eco-friendly strategy for cassava mite management. Compared to breeding a novel mite-resistant cassava plant, identification of the resistance level of existing cassava germplasms is much more convenient and efficient. Several organizations like the International Center for Tropical Agriculture (CIAT) and the International Institute for Tropical Agriculture (IITA) have made tremendous labors to evaluate cassava resistance to insect pests like CGM, whitefly, and thrip (Bellotti et al., 2012)—for example, MEcu 72, a cassava genotype from CIAT, has been identified as resistant to *Aleurotrachelus socialis* (Bellotti and Arias, 2001; Carabalí et al., 2010a; Carabalí et al., 2010b), *Bemisia tuberculata* (Barilli et al., 2019), and *B. tabaci* (Omongo et al., 2012). In addition, some genotypes from South America and Africa posed different levels of resistance to *B. tabaci* (Omongo et al., 2012; Ochwo-Ssemakula et al., 2019). In a similar study at the IITA in Nigeria, two cassava genotypes supported the lowest number of whiteflies (Ariyo et al., 2005). By comparison, studies focusing on identifying mite-resistant materials in cassava populations are relatively limited, and most studies were focused on screening resistant materials against cassava green mite (CGM). Over 300 accessions from the CIAT Columbia germplasm were shown to have some degree of resistance to CGM (Bellotti et al., 1999). At IITA Tanzania, 58 clones were observed to have a distinct resistance level to CGM (Bellotti et al., 2012). In addition, evaluations of cassava resistance to different insect pests were also conducted. Parsa et al. (2015) performed 89 field evaluations of cassava landraces resistant to several insect pests and found that 129 landraces were highly resistant to thrips, while 33 landraces were resistant to CGM, and 19 landraces were resistant to whiteflies.

CGM sporadically emerged in China (Lu et al., 2014), while the TSSM is the predominant cassava mite; therefore, it is more imperative and practical to develop a TSSM-resistant variety. However, field identifications of cassava resistant to *Tetranychus* mite were rather limited compared with those focused on CGM, which was hindered by the low population in certain cassava-planting regions, for example, the CIAT in Colombia (Bellotti et al., 2012). Nevertheless, laboratory identifications still have been carried out, and mite mortality and hatching rate were used as key indexes for evaluating the resistance of cassava plants—for instance, while fed on the resistant cultivars MBra 12 and MCol 1434, the mortality of TSSM larvae and nymphs was 68% and 50% higher than those on the TSSM-susceptible cultivar MCol 22, respectively. Moreover, the hatching rate and the survival rate of larvae were significantly lower on resistant cultivar MCol 1351 than on MCol 22 (Bellotti et al., 2012). On the contrary, identification of cassava germplasm resistant to

Tetranychus species is recommended to be conducted in regions with high mite populations. Asian countries that cultivated cassava and suffered huge economic losses provide a good research opportunity (Bellotti et al., 2012). Based on leaf morphology, secondary metabolites, and proteomic analysis, Yang et al. (2020) found that the cultivar Xinxuan 048 exhibited high resistance to *T. cinnabarinus* under both greenhouse and field conditions.

However, as far as we know, there is a lack of study regarding either the laboratory or field screening of TSSM-resistant cassava varieties from large populations. To fill this knowledge gap, in the present study, 202 cassava germplasms including all the main cultivars in China were subjected to three rounds of large-scale greenhouse screening. Furthermore, the identified resistant and susceptible varieties from the preliminary screening were validated for their field performance at three different regional sites in five consecutive planting seasons. We expect to screen cassava germplasms with stable mite-resistant performance and provide promising materials for either mite control or future breeding programs of mite resistance.

Materials and methods

Cassava germplasms

A total of 202 cassava germplasms were derived from the National Cassava Germplasm Nursery of China, Chinese Academy of Tropical Agricultural Sciences (CATAS). Cassava stems of about 20 cm in length were vertically planted with nutritive soil (equal quantity of soil, peat, and perlite) in pots and grown in a greenhouse for TSSM resistance screening. The light/dark photoperiod was set as 14/10 h, the temperature was maintained at 28 ± 1°C, and the relative humidity was kept at 75 ± 5%.

Laboratory rearing of TSSM

TSSM rearing was conducted based on our previous study (Liang et al., 2017). Healthy adults were maintained by the Environment and Plant Protection Institute, CATAS, and reared on the back of healthy cassava leaves of BRA900 cultivars at 28 ± 1°C, 75 ± 5% relative humidity, and L14:D10 photoperiod. A water-saturated blotting paper strip was wrapped around the leaf margin to prevent the escape of mites and to keep the leaves fresh. The leaves were replaced every 2 to 3 days.

Identification method of cassava resistance to TSSM

Identification of cassava resistance to TSSM was based on the leaf damage symptoms caused by TSSM. The mite damage symptoms were classified into five scales (Table 1 and Figure 1A) and first evaluated based on the leaf damage rates. The leaf damage rate was precisely calculated using Leaf Image Analyser (YMJ-E, Daji Co., Ltd., Hangzhou, China) (Supplementary Figure S1). After that, the mite damage index (MDI) was calculated according to the equation shown below:

$$\mathrm{MDI} = \frac{\sum (S \times N_s)}{N \times 5} \times 100$$

where S indicates mite damage scale, N_s indicates the number of damaged leaves at a certain damage scale, and N indicates the total number of investigated leaves. Finally, the six mite resistance levels were identified based on the MDI ranges (%), which were listed as HR (with MDI ranging from 0.1% to 12.5%), R (with MDI ranging from 12.6% to 37.5%), moderately resistant (MR, with MDI ranging from 37.6% to 62.5%), susceptible (S, with MDI ranging from 62.6% to 87.5%), and highly susceptible (with MDI beyond 87.5%).

Greenhouse study to identify the cassava germplasms resistant to TSSM

Greenhouse identification was conducted at the Key Laboratory of Integrated Pest Management on Tropical Crops, CATAS. A total of 202 cassava germplasms, with each germplasm consisting of three replicates and each replicate

TABLE 1 Leaf damage scale classification and definition.

Leaf damage scale	Definition
D_0	No leaf damage
D_1	Minor leaf damage with sporadic white spots; the damaged area accounts for 0.1%–25.0% of the whole leaf
D_2	Minor coherent mite feeding marks; the leaf's damaged area accounts for 25.1%–50.0% of the whole leaf
D_3	The damaged area covers most of the leaf, the leaf appearance seems chlorisis, and the damage area accounts for 50.1%–75.0% of the whole leaf
D_4	The damaged area covers the entire leaf, the leaf demonstrates severe chlorosis symptoms, and the damaged area was beyond 75.1% of the whole leaf.

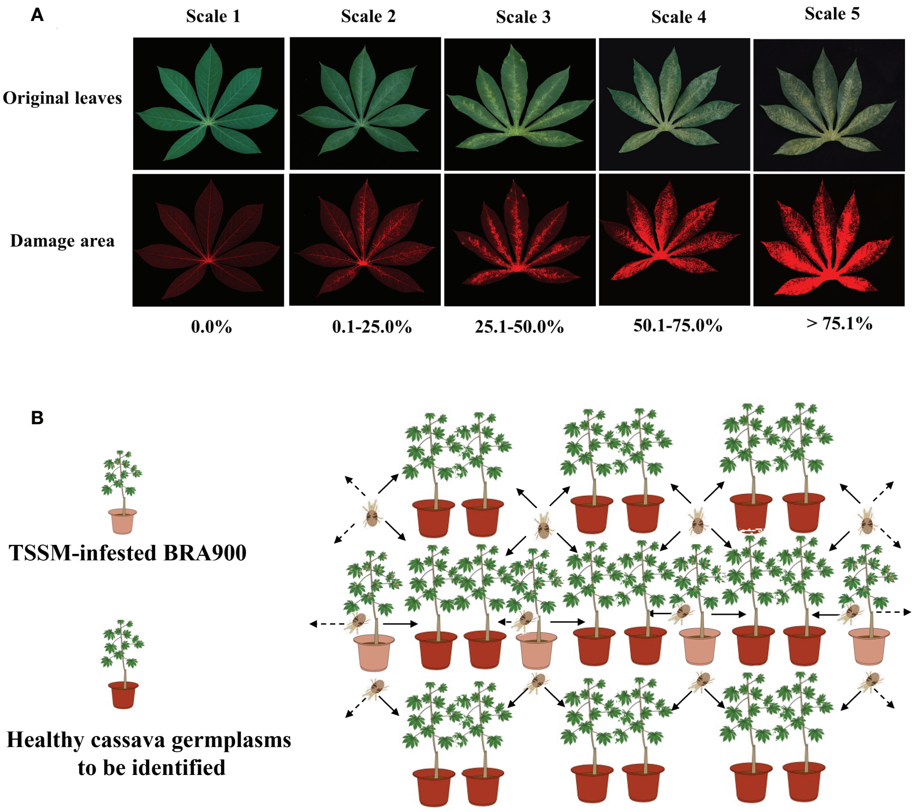


FIGURE 1
Methodology of identification of cassava germplasm resistance to two-spotted spider mite in greenhouse. **(A)** Classification of mite damage scales of cassava leaves. The upper panel indicated the original leaf images, and the lower panel indicated the mite damage area that was visualized by specialized leaf image analysis software. **(B)** Sketch map of mite inoculation in greenhouse identification.

consisting of six plants, were simultaneously used for each experimental setting, and a completely random design was used in glasshouse screening. All cassava germplasms were planted in pots in a greenhouse as previously described. After 3 months, the cassava germplasms were used to evaluate resistance to TSSM. For artificial TSSM infestation, a single cassava variety BRA900 which had previously been infested with identical adult TSSM per plant (approximately 500–600 mites) was placed between the cassava test pots, and it was made sure that the infested plants attached to the healthy ones so that the mites were allowed to move naturally between plants. In this setup, six cassava plants were exposed to two infested plants (Figure 1B). After 8 days post-mite infestation (dpi), the three most seriously infested mature leaves (judged from the phenotypical symptoms) were sampled and evaluated in terms of TSSM resistance level as per the method mentioned above; therefore, there were 18 leaves for each replicate of each germplasm. The greenhouse identification was conducted for three consecutive rounds (in the year 2015), and each round of experiment took about 4 months.

The effect of identified mite-resistant and mite-susceptible cassava germplasms on the reproduction and development of TSSM

The cassava germplasms which were identified to be resistant and susceptible to TSSM were selected to evaluate their performance on the reproduction and development of TSSM. Fifty female adults (1 day old) were placed on the leaf back of each cassava germplasm. The mortality was recorded every day until 8 dpi. In addition, eggs laid within 24 h remained on the leaf. For TSSM development observation, the individual newly emerged larva was placed on the surface of the cassava leaf grids, which was divided by using water-saturated blotting paper strip. The developmental duration of eggs, protonymph, deutonymph, and female adults of F_0 TSSM was investigated every 12 h, and at least 50 tested mites were observed per cassava germplasm. Moreover, the fecundity and the egg hatchability of a single female adult were observed on a leaf from the cassava germplasm until the adult died. The average egg number of each female together with the hatching rate of F_1 TSSM were recorded using a microscope. Fifty individuals divided into three groups were observed for each cassava germplasm.

Field identification of cassava variety resistant to TSSM

Experiments were performed in the field located at Danzhou, Wuming, and Baoshan, respectively (the geographic information and soil properties of these three regions are listed in Supplementary Table S1). These three experiment sites were the perennial epidemic area of TSSM; hence, the TSSM population was allowed to spontaneously accumulate in the field test without artificial inoculation. Weather elements, *i.e.*, rainfall, relative humidity, and temperature, were analyzed according to the data recorded by the weather stations located in the experiment sites (Supplementary Figures S2–S7).

Before formal field identification, a mite population survey was first conducted on identified "HS" cassava variety BRA900 at the three experiment sites (in the year 2016); the purpose is to confirm the precise inspection time for identification of cassava resistance to TSSM. The survey was carried out 1 month after planting until harvest and on the 5th and 20th of each month (twice a month). The months of mite population peak at those three regions (about 2 months after planting) were recorded and considered as the propriate time for further mite resistance identification (six times for each planting season). In addition, the survey was continuously carried on with the mite resistance identification to ensure the reliability of the identification results in each tested year.

The consecutive 5 years of field identifications were carried out from 2017 to 2021, and 10 months was required from planting to harvest (the schedule can be seen in Supplementary Table S2). The treatment plots for each variety were 8 m × 2 m (16 m^2) in a randomized complete block design (Supplementary Figure S8A). Cassava stem segments were planted at a distance of 1.0 m between rows and 0.8 m between plants in the row, and each plot consisted of three rows (30 cassava plants) (Supplementary Figure S8B). In addition, each variety consisted of four replicates (plots), that is, 120 plants for each variety, and BRA900 was planted three rows in the buffer zone (Supplementary Figure S8A). During the identification process, 10 plants from each plot were used, and then three most seriously damaged mature leaves (leaf with ambiguous symptom or coexistence with other pests other than TSSM was not sampled; only that with typical or explicit TSSM symptom was selected) from the top, middle, and basal canopies of the plant were used for TSSM resistance identification (that was nine leaves for a single plant, approximately 90 leaves for a plot, and 360 leaves for a variety). Once the 10 plants were selected for the first time, subsequent sampling was also performed on these same plants, but the leaf sampling was random and depended on the mite damage phenotype. The mite damage scale of each sampled leaf and the number of leaves that correspond to the mite damage scale were recorded, and the final mite resistance level of each variety was calculated based on the average of 5 years of MDI analysis. Furthermore, during the identification period, no acaricide was allowed to be sprayed, while the application of fungicide (50% carbendazim wettable powder) or germicide (2% kasugamycin aqueous solution) was encouraged in case of occurrence of cassava disease. In addition, acaricide treatments were conducted in parallel with the above-mentioned mite resistance identification test, the identical varieties and plot sets were performed, the 43% bifenazate suspension concentrate was used for TSSM control as it is recognized as an effective acaricidal compound for Tetranychidae mite control in a previous study (Liang et al., 2018) and was harmless to natural enemies (Ochiai et al., 2007). The acaricide applications were calendar-based, on the 20th of each month, and sprayed 1 month after planting until 1 month before harvest, respectively (total of eight times for the whole planting season). After 10 months from planting, the yield of each variety was measured (total fresh root tube yield per plot converting to yield per hectare) in either TSSM resistance identification (acaricide-free) and acaricide application tests (yield test was performed for one time with three replicates).

Statistical analyses

The data were analyzed using SPSS (IBM v.25). When analyzing the effects of cassava varieties on the development and reproduction of TSSM, one-way analysis of variance (ANOVA) with Tukey's honestly significant difference multiple-comparison test was used for statistical analysis. $P<0.05$ was considered as a significant difference. All data were firstly subjected to a homogeneity test and were log- or square root-transformed if they did not meet the assumptions of normality and homoscedasticity. Moreover, a non-parametric method (Kruskal–Wallis test for independent samples) was applied when combining the three rounds of greenhouse screening results due to the utilization of categorical/qualitative data. In addition, for the field validation, a generalized linear mixed model (GLMM) was used to analyze the multiple effects such as experiment sites, acaricide application, or cassava varieties on the mite damage symptom or yield. GLMM was implemented by SPSS GENLINMIXED, with a robust estimation method for standard errors (Huber-White sandwich estimator) to account for heterogeneity of variances (Bolker et al., 2009).

Additive main effects and multiplicative interaction (AMMI) model (Gupta et al., 2021) was used to analyze the variety–environment interaction and evaluate the stability of mite resistance of each cassava variety by using Data Processing

System software v. 9.50 (Tang and Zhang, 2013). Firstly, the total variation was decomposed into variety main effect (V), environment main effect (E), and variety–environment interaction effects (VEI), and then VEI values were subjected to principal component analysis, and several significant interaction principal component axes (IPCA) were obtained. Usually, IPCA1 and IPCA2 were used, where IPCA1 represents responses of variety that are proportional to the environments, which are associated with the variety × environment interaction, while IPCA2 provides information about cultivation locations that are not proportional to the environments, indicating that those are responsible of the variety × environment crossover interaction (Ajay et al., 2020). Moreover, the stability parameters (Dv and De) were available by calculating the Euclidean distance between each variety (V) or environment (E) point in the significant IPCA space and the coordinate origin. These parameters were used to evaluate the stability of the varieties in three different field test regions. The equations for Dv and De are listed below:

$$Dv = \sqrt{\sum_{k}^{\mathrm{m}} IPCA_{vk}^{2}}$$

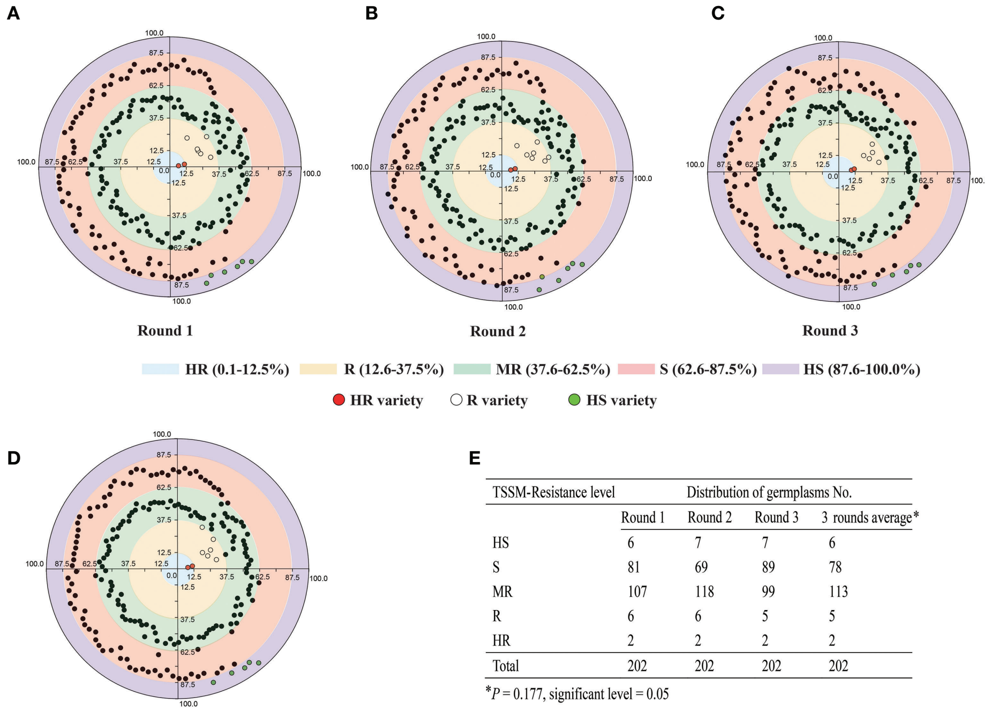
$$De = \sqrt{\sum_{k}^{\mathrm{m}} IPCA_{ek}^{2}}$$

where m represents the number of significant IPCAs in the model, and $\mathrm{IPCA_{vk}}$ and $\mathrm{IPCA_{ek}}$ represent the values of vV and E on the *k*-th of IPCA, respectively.

Results

Large-scale identification of cassava germplasm resistant to TSSM under greenhouse condition

A total of 202 cassava germplasms, including all the main cultivars in China, were subjected to three rounds of TSSM resistance identification under greenhouse condition. As shown



TSSM-Resistance level	Distribution of germplasms No.			
	Round 1	Round 2	Round 3	3 rounds average*
HS	6	7	7	6
S	81	69	89	78
MR	107	118	99	113
R	6	6	5	5
HR	2	2	2	2
Total	202	202	202	202

*$P = 0.177$, significant level = 0.05

FIGURE 2
Three rounds of greenhouse identification of 202 cassava germplasm resistance to two-spotted spider mite. **(A)** First round of greenhouse identification. **(B)** Second round of greenhouse identification. **(C)** Third round of greenhouse identification. **(D)** Average of three rounds of greenhouse identifications. The different color zones indicated the different resistant levels; moreover, the values on the axes indicated the mite damage index ranges that distinguish different levels of resistance. In addition, the identified HR, R, and HS varieties were marked as red, white and green circles; other MR or S varieties were marked as black circles, respectively. **(E)** Summary data of the three rounds of greenhouse identification. The data was analyzed separately (for each round) or combined (three round average). The asterisk indicates that there is no significant difference (P = 0.177) of resistance levels among the three rounds of screening, and the results can be combined for analysis as validated by the non-parametric test method (Kruskal–Wallis test for independent samples).

in Figures 2A–E and Supplementary Data 1, the majority of the cassava germplasms were distributed in the S and MR regions. Nevertheless, several germplasms may shift from one resistance level to another level. More specifically, the first-round assay identified six "HS" germplasms, 81 "S" germplasms, 107 "MR" germplasms, six "R" germplasms, and two "HR" germplasms (Figures 2A, E); the second-round assay identified seven "HS" germplasms, 69 "S" germplasms, 118 "MR" germplasms, six resistance "R" germplasms, and two "HR" germplasms (Figures 2B, E); the third-round assay identified seven "HS" germplasms, 89 "S" germplasms, 99 "MR" germplasms, five "R" germplasms, and two "HR" germplasms (Figures 2C, E). In addition, there was no significant difference (P = 0.177) of resistance levels among the three rounds of screening, and the results can be combined for analysis (Figure 2E). To sum up, on the average, three rounds of assays identified six "HS" germplasms, 78 "S" germplasms, 113 "MR" germplasms, five "R" germplasms, and two "HR" germplasms (Figures 2D, E). Furthermore, C1115 and MIANDIAN are two varieties that were always identified to be "HR", while SC5, SC9, SC15, COLUMBIA-4D, and LIMIN are five varieties that were always identified to be "R". Moreover, KU50, BREAD, SC205, TMS60444, and BRA900 are five varieties that were always identified to be "HS". Hence, those 12 cassava varieties, which exhibited stable resistant performance in greenhouse screening, were used to investigate their capability in affecting the reproduction and development of TSSM in the laboratory.

The effect of identified mite-resistant and mite-susceptible cassava germplasms on the reproduction and development of TSSM

To examine the effect of 12 cassava varieties on TSSM, the mortality, reproduction, and development of TSSM were analyzed. The results speculated that the mortality to TSSM (F_0 generation) was significantly different among the 12 cassava varieties (Figure 3A). The TSSM showed very low mortalities when fed on "HS" varieties such as KU50, BREAD, SC205, TMS60444, and BRA900 (the cumulative mortality within 8 days were all below 10%). On the contrary, mite fed on "R" varieties,

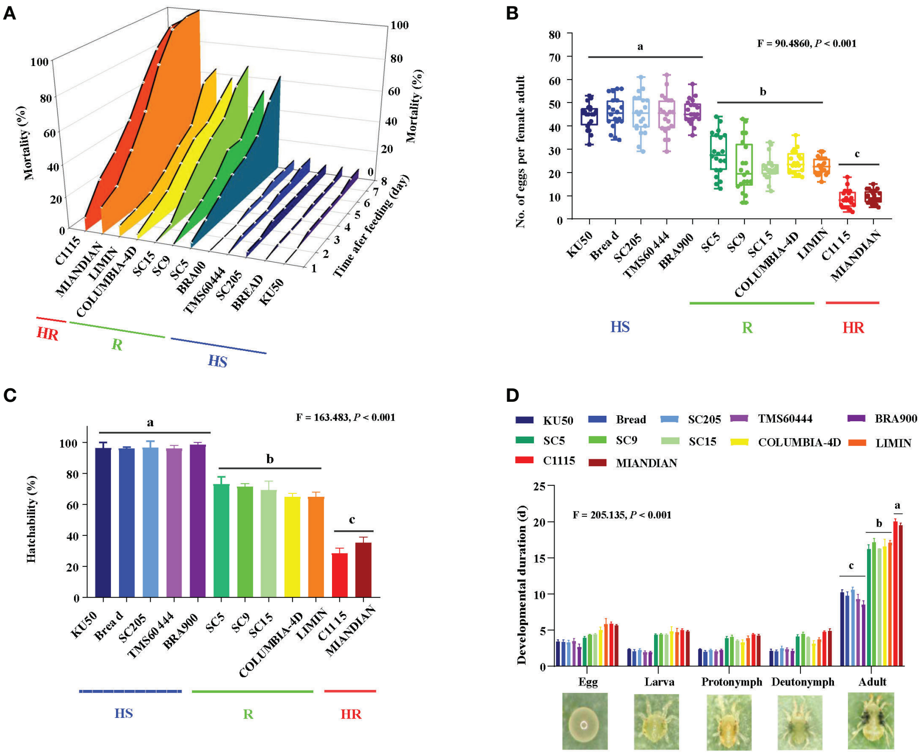


FIGURE 3
Effect on the reproduction and development of two-spotted spider mite (TSSM) while fed on different cassava varieties. **(A)** Mortalities of TSSM at 8 days post-infestation, **(B)** number of eggs per female adult, **(C)** hatchability, and **(D)** developmental duration. All data were first subjected to a homogeneity test and were log- or square root-transformed if they did not meet the assumptions of normality and homoscedasticity. Different letters indicate significant differences among batches of HR, R, and HS varieties; all analyses were based on one-way analysis of variance with Tukey's honestly significant difference multiple comparison test (P< 0.05). The F- and P-values were presented in each panel.

i.e., SC5, SC9, SC15, COLUMBIA-4D, and LIMIN, exhibited very high mortalities (the cumulative mortality within 8 days ranged from 52.45% to 67.26%). Most notably, the "HR" varieties C1115 and MIANDIAN presented the most robust lethal effect to TSSM. The mortalities of TSSM on these two varieties sharply increased after feeding and suffered 100% death within 8 dpi (Figure 3A). The phenomenon that cassava varieties with different resistant levels possessed significantly different capacities in inhibiting TSSM reproduction can also be seen in the aspect of fecundity (*F* = 90.486, *P*< 0.001) and hatchability (*F* = 163.483, *P*< 0.001). The average number of eggs per female adult on "HS", "R", and "HR" varieties were approximately 45.18, 23.86, and 9.10, respectively (Figure 3B). In addition, the average hatchability of TSSM on "HS", "R", and "HR" varieties was approximately 96.73%, 68.73%, and 31.82%, respectively (Figure 3C). However, the results were reversed in terms of development; both the "HR" and "R" varieties might significantly prolong the developmental duration of TSSM in each stage (*i.e.*, egg, larva, protonymph, and deutonymph). The duration from egg to adults was 19.75 and 16.64 days, which were significantly longer compared with the HS varieties (9.81 days) (*F* = 205.135, *P*< 0.001) (Figure 3D). The abovementioned results suggested that resistant cassava varieties may significantly impede the normal reproduction and development of TSSM.

Field identification of cassava varieties resistant to TSSM

Cassava varieties that possessed ideal resistance to TSSM in the laboratory were used for further field identification and validation; for comparison, the "HS" varieties were also used for field test. The field experiments were carried out in three major production provinces in China (Danzhou City, Hainan

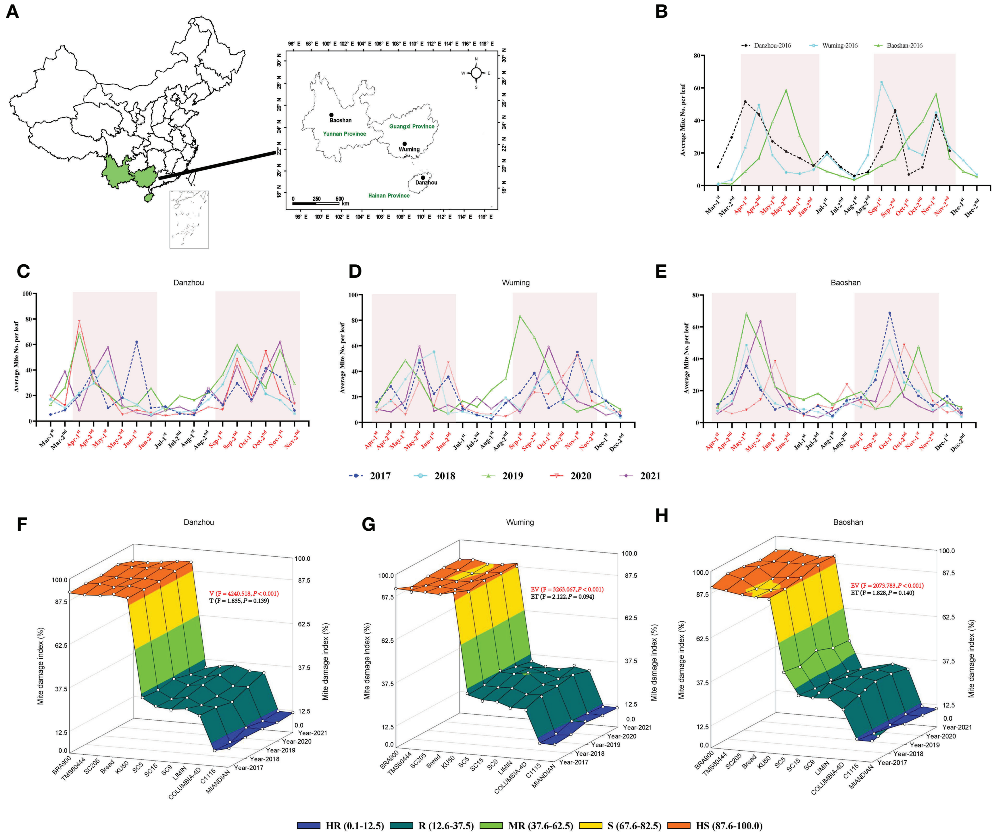


FIGURE 4
Five years of field validation of cassava varieties resistant to two-spotted spider mite (TSSM). **(A)** Geographic information of the three different sites (Danzhou, Wuming, and Baoshan) for field validation in China. **(B)** Population dynamic of TSSM at Danzhou, Wuming, and Baoshan in the year of 2016. **(C–E)** Population dynamic of TSSM at Danzhou **(C)**, Wuming **(D)**, and Baoshan **(E)** from 2017 to 2021. The shaded boxes indicate the TSSM population peak time frame of the three sites. **(F–H)** Field validation results of TSSM resistance of 12 cassava varieties at Danzhou **(F)**, Wuming **(G)**, and Baoshan **(H)**. The different color zones indicated the different resistant levels. The generalized linear mixed model was used to analyze the multiple effects such as experimental varieties and experiment time on the mite damage index, the *F*- and *P*-values were indicated within panels **(F–H)**. Furthermore, the *P*-values that represent statistical difference were marked in red (significance level = 0.05).

Province, Wuming City, Guangxi Province, and Baoshan City, Yunnan Province) (Figure 4A). During the experimental period, the weather elements were recorded and analyzed. In general, the monthly average temperature, rainfall, and humidity were ranked as follows: Danzhou > Wuming > Baoshan; Danzhou and Wuming were relatively similar in weather elements (Supplementary Figures S2–S7), where these two sites were "wetter" than Baoshan.

Before formal field identification, a mite population survey was first conducted on "HS" cassava variety BRA900 (in the year 2016). The results showed that the mite population was correlated to the weather condition. For the years with too much rainfall (*i.e.*, 2020 and 2021), the mite population was relatively lower than the other years and *vice versa* (Supplementary Figures S2–S7). In addition, although the three sites presented different dynamics of the mite population, there were two mite population peak periods during every planting season, which were from April to June and from September to November (Figure 4B). Furthermore, when conducting formal field identification in the following 5 years (from 2017 to 2021), the mite population peaks were also confined to those periods (Figures 4C–E). Based on the preliminary and formal surveys of mite population, the field identification for each variety was also conducted six times per planting season (on the 20th of April, May, June, September, October, and November).

The results of 5 years of field identification of a cassava variety resistant to TSSM are depicted in Figures 4F–H and Supplementary Data 2. The TSSM resistance levels of 12 cassava varieties in Danzhou was most stable, as for every tested year the identified resistance level of each variety in the field was exactly consistent with the results acquired from greenhouse analysis, namely as follows: C1115 and MIANDIAN were also identified as "HR" varieties; SC5, SC9, SC15, COLUMBIA-4D, and LIMIN were also identified as 'R' varieties, while KU50, BREAD, SC205, TMS60444 and BRA900 were also identified as "HS" varieties under field conditions (Figure 4F). The field performance in Wuming was also supposed to be "consistent" with the lab results but with minor exceptions—for example, in the years 2020 and 2021, the variety SC205 was "S", while it was "HS" in the rest of the years (from 2017 to 2019). In addition, the resistance level of SC15 in the year 2019 was identified as "MR", while it was "R" in other years. A similar fluctuation of TSSM resistance can also be seen in the experiment conducted in Baoshan [*i.e.*, SC205 (2018) and BREAD (2017 and 2018)] (Figure 4G). However, it was noticeable that SC5 was identified to be "MR" variety for five consecutive years, while it was supposed to be "R" in the greenhouse identification (Figure 4H). Some cassava varieties, like C1115, MIANDIAN, COLUMBIA-4D, LIMIN, KU50, TMS60444, and BRA900, exhibited a stable and consistent resistance level across all the five years tested. In addition, the GLMM analysis for the three sites showed that a significant difference of MDI was observed between resistant and susceptible varieties, and the overall MDIs in Danzhou were significantly higher compared with those in Baoshan (Table 2), while the results were relatively stable across the five experiment years (Figures 5F–H, Table 2).

To evaluate the capacity in reducing the losses of the yield, those 12 cassava varieties were subjected to estimation of yield under either acaricide-free or acaricide application conditions. In general, for all the varieties, it was common that the acaricide application groups all showed a significantly higher yield compared with those in the acaricide-free groups. The yields for acaricide application ranged from 23.12 to 42.36 tons/ha (Figures 5A–L), depending on the tested varieties, in which SC5 showed the highest yield, while TMS60444 showed the lowest yield. Conversely, without acaricide application, the yields decreased significantly; specifically, the "HR" varieties could maintain about 60%–70% of the yield (Figures 5A, B), and the "R" varieties could maintain about 50%–60% of the yield (Figures 5C–G). However, the "HS" varieties suffered the most remarkable drop in production, with approximately 70%–90% reduction of yield (Figures 5H–L). In addition, the GLMM analysis showed that, for each variety, the acaricide application always presented a significantly higher yield (all *P*-values lower than 0.05), while most of the experiment sites and years presented a statistical difference, depending on the TSSM resistance level (all the resistant varieties showed a significant difference in the three sites). These results indicated that the higher the TSSM resistant level, the better the performance in maintaining the yield.

The stability and the adaptability of the 12 cassava varieties in different regions were also examined by using the AMMI model. The results speculated that variety (V), environment (E), and variety–environment interaction (VEI) would extremely significantly influence either MDI (Supplementary Table 3) or the cassava yield (Supplementary Table 4). Furthermore, in the AMMI bi-plot, the average MDI or yield was set in the X-axis, while the corresponding IPCA1s were set in the Y-axis. In the bi-plot chart, the closer the variety to the Y-axis, the more stable that the variety showed in MDI or yield and the lesser were the variation of the region sites in the TSSM resistance identification. As suggested by Figure 6, on one hand, when focused on MDI, the stability of the 12 cassava varieties was ranked as follows: BRA900 > SC9 > MIANDIAN > TMS60444 > C1115 > COLUMBIA-4D > SC15 > KU50 > LIMIN > SC205 > BREAD > SC5 (Figure 6A, Supplementary Table S5); on the other hand, when focused on yield, the stability of the 12 cassava varieties were ranked as follows: C1115 > MIANDIAN > COLUMBIA-4D > SC15 > LIMIN > SC9 > BRA900 > SC205 > BREAD > KU50 > TMS60444 > SC5 (Figure 6B, Supplementary Table S6). Moreover, in those two situations, Wuming was the region with the lowest variation for TSSM resistance identification, followed by Danzhou, while Baoshan showed a higher variation than the former two regions (Supplementary Tables S7, S8).

TABLE 2 Generalized linear mixed model (GLMM) evaluating the effect of experiment sites, cassava varieties, and experiment years on mite damage index.

Variables	Factors	Estimate	SE	t	*P*	95% CI	
						Upper	Lower
Experiment sites	Danzhou	0.940	0.3162	2.971	**0.003**	0.315	1.564
	Wuming	-0.008	0.3162	0.000	1.000	-.624	0.624
	Baoshan	0[a]					
Cassava varieties	SC5	-54.850	0.6325	-86.723	**0.000**	-56.099	-53.601
	SC9	-56.160	0.6325	-88.795	**0.000**	-57.409	-54.911
	SC15	-56.291	0.6325	-89.001	**0.000**	-57.540	-55.042
	COLUMBIA-4D	-58.566	0.6325	-92.599	**0.000**	-59.815	-57.317
	LIMIN	-58.378	0.6325	-92.302	**0.000**	-59.627	-57.130
	MIANDIAN	-78.429	0.6325	-124.004	**0.000**	-79.678	-77.180
	C1115	-78.645	0.6325	-124.345	**0.000**	-79.894	-77.396
	KU50	1.376	0.6325	2.176	**0.029**	0.127	2.625
	Bread	0.713	0.6325	1.127	0.261	-0.536	1.962
	SC205	0.330	0.6325	0.522	0.602	-0.919	1.579
	TMS60444	0.727	0.6325	1.150	0.252	-0.522	1.976
	BRA900	0[a]					
Experiment years	Year-2017	-0.101	0.3571	-.284	0.777	-0.807	0.604
	Year-2018	0.015	0.3774	0.040	0.968	-0.730	0.761
	Year-2019	1.019	0.4615	2.208	0.092	0.108	1.930
	Year-2020	0.539	0.3899	1.383	0.168	-0.231	1.309
	Year-2021	0[a]					

[a]This factor is redundant in GLMM analysis, so it is set to zero and as reference during pairwise comparison.
The bold values indicate the P values were lower than 0.05 and showed statistical significance.

Discussion

In the present study, we made considerable efforts in identifying the major cassava germplasms resistant to TSSM in China, under both greenhouse and field conditions. Finally, six cassava varieties with stable resistant levels across different planting environments were identified from 202 tested germplasms. In a previous study, four cassava varieties were identified to be resistant to a sibling mite species, the carmine spider mite, *Tetranychus cinnabarinus*, under laboratory conditions (Lu et al., 2017), but there was a lack of further field validation. Another study had also identified two cassava varieties presenting resistance to *T. cinnabarinus* under both greenhouse and field trials (Yang et al., 2020); however, the experiments were only performed in a single site and a quite limited planting season. As far as we know, this is the first attempt to conduct a large screening of cassava resistance to *Tetranychus* species. In this study, both the environmental and temporal stability were taken into account, which might largely ensure the reliability of the results.

This study provides a reliable and easy-to-handle method for screening cassava germplasm resistance to mites. In our opinion, this method is based on two factors: first, the mite damage symptom, which indicates the resistant level of cassava germplasm, must be correctly distinguished by relying on a measurable strategy rather than experience; and second, the sampled leaves should ultimately represent the actual resistance level of the overall sample, in which a quantitative method is recommended. For the first point, we have developed a computer-aided visual quantification method to determine the mite damage scale. This method can also be seen in a study regarding assessing cucumber leaf damage caused by TSSM (Uygun et al., 2020). Nevertheless, most of the studies still used the traditional way of monitoring leaf symptoms caused by pest—for example, a 1–6 damage scale was used to define the CGM resistance level by the CIAT, and 72 cultivars among the 300 cultivars consistently demonstrated lower than 3.0 of the damage ratings, indicating low to moderate CGM resistance (Bellotti et al., 2012). Hussey and Parr (1963) established a leaf damage index with 0–5 scale based on the visual evaluation of leaf infestation caused by TSSM. In another study, chlorophyll content was used to evaluate mite damage (Iatrou et al., 1995). This study offered a novel means of distinguishing the severity of TSSM damage symptoms on cassava. In addition, resistance level judgment only relies on symptoms rather than on a quantitative method, which seems empirical and not precise. Thus, parameters such as damage index or resistance index were introduced to quantitatively calculate the exact pest resistance levels of several crops like cowpea (Jackai, 1982) and potato (Fathi, 2014). Once again, for the first time, we

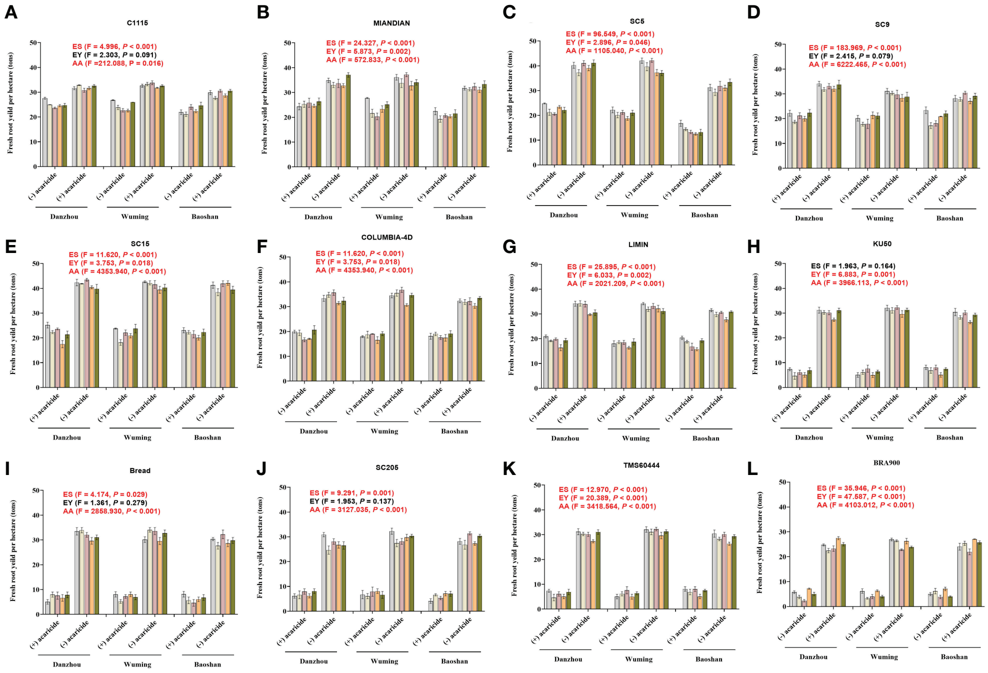


FIGURE 5
The capacity of 12 cassava varieties in reducing the yield losses during the field validation. **(A)** C1115. **(B)** MIANDIAN **(C)** SC5. **(D)** SC9. **(E)** SC15. **(F)** COLUMBIA-4D. **(G)** LIMIN. **(H)** BRA900. **(I)** SC205. **(J)** BREAD. **(K)** TMS60444. **(L)** BRA900. "+" indicates acaricide application, while "-" indicates without acaricide application. Generalized linear mixed model was used to analyze the multiple effects such as acaricide application, experiment sites, and years of experiments on the cassava yield. The *F*- and *P*-values are indicated within panels **(A–L)**. Moreover, the *P*-values that represent statistical difference are marked in red (significance level = 0.05).

developed a MDI-based approach that makes it more accurate and easier to evaluate cassava resistance to TSSM. With this method, stable TSSM resistance cassava varieties were excavated from 202 cassava core germplasms in China.

A total of 202 cassava germplasms, including all the main cultivars in China, were subjected to three rounds of TSSM resistance identification in 2016. On the whole, the distribution of the germplasms with different resistant levels fit the "spindle type", as the "HR" and "HS" varieties were relatively rare, two varieties were "HR", five varieties were "R", and five varieties were "HS". In comparison, most of the germplasms belonged to "MR" or "S" levels. In addition, the resistance levels of several

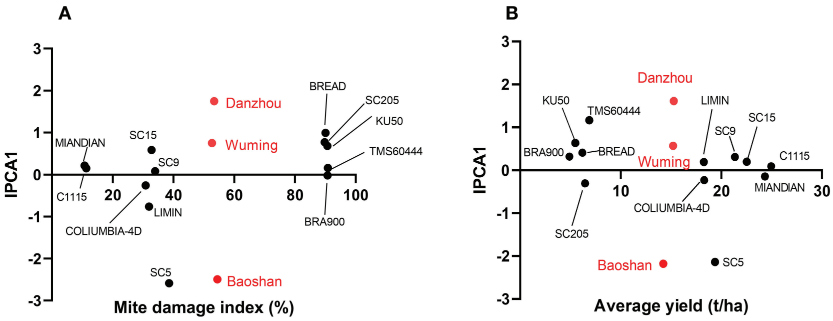


FIGURE 6
Stability analysis of the 12 cassava varieties and discrimination of regional sites by Additive main effects and multiplicative interaction biplot. **(A)** Mite damage index-based stability analysis of the 12 cassava varieties. **(B)** Yield-based stability analysis of the 12 tested cassava varieties. The 12 cassava varieties are marked with black circles, while the three regional sites are marked with red circles.

germplasms were flexible or unstable in different rounds of identification, as one germplasm would shift from one resistance level to another level, while the 12 varieties mentioned above were steadily kept at their fixed resistant levels in any round of identification. The phenomenon of the scarcity of resistant germplasm resource is common in the insect pest resistance identification of other plant species, as cases can be found in the screening of soybean (334 genotypes) resistant to aphid (*Aphis glycines*) (Bhusal et al., 2013), pepper (50 accessions) resistant to green peach aphid (*Myzus persicae*) (Frantz et al., 2004), cotton (over 400 cultivars) resistant to whitefly (*Bemisia tabaci*) (Li et al., 2016), and maize (38 genotypes) resistant to fall armyworm (*Spodoptera frugiperda*) (Soujanya et al., 2022). In those studies, no more than seven resistant genotypes/accessions/cultivars were identified. In addition, the identification of cassava resistance to insect pest also showed this similar phenomenon—for example, field identification of about 5,500 genotypes was performed in Colombia. Approximately 75% are susceptible to whitefly (*Aleurodicus socialis*). Moreover, over 5,000 landrace cultivars in the CIAT cassava germplasm bank had been assessed, but only 6% were identified as being with low to moderate resistance to CGM. Despite the fact that considerable effort has been concentrated on screening pest-resistant cassava genotypes, few insect pest-resistant, commercial varieties are being cultivated during the past decades (Amelework and Bairu, 2022). Nevertheless, in this study, seven stably resistant varieties were identified; in particular, the good news is that although the opportunity to get highly resistant materials is usually rare during germplasm screening, we are still lucky to get two highly resistant varieties. In addition, the resistant varieties identified here are totally different from those reported in other studies, and some of them are landraces in China. Collectively, these varieties could be used as good materials for germplasm exchange or creation and will benefit future breeding programs for better management of TSSM.

The resistant cassava varieties significantly inhibited the reproduction and development of TSSM. When fed on resistant cassava varieties, the survival, oviposition, and hatchability of TSSM were all significantly inhibited, while the developmental durations were dramatically extended. Most notably, those adverse effects on TSSM were differentiated by the 12 varieties with distinct resistant levels. This phenomenon can explain the contrasting TSSM infestation phenotype of different varieties in the greenhouse as well as in the field, such that the higher the resistance level, the stronger the inhibition to TSSM. Similar results mentioned above can also be found in several pest–crop interaction studies, *i.e.*, cotton genotype and silverleaf whitefly (Miyazaki et al., 2013), cassava varieties and papaya mealybug (*Paracoccus marginatus*) (Chen et al., 2022b) or *T. cinnabarinus* (Lu et al., 2017), rubber tree germplasms, and *Eotetranychus sexmaculatus* (Lu et al., 2016).

As a general rule, identification of crop resistance to insect pest should undergo both greenhouse and field tests; germplasms that presented excellent and consistent resistance performance can be considered as promising materials for pest control or for further breeding programs (Miyazaki et al., 2013). In the present study, the three rounds of greenhouse identification as well as the distinct effect on TSSM development and reproduction might ensure the stability and the reliability of the 12 cassava varieties to a great extent, although the resistance level of certain varieties seemed to fluctuate in certain seasons. Generally speaking, the resistance performance of the resistant or susceptible varieties in the field was quite consistent with those in the greenhouse. By contrast, fewer varieties showed a comparable performance of pest resistance in the field test compared with those in the laboratory test—for instance, there were nine aphid-resistant soybean genotypes identified in the greenhouse, but only two genotypes were identified as resistant in the field test (Bhusal et al., 2013). Zhu et al. (2018) conducted identification of resistance to *B. tabaci* using 550 cotton genotypes in greenhouse and field experiments, although the greenhouse test identified 100 resistant and susceptible genotypes, there were only 42 genotypes that showed identical resistance performance in the field test. This inconsistency may be due to the change of experimental environment—for example, the faba bean (*Vicia faba*) resistance to weed or fungi under multi-environments exhibited distinct resistance levels (Rubiales et al., 2014). In another study of *Botrytis fabae* resistance identification, field validation also revealed the instability of the resistance performance across different environments (Villegas-Fernández et al., 2009). Similar results can also be found in the identification of cassava resistance to GCM—for instance, though 300 cultivars with low to moderate resistance to CGM had been identified by the CIAT in Columbia (in the tropical lowland that possessed a prolonged dry season and endured high CGM populations) (Bellotti, 2008), only 72 cultivars were consistently demonstrated to have the same resistance level in Brazil (primarily in the northeastern semiarid regions) (Bellotti et al., 2012). This phenomenon is probably due to the environmental variability in the field, compared with the stable and normalized culture condition in the greenhouse. Abiotic stress in the field like drought, chilling, loss of applied fertilizers, and waterlogging might hinder the normal physiological development and cause the deterioration of pest resistance. However, in this study, only a minor inconsistency of resistance performance was found between greenhouse and field experiments, indicating that the resistance level of most tested varieties was stable and not inclined to be affected by environmental factors.

In this study, we state that the delicate experiment design (three rounds of greenhouse trials, five consecutive years of field validation, and three different experiment sites) ensures to get stably resistant materials. In addition, by employing the AMMI model, we also found that environment factor will significantly influence the TSSM resistance, as the variety SC5 was identified as resistant at Danzhou and Wuming but was moderately resistant at Baoshan, while the rest of the 11 varieties showed a consistent resistance performance across these three different sites. It is interesting to decipher in a future study why only SC5 exhibited an environment-dependent manner. Once again, as highlighted in the present study, most of the tested varieties showed equal resistance performance in Danzhou, Wuming, and Baoshan. As these three cities are the major areas of cassava cultivation in China, the resistant varieties identified here, to a certain extent, could probably accommodate various cassava planting environments in China; however, more regional tests are still needed to verify this hypothesis.

The yield of the 12 cassava varieties can reflect their resistance levels in the field. As acaricide was applied eight times throughout the planting season and covered all the mite population peak period of different experiment sites, we assumed that the yield we tested can represent the veritable yield of each of the tested varieties. Although C1115 and MIANDIAN were identified as "HR" varieties, they still suffered about 30% of yield losses without acaricide application. Comparatively, the "R" varieties can maintain less of the yield (50%–60%), and the TSSM caused significant yield loss to the "HS" varieties (over 80%). Interestingly, SC5 was identified as "MR" varieties in Baoshan. As expected, the yield losses were higher than those in Danzhou and Wuming, where it was identified as "R" varieties. There were rare but still some reports on insect- or mite-resistant cultivars being released to the field for pest control and to achieve a good profit. A selected number of moderate CGM-resistant cultivars had been introduced to growers by breeders and entomologists (Bellotti, 2008). Moreover, a cultivar named "Nataima-31" had been cultivated in Tolima, Colombia. This cultivar can attain a high yield of 33 t/ha (34% higher than the regional famers' variety) without pesticide applications, and now this cultivar is being grown commercially in different regions of Colombia, Ecuador, and Brazil (CIAT, 2007). The planting area of the resistant varieties SC9, SC15, and LIMIN is expanding in China (Qin et al., 2017), which were promising main cultivars in TSSM control. Although C1115, MIANDIAN, and COLUMBIA-4D were not commercially released, they can also be considered as good material in breeding programs of mite resistance. Moreover, those varieties with distinct resistance to TSSM can promote the omics study, especially for mining the pest resistance gene, or to probe markers of pest resistance, which will, in turn, accelerate the progress of resistance breeding.

Conclusion

In conclusion, a quantifiable identification method was used to identify cassava resistance to TSSM, and based on this method, a panel of TSSM-resistant varieties were identified under greenhouse and field conditions. This study provides promising materials for effective mite control or as good materials for deciphering the mite resistance mechanism as well as benefiting for future breeding programs of mite resistance.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Author contributions

XL, QC, and YL planned and designed the research and experiments. KL, CW, MW, XY, YQ, YZ, and YG performed the laboratory experiments and analyzed the data. XL, QC, and YL wrote and edited the paper. XL and QC acquired the funds. All authors contributed to the article and approved the submitted version.

Funding

The authors acknowledge financial support from the National Key R&D Program of China (2020YFD1000603), China Agriculture Research System (CARS-11-HNCQ), NanFeng Earmarked Fund of the Ministry of Agriculture and Rural Affairs of P.R.C (NFZX-2021), and Hainan Major Science and Technology Project (no. ZDKJ202002).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1054909/full#supplementary-material

References

Ajay, B. C., Bera, S. K., Singh, A. L., Kumar, N., Gangadhar, K., and Kona, P. (2020). Evaluation of genotype × environment interaction and yield stability analysis in peanut under phosphorus stress condition using stability parameters of AMMI model. *Agric. Res.* 9 (4), 477–486. doi: 10.1007/s40003-020-00458-3

Amelework, A. B., and Bairu, M. W. (2022). Advances in genetic analysis and breeding of cassava (*Manihot esculenta* crantz): A review. *Plants* 11 (12), 1617. doi: 10.3390/plants11121617

Ariyo, O. A., Dixon, A. G. O., and Atiri, G. I. (2005). Whitefly *Bemisia tabaci* (Homoptera: Aleyrodidae) infestation on cassava genotypes grown at different ecozones in Nigeria. *J. Econ. Entomol.* 98 (2), 611–617. doi: 10.1093/jee/98.2.611

Barilli, D. R., Wengrat, A. P. G., d., S., Guimarães, A., Pietrowski, V., Ringenberg, R., et al. (2019). Resistance of cassava genotypes to *Bemisia tuberculata*. *Arthropod-Plant Int.* 13, 663–669. doi: 10.1007/s11829-019-09694-z

Bellotti, A. C. (2008). *Cassava pests and their management: encyclopedia of entomology*. Ed. J. L. Capinera (Dordrecht: Springer Netherlands), 764–794.

Bellotti, A. C., and Arias, B. (2001). Host plant resistance to whiteflies with emphasis on cassava as a case study. *Crop Prot.* 20, 813–823. doi: 10.1016/S0261-2194(01)00113-2

Bellotti, A., Herrera Campo, B. V., and Hyman, G. (2012). Cassava production and pest management: present and potential threats in a changing environment. *Trop. Plant Biol.* 16, 39–72. doi: 10.1007/s12042-011-9091-4

Bellotti, A. C., Smith, L., and Lapointe, S. L. (1999). Recent advances in cassava pest management. *Annu. Rev. Entomol.* 44 (1), 343–370. doi: 10.1146/annurev.ento.44.1.343

Bhusal, S. J., Jiang, G.-L., Tilmon, K. J., and Hesler, L. S. (2013). Identification of soybean aphid resistance in early maturing genotypes of soybean. *Crop Sci.* 53 (2), 491–499. doi: 10.2135/cropsci2012.06.0397

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., et al. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends. Ecol. Evol.* 24 (3), 127–135. doi: 10.1016/j.tree.2008.10.008

Carabalí, A., Bellotti, A. C., Montoya-Lerma, J., and Fregene, M. (2010a). Manihot *flabellifolia pohl*, wild source of resistance to the whitefly *Aleurotrachelus socialis* bondar (Hemiptera: Aleyrodidae). *Crop Prot.* 29 (1), 34–38. doi: 10.1016/j.cropro.2009.08.014

Carabalí, A., Bellotti, A. C., Montoya-Lerma, J., and Fregene, M. (2010b). Resistance to the whitefly, *Aleurotrachelus socialis*, in wild populations of cassava, *Manihot tristis*. *J. Insect. Sci.* 10 (1), 170–178. doi: 10.1673/031.010.14130

Chen, Q., Liang, X., Wu, C., Gao, J., Chen, Q., and Zhang, Z. (2019). Density threshold-based acaricide application for the two-spotted spider mite *Tetranychus urticae* on cassava: from laboratory to the field. *Pest. Manage. Sci.* 75 (10), 2634–2641. doi: 10.1002/ps.5366

Chen, Q., Liang, X., Wu, C., Liu, Y., Liu, X., Zhao, H., et al. (2022a). Overexpression of leucoanthocyanidin reductase or anthocyanidin reductase elevates tannins content and confers cassava resistance to two-spotted spider mite. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.994866

Chen, Q., Liu, X.-Q., Liang, X., Liu, Y., Wu, C.-L., Xu, X.-L., et al. (2022b). Resistant cassava cultivars inhibit the papaya mealybug *Paracoccus marginatus* population based on their interaction: from physiological and biochemical perspectives. *J. Pest. Sci.* doi: 10.1007/s10340-022-01550-z

Chen, W., Wu, F., and Zhang, J. (2016). Potential production of non-food biofuels in China. *Renew. Energ.* 85, 939–944. doi: 10.1016/j.renene.2015.07.024

CIAT (2007). *Annual report cassava program* (Cali: CIAT).

Fathi, S. A. A. (2014). Screening of the susceptibility of newly released genotypes of potato to thrips infestation under field conditions in northwest Iran. *Crop Prot.* 62, 79–85. doi: 10.1016/j.cropro.2014.04.006

Frantz, J., Gardner, J., Hoffmann, M., and Jahn, M. (2004). Greenhouse screening of capsicum accessions for resistance to green peach aphid (*Myzus persicae*). *Hort. Sci.* 39, 1332–1335. doi: 10.21273/HORTSCI.39.6.1332

Gupta, P., Dhawan, S. S., Lal, R. K., Chanotiya, C. S., and Mishra, A. (2021). Genotype selection over years using additive main effects and multiplicative interaction (AMMI) model under the ascendancy of genetic diversity in the genus ocimum. *Ind. Crop Prod.* 161, 113198. doi: 10.1016/j.indcrop.2020.113198

Hussey, N. W., and Parr, W. J. (1963). The effect of glasshouse red spider mite (*Tetranychus urticae* Koch) on the yield of cucumbers. *J. Pomolog. Horti. Sci.* 38 (4), 255–263. doi: 10.1080/00221589.1963.11514076

Iatrou, G., Cook, C. M., Stamou, G., and Lanaras, T. (1995). Chlorophyll fluorescence and leaf chlorophyll content of bean leaves injured by spider mites (Acari: Tetranychidae). *Exp. Appl. Acarol.* 19 (10), 581–591. doi: 10.1007/BF00048813

Jackai, L. E. (1982). A field screening technique for resistance of cowpea (Vigna unguiculata) to the pod-borer *Maruca testulalis* (Geyer)(Lepidoptera: Pyralidae). *B. Entomol. Res.* 72 (1), 145–156. doi: 10.1017/S0007485300050379

Jiang, D., Hao, M., Fu, J., Tian, G., and Ding, F. (2019). Estimating the potential of energy saving and carbon emission mitigation of cassava-based fuel ethanol using life cycle assessment coupled with a biogeochemical process model. *Int. J. Biometeorol.* 63 (5), 701–710. doi: 10.1007/s00484-017-1437-7

Khalighi, M., Dermauw, W., Wybouw, N., Bajda, S., Osakabe, M., Tirry, L., et al. (2016). Molecular analysis of cyenopyrafen resistance in the two-spotted spider mite *Tetranychus urticae*. *Pest. Manage. Sci.* 72 (1), 103–112. doi: 10.1002/ps.4071

Leeuwen, T. V., Vanholme, B., Pottelberge, S. V., Nieuwenhuyse, P. V., Nauen, R., Tirry, L., et al. (2008). Mitochondrial heteroplasmy and the evolution of insecticide resistance: Non-mendelian inheritance in action. *P. Natl. Acad. Sci. U.S.A.* 105 (16), 5980–5985. doi: 10.1073/pnas.0802224105

Liang, X., Chen, Q., Lu, H., Wu, C., Lu, F., and Tang, J. (2017). Increased activities of peroxidase and polyphenol oxidase enhance cassava resistance to tetranychus urticae. *Exp. Appl. Acarol.* 71 (3), 195–209. doi: 10.1007/s10493-017-0125-y

Liang, X., Chen, Q., Wu, C., and Zhao, H. (2018). The joint toxicity of bifenazate and propargite mixture against *Tetranychus urticae* Koch. *Int. J. Acarol.* 44, 35–40. doi: 10.1080/01647954.2017.1398276

Li, J., Zhu, L., Hull, J. J., Liang, S., Daniell, H., Jin, S., et al. (2016). Transcriptome analysis reveals a comprehensive insect resistance response mechanism in cotton to infestation by the phloem feeding insect *Bemisia tabaci* (whitefly). *Plant Biotechnol. J.* 14 (10), 1956–1975. doi: 10.1111/pbi.12554

Lu, F., Chen, Q., Chen, Z., Lu, H., Xu, X., and Jing, F. (2014). Effects of heat stress on development, reproduction and activities of protective enzymes in *Mononychellus mcgregori*. *Exp. Appl. Acarol.* 63 (2), 267–284. doi: 10.1007/s10493-014-9784-0

Lu, F., Chen, Z., Lu, H., Liang, X., Zhang, H., Li, Q., et al. (2016). Effects of resistant and susceptible rubber germplasms on development, reproduction and protective enzyme activities of *Eotetranychus sexmaculatus* (Acari: Tetranychidae). *Exp. Appl. Acarol.* 69 (4), 427–443. doi: 10.1007/s10493-016-0049-y

Lu, F., Liang, X., Lu, H., Li, Q., Chen, Q., Zhang, P., et al. (2017). Overproduction of superoxide dismutase and catalase confers cassava resistance to *Tetranychus cinnabarinus*. *Sci. Rep.* 7 (1), 40179. doi: 10.1038/srep40179

Miyazaki, J., Stiller, W. N., and Wilson, L. J. (2013). Identification of host plant resistance to silverleaf whitefly in cotton: Implications for breeding. *Field. Crop Res.* 154, 145–152. doi: 10.1016/j.fcr.2013.08.001

Ochiai, N., Mizuno, M., Mimori, N., Miyake, T., Dekeyser, M., Canlas, L. J., et al. (2007). Toxicity of bifenazate and its principal active metabolite, diazene, to tetranychus urticae and panonychus citri and their relative toxicity to the predaceous mites, *Phytoseiulus persimilis* and *Neoseiulus californicus*. *Exp. Appl. Acarol* 43 (3), 181–197. doi: 10.1007/s10493-007-9115-9

Ochwo-Ssemakula, M., Catherine, G., and Sseruwagi, P. (2019). Whitefly resistance in African cassava genotypes. *Afr. Crop Sci. J.* 27 (2), 213–228. doi: 10.4314/acsj.v27i2.7

Omongo, C. A., Kawuki, R., Bellotti, A. C., Alicai, T., Baguma, Y., Maruthi, M. N., et al. (2012). African Cassava whitefly, *Bemisia tabaci*, resistance in African and south American cassava genotypes. *J. Integr. Agr.* 11 (2), 327–336. doi: 10.1016/S2095-3119(12)60017-3

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

Parsa, S., Medina, C., and Rodríguez, V. (2015). Sources of pest resistance in cassava. *Crop Prot.* 68, 79–84. doi: 10.1016/j.cropro.2014.11.007

Qin, Y., Djabou, A. S., An, F., Li, K., Li, Z., Yang, L., et al. (2017). Proteomic analysis of injured storage roots in cassava (*Manihot esculenta* crantz) under postharvest physiological deterioration. *PloS One* 3, e0174238. doi: 10.1371/journal.pone.0174238

Rubiales, D., Flores, F., Emeran, A. A., Kharrat, M., Amri, M., Rojas-Molina, M. M., et al. (2014). Identification and multi-environment validation of resistance against broomrapes (*Orobanche crenata* and *Orobanche foetida*) in faba bean (*Vicia faba*). *Field. Crop Res.* 166, 58–65. doi: 10.1016/j.fcr.2014.06.010

Soujanya, P. L., Sekhar, J. C., Yathish, K. R., Karjagi, C. G., Rao, K. S., Suby, S. B., et al. (2022). Leaf damage based phenotyping technique and its validation against fall armyworm, *Spodoptera frugiperda* (J. e. smith), in maize. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.906207

Tang, Q.-Y., and Zhang, C.-X. (2013). Data processing system (DPS) software with experimental design, statistical analysis and data mining developed for use in entomological research. *Insect Sci.* 20 (2), 254–260. doi: 10.1111/j.1744-7917.2012.01519.x

Tan, Y., Li, C., and Zeng, H. (2018). Analysis of cassava production and trade development in China. *Global Agricult.* 10, 163–168. doi: 10.13856/j.cn11-1097/s.2018.10.024

Uygun, T., Ozguven, M. M., and Yanar, D. (2020). A new approach to monitor and assess the damage caused by two-spotted spider mite. *Exp. Appl. Acarol* 82 (3), 335–346. doi: 10.1007/s10493-020-00561-8

Van Pottelberge, S., Khajehali, J., Van Leeuwen, T., and Tirry, L. (2009). Effects of spirodiclofen on reproduction in a susceptible and resistant strain of tetranychus urticae (Acari: Tetranychidae). *Exp. Appl. Acarol* 47 (4), 301–309. doi: 10.1007/s10493-008-9226-y

Vásquez-Ordóñez, A., and Parsa, S. (2014). A geographic distribution database of mononychellus mites (Acari, tetranychidae) on cassava (*Manihot esculenta*). *Zookeys* 407, 1–8. doi: 10.3897/zookeys.407.7564

Villegas-Fernández, A. M., Sillero, J. C., Emeran, A. A., Winkler, J., Raffiot, B., Tay, J., et al. (2009). Identification and multi-environment validation of resistance to *Botrytis fabae* in *Vicia faba*. *Field. Crop Res.* 114 (1), 84–90. doi: 10.1016/j.fcr.2009.07.005

Wu, X., Xu, J., Ma, Q., Ahmed, S., Lu, X., Ling, E., et al. (2022). Lysozyme inhibits postharvest physiological deterioration of cassava. *J. Integr. Plant Biol.* 64 (3), 621–624. doi: 10.1111/jipb.13219

Yang, Y., Luo, X., Wei, W., Fan, Z., Huang, T., and Pan, X. (2020). Analysis of leaf morphology, secondary metabolites and proteins related to the resistance to *Tetranychus cinnabarinus* in cassava (*Manihot esculenta* crantz). *Sci. Rep.* 10 (1), 14197. doi: 10.1038/s41598-020-70509-w

Zhu, L., Li, J., Xu, Z., Hakim,, Sijia, L., li, S., et al. (2018). Identification and selection of resistance to *Bemisia tabaci* among 550 cotton genotypes in thefield and greenhouse experiments. *Front. Agr. Sci. Eng.* 5 (2), 236–252. doi: 10.15302/J-FASE-2018223

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 16 December 2022
DOI 10.3389/fpls.2022.1071156

OPEN ACCESS

EDITED BY
Rodomiro Ortiz,
Swedish University of Agricultural Sciences, Sweden

REVIEWED BY
Just Jensen,
Aarhus University, Denmark
Maryke T. Labuschagne,
University of the Free State,
South Africa

*CORRESPONDENCE
Marcos Deon Vilela de Resende
marcos.resende@embrapa.br
Eder Jorge de Oliveira
eder.oliveira@embrapa.br

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 15 October 2022
ACCEPTED 02 December 2022
PUBLISHED 16 December 2022

CITATION
de Andrade LRB, Sousa MBe, Wolfe M, Jannink J-L, de Resende MDV, Azevedo CF and de Oliveira EJ (2022) Increasing cassava root yield: Additive-dominant genetic models for selection of parents and clones. *Front. Plant Sci.* 13:1071156. doi: 10.3389/fpls.2022.1071156

Increasing cassava root yield: Additive-dominant genetic models for selection of parents and clones

Luciano Rogério Braatz de Andrade[1], Massaine Bandeira e Sousa[2], Marnin Wolfe[3], Jean-Luc Jannink[4,5], Marcos Deon Vilela de Resende[6,7,8]*, Camila Ferreira Azevedo[8] and Eder Jorge de Oliveira[2]*

[1]Department of Crop Science, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil, [2]Embrapa Mandioca e Fruticultura, Cruz das Almas, Bahia, Brazil, [3]Department of Crop, Soil and Environment Sciences, Auburn University, Auburn, AL, United States, [4]Section on Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY, United States, [5]United States Department of Agriculture – Agriculture Research Service, Plant, Soil and Nutrition Research, Ithaca, NY, United States, [6]Department of Forestry Engineering, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil, [7]Embrapa Florestas, Colombo, Paraná, Brazil, [8]Department of Statistics, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil

Genomic selection has been promising in situations where phenotypic assessments are expensive, laborious, and/or inefficient. This work evaluated the efficiency of genomic prediction methods combined with genetic models in clone and parent selection with the goal of increasing fresh root yield, dry root yield, as well as dry matter content in cassava roots. The bias and predictive ability of the combinations of prediction methods Genomic Best Linear Unbiased Prediction (G-BLUP), Bayes B, Bayes Cπ, and Reproducing Kernel Hilbert Spaces with additive and additive-dominant genetic models were estimated. Fresh and dry root yield exhibited predominantly dominant heritability, while dry matter content exhibited predominantly additive heritability. The combination of prediction methods and genetic models did not show significant differences in the predictive ability for dry matter content. On the other hand, the prediction methods with additive-dominant genetic models had significantly higher predictive ability than the additive genetic models for fresh and dry root yield, allowing higher genetic gains in clone selection. However, higher predictive ability for genotypic values did not result in differences in breeding value predictions between additive and additive-dominant genetic models. G-BLUP with the classical additive-dominant genetic model had the best predictive ability and bias estimates for fresh and dry root yield. For dry matter content, the highest predictive ability was obtained by G-BLUP with the additive genetic model. Dry matter content exhibited the highest heritability, predictive ability, and bias estimates compared with other traits. The prediction methods showed similar selection gains with approximately 67% of the phenotypic selection gain. By shortening the breeding cycle time by 40%, genomic selection may overcome phenotypic selection by 10%, 13%, and 18% for fresh root yield, dry root yield, and dry matter content, respectively, with a selection proportion of 15%. The most suitable genetic model for each trait allows for genomic selection optimization in cassava with high selection gains, thereby accelerating the release of new varieties.

KEYWORDS

genomic selection, non-additive effects, dominance, breeding, breeding values, genotypic values

1 Introduction

Cassava (*Manihot esculenta* Crantz) has great social and economic importance for Brazilian agriculture, where nearly 18.2 million tons were produced across 1.2 million hectares in 2020 (FAO, 2022). Most of the planted area is within small farms where the product is destined for on-farm consumption or local sales. However, with the starch price rising, there is a trend of increasing industry involvement in intensive cassava production. Although almost the entire plant can be used for human and animal consumption, farmers have chiefly focused on root production.

Cassava can be propagated by seeds or vegetatively by stem pieces (cuttings), with the former generally limited to breeding programs for allele recombination and generation of new hybrid combinations and the latter the most common method used by farmers for multiplication and root production (Ceballos et al., 2012). Once the F_1 population is obtained, the hybrids are evaluated and selected regularly through several stages. Selection intensity and the evaluated traits depend on the amount of propagation material and the evaluation potential in different environments. According to Barandica et al. (2016), until the 21st century, hybrid selection in early-phase breeding programs was performed visually without extensive phenotypic data collection. Therefore, until relatively recently, inheritance knowledge about relevant traits was very limited (Calle et al., 2005; Zacarias and Labuschagne, 2010; Ceballos et al., 2012; Tumuhimbise et al., 2014; Oliveira et al., 2015a).

In several phases of the breeding program, vegetative propagation allows the maintenance of high heterozygosity and phenotypic plasticity expression for several traits (Oliveira et al., 2015a). In addition, it allows hybrids to be evaluated and selected in different locations and crop seasons (Barandica et al., 2016), thus allowing the separation of genetic and environmental effects, through the effects of the genotype by environment interactions (Ceballos et al., 2016a; Bakare et al., 2022). Due to vegetative propagation and the high heterozygosity of the parents (Ceballos et al., 2016a), genetic variability within families represents approximately 90% of total genetic variability (Ceballos et al., 2016b), supporting the idea that elite clones can be obtained within any family.

One hypothesis that may explain this high intra-family variability is the presence of non-additive genetic effects, especially for yield traits (Calle et al., 2005; Jaramillo et al., 2005; Zacarias and Labuschagne, 2010; Parkes et al., 2013; Tumuhimbise et al., 2014). While the non-additive effects hamper clone and parent selection, they allow for exploration of heterosis, as the best hybrids can be multiplied by vegetative propagation and then be release as new varieties (Parkes et al., 2013). However, the low correlation between root yield performance in the initial and final stages of the breeding program prevents the early and accurate selection of the best hybrids in clonal evaluation trials (Barandica et al., 2016). As a result, large seedling populations are evaluated annually and selected for the next stages (Ceballos et al., 2012), with the goal of identifying the most promising genotypes in advanced phases of the breeding program. This greatly increases the costs of the variety development pipeline, as phenotypic measurements demand suitable infrastructure, skilled labor, and consequently large amounts of financial resources.

Progress in genotyping, especially in reducing costs and increasing marker density, is revolutionizing marker applications in plant breeding (Fergunson et al., 2012). Since Meuwissen et al. (2001), there have been high expectations of genomic selection implementation in multiple breeding programs, due to possible selection gain in situations where traditional evaluation methods are expensive, laborious, and/or inefficient (Crossa et al., 2013). In genomic selection, breeding populations are phenotyped and genotyped with high genomic coverage markers in order to allow prediction methods to predict genomic estimated breeding values (GEBVs) of each clone (Fergunson et al., 2012). According to Crossa et al. (2013), genomic selection can predict clones' breeding values to accelerate recombination and their genotypic values as a means of targeting clones for advancement in the breeding pipeline.

For cassava, there is an expectation of genomic selection use for early selection in seedling trials as an alternative method to select traits that are difficult to measure or that demand high experimental accuracy (Oliveira et al., 2012), such as fresh root yield (FRY) and starch yield. In general, yield traits have predominantly non-additive effects (Jaramillo et al., 2005; Zacarias and Labuschagne, 2010; Parkes et al., 2013;

Tumuhimbise et al., 2014) and low correlation of the phenotypic values obtained at initial phases (seedling and clonal evaluation trials) with those of advanced trials (uniform yield trials) (Barandica et al., 2016). Another trait of great importance in cassava is the dry matter content (DMC) in roots; its genetic heritability has predominantly been associated with additive effects (Jaramillo et al., 2005; Parkes et al., 2013; Tumuhimbise et al., 2014; Wolfe et al., 2016a), and high correlation between the different breeding program stages (Barandica et al., 2016). As a result, clone and parent selection in the seedling trials is less accurate for yield traits than for DMC. However, early selection for DMC may also increase breeding efficiency, even though phenotyping in seedling trials is time-consuming and laborious. This is because seedling trials involve the evaluation of thousands of clones, and there is limited root production per clone, which prevents the use of a simple method of evaluation (specific gravimetry).

When only the additive effects are considered in the parent selection, the progeny mean is equal to the mean of the parents' breeding values; however, dominant effects prediction allows for heterosis exploration through parent complementarity (Almeida Filho et al., 2016). The genomic prediction of non-additive effects incorporated into genetic models increases the accuracy in parent and clone selection for low inheritance traits, as was observed in interspecific hybrid selection in *Eucalyptus* (Tan et al., 2018), intraspecific hybrids of *Pinus taeda* (Almeida Filho et al., 2016), maize (Lyra et al., 2019), inbred lines and crossbreed selection in Landrace and Yorkshire pigs (Esfandyari et al., 2016), and in clone selection of cassava (Wolfe et al., 2016a).

Genomic selection was also efficiently applied for predicting resistance to cassava mosaic disease, which displays a predominantly additive inheritance (Parkes et al., 2013; Tumuhimbise et al., 2014). In two years (annual breeding cycle), the allelic frequency of the marker with the greatest effect on cassava mosaic disease resistance rapidly increased from 44% to 66% (Wolfe et al., 2016b), much faster than the five or six years required in a conventional breeding cycle. Oliveira et al. (2012) noted that the two-year breeding cycle may have resulted in genetic gains higher than the conventional breeding cycle, of 56.9% and 39.92% for FRY and DMC, respectively.

Other important genomic selection goals are breeding population size reduction, time required to develop a new variety, and the ability to grow breeding populations outside the variety's recommended location, allowing selection for biotic and abiotic disturbances outside the endemic region (Fergunson et al., 2012). New prediction methodologies are consistently being published (Meuwissen et al., 2001; Park and Casella, 2008; Habier et al., 2011; Legarra et al., 2011; Azevedo et al., 2015; Wolfe et al., 2021). Application of the appropriate methodology to a trait of interest may increase selection gains and simultaneously reduce the work required in phenotypic evaluations, which are mostly high in cost and low in yield (Fergunson et al., 2012). Wolfe et al. (2016a) have noted that a non-additive genomic relationship matrix may contribute to increased efficiency and yield in clone selection for traits with low heritability and/or that are difficult to measure.

Several studies have explored the efficiency of additive models of genomic selection. However, few have addressed the efficiency of dominant effects incorporated in genetic models for cassava breeding. Therefore, the objective of this work was to infer the efficiency of the G-BLUP, Bayes B, and RKHS genomic prediction methods with different genetic models for clone and parent selection to increase FRY, dry root yield (DRY) and DMC. Breeding program stages and genomic selection that may increase the efficiency of cassava breeding programs are also discussed.

2 Material and methods

2.1 Training population

The training population included 888 accessions belonging to the Cassava Germplasm Bank of Embrapa Cassava and Fruits (Cruz das Almas, Bahia, Brazil). This germplasm comprised 835 landraces and 53 improved varieties. One hundred and eighty accessions were characterized as sweet cassava (< 50 ppm of cyanogenic compounds), 136 as containing intermediary cyanide content (50–100 ppm cyanogenic compounds), 560 as bitter cassava (> 100 ppm cyanogenic compounds), and 12 as unclassified. These accessions were collected from all 26 Brazilian states, with every state represented by at least one genotype. The genotypes were evaluated in the cities of Cruz das Almas and Laje in the state of Bahia, Brazil, in 21 trials over a six-year period (2011 to 2016).

2.2 Phenotypic data collection

For most experiments, 15–20 cm stem cuttings were planted in double lines during the rainy season in the region (May–July). The experimental plot consisted of two rows of eight plants per row. The rows were 0.9 m apart, while plants in the same row were 0.8 m apart, with 11.52 m^2 per plot. All recommended cassava cultural practices were employed (as in Souza et al., 2006). Trials were harvested 11–12 months after planting. The traits measured to estimate genomic selection efficiency were: 1) fresh root yield (FRY) at plot level (16 plants) and then adjusted to $t.ha^{-1}$, 2) dry matter content in the roots (DMC), according to Kawano et al. (1987), where approximately 5 kg of roots were weighed in a hanging scale (WA) and then, the same sample was weighed with the roots submerged in water (WW). DMC was estimated utilizing the following formula: $DMC(\%) = (\frac{WA}{WA-WW} x158.3) - 142$ and 3) dry root yield (DRY) in to $t.ha^{-1}$, estimated per plot by multiplying the FRY and DMC.

A joint analysis of 21 trials with complete randomized block design or augmented block design were used to obtain the phenotypic data. Three replicates were used in the complete randomized block design, while in the augmented block design, 10–16 replicates of the common checks were used, with equal distribution of accession number per block. Improved clones (9602-02, 9607-07, 9824-09, 9655-02) and improved varieties (BRS Dourada, BRS Gema de Ovo, and BRS Novo Horizonte) were used as checks in different field trials. More details from the phenotypic dataset could be seen in Table S1 and S2.

Due to unbalanced trials, we obtained the BLUP and deregressed BLUP (Garrick et al., 2009) for each clone. The BLUPs were obtained by the following mixed linear model: y_{ijl}=μ+c_i+β_j+$r_{l(j)}$+ϵ_{ijl} in which y_{ijl} is the vector of phenotypic observations; c_i is the clone random effect with $c_i{}^{\sim}N(0, I\hat{\sigma}_c^2)\beta_j$ is the combination of location and year, assumed as fixed effect; $r_{l(j)}$ s the replication nested within location and year, assumed as random effect with $r_{j(l)}{}^{\sim}N(0, I\hat{\sigma}_r^2)$and ϵijl is the residual with $\epsilon_{ijl}{}^{\sim}N(0, I\hat{\sigma}_e^2)$The deregressed BLUPs were estimated by: *dereg ressed* $BLUP = \frac{BLUP}{1-\frac{PEV}{\hat{\sigma}_c^2}}$Garrick et al., 2009), where the PEV is the prediction error variance of each clone and $\hat{\sigma}_c^2$s the clonal variance component. The package lme4 (Bates et al., 2015) in R software version 3.5.2 (R Core Development Team, 2018) was used to obtain the BLUPs and deregressed BLUPs for each clone.

2.3 Genotyping and SNP quality control

DNA was extracted from cassava leaves following the CTAB (cetyltrimethylammonium bromide) protocol described by Doyle and Doyle (1987). To evaluate DNA integrity and standardize its concentration, 1.0% (w/v) agarose gels were stained with ethidium bromide (1.0 mg L^{-1}) for visual comparison of a series of DNA phage Lambda (Invitrogen) concentrations. The DNA samples were sent to the Genomic Diversity Facility at Cornell University (http://www.biotech.cornell.edu/brc/genomic-diversity-facility) for genotyping-by-sequencing (GBS) (Hamblin and Rabbi, 2014). Genotypic data were selected using a minimum call rate of 0.90 and the missing markers were imputed by Beagle 4.1 software (Browning and Browning, 2016). Finally, SNPs with minor allele frequency (MAF) > 0.05 were retained. After applying marker quality control, 48,655 SNPs were selected for genomic prediction.

2.4 Genomic selection methods and genetic models

The genomic best linear unbiased prediction (G-BLUP), Reproducing Kernel Hilbert Spaces (RKHS), and Bayes B prediction methods were evaluated, considering the additive (A) and additive-dominant (A+D) genetic models, except RKHS, which predicts genetic effects based on non-parametric—and thus neither additive nor dominance—covariances. The additive-dominant genetic model of G-BLUP is expressed: y_d=$J\mu$+Za+Hd+ϵ where yd is the deregressed BLUP vector; μ is the general mean; a is the additive effect vector, random $a{}^{\sim}N(0, G\hat{\sigma}_a^2)$d is the dominant deviation effect vector, random $d{}^{\sim}N(0, D\hat{\sigma}_d^2)$ϵ is the residual effect vector, $\epsilon{}^{\sim}N(0, I\hat{\sigma}_e^2)$J, Z and H are the incidence matrices for μ, a and d, respectively, as *COV* (*a*,*d*)=0 The additive relationship matrix G was: $G = \frac{ZZ'}{2\sum\ p_i(1-p_i)}$in which *Z* is the marker matrix (-1, 0 and 1) and p_i is the major allele frequency of *i* marker. Two additive-dominant genetic models were tested for the G-BLUP method, the Classical (Vitezica et al., 2013) and the Genotypic (Su et al., 2012), differing in the parameterization of the genomic relationship matrix due to dominance. The Classical dominant relationship matrix was parameterized by the following Vitezica et al. (2013):

$$H = \begin{cases} if\ \ MM : -q^2 \\ if\ \ Mm : 2pq \\ if\ \ mm : -p^2 \end{cases}$$

$$D = \frac{HH'}{2\sum\ p_iq_i(1-p_iq_i)}.$$

The Genotypic dominant relationship matrix was estimated by the following equation (Su et al., 2012):

$$H^* = \begin{cases} if\ \ MM : -2pq \\ if\ \ Mm : p^2+q^2, \\ if\ \ mm : -2pq \end{cases}$$

$$D^* = \frac{H^*H^{*\prime}}{2\sum\ p_iq_i(1-p_iq_i)},$$

For the Bayes B method, the complete conditional prior distribution was used: $y_{d_i}|a_j, d_j, Z_{i\times j}, H_{i\times j}{}^{\sim}N(\mu+\sum_i Z_{i\times j}a_j+\sum_i H_{i\times j}d_j, \hat{\sigma}_e^2)$in which yd is the deregressed BLUP vector; μ is the general mean; a_j nd d_j re the additive and dominant marker effects, both random $a_j|\hat{\sigma}_{a_j}^2{}^{\sim}N(0, I\hat{\sigma}_{a_j}^2)$ $d_j|\hat{\sigma}_{d_j}^2{}^{\sim}N(0, I\hat{\sigma}_{d_j}^2)$ and *COV*(a_i,d_i)=0 Z and H are the incidence matrix of a_j and d_j respectively.

The model of the RKHS method was: y_d=$J\mu$+Xg+ϵ where yd is the deregressed BLUP vector; μ is the general mean; g is the genotypic effect vector, random $g{}^{\sim}N(0, K\hat{\sigma}_g^2)$ϵ is the residual effect vector, $\epsilon{}^{\sim}N(0, I\hat{\sigma}_e^2)$J and X are the incidence matrix of μ and g, respectively. K is a gaussian matrix estimated by: $K = exp(\frac{-hD}{median(D)})$h is the reduction coefficient to K values (in this work h was equal to 1), and D is the Euclidian distance of *Z* codified marker matrix (Gianola et al., 2006; Crossa et al., 2010).

The 5-fold cross-validation with three repetitions was performed to estimate the following parameters: 1) predictive ability ($\hat{r}_{\hat{y}y} = C\hat{O}R(Pr\hat{e}d_{Val}, BLUP_{Val})$) in which $Pr\hat{e}d_{Val}$ are the genomic estimated breeding values (GEBVs) for additive genetic models, or genomic estimated genotypic values (GEGVs) for additive-dominant and RKHS models, and $BLUP_{Val}$ are the BLUPs from the validation population; 2) bias ($\hat{b} = \hat{COV}(\hat{Pred}_{Train}, BLUP_{Train})/\hat{\sigma}^2_{Pred_{Train}}$)in which $\hat{Pred}_{Train}$are the genomic estimated breeding values (GEBVs) for additive genetic models, or genomic estimated genotypic values (GEGVs) for additive-dominant genetic models, of the training population, $BLUP_{Train}$ are the BLUPs from the training population, $\hat{\sigma}^2_{Pred_{Train}}$s the variance of the GEBVs for additive genetic models, or genomic estimated genotypic values (GEGVs) for additive-dominant genetic models of the training population; 3) broad-sense genomic heritability ($\hat{H}^2 = \hat{\sigma}^2_g/(\hat{\sigma}^2_g + \hat{\sigma}^2_e)$)in which $\hat{\sigma}^2_g$s the genomic variance, $\hat{\sigma}^2_e$s the residual variance; 4) narrow-sense genomic heritability ($\hat{h}^2 = \hat{\sigma}^2_a/(\hat{\sigma}^2_g + \hat{\sigma}^2_e)$)which $\hat{\sigma}^2_a$s the additive genomic variance, $\hat{\sigma}^2_g$s the genomic variance, $\hat{\sigma}^2_e$s the residual variance. For each replicate of the cross-validation process, the population was split into five equal folds. Five genomic predictions were performed per fold used as test set (no phenotypes) each fold was predicted by the remaining four-folds training set (with phenotypes).

The *sommer* R package (Covarrubias-Pazaran, 2016) was used to fit the G-BLUP and RKHS models, while the *BGLR* R package (Perez and De Los Campos, 2014) was used to fit the Bayes B model. All methods were performed using R software version 3.5.2 (R Core Development, 2018). For Bayes B method, we ran 20,000 Markov Chain Monte Carlo (MCMC) iterations with the burn-in of the initial 4,000 iterations and thinning of 10, we applied different priori for π for each trait and genetic model, these values were previously estimated by Bayes Cπ (Table S3).

The training-validation partitions of the population used in cross-validation were set up to be identical across prediction models, using the set.seed() function of R software version 3.5.2 (R Core Development, 2018). The residual variances of Markov Chain Monte Carlo (MCMC) of the Bayes B method were used to evaluated the MCMC convergency by the Raftery and Lewis's convergence diagnostic (Raftery and Lewis, 1992) applied in coda R package (Plummer et al., 2006).

2.5 Analysis of variance and Tukey's multiple comparison test

Analysis of variance was performed to estimate the effects of the genomic selection methods for predictive ability and bias estimates for DMC, FRY, and DRY. These analyses were performed using the *lme4* R package (Bates et al., 2015).

The following mixed model was used to estimate the efficiency of the genomic selection methods: $y_{ijk}=m_i+s_{jk}+e_{ijk}$ which y is the dependent variable, as predictive ability and bias; mi is the mean of the genomic selection method I, assumed as fixed effect; s_{jk} is the effect of cross validation of the replication j and fold k, assumed as random effect $s^{\sim}(0, \hat{\sigma}^2_{cv})$ and eijk is the residual effect of the i genomic selection method of the j replication and k fold, $e^{\sim}(0, \hat{\sigma}^2_e)$The genomic prediction means were submitted to the Tukey multiple comparison test implemented in the *emmeans* R package (Russel, 2018).

2.6 Cohen's Kappa coefficient

The Cohen's Kappa coefficient (Cohen, 1960) was used to analyze the coincidence of clone selection by the different genomic selection methods, considering a selection proportion (SP) amplitude ranging from 5–30%. The coincidence selection was performed using a binary code and the selected and unselected individuals received code "1" and "0", respectively. The Kappa coefficient and coincidences selection index were calculated using R.

3 Results

3.1 Efficiency of the genomic selection methods and genetic models

In general, the inclusion of the dominant genetic effects increased the genomic variance explained by the markers (Table 1), and reduced the genomic additive variance and residuals (Table 1 and Figure S1). Smaller changes in the broad-sense genomic heritability were observed for DMC, except for the Bayes B method, which demonstrated the highest broad-sense genomic heritability among the prediction methods with an additive-dominant genetic model.

Insert Table 1

A predominance of additive effects for DMC was identified with the G-BLUP method (Table 1 and Figure S1), while for FRY and DRY the dominant effects prevail. The Bayes B method showed the highest estimates of broad-sense genomic heritability and genomic variance components. However, the variation of the broad-sense genomic heritabilities between traits was smaller, suggesting a relatively large proportion of dominance variance. Even with the highest broad-sense genomic heritability, the Bayes B A+D method exhibited smaller narrow-sense genomic heritability than the G-BLUP A +D method, regardless of the dominant relationship matrix used (Table 1). However, all the additive-dominant genetic models overestimated the broad-sense genomic heritability because it was higher than the phenotypic heritability (0.337, 0.351, and 0.545 for FRY, DRY, and DMC, respectively).

The additive-dominant genetic models showed higher predictive ability than additive models and RKHS method for yield traits (FRY and DRY, Figure 1). The highest predictive ability was demonstrated by the G-BLUP A+D classical method (average of 0.484 for FRY and 0.492 for DRY), followed by Bayes B A+D (average of 0.479 for FRY and 0.488 for DRY). In addition, the predictions of dominant effects in genetic models for yield traits reduced the bias estimate, with the smaller bias at

TABLE 1 Means of the genetic parameters estimated by different genomic prediction methods for fresh root yield (FRY), dry root yield (DRY), and dry matter content (DMC) in roots of cassava.

Traits / Prediction methods	Genetic parameters					
Fresh root yield	$\hat{h}^2$	$\hat{H}^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_g^2$	$\hat{\sigma}_d^2$	$\hat{\sigma}_e^2$
G-BLUP A[1]	0.347	–	17.0	–	17.0	32.0
G-BLUP A+D Classical[2]	0.139	0.386	6.4	11.5	17.9	28.5
G-BLUP A+D Genotypic[3]	0.053	0.400	2.6	16.8	19.4	29.1
RKHS	–	0.520	–	–	31.6	29.0
Bayes B A[4]	0.582	–	43.9	–	43.9	31.2
Bayes B A+D[5]	0.257	0.734	26.2	49.2	75.4	26.9
Dry root yield	$\hat{h}^2$	$\hat{H}^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_g^2$	$\hat{\sigma}_d^2$	$\hat{\sigma}_e^2$
G-BLUP A[1]	0.332	–	1.39	–	1.39	2.81
G-BLUP A+D Classsical[2]	0.175	0.369	0.71	0.79	1.49	2.55
G-BLUP A+D Genotypic[3]	0.096	0.381	0.40	1.20	1.60	2.59
RKHS	–	0.504	–	–	2.61	2.57
Bayes B A[4]	0.571	–	3.69	–	3.69	2.74
Bayes B A+D[5]	0.262	0.728	2.32	4.15	6.46	2.39
Dry matter content	$\hat{h}^2$	$\hat{H}^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_g^2$	$\hat{\sigma}_d^2$	$\hat{\sigma}_e^2$
G-BLUP A[1]	0.517	–	2.10	–	2.10	
G-BLUP A+D Classical[2]	0.477	0.522	1.92	0.18	2.10	1.92
G-BLUP A+D Genotypic[3]	0.457	0.525	1.86	0.27	2.13	1.92
RKHS	–	0.504	–	–	2.61	2.57
Bayes B A[4]	0.673	–	4.04	–	4.04	1.95
Bayes B A+D[5]	0.325	0.792	2.73	3.99	6.72	1.75

$\hat{h}^2$, narrow-sense genomic heritability; $\hat{H}^2$, broad-sense genomic heritability; $\hat{\sigma}_a^2$, additive genomic variance; $\hat{\sigma}_d^2$, dominant genomic variance; $\hat{\sigma}_g^2$, genomic variance; $\hat{\sigma}_e^2$, residual variance. [1]G-BLUP with additive model; [2]G-BLUP with additive-dominant model, classical dominant relationship matrix (Vitezica et al., 2013); [3]G-BLUP with additive-dominant model, genotypic dominant relationship matrix (Su et al., 2012); [4]Bayes B with additive model; [5]Bayes B with additive-dominant model.

Bayes B method (Figure 1). The RKHS method showed the highest bias estimates for all traits.

3.2 Analysis of variance and Tukey's multiple comparison test of the different genomic selection methods

Significant differences between the genomic selection methods with different genetic models were identified for predictive ability and bias for all agronomic traits except the predictive ability of DMC (Table 2). Although there were no significant differences in the predictive ability between the genomic selection methods with additive-dominant models, the G-BLUP A+D classical method showed the highest predictive ability for FRY (0.483) and DRY (0.492) (Table 2). Bayes B A+D and RKHS methods did not show significant differences for predictive ability in comparison with G-BLUP A+D classical method for DRY. On the other hand, for FRY only Bayes B A+D and G-BLUP A+D genotypic methods did not show significant differences with the G-BLUP A+D classical method.

Among the methods with non-additive effects, the G-BLUP A+D classical was significantly different from the RKHS method for DRY but not for FRY. As the RKHS method can predict additive and partial epistatic effects (Gianola et al., 2006; Crossa et al., 2010), it is possible that the epistatic effects were more important for FRY than DRY, as the RKHS method did not show a significant difference with the additive genetic models G-BLUP A and Bayes B A (Table 2).

DMC showed the highest phenotypic heritability and predictive ability of traits. However, there was no improvement

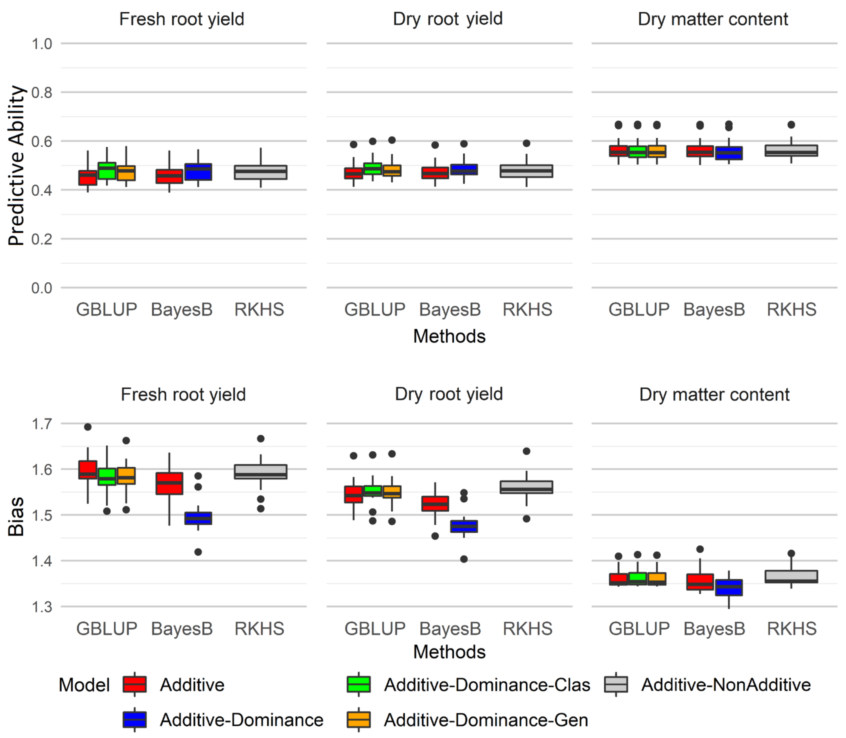


FIGURE 1
SP: selection proportion; SD GS: genomic selection differential; SD PS: phenotypic selection differential; GB/PB: ratio between the breeding cycle assisted by genomic selection and conventional breeding cycle; Efficiency = *SD GS/[SD PS×(GB/PB)]* Boxplots of predictive ability and bias for different genomic selection methods (G-BLUP, Bayes B, and RKHS) with additive and additive-dominant genetic models for fresh root yield (FRY), dry root yield (DRY), and dry matter content (DMC). GBLUP, genomic best linear unbiased prediction; RKHS, reproducing kernel Hilbert spaces.

in predictive ability when the additive-dominant genetic models were used to predict this trait, which reinforced the theory that DMC in cassava has a high influence from additive effects. On the other hand, for FRY and DRY, the additive-dominant models demonstrated increased predictive ability, suggesting a greater importance of dominant effects for these traits in cassava.

TABLE 2 Analysis of variance (ANOVA) and Tukey's multiple comparison test ($p \leq 0.05$) for prediction parameters of different genomic selection methods for fresh root yield (FRY), dry root yield (DRY), and dry matter content (DMC) in cassava.

ANOVA	DF	Fresh root yield			Dry root yield			Dry matter content	
		$\hat{r}_{\hat{y}y}$	$\hat{b}$		$\hat{r}_{\hat{y}y}$	$\hat{b}$		$\hat{r}_{\hat{y}y}$	$\hat{b}$
Methods	5	21.91*	51.76*		10.95*	50.35*		2.07	9.28*
Tukey multiple comparison test									
Bayes B A		0.458C	1.568B		0.474C	1.522B		0.566A	1.357B
Bayes B A+D		0.479AB	1.497A		0.488AB	1.477A		0.561A	1.340A
G-BLUP A		0.457C	1.598C		0.474C	1.547C		0.567A	1.360B
G-BLUP A+D Classical		0.483A	1.580BC		0.492A	1.552C		0.564A	1.362B
G-BLUP A+D Genotypic		0.474B	1.582BC		0.485AB	1.550C		0.565A	1.361B
RKHS		0.476AB	1.590C		0.482BC	1.560C		0.567A	1.366B

$\hat{r}_{\hat{y}y}$, predictive ability; $\hat{b}$, bias, DF, degrees of freedom. *significant by chi-square test ($p \leq 0.05$). Upper case letters means significant difference between genomic selection ethods for the Tukey multiple comparison test ($p \leq 0.05$) : predictive ability; : bias, DF: degrees of freedom. *significant by chi-square test (p≤0.05). Upper case letters means significant difference between genomic selection methods for the Tukey multiple comparison test (p≤0.05).

3.3 Expected genetic gains from different genomic prediction methods through different selection proportion

Although significant differences were detected between genomic prediction methods with different genetic models by ANOVA and Tukey's mean test (Table 2), the expected genetic gains for genomic prediction were still smaller than those obtained by phenotypic selection, with expected selection gains equivalent to 67.5%, 67.1%, and 69.4% of the phenotypic selection for FRY, DRY, and DMC, respectively (Figure 2). Although selection gains with genomic predictions were similar for all traits, the non-additive genetic models, such as Bayes B A+D, RKHS, and G-BLUP A+D classical and genotypic, increased the gain by an average of 0.69 t/ha for FRY and 0.24 t/ha for DRY in comparison with the additive genetic models. For DMC, the differences between the selection gains of genomic prediction methods were lower (average of 0.04%), because there was no significant difference between the clone prediction methods for this trait (Table 2). Moreover, the selection differential for DMC in the roots was lower than for yield traits due to the smaller trait amplitude (17–38%).

There was a great uniformity in the differences between the selection gains of the phenotypic BLUP and the predicted gains in the different selection proportions, with a mean difference of selection gain of 6.18% and 7.79% of the Bayes B A+D model for FRY and DRY, respectively (Figure 2 and Table S2). For DMC, there were lower gains differences between the phenotypic selection and genomic prediction, with the largest difference observed in the G-BLUP A method (average of 1.40% of genetic gain) compared to others (Figure 2 and Table S4).

The genomic expected selection gain and its relative efficiency to phenotypic expected selection gain were calculated. According to Oliveira et al. (2012), the conventional breeding cycle of cassava is at least four years due to the need to include phenotypic information from a minimum of four breeding phases (clonal evaluation trial,

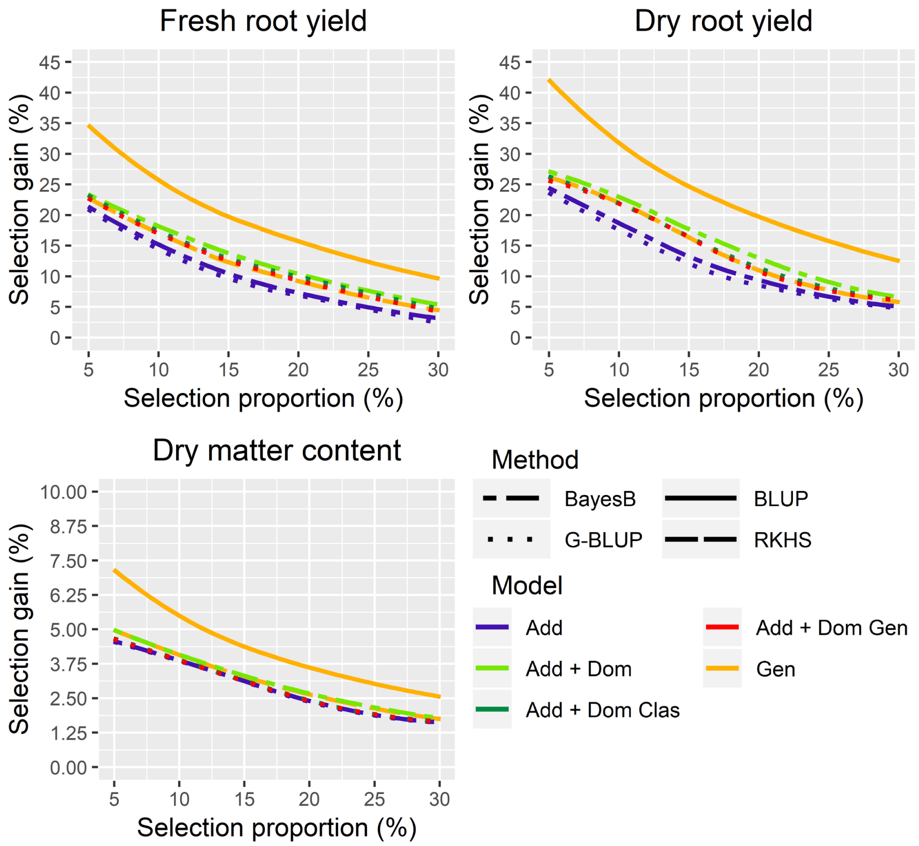


FIGURE 2
Expected selection gains for combinations of different genomic prediction methods and genetic models for fresh (FRY) and dry root yield (DRY) and dry matter content (DMC) in the roots of cassava, considering a selection proportion ranging from 5 to 30%. G-BLUP: genomic best linear unbiased prediction method; BLUP: phenotypic best linear unbiased prediction method; RKHS: reproducing kernel Hilbert spaces method; Add: additive; Add + Dom, Additive and dominant genetic model; Add + Dom Clas, Additive and dominant classical genetic model; Add + Dom Gen, Additive and dominant genotypic genetic model; Gen, genotypic model.

preliminary yield trial, advanced yield trial, and uniform yield trial). The efficiency and the selection gains per time unit to simulate early selection assisted by genomic selection were calculated. The efficiency was determined by comparing time required to recombine the selected clones as parents in a conventional breeding program vs. one assisted by genomic selection.

Genomic selection based on the G-BLUP A+D classical method for FRY and G-BLUP A for DMC was more efficient than phenotypic selection when the breeding cycle was ≤ 0.60 of the conventional breeding cycle (Table 3 and Figure S2). However, for DRY the genomic selection was more efficient than phenotypic selection only with a selection proportion of 5–15%.

Reducing breeding cycle time by 60% using genomic selection could result in gains of 65%, 69%, and 77% over those provided by phenotypic selection for FRY, DRY, and DMC, respectively, in a selection proportion of approximately 15% of the best clones (Table 3). If the breeding cycle was reduced to 20% of the conventional breeding cycle (four years to ten months), the genetic gains would be 163%, 170%, and 183% over those provided by phenotypic selection for FRY, DRY, and DMC, respectively.

The selection proportion affected significantly the relative efficient of the genomic prediction only in breeding cycle time reductions biggest then 35% of the conventional Cassava breeding cycle (Table 3).

4 Discussion

4.1 Phenotypic and genomic heritability and its implications for genomic selection

According to Oliveira et al. (2015b), heritability estimates can assist selection strategies in increasing genetic gain, as well as defining the breeding method and experimental design. Given the broad- and narrow-sense genomic heritability, the G-BLUP A+D classical method showed that cassava yield traits demonstrate a predominance of dominant effects. In addition, the broad-sense genomic heritability of G-BLUP A+D was closer to the phenotypic heritability values (0.337 for FRY, 0.351 for DRY, and 0.545 for DMC [Table 1]). Stability of FRY and DRY are important agronomic attributes for any cassava variety to ensure high market competitiveness in the starch industry, especially as there is a minimum acceptable DMC threshold for processing the raw material. Roots with DMC index below this threshold are not processed by the starch industry due to the high industrial cost and low starch yield.

Knowledge about trait heritability and variation gained during field evaluation in different environments may assist in optimizing selection of cassava breeding programs, with the goal of developing new cassava varieties with higher starch yield stability. Optimizing the selection proportion and evaluated traits in each breeding phase can maximize the probability of selecting the best clone. This is because low heritability traits such as FRY and starch yield are generally evaluated in the final breeding phases due to greater stem cutting availability (more plants per plot across multiple locations).

TABLE 3 Relative efficiency of genomic selection compared to phenotypic selection using different selection proportions with the G-BLUP A+D classical method for fresh root yield (FRY), dry root yield (DRY), and dry matter content (DMC) in cassava.

	Fresh root yield						Dry root yield						Dry matter content					
Method	G-BLUP Additive Dominant Classical						G-BLUP Additive Dominant Classical						G-BLUP Additive					
SP (%)	5%	10%	15%	20%	25%	30%	5%	10%	15%	20%	25%	30%	5%	10%	15%	20%	25%	30%
SD GS	23.8	17.2	13.0	9.9	7.4	4.7	25.8	22.0	16.7	11.3	8.4	5.8	4.6	3.9	3.1	2.4	1.9	1.6
SD PS	35.1	25.4	19.8	15.7	12.4	9.6	42.8	31.7	24.7	19.7	15.8	12.5	7.3	5.5	4.4	3.6	3.0	2.5
GB/PB	Efficiency						Efficiency						Efficiency					
1.00	0.68	0.68	0.66	0.63	0.60	0.49	0.60	0.69	0.68	0.57	0.53	0.47	0.64	0.71	0.71	0.67	0.62	0.62
0.80	0.85	0.84	0.82	0.79	0.75	0.61	0.75	0.87	0.84	0.72	0.66	0.59	0.79	0.88	0.89	0.83	0.78	0.78
0.65	1.04	1.04	1.01	0.97	0.92	0.75	0.93	1.07	1.04	0.88	0.82	0.72	0.98	1.09	1.09	1.02	0.96	0.96
0.60	1.13	1.13	1.10	1.06	1.00	0.81	1.00	1.16	1.13	0.95	0.88	0.78	1.06	1.18	1.18	1.11	1.04	1.04
0.40	1.70	1.69	1.65	1.58	1.50	1.22	1.51	1.74	1.69	1.43	1.33	1.17	1.59	1.77	1.77	1.66	1.56	1.56
0.20	3.40	2.70	2.63	2.53	2.40	1.95	2.41	2.78	2.70	2.29	2.12	1.87	2.54	2.83	2.83	2.66	2.49	2.49

SP, selection proportion; SD GS, genomic selection differential; SD PS, phenotypic selection differential; GB/PB, ratio between the breeding cycle assisted by genomic selection and conventional breeding cycle; Efficiency = SD GS/[*SD PS*×(GB/PB)].

Wolfe et al. (2016b) also related the predominance of additive and dominant deviation effects for DRY and FRY, respectively. They found similar estimates of broad- and narrow-sense heritability for the first genomic selection cycle of IITA population using the G-BLUP A+D method (0.12 and 0.35 for narrow- and broad-sense heritability, respectively, for FRY, and 0.47 and 0.52 for narrow and broad-sense heritability, respectively, for DMC). Wolfe et al. (2016b) found that the appropriate genetic model for DMC was the additive-dominant, while in the present study the additive-dominant models obtained similar results to the other models'. The reduction of the variance explained by the additive component was noted previously in cassava (Wolfe et al., 2016b) and other species such as *Pinus taeda* L. (Muñoz et al., 2014) and hybrids of *Eucalyptus urophylla* and *E. grandis* (Bouvet et al., 2016). Several authors reported that during prediction using additive genetic models, part of the dominant deviation was predicted along with the additive effects; however, when using additive-dominant genetic models, this dominant deviation predicted by the additive effects is then computed by the dominant variance (Zuk et al., 2012; Vitezica et al., 2013; Muñoz et al., 2014; Wolfe et al., 2016a). According to Vitezica et al. (2013) genetic models with assumptions of additive and dominant deviation effects result in better genomic predictions.

4.2 Efficiency of cassava selection considering different genomic prediction methods and genetic models

There were significant differences in predictive ability between the methods for FRY and DRY, mainly due to different genetic models (additive and non-additive). The additive-dominant genetic models showed higher predictive ability than additive genetic models for FRY and DRY. Among the genomic selection methods, the G-BLUP A+D classical (Vitezica et al., 2013) had high predictive ability and low bias, statistically similar to other additive-dominant genetic models. Therefore, the additive-dominant genetic models allow for exploration of part of the non-additive effects by increasing cassava clone selection accuracy. Other authors have evaluated relationship matrices with classical and genotypic dominant models and verified the lack of differences in the genomic predictions of these matrices, although the broad-sense heritability has been somewhat lower in the matrix (H^*) of genotypic dominant (Vitezica et al., 2013; Wolfe et al., 2016a). However, the correlation between the additive and dominant parameters was higher in the G-BLUP genotypic method in comparison with the classical one (Wolfe et al., 2016a), which corroborates the correlations found in this work.

Dominance effects occurs due to the interaction between alleles at the same locus and its main benefits are expected in crossbreeding, since dominance has been suggested as one of the genetic mechanisms explaining heterosis (Shull, 1908). Indeed, hybrid vigor for yield components in cassava over better-parent values has been reported (Parkes, et al., 2020), indicating that heterosis should be explored in order to develop superior cassava genotypes. Therefore, genomic predictions for traits such as FRY and DRY must be based on the assumption that non-additive effects are an important component that should be considered in the predictions to optimize crossing designs, such as in mate-pair allocation (Almeida Filho et al., 2019). As a further step, the role of dominance effects on the genetic architecture of FRY and DRY should be evaluated in the breeding population generated in this study.

For DMC, there was no significant difference between genetic models for predictive ability, although the additive genetic models showed better prediction abilities than additive-dominant models. Similar results were reported in the first genomic selection cycle of IITA cassava population (Wolfe et al., 2016a). According to Denis and Bouvet (2013) the decision of which genetic model to use in genomic selection depends on the training population as well the traits under selection. Specifically, in cassava, the genomic prediction of FRY and DRY would be more efficient if applying additive-dominant genetic models, while for DMC the additive models are satisfactory. Among genetic models, additive gene action had highest response to selection. Therefore, in the case of DMC, the population improvement focused on genetic additive effects can achieve large medium-to-long term genetic gains.

Incorporating non-additive effects into the genetic model reduces the additive genomic variance and the bias of the GEBVs, as well increasing the accuracy for selecting the best parent (Vitezica et al., 2013; Wolfe et al., 2016a). On the other hand, some simulated studies did not find any differences in predictive ability of GEBVs between additive and additive-dominant genetic models (Almeida Filho et al., 2016; Heidaritabar et al., 2016). Therefore, it is expected that non-additive effects prediction may increase the genetic gains for yield traits of new cassava varieties in the breeding programs (Muñoz et al., 2014; Wolfe et al., 2016a).

The expected selection gains for FRY and DRY were high due to the training population being composed of germplasm accessions with high genetic variability (Figure S2) for several traits, including yield traits (Oliveira et al., 2015a; Oliveira et al., 2016). Within a group of clones that deviated from the FRY and DRY mean (Figure S2), some were from high-yield, improved varieties (FRY potential of > 30 t/ha) such as BRS Novo Horizonte, BRS Poti Branca, BRS Kiriris, and BRS Tapioqueira.

According to the Bayes B A+D method, if the cassava breeding cycle was reduced by 40%, the genomic selection gains would be on average 12.48% and 11.92% higher than phenotypic selection gains for FRY and DRY, respectively. A similar observation was made for DMC (22.16%). Oliveira et al. (2012) reported that with a 25% reduction in breeding cycle time, the relative efficiency of RR-BLUP genomic prediction was

4.6%, 15.96%, and -7.05% for FRY, starch yield, and DMC, respectively. According to these authors, higher selection gains may be achieved by reducing the time required to identify and recombine the parents in the breeding cycle. These results may assist in the planning of genomic selection implementation to increase the frequency of new cassava varieties with good agronomic traits and adaptations to new biotic and abiotic stress challenges.

Reducing the selection proportion is not feasible as reducing breeding cycle time to improve genetic gain, but it may be the next milestone to improve genetic gain in cassava breeding, by increasing the number of clones evaluated in earlier stages by genomic prediction.

4.3 Potential application of genomic selection in cassava breeding

A previously recommended method for clone and parent selection in seedling trial phases was assessment of the harvest index (the ratio of FRY and the biomass yield), used for FRY indirect selection in seedling nursery trials and clonal evaluation trials (Kawano et al., 1998) due to its high correlation (0.730). However, when analyzing the historical data (2000–2013) of the cassava breeding program at the International Center for Tropical Agriculture (CIAT), Barandica et al. (2016) found very low correlation between FRY and harvest index (0.11). In addition, the harvest index assessment is more labor-intensive than the FRY evaluation alone. According to Barandica et al. (2016) the correlation of FRY between the clonal evaluation trials and the uniform yield trials was 0.29, while in the present study the correlation between the GEBVs and the uniform yield trials for the G-BLUP A+D classical method was 0.483. In future studies, the efficiency of genomic selection in the seedling trial phase for FRY could be better understood by determining realized genetic gains.

Ceballos et al. (2016a) stated that one issue in the selection of good parents is the high intra-family genetic variability due to high heterozygosity. Thus, new cassava varieties may derive from crosses between parents with low agronomic performance. Indeed, Kawano et al. (1998) evaluated almost 327,000 clones from 4,130 crosses during 14 years of research, and among all those evaluated clones only three were officially released as new varieties.

Commonly the standard methods used for parent selection are the *per se* performance and, less commonly, general combining ability (Ceballos et al., 2004). Unfortunately, there is no linear relationship between the *per se* performance and the progeny's breeding values due to dominant deviation (Ceballos et al., 2004). In addition, the diallel analysis in cassava breeding programs is problematic because the crosses in cassava are laborious and usually imbalanced due to issues of flowering synchronization (Ceballos et al., 2017), the considerable unpredictability of the flowering season (between four to ten months after planting), and the time for seed maturity after harvest demanding at least one year to obtain the seeds of controlled crossings (Ceballos et al., 2004). Su et al. (2012) reported that genetic models with additive and non-additive effects prediction might allow for exploitation of specific combining ability. Therefore, applying genomic selection with genetic models that consider both genetic effects may be a faster alternative for selecting clones for advancement in the breeding pipeline, parents for crossings, inheritance studies, and variation of traits at the different stages of the cassava breeding program.

Another strategy for selecting promising parents is the pedigree-based best linear unbiased prediction (P-BLUP) method for predicting breeding values (Ceballos et al., 2016a). This strategy attempts to estimate breeding values after obtaining clone phenotypic data. According to Piepho et al. (2008) this method allows for dissection of the genotypic value in additive and non-additive effects, and it can be approached by identity-by-descent (P-BLUP) or identity-by-state (G-BLUP) information. However, this method requires a large amount of kinship information, which is not always available once several crosses have been carried out between germplasm accessions with no kinship data available. Nevertheless, the lack of kinship/pedigree information can be efficiently compensated for by identity-by-state (IBS) performed using an additive relationship matrix proposed by VanRaden (2007), as Hayes et al. (2009) considered the additive genomic relationship matrix as accurate as the kinship matrix. Bouvet et al. (2016) found that the G-BLUP prediction method had higher predictive ability than P-BLUP in several genetic models. According to Zhang et al. (2015) the GEBVs may be even more accurate when using a genetic architecture-enhanced relationship matrix for each trait, with the parametrization of relationship matrix composed by markers with high effect for the trait.

Using G-BLUP for breeding value estimation at preliminary, advanced and uniform yield trials, we assume that there is a genetic correlation between clones due to relationship-by-state (Piepho et al., 2008). Since cassava is vegetatively propagated, the additive genomic matrix may be used as a genetic covariance matrix for selecting promising parents by applying mixed models in the different breeding phases (such as the clonal evaluation trial, preliminary yield trial, advanced yield trial, and uniform yield trial). As the correlation between breeding values vs. genotypic values is not perfect (0.716 of selection coincidence at 13.3% selection proportion for FRY, and 0.690 of selection coincidence at 8.3% selection proportion for DRY), the coincidence in the selection of clones to be used as parents and for advancement in the breeding program tends to be low. Therefore, by using the genetic covariance matrix in mixed models, the selection of parents with high potential to generate promising clones would be performed based on their breeding values even if the clones had low genotypic value and/or low *per se* performance. This strategy can increase the parent selection accuracy, estimate the narrow-sense heritability, and predict the GEBVs and GEGVs across the field trials, assisting parent and clone selection, respectively.

5 Conclusions

The genetic variances for FRY and DRY were largely derived from dominance deviations, while DMC was predominantly additive. Identification of the best genetic model allows breeders to achieve higher genetic gains in the cassava breeding program. Genomic selection can be used to assist in breeding value prediction and the selection of outstanding parents at early breeding steps, as well as to identify and select the genotypic value of good clones for advancement in the breeding pipeline. Genomic selection may achieve higher genetic gains by reducing the breeding cycle time by at least 40%.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://figshare.com/ , doi.org/10.6084/m9.figshare.21330972.

Author contributions

Study conception and design: LA, MR, EO. Data collection: LA, MS, EO. Analysis and interpretation of results: LA, MW, J-LJ, MR, EO. Draft manuscript preparation: LA, MS, MW. Final manuscript revision: MS, MW, CA, MR, EO. All authors reviewed the results and approved the final version of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

MS: Funarbe (Fundação Arthur Bernardes). Grant number: 4986. LA: Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Grant number: 001 - 88882.157009/2017-01. EO: CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). Grant number: 442050/2019-4 and 303912/2018-9. FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia). Grant number: Pronem 15/2014. This work was partially funded by UK's Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation. Grant INV-007637. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. The funder provided support in the form of fellowship and funds for the research, but did not have any additional role in the study design, data collection and analysis, decision to publish, nor preparation of the manuscript.

Acknowledgments

Last, we would like to pay our gratitude and our respects to our colleague, Professor Fabyano Fonseca e Silva. After helping the genomic analysis and performed a deep revision of the manuscript, Professor Fabyano passed away in August, 2021.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1071156/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Variance components and standard deviation of genomic effects predicted by the genetic models of different genomic prediction method for fresh (FRY) and dry root yield (DRY), dry matter content (DMC) in cassava. Bayes B A: Bayes B method with additive genetic model; Bayes B A+D: Bayes B method with additive-dominant genetic model; G-BLUP A: G-BLUP method with additive genetic model; G-BLUP A+D Cla: G-BLUP method with additive-dominant classical genetic model; G-BLUP A+D Gen: G-BLUP method with additive-dominant genotypic genetic model; RKHS: reproducing kernel Hilbert spaces..

SUPPLEMENTARY FIGURE 2
Genetic gains by reducing the cassava breeding cycle takes into account the genomic (GB) and phenotypic breeding (PB) using the genomic prediction method and genetic models with higher predictive ability. G-BLUP A+D classical method for fresh root yield (FRY) and dry root yield (DRY) and G-BLUP A for dry matter content (DMC).

References

Almeida Filho, J. E., Guimarães, J. F. R., Silva, F. F., Resende, M. D. V., Muñoz, P., Kirst, M., et al. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity* 117, 33–41. doi: 10.1038/hdy.2016.23

Almeida Filho, J. E., Guimarães, J. F. R., Silva, F. F., Resende, M. D. V., Muñoz, P., Kirst, M., et al. (2019). Genomic prediction of additive and non-additive effects using genetic markers and pedigrees. *G3 (Bethesda).* 8; 9 (8), 2739–2748. doi: 10.1534/g3.119.201004

Azevedo, C. F., Resende, M. D. V., Silva, F. F., Viana, J. M. S., Valente, M. S. F., Resende C.OMMAJr., M. F. R., et al. (2015). Ridge, lasso and Bayesian additive-dominance genomic models. *BMC Genet.* 16 (105), 1–13. doi: 10.1186/s12863-015-0264-2

Bakare, M. A., Kayondo, S. I., Aghogho, C. I., Wolfe, M. D., Parkes, E. Y., Kulakow, P., et al. (2022). Exploring genotype by environment interaction on cassava yield and yield related traits using classical statistical methods. *PLoS One* 17 (7), e0268189. doi: 10.1371/journal.pone.0268189

Barandica, O. J., Pérez, J. C., Lenis, J. I., Calle, F., Morante, N., Pino, L., et al. (2016). Cassava breeding II: phenotypic correlations through the different stages of selection. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01649

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67 (1), 1–48. doi: 10.18637/jss.v067.i01

Bouvet, J. M., Makouanzi, G., Cros, D., and Vigneron, P. H. (2016). Modelling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. *Heredity* 116, 146–157. doi: 10.1038/hdy.2015.78

Browning, B. L., and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *Am. J. .Hum. Genet.* 98 (1), 116–126. doi: 10.1016/j.ajhg.2015.11.020

Calle, F., Perez, J. C., Gaitán, W., Morante, N., Ceballos, H., Llano, G., et al. (2005). Diallel inheritance of relevant traits in cassava (*Manihot esculenta* crantz) adapted to acid-soil savannas. *Euphytica* 144, 177–186. doi: 10.1007/s10681-005-5810-y

Ceballos, H., Iglesias, C. A., Pérez, J. C., and Dixon, A. G. (2004). Cassava breeding: opportunities and challenges. *Plant Mol. Biol.* 56 (4), 503–516. doi: 10.1007/s11103-004-5010-5

Ceballos, H., Jaramillo, J. J., Salazar, S., Pineda, L. M., Calle, F., Setter, T., et al (2017). Induction of flowering in cassava through grafting. *J. Plant Breed Crop Sci.* 9, 19–29. doi: 10.5897/JPBCS2016.0617

Ceballos, H., Kulakow, P., and Hershey, C. (2012). Cassava breeding: current status, bottlenecks and the potential of biotechnology tools. *Trop. Plant Biol.* 5, 73–87. doi: 10.1007/s12042-012-9094-9

Ceballos, H., López-Lavalle, L. A. B., Calle, F., Morante, N., Ovalle, T. M., and Hershey, C. (2016b). Genetic distance and specific combining ability in cassava. *Euphytica* 210 (1), 79–92. doi: 10.1007/s10681-016-1701-7

Ceballos, H., Pérez, J. C., Barandica, O. J., Lenis, J. I., Morante, N., Calle, F., et al. (2016a). Cassava breeding I: the value of breeding value. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01227

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educ. Psychol. Meas* 20 (1), 37–46. doi: 10.1177/001316446002000104

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the r package sommer. *PLoS One* 11 (6), 1–15. doi: 10.1371/journal.pone.0156744

Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186 (2), 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., et al. (2013). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112, 48–60. doi: 10.1038/hdy.2013.16

Denis, M., and Bouvet, J. M. (2013). Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. *Tree Genet. Genomes* 9 (1), 37–51. doi: 10.1007/s11295-012-0528-1

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bull.* 19, 11–15.

Esfandyari., H., Bijma, P., Henryon, M., Christensen, O. F., and Sørensen, A. C. (2016). Genomic prediction of crossbred performance based on purebred landrace and Yorkshire data using a dominance model. *Genet. Sel. Evol.* 48 (40), 1–9. doi: 10.1186/s12711-016-0220-2

Ferguson, M., Rabbi, I., Kim, D. J., Gedil, M., Lopez-Lavalle, L. A. B., and Okogbenin, E. (2012). Molecular markers and their application to cassava breeding: past, present and future. *Trop. Plant Biol.* 5 (1), 95–109. doi: 10.1007/s12042-011-9087-0

Food and Agriculture Organization of the United Nations (2022). *FAOSTAT statistical database* (Rome: FAO).

Garrick, D. J., Taylor, J. F., and Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* 41, 55. doi: 10.1186/1297-9686-41-55

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173 (3), 1761–1776. doi: 10.1534/genetics.105.049510

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinf.* 12 (186), 1–12. doi: 10.1186/1471-2105-12-186

Hamblin, M. T., and Rabbi, I. Y. (2014). The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: a study in cassava. *Crop Sci.* 54 (6), 2603–2608. doi: 10.2135/cropsci2014.02.0160

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91 (1), 47–60. doi: 10.1017/S0016672308009981

Heidaritabar, M., Wolc, A., Arango, J., Zeng, J., Settar, P., Fulton, J., et al. (2016). Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers. *J. Anim. Breed. Genet.* 133, 334–346. doi: 10.1111/jbg.12225

Jaramillo, G., Morante, N., Pérez, J. C., Calle, F., Ceballos, H., Arias, B., et al. (2005). Diallel analysis in cassava adapted to the midaltitude valleys environment. *Crop Sci.* 45 (3), 1058–1063. doi: 10.2135/cropsci2004.0314

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

Kawano, K., Narintaraporn, K., Narintaraporn, P., Sarakarn, S., Limsila, A., Limsila, J., et al. (1998). Yield improvement in a multistage breeding program for cassava. *Crop Sci.* 38 (2), 325–332. doi: 10.2135/cropsci1998.0011183X003800020007x

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., and Fritz, S. (2011). Improved lasso for genomic selection. *Genet. Res.* 93 (1), 77–87. doi: 10.1017/S0016672310000534

Lyra, D. H., Galli, G., Alves, F. C., Granato, I. S. C., Vidoti, M. S., Sousa, M. B., et al. (2019). Modeling copy number variation in the genomic prediction of maize hybrids. *Theor. Appl. Genet.* 132 (1), 273–288. doi: 10.1007/s00122-018-3215-2

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4), 1819–1829. doi: 10.1093/genetics/157.4.1819

Muñoz, P. R., Resende, M. F. R., Gezan, S. A., Resende, M. D. V., De Los Campos, G., Kirst, M., et al. (2014). Unraveling additive from nonadditive effects using genomic kinship matrices. *Genetics* 198 (4), 1759–1768. doi: 10.1534/genetics.114.171322

Oliveira, E. J., Aud, F. F., Morales, C. F. G., Oliveira, S. A. S., and Santos, V. S. (2016). Non-hierarchical clustering of *Manihot esculenta* crantz germplasm based on quantitative traits. *Cienc. Agron.* 47 (3), 548–555. doi: 10.5935/1806-6690.20160066

Oliveira, E. J., Filho, O. S., and Santos, V. S. (2015b). Classification of cassava genotypes based on qualitative and quantitative data. *Genet. Mol. Res.* 14 (1), 906–924. doi: 10.4238/2015

Oliveira, E. J., Resende, M. D. V., Santos, V. S., Ferreira, C. F., Oliveira, G. A. F., Silva, M. S., et al. (2012). Genome-wide selection in cassava. *Euphytica* 187 (2), 263–276. doi: 10.1007/s10681-012-0722-0

Oliveira, E. J., Santana, F. A., Oliveira, L. A., and Santos, V. S. (2015a). Genotypic variation of traits related to quality of cassava roots using affinity propagation algorithm. *Sci. Agric.* 72 (1), 53–61. doi: 10.1590/0103-9016-2014-0043

Park, T., and Casella, G. (2008). The Bayesian lasso. *J. Am. Stat. Assoc.* 103 (482), 681–686. doi: 10.1198/016214508000000337

Parkes, E., Aina, O., Kingsley, A., Iluebbey, P., Bakare, M., Agbona, A., et al. (2020). Combining ability and genetic components of yield characteristics, dry matter content, and total carotenoids in provitamin a cassava F1 cross-progeny. *Agronomy* 10, 1850. doi: 10.3390/agronomy10121850

Parkes, E. Y., Fregene, M., Dixon, A., Boakye-Peprah, B., and Labuschagne, M. T. (2013). Combining ability of cassava genotypes for cassava mosaic disease and cassava bacterial blight, yield and its related components in two ecological zones in Ghana. *Euphytica* 194 (1), 13–24. doi: 10.1007/s10681-013-0936-9

Perez, P., and De Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198 (2), 483–495. doi: 10.1534/genetics.114.164442

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161 (1), 209–228. doi: 10.1007/s10681-007-9449-8

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* 6 (1), 7–11.

Raftery, A. E., and Lewis, S. M. (1992). Practical Markov chain Monte Carlo: Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statist. Sci.* 7 (4), 493–497. doi: 10.1214/ss/1177011143

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

Russell, L. (2018) *Emmeans: estimated marginal means, aka least-squares means*. Available at: https://CRAN.R-project.org/package=emmeans.

Shull, G. H. (1908). The composition of a field of maize. *J. Hered*, 296–301. doi: 10.1093/jhered/os-4.1.296

Souza, L. S., Farias, A. R., Mattos, P. L. P., and Fukuda, W. M. G. (2006). Aspectos socioeconômicos e agronômicos da mandioca. Embrapa Mandioca e Fruticultura Tropical: Cruz das Almas (BA).

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S. (2012). Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One* 7 (9), e45293. doi: 10.1371/journal.pone.0045293

Tan, B., Grattapaglia, D., Wu, H. X., and Ingvarsson, P. K. (2018). Genomic kinships reveal significant dominance effects for growth in hybrid *Eucalyptus*. *Plant Sci.* 267, 84–93. doi: 10.1016/j.plantsci.2017.11.011

Tumuhimbise, R., Melis, R., and Shanahan, P. (2014). Diallel analysis of early storage root yield and disease resistance traits in cassava (*Manihot esculenta* crantz). *Field Crops Res.* 167, 86–93. doi: 10.1016/j.fcr.2014.07.006

Vanraden, P. M. (2007). Genomic measures of kinship and inbreeding. *Interbull Anal. Meet. Proc.* 37, 33–36. doi: 10.3168/jds.2007-0980

Vitezica, Z. G., Varona, L., and Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* 195 (4), 1223–1230. doi: 10.1534/genetics.113.155176

Wolfe, M. D., Chan, A. W., Kulakow, P., Rabbi, I., and Jannink, J. L. (2021). Genomic mating in outbred species: predicting cross usefulness with additive and total genetic covariance matrices. *Genetics* 219 (3), iyab122. doi: 10.1093/genetics/iyab122

Wolfe, M. D., Kulakow, P., Rabbi, I. Y., and Jannink, J. L. (2016a). Marker-based estimates reveal significant non-additive effects in clonally propagated cassava (*Manihot esculenta*): implications for the prediction of total genetic value and the selection of varieties. *G3* 6 (11), 3497–3506. doi: 10.1534/g3.116.033332

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., et al. (2016b). Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *Plant Genome* 9 (2), 1–13. doi: 10.3835/plantgenome2015.11.0118

Zacarias, A. M., and Labuschagne, M. T. (2010). Diallel analysis of cassava brown streak disease, yield and yield related characteristics in Mozambique. *Euphytica* 176 (3), 309–320. doi: 10.1007/s10681-010-0203-2

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, H., et al. (2015). Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. *G3* 5 (4), 615–627. doi: 10.1534/g3.114.016261

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* 109 (4), 1193–1198. doi: 10.1073/pnas.1119675109

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 December 2022
DOI 10.3389/fpls.2022.1042379

A promoter toolbox for tissue-specific expression supporting translational research in cassava (*Manihot esculenta*)

Wolfgang Zierer[1*‡], Ravi Bodampalli Anjanappa[2†‡], Christian Erwin Lamm[1], Shu-Heng Chang[3], Wilhelm Gruissem[2,3] and Uwe Sonnewald[1]

[1]Biochemistry, Department of Biology, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany, [2]Plant Biotechnology, Department of Biology, Eidgenössische Technische Hochschule (ETH) Zurich, Zurich, Switzerland, [3]Advanced Plant Biotechnology Center, National Chung Hsing University, Taichung, Taiwan

Check for updates

OPEN ACCESS

EDITED BY
Xiaofei Zhang,
International Center for Tropical Agriculture (CIAT), Colombia

REVIEWED BY
Karl Kunert,
University of Pretoria, South Africa
Kan Wang,
Iowa State University, United States

*CORRESPONDENCE
Wolfgang Zierer
wolfgang.zierer@fau.de

†PRESENT ADDRESS
Ravi Bodampalli Anjanappa,
Plant Biochemistry, Department of Biology, Eidgenössische Technische Hochschule, Zurich, Switzerland

‡These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 12 September 2022
ACCEPTED 18 November 2022
PUBLISHED 20 December 2022

CITATION
Zierer W, Anjanappa RB, Lamm CE, Chang S-H, Gruissem W and Sonnewald U (2022) A promoter toolbox for tissue-specific expression supporting translational research in cassava (*Manihot esculenta*). *Front. Plant Sci.* 13:1042379. doi: 10.3389/fpls.2022.1042379

There is an urgent need to stimulate agricultural output in many tropical and subtropical countries of the world to combat hunger and malnutrition. The starchy crop cassava (*Manihot esculenta*), growing even under sub-optimal conditions, is a key staple food in these regions, providing millions of people with food. Cassava biotechnology is an important technique benefiting agricultural progress, but successful implementation of many biotechnological concepts depends on the availability of the right spatiotemporal expression tools. Yet, well-characterized cassava promoters are scarce in the public domain. In this study, we investigate the promoter activity and tissue specificity of 24 different promoter elements in stably transformed cassava plants. We show that many of the investigated promoters, especially from other species, have surprisingly low activity and/or tissue specificity, but feature several promoter sequences that can drive tissue-specific expression in either autotrophic-, transport- or storage tissues. We especially highlight *pAtCAB1*, *pMePsbR*, and *pSlRBCS2* as strong and specific source promoters, *pAtSUC2*, *pMeSWEET1-like*, and *pMeSUS1* as valuable tools for phloem and phloem parenchyma expression, and *pStB33*, *pMeGPT*, *pStGBSS1*, as well as *pStPatatin Class I*, as strong and specific promoters for heterotrophic storage tissues. We hope that the provided information and sequences prove valuable to the cassava community by contributing to the successful implementation of biotechnological concepts aimed at the improvement of cassava nutritional value and productivity.

KEYWORDS

cassava, biotechnology, promoter, storage root, parenchyma, phloem, xylem, tissue

Highlight

Providing expression tools for biotechnological applications by characterizing twenty-four promoter sequences in stably transformed cassava plants.

Introduction

According to the latest Food and Agricultural Organization of the United Nations report (FAO et al., 2022), it is estimated that between 702 and 828 million people were affected by hunger in 2021 worldwide. The report states that most of the world's undernourished people live in Asia (425 million people; approximately 9.1% of total population), while Africa is the region where the prevalence is the highest. 278 million people in Sub-Saharan Africa (SSA) suffer from chronic hunger. This is approximately 20% of the entire population (FAO et al., 2022). In addition, 399 million people are moderately food insecure, meaning that they don´t have regular access to sufficient food, even though they aren´t necessarily suffering from chronic hunger. The food insecurity situation has grown worse in the past years, mainly due to climate shocks, conflicts and economic slowdowns. The report concludes that it is equally important to diversify the economy and to simulate agricultural output (FAO et al., 2020).

The woody shrub cassava (*Manihot esculenta*) assumes a central role in (sub-)tropical countries, as one of the most important staple food crops. Especially in SSA, the crop is almost exclusively grown by smallholder farmers with limited resources for agricultural inputs, like industrial fertilizer. Even on poor soil, cassava can generate reasonable yields, is water efficient, and can withstand prolonged periods of drought. These factors, together with its flexible harvest time, make the crop very suitable for staple food production in a low input environment.

Half of the global annual cassava yield is produced in SSA, with Nigeria being by far the largest producer. Publically available FAO data (https://www.fao.org/faostat/en/#data/QCL; Inputs = Nigeria, Yield, Area harvested, 2000 – 2020) show that the cassava farming area in Nigeria has doubled between the years 2011 and 2012 and after, from 3 million to 6 million hectare. Around the same time, yield per area has dropped from approximately 10 metric tons per hectare to approximately 8 metric tons per hectare. These data show an overall increase in yield for Nigeria, which is mostly attributed to increased land use but not to increased productivity per area, which has in fact declined. Increasing the countries total cassava yield by further increasing land use does not seem to be a suitable solution for the individual smallholder farmers. Ideally, increases in yield would come from improvements to productivity per area, or in other words, more efficient farming methods and more high-yielding cassava varieties.

Alongside cassava breeding, biotechnology might be one of the tools that can contribute to increasing cassava yield. In the recent years, several biotechnological improvements have been realized in this important crop, especially concerning nutritional improvements and virus resistance. Some notable examples include the improvement of cassavas vitamin B6 content, iron and zinc content, and plant resistance to cassava mosaic virus and cassava brown streak virus (Li et al., 2015; Narayanan et al., 2019; Narayanan et al., 2021). A recent report from Chavarriaga-Aguirre et al. (2016) summarizes several of these improvements. However, the authors rightfully note that all transgenic cassava plants are stuck in the proof of concept stage and more translational research needs to happen to get these plants into the hands of farmers. Transgenic concepts need to move out of the laboratory into the field and be tested in multi-year and multi-location trials (Chavarriaga-Aguirre et al., 2016).

The "Cassava Source-Sink" project (https://cass-research.org/) focusses on cassava translational research and aims to improve cassava yield through breeding and biotechnology. Yield traits are typically polygenic traits, depending on the interaction of many genes. Flux through biochemical pathways is often coordinated with that of competing pathways, therefore, effective metabolic engineering will only be achieved by controlling multiple genes of the same, or interconnected, pathways (Halpin, 2005). Recent advances in cloning technologies (e.g. Golden Gate) and declining prices for DNA synthesis are supporting multigene approaches. Sonnewald et al. (2020) recently outlined a strategy towards cassava yield improvement by combining metabolic source-, transport-, and sink- improvements into transgenic cassava plants with subsequent field performance testing. However, it has to be noted that realizing such transgenic multigene approaches and their translation into the field comes with additional challenges like complicated international logistics, high regulatory effort, and long time-lines for cassava transformation and field-testing.

Another challenge for the translation of transgenic yield improvements to cassava is the availability of established expression tools, especially tissue-specific promoter elements. Since the successful implementation of a transgenic concept often needs very cell-/tissue-specific promoters or a combination of several promoters with a particular strength and specificity, the characterization of such promoters becomes essential. This is especially true for yield traits, where likely more than one gene needs to be transferred. To name just three examples from a large body of literature: (i) Root growth, drought resistance and overall yield could be improved by specifically expressing a cytokinine oxidase in the root elongation zone in thale cress, tobacco, barley, and chickpea. Due to the cell-/tissue-specific expression, the inhibitory effect of cytokinine on side-root formation was removed without negatively affecting the elongation root growth from the root apical meristem, leading to an overall larger root system (Werner et al., 2010; Ramireddy et al., 2018; Khandal et al., 2020). (ii) In field-grown maize, yield improvements could be demonstrated by expressing a trehalose-6-phosphatase specifically in maize ears, leading to an increased assimilate supply for this specific plant part (Nuccio et al., 2015). (iii) Recently, a couple of successful multigene stack approaches for yield improvement have been published for thale cress, tobacco, or potato, each requiring at least three well-performing promoters (Jonik et al., 2012; Kromdijk et al., 2016; South et al., 2019).

Cassava promoters, which have in fact been tested and confirmed in cassava itself, are quite scarce. Due to the difficult and lengthy cassava transformation process, cassava promoters have often been characterized by using heterologous expression systems in the past [e.g. Arango et al. (2010); Suhandono et al. (2014)], potentially resulting in incorrect promoter assessments. Unfortunately, there seems to be only a small amount of literature characterizing cassava promoters with stably transformed cassava plants (Zhang et al., 2003; Beltran et al., 2010; Koehorst-van Putten et al., 2012; Oyelakin et al., 2015; Wilson et al., 2017; Mehdi et al., 2019). Zhang et al. (2003); Beltran et al. (2010), and Oyelakin et al. (2015) have described promoter sequences from *Manes.12g132900* and *Manes.12g062400*, from a glutamic-acid-rich protein Pt2L4, or from the cassava vein mosaic virus, respectively. However, all four promoters displayed a rather ubiquitous expression pattern with slight preference for particular tissues. Several promoters have also been analyzed in the frame of a global cassava expression study (Wilson et al., 2017), although unfortunately not in great detail. Mehdi et al. (2019) has analyzed the specificity of the thale cress *SUC2/SUT1* promoter in cassava *via* stably transformed promoter-GFP plants, demonstrating its phloem companion cell specificity. While the leave vasculature was not visible in the *pSUC2::GFP* plants, presumably because of the detection limit, the *pSUC2::GUS* plants presented here, confirm its activity along the entire phloem.

Since storage roots are the prime product of cassava, storage root specific promoters are particularly useful for cassava trait improvement and multiple studies have highlighted the specificity of the potato *Patatin Class I* promoter [e.g. Ihemere et al. (2006); Zidenga et al. (2012); Vanderschuren et al. (2014); Gaitan-Solis et al. (2015); Li et al. (2015); Zhou et al. (2017); Beyene et al. (2018); Wang et al. (2018); Narayanan et al. (2019)]. In addition, Koehorst-van Putten et al. (2012) suggested *pMeGBSS1* as a storage-root specific promoter for cassava, based on the analysis of promoter-luciferase plants. Unfortunately, storage root specificity for the *MeGBSS1* promoter sequence could not be confirmed in this study. More well described promoters are needed to support cassava biotechnology approaches. In addition to promoter sequences specific for autotrophic tissues, promoters specific for heterotrophic tissues like phloem or storage parenchyma, or promoters with very cell-specific expression patterns, will be most valuable.

In this study, we share our findings about the promoter activity and specificity of 24 promoter elements in total. Initially, we characterized 10 promoters with a combination of expression data from field-grown, multigene construct lines, as well as dedicated promoter-gus plants and discovered a surprisingly low activity and/or specificity for the majority of these promoters. Consequently, we tested 14 additional promoter sequences *via* stably transformed promoter-gus plants with the goal to obtain a selection of tissue-specific promoters for autotrophic-, transport-, and heterotrophic storage tissues. We recommend a subset of tissue-specific promoters in the hope that these tools will also help other groups to improve their cassava research and translational work.

Results

Activity and tissue specificity of ten promoters in transgenic, field-grown cassava plants

In an attempt to improve cassava yield by altering different parts of cassava metabolism simultaneously, transgenic cassava plants expressing various combinations of metabolically active genes, altering photosynthetic-, transport-, and storage metabolism, were created and field-tested at NCHU experimental station in Taichung, Taiwan. The plants contained one of seven different multigene constructs, each construct combining three to six different target genes, with the respective target genes always being controlled by the same promoter (Figure 1A). The ten promoter elements used in these constructs were untested for their performance in cassava prior to their use and were initially selected due to their described activity in other plant species. The promoters were expected to mediate specific expression of target genes for autotrophic (also called “source” tissues, following the carbohydrate-based definition) or heterotrophic (also called “sink” tissues, following the carbohydrate-based definition) tissues (Table 1).

Over 400 field-grown plants, representing 7 different constructs and 84 transgenic events were analyzed for their transgene expression, to get an insight into the promoter performance controlling the respective expression. Cassava source leaves, stems, and storage root samples were analyzed *via* quantitative RT-PCR and the results were summarized for each promoter (Figure 1B).

A very strong source leaf expression, although with large variation, was observed for the transcripts controlled by the promoters of *pSlRBCS2* (739 bp) and *pAtRBCS1A* (1175 bp). Their transcripts were approximately 4-5 times more abundant than the transcript controlled by the next strongest leaf promoter *pAtCAB1* (779 bp). High abundance in source leaves was also observed for the transcripts controlled by *pAtGAPA* (1008 bp) and *pStLS1* (1497 bp). Moderate transcript abundance in source leaves was detected for *pMeGBSS1* (1163 bp) and *pAtRBCS3B* (800 bp). Low levels in source leaves were found for the transcripts controlled by *pStSSS3* (1015 bp) and *pAtFBA2*

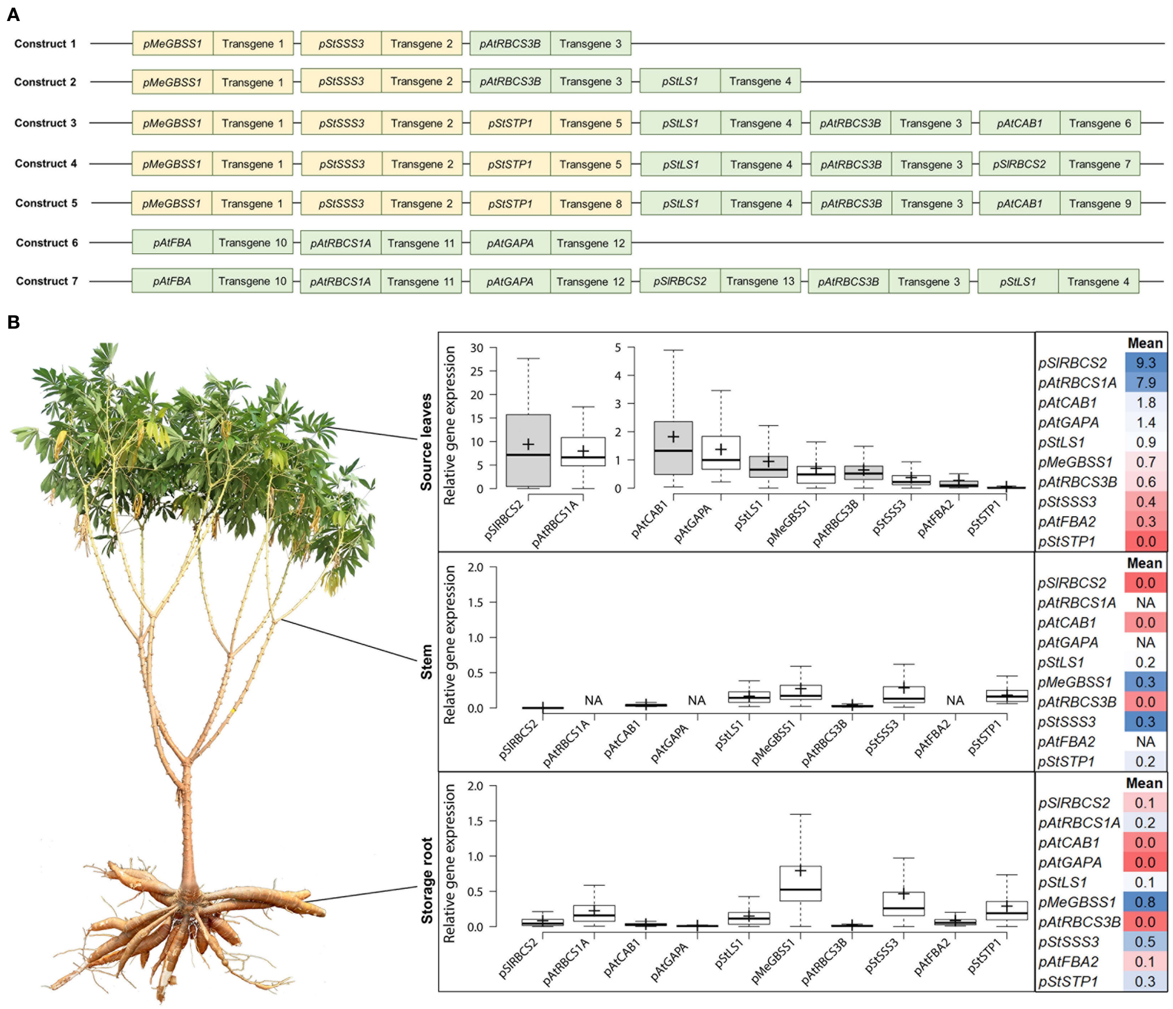


FIGURE 1
Summary of the approximate promoter activity of ten different promoters in source leaves, stem, and storage root. **(A)** Composition of the seven multigene constructs analyzed for gene expression of the individual target genes (target genes not shown). Orange indicates promoter choice for desired expression in heterotrophic tissues, green indicates promoter choice for desired expression in autotrophic tissues. **(B)** The relative gene expression (normalized to *MeGAPDH*) of different transcripts was determined and the data was used to infer the approximate activity of the promoter element controlling its expression. Field-grown cassava plants were used to sample fully exposed source leaves (in the afternoon), stem pieces at the lower end of the first branching point, and storage root material from the two thickest storage roots per plant.

(1000 bp), while no transcripts were found for *pStSTP1* (2081 bp).

In the heterotrophic organs, moderate to low levels were determined for the transcripts controlled by *pMeGBSS1*, *pStSSS3*, and *pStSTP1*. Low levels were found for *pAtRBCS1A* and residual levels were found in heterotrophic tissues for the transcripts controlled by the leaf-promoters *pSlRBCS2*, *pStLS1*, and *pAtFBA2*.

Based on the transcript abundance observed in the different organs (Figure 1), the promoters of *pSlRBCS2* and *pAtRBCS1A* appear to be very active in source leaves, although *pAtRBCS1A* seems to have a low-level activity in sink organs, as well. The promoters of *AtCAB1* and *AtGAPA* were characterized by high and very specific source leaf expression, while the promoters of *pAtRBCS3B* and *pAtFBA2* appeared rather weak. The promoter of *StLS1* also showed weak activity in source leaves with additional residual activity in sink organs.

The promoters of *MeGBSS1*, *StSSS3*, and *StSTP1* were expected to be specific for heterotrophic organs. However, the abundance of transcripts controlled by the promoters of *MeGBSS1* and *StSSS3* was comparable between the three tissues tested. Only *pStSTP1* seems to have a specific activity for heterotrophic organs (Figure 1). According to the PCR results, all three of these promoter sequences resulted in rather weak activity.

TABLE 1 Summary of 10 promoters assayed via their GUS staining pattern and/or transgene expression in field-grown, transgenic cassava plants.

Name	Code	Source organism	Gene/Closest NCBI identifier	Reference	Expected tissue	Analyzed by	Results in Fig.
CHLOROPHYLL A/B-BINDING PROTEIN 1	*AtCAB1*	*Arabidopsis thaliana*	*At1g29930*	Mitra et al., 2009; Engler et al., 2014	Autotrophic tissues	Histology/ qPCR	1 & 3
FRUCTOSE-BISPHOSPHATE ALDOLASE 2	*AtFBA2*	*Arabidopsis thaliana*	*AT4G38970*	Lu et al., 2012	Autotrophic tissues	qPCR	1
GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE SUBUNIT A	*AtGAPA*	*Arabidopsis thaliana*	*AT3G26650*	Shih et al., 1992	Autotrophic tissues	qPCR	1
LEAF-SPECIFIC 1	*StLS1*	*Solanum tuberosum*	*X04753.1*	Stockhaus et al., 1987; Engler et al., 2014	Autotrophic tissues	Histology/ qPCR	1 & 3
RIBULOSE BISPHOSPHATE CARBOXYYLASE SMALL SUBUNIT 1A	*AtRBCS1A*	*Arabidopsis thaliana*	*AT1G67090*	Dedonder et al., 1993	Autotrophic tissues	qPCR	1
RIBULOSE BISPHOSPHATE CARBOXYYLASE SMALL SUBUNIT 2	*SlRBCS2*	*Solanum lycopersicum*	*X66069.1*	Kyozuka et al., 1993; Engler et al., 2014	Autotrophic tissues	qPCR	1
RIBULOSE BISPHOSPHATE CARBOXYYLASE SMALL SUBUNIT 3	*AtRBCS3B*	*Arabidopsis thaliana*	*At5g38410*	Dedonder et al., 1993; Engler et al., 2014	Autotrophic tissues	Histology/ qPCR	1 & 3
STARCH PHOSPHORYLASE 1	*StSTP1*	*Solanum tuberosum*	*X73684.1*	Sonnewald et al., 1995	Heterotrophic tissues	Histology/ qPCR	1 & 3
GRANULE-BOUND STARCH SYNTHASE 1	*MeGBSS1*	*Manihot esculenta*	*Manes.02G001000*	Koehorst-van Putten et al., 2012	Heterotrophic tissues	Histology/ qPCR	1 & 3
SOLUBLE STARCH SYNTHASE 3	*StSSS3*	*Solanum tuberosum*	*X95759.1*	Abel et al., 1996	Heterotrophic tissues	Histology/ qPCR	1 & 3

Promoters expected to be specific for photosynthetic tissues are highlighted in light green and promoters expected to be specific for heterotrophic storage organs are highlighted in orange.

Histological characterization of the analyzed cassava tissues

For six (*pAtCAB1*, *pStLS1*, *pAtRBCS3*, *pMeGBSS1*, *pStSSS3*, *pStSTP1*) of the ten promoters included in the multigene construct plants and tested for their gene expression in the field (Figure 1), dedicated promoter-GUS plants were created, as well (Figure 2). For the analysis of promoter-GUS cassava plants, up to seven different tissues have been sampled and subjected to staining and microscopy: Emerging leaves, developing leaves, fully developed leaves, petioles, upper stem sections, lower stem sections, storage root sections, and fibrous roots (Figure S1). Emerging and developing leaves are characterized by brownish color and were termed “sink” leaves (defined as leaves that have a net import of carbon), while green, fully expanded leaves were considered “source” leaves (defined as leaves with net export of carbon).

The respective reporter plants displayed GUS staining in different tissues and cell types. To define these cell types, counterstaining with toluidine blue was performed and the results summarized in Figure S1. Source- and sink leaves can easily be divided into vascular bundles and mesophyll cells (Figures S1A, B). In petioles, the collenchyma, the sclerenchyma, the phloem, protoxylem/xylem parenchyma, pith parenchyma, and the pith cells can be differentiated from outside to inside (Figure S1C). Stem tissues are characterized by collenchyma, sclerenchyma, phloem, vascular cambium, and varying degrees of secondary xylem and pith tissue, depending on the position of the stem (Figures S1D, E). Especially the lower, heterotrophic stem tissues display increasing levels of secondary xylem tissues, consisting of xylem fibers, water-transporting xylem vessels, and starch-storing xylem parenchyma cells (Figure S1E). Storage roots have periderm tissue, the cork cambium, phelloderm/phloem parenchyma, phloem, vascular cambium, and xylem cells from outside to inside. Alongside xylem vessels, the xylem tissue is mostly dominated by starch-storing xylem parenchyma cells in storage roots (Figure S1F). Lower stems and storage roots, both heterotrophic starch-storing tissues, are overall similar and both tissues are characterized by many vascular rays, ensuring the connection of assimilate- and water transport systems, despite the increasing distance through the formation of secondary xylem during secondary growth (Figures S1E, F).

Analysis of promoter-GUS plants matching the field-tested multigene construct plants

In the *pAtCAB1::GUS* events, staining was observed in the mesophyll of source leaves, sink leaves and newly emerging leaves (Figures 2A1–C1). Petiole and upper stem cross-sections displayed staining in collenchyma, outer parenchyma, and protoxylem areas (Figures 2D1, E1). Lower stem sections, storage roots and fibrous roots were completely devoid of GUS staining (Figures 2F1–H1). Therefore, the chosen promoter element of *AtCAB1* can drive expression in autotrophic cassava tissues without activity in heterotrophic plant parts. To determine the approximate expression strength of these promoter elements in the reporter plants, we determined the relative expression level of the different lines and compared them to the relative expression levels of *pCaMV35* as determined in three *pCaMV35S::GUS* lines. The promoter of *CaMV35S* is ubiquitously active in cassava as well (Figure S2) and its expression strength was used as a tangible reference point throughout the study. The promoter of *AtCAB1* showed approximately 25% to 45% activity compared to the promoter element of *CaMV35S*, respectively (Figure 2I1). Since *pCaMV35S* is a well-documented, strong promoter, the promoter of *AtCAB1* can drive specific and reasonably strong expression in the autotrophic tissues of cassava, which is in line with the field expression results displayed in Figure 1.

The promoter of *pStLS1* displayed the expected staining in the source- and sink leaves (Figures 2A2, B2). However, it also displayed staining in phloem and xylem tissues of petioles, stems and storage roots (Figures 2C2, E2). Only the fibrous roots were devoid of staining (Figure 2F2). This staining pattern matches the expression results (Figure 1), demonstrating that *pStLS1* has activity in both source- and sink tissues in cassava.

Similar to the low transcript levels observed for *pAtRBCS3B* (Figure 1), rather faint staining patterns were observed for *pAtRBCS3B::GUS*. Staining was seen in source- and sink leaves (Figures 2A3, B3), as well as, unexpectedly, in the storage root cambium region (Figure 2E3). It seems that *pAtRBCS3B* is not a good promoter for strong or specific expression in cassava.

The promoter of *pMeGBSS1* was expected to be sink specific (Koehorst-van Putten et al., 2012). However, activity in both source- and sink tissue was observed. Source leaves (Figure 2A4) and sink leaves (Figure 2B4) were stained, and strong staining was seen in the phloem area of the petiole (Figure 2C4). Besides the stem pith, all cell types of stems and storage roots were stained (Figures 2D4, E4). Fibrous roots were devoid of staining (Figure 2F4). Although the staining in stems and storage roots appears to be stronger compared to the other tissues, the inferred promoter activity from the expression results (Figure 1) suggest a rather equal activity between source- and sink. In any case, the promoter was not storage root specific in our experiments.

In contrast to the expression results (Figure 1), indicating comparable source- and sink activity, the *pStSSS3::GUS* plants displayed a staining specific for heterotrophic tissues. Staining was observed in the protoxylem of petioles (Figure 2C5) and stems (Figure 2D5), in the phloem and xylem areas of the storage root (Figure 2E5), but not in the fibrous roots (Figure 2F5). Since the promoters used in the multigene constructs (Figure 1) can potentially be influenced by neighboring promoters, *pStSSS3* might indeed be specific for heterotrophic organs. However, it does not seem to display a strong activity.

The promoter of *StSTP1* displayed a weak but sink-specific behavior in the multigene construct plants (Figure 1). A matching staining pattern was observed in the promoter-GUS plants. The activity in source- and sink leaves was confined to the vasculature (Figures 2A6, B6). Petioles showed staining in the protoxylem and outside the sclerenchyma (Figure 2C6). While fibrous roots displayed no staining besides the root tip (Figure 2F6), most staining was observed in the stems (Figure 2D6) and storage roots (Figure 2E6). Although the activity seems limited, *pStSTP1* can mediate a rather sink specific expression.

Characterization of additional promoter sequences mediating higher tissue specificity

While some of the ten promoter elements tested during field trials could mediate a specific expression pattern for autotrophic tissues (e.g. *pSlRBCS2*, *pAtCAB1*) and some of them could mediate a rather specific expression pattern for heterotrophic organs (e.g. *pStSSS3*, *pStSTP1*), none of them appeared to be particularly strong and specific for the sink tissues. Therefore, we searched for additional promoter candidates in the literature or RNA sequencing datasets, with a particular focus on promoters with potential transport and heterotrophic storage tissues specificity and created additional reporter lines for 14 promoter-GUS constructs in an effort to identify a complete set of tissue-specific promoters for source-, transport- and sink tissues (Table 2).

Identification of additional promoters specific for autotrophic tissues

Two additional promoter-GUS constructs with an expected specificity for autotrophic tissues were created, the promoter of the cytosolic fructose-1,6-bisophosphatase *StFBPase*$_{cyt}$ (1716 bp) and the promoter of *MePsbr* (2019 bp) were chosen. The promoter of StFBPase$_{cyt}$ was chosen due to its previously demonstrated specificity for leaf mesophyll cells (Ebneth (1996), patents EP0938569, US6229067) and the promoter of *MePsbR* was chosen due to the high and leaf specific transcript levels of *MePsbR* in an RNA sequencing dataset (Kuon et al., 2019).

In contrast to the expected mesophyll-specific staining pattern, *pStFBPase*$_{cyt}$ showed considerable staining in the phloem- and cambium areas of stems (Figure S3D) and storage roots (Figure S3E), in addition to staining in the mesophyll of source (Figure S3A) and sink leaves (Figure S3B).

However, a very specific staining pattern was found for *pMePsbR* (Figure 3). Here, staining was observed in the mesophyll of source leaves, sink leaves and newly emerging

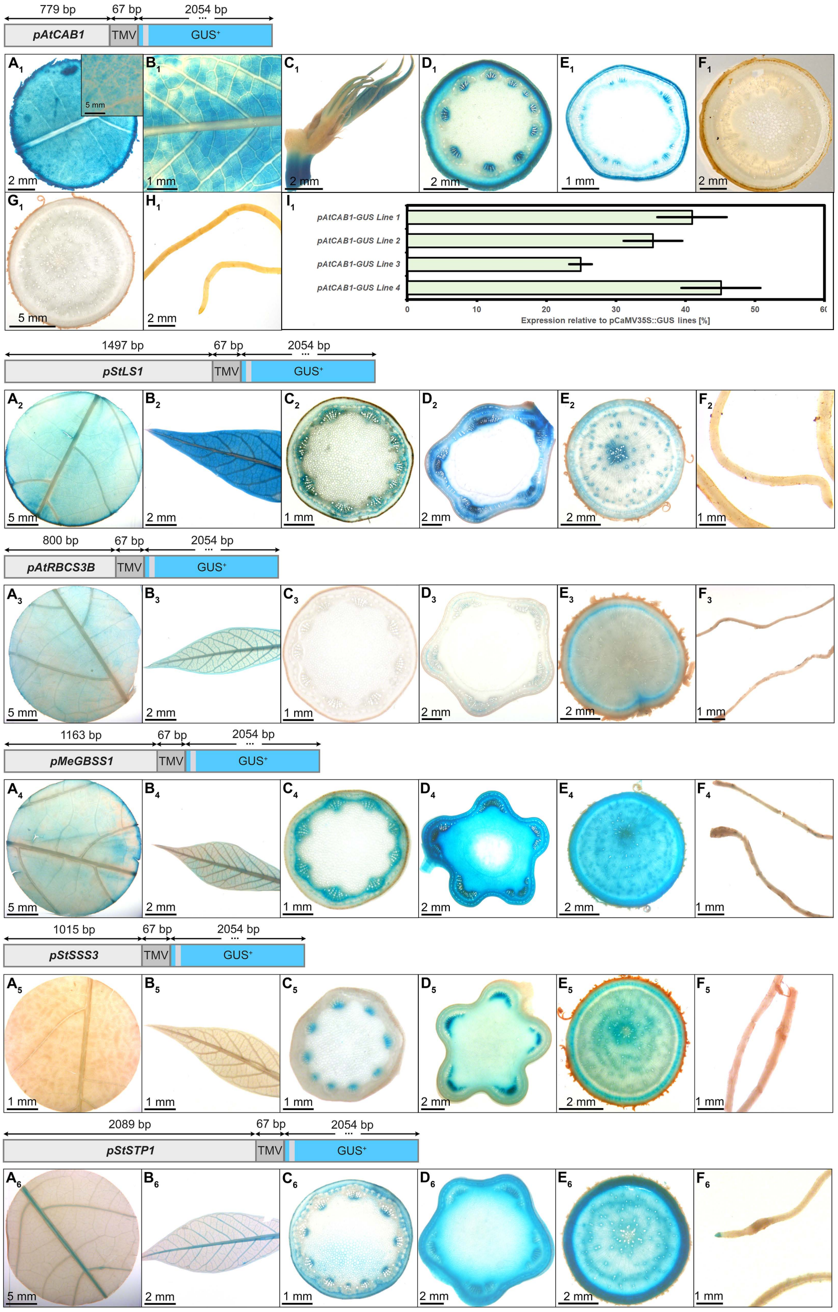


FIGURE 2

Representative GUS staining pattern of at least three *events* from *pAtCAB1::GUS*, *pStLS1::GUS*, *pAtRBCS3B::GUS*, *pMeGBSS1*, *pStSSS3*, *and pStSTP1* promoter-reporter plants. *pAtCAB1::GUS* = **A_1**) Source leaf (Inlay = Close-up), **B_1**) Sink leaf, **C_1**) Emerging leaves, **D_1**) Petiole cross-section, **E_1**) Upper stem cross-section, **F_1**) Lower stem cross-section, **G_1**) Storage root cross-section, **H_1**) Fibrous roots, **I_1**) GUS expression levels of four *pAtCAB1::GUS* lines relative to three *pCaMV35S::GUS* lines in %. *pStLS1::GUS* = **A_2**) Source leaf, **B_2**) Sink leaf, **C_2**) Petiole cross-section, **D_2**) Upper stem cross-section, **E_2**) Storage root cross-section, **H_2**) Fibrous roots. *pAtRBCS3B::GUS* = **A_3**) Source leaf, B_3) Sink leaf, **C_3**) Petiole cross-section, **D_3**) Upper stem cross-section, **E_3**) Storage root cross-section, **H_3**) Fibrous roots. p*MeGBSS1::GUS* = **A_4)** Source leaf, **B_4**) Sink leaf, **C_4**) Petiole cross-section, **D_4**) Upper stem cross-section, **E_4**) Storage root cross-section, **H_4**) Fibrous roots. p*StSSS3::GUS* = **A_5**) Source leaf, **B_5**) Sink leaf, **C_5**) Petiole cross-section, **D_5**) Upper stem cross-section, **E_5**) Storage root cross-section, **H_5**) Fibrous roots. p*StSTP1::GUS* = **A_6**) Source leaf, **B_6**) Sink leaf, **C_6**) Petiole cross-section, **D_6**) Upper stem cross-section, **E_6**) Storage root cross-section, **H_6**) Fibrous roots. Plants were either grown on the field at NCHU experimental station Taichung, Taiwan or in a greenhouse in Erlangen, Germany. Tissues from approximately 3-month old cassava plants were used.

TABLE 2 Summary of 14 promoters assayed via their GUS expression and/or GUS staining pattern in transgenic cassava plants grown in the greenhouse.

Name	Code	Source organism	Gene/Closest NCBI identifier	Reference	Expected tissue	Analyzed by	Results in Fig.
CYTOSOLIC FRUCTOSE-1,6-BISPHOSPHATASE	*StFBPase*$_{cyt.}$	*Solanum tuberosum*	*LOC102589275*	Ebneth, 1996	Source leaf mesophyll	Histology	S2
PHOTOSYSTEM II SUBUNIT R	*MePsbR*	*Manihot esculenta*	*Manes.15G102500*	This study	Autotrophic tissues	Histology + qPCR	4
BIDIRECTIONAL SUGAR TRANSPORTER SWEET 1	*MeSWEET1*	*Manihot esculenta*	*Manes.18G086400*	This study	Phloem and phloem parenchyma	Histology	8
COMMELINA YELLOW MOTTLE VIRUS	*CoYMV*	*Commelina yellow mottle virus*	*X52938.1*	Medberry et al., 1992	Phloem tissues	Histology	7
GALACTINOL SYNTHASE 1	*GolS1*	*Cucumis melo*	*AF249912.2*	Haritatos et al., 2000	Loading phloem	Histology	6
SUCROSE SYNTHASE 1	*MeSUS1*	*Manihot esculenta*	*Manes.03g044400*	This study	Phloem and phloem parenchyma	Histology + qPCR	9
SUCROSE-PROTON SYMPORTER 2	*AtSUC2*	*Arabidopsis thaliana*	*AT1G22710*	Truernit and Sauer, 1995	Phloem companion cells	Histology	5
B33 GENE	*StB33*	*Solanum tuberosum*	*X14483.1*	Rocha-Sosa et al., 1989	Heterotrophic tissues	Histology + qPCR	11
DISCORIN 3 SMALL SUBUNIT	*DjDio3*	*Discorea japonica*	*GU324672.1*	Arango et al., 2010	Heterotrophic tissues	Histology	S3
GLUCOSE-6-PHOSPHATE/ PHOSPHATE TRANSLOCATOR	*MeGPT*	*Manihot esculenta*	*Manes.16G010700*	This study	Heterotrophic tissues	Histology + qPCR	13
GRANULE-BOUND STARCH SYNTHASE 1	*StGBSS1*	*Solanum tuberosum*	*X58453.1*	Van der Steege et al., 1992	Heterotrophic tissues	Histology + qPCR	12
PATATIN CLASS 1	*StPat*	*Solanum tuberosum*	*GQ352473*	Bevan et al., 1986	Heterotrophic tissues	Histology + qPCR	10
MADS-BOX PROTEIN SRD1	*IbSRD1*	*Ipomoea batatas*	*ACN39597.1*	Noh et al., 2010; Noh et al., 2012	Cambium and metaxylem	Histology	14
CAULIFLOWER MOSAIC VIRUS 35S	*CaMV35S*	*Cauliflower mosaic virus*	*MT233541.1*	Engler et al., 2014	All tissues	Histology + qPCR	S2

Promoters expected to be specific for photosynthetic tissues are highlighted in light green, promoters expected to be specific for phloem tissues are highlighted in yellow, promoters expected to be specific for heterotrophic storage organs are highlighted in light red, promoters expected to be specific for dividing tissues are highlighted in light blue, and promoters expected to be active ubiquitously are highlighted in light grey.

leaves (Figures 3A–C). Petiole cross-sections displayed labeling in most cell types beside sclerenchyma and pith tissue (Figure 3D) and upper stem sections displayed staining in the pith parenchyma, the phloem and cambium area, and the collenchyma (Figure 3E). The heterotrophic lower stem sections, storage roots and fibrous roots were completely devoid of GUS staining (Figures 3F–H). Therefore, the chosen promoter sequence for *MePsbR* can drive specific expression in autotrophic cassava tissues. To determine the approximate expression strength of *pMePbsbR*, we determined the relative expression level of the different lines and compared them to the relative expression levels of *pCaMV35* as determined in three *pCaMV35S::GUS* lines. The promoter of *MePsbR* showed approximately 15% to 35% activity compared to the promoter element of *CaMV35S* (Figure 3I). Similar expression levels were obtained for *pAtCAB1* and since *pCaMV35S* is a well-documented, strong promoter, the promoters of *MePsbR* can drive specific and reasonable strong expression in the autotrophic tissues of cassava.

Identification of promoter sequences with predominant activity in phloem tissues

The promoter of *AtSUC2* (946 bp) was selected and expected to be phloem specific in cassava, since this promoter has been used as a phloem-specific tool in numerous studies in different species over the years [recently reviewed in Stadler and Sauer (2019)]. Indeed, *pAtSUC2::GUS* lines displayed pronounced staining in the minor and major veins of source leaves (Figure 4A), sink leaves (Figure 4B), newly developing leaves (Figure 4C), as well as a dotted staining in the phloem area of petioles (Figure 4D), upper stem (Figure 4E), lower stem

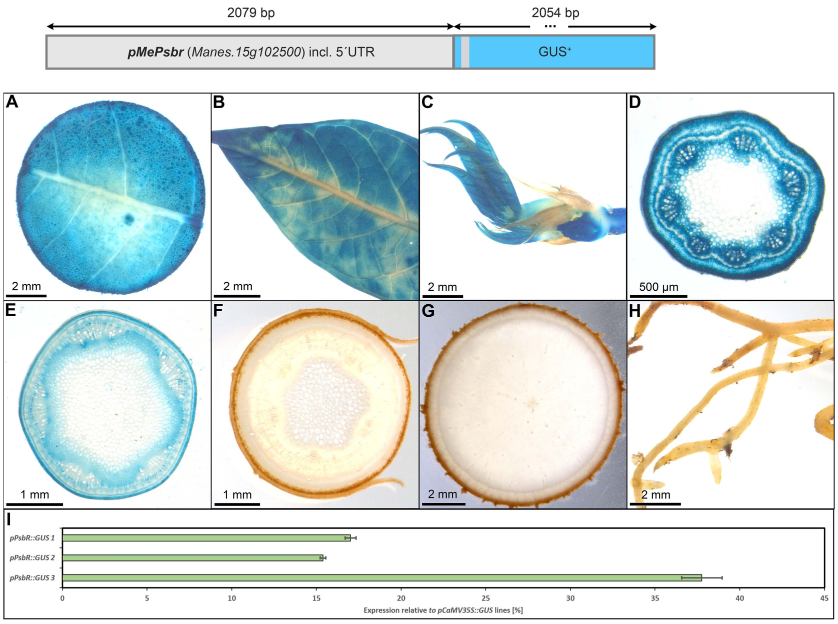


FIGURE 3
Representative GUS staining pattern of three *pMePsbr::GUS* promoter-reporter lines. **(A)** Source leaf, **(B)** Sink leaf, **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section, **(H)** Fibrous roots, **(I)** GUS expression levels of three *pPsbR::GUS* lines relative to three *pCaMV35S::GUS* lines in %. Bars represent mean values with standard deviation (n=4).

(Figure 4F), and storage roots (Figure 4G). The dotted GUS staining in the phloem is very likely resulting from the staining of phloem companion cells. The vasculature of fibrous roots and the root tips also displayed GUS staining (Figure 4H). In addition, some staining was observed in protoxylem and xylem parenchyma areas (Figure 4D, F). These results demonstrate that *pAtSUC2* is well-suited to drive phloem companion cell specific expression also in cassava.

Two additional, well-known phloem promoters were chosen for testing in cassava: A 3000 bp-long promoter sequence of *Cucumis melo* driving expression of the *GALACTINOL SYNTHASE1* [*pGolS1*; Haritatos et al. (2000)] and a 1040 bp-long sequence from *Commelina Yellow Mottle Virus* (Medberry et al., 1992). The former sequence was previously described to have specific activity for the loading phloem, since GUS staining was specifically observed in the smallest veins of the source leaves (Haritatos et al., 2000). The later promoter sequence was described as a promoter with high-level expression, specific to phloem cells, as well as phloem-associated cells (Medberry et al., 1992). In addition, GUS staining was seen in phloem unloading tissues, like the tapetum (Medberry et al., 1992).

GUS staining of transgenic *pCmGolS1::GUS* plant lines revealed specific staining of minor veins in the source leaves in cassava (Figure 5A), matching the results obtained in previous publications (Haritatos et al., 2000). The majority of lines also displayed slightly patchy staining in the veins of sink and newly emerging leaves (Figures 5B, C), staining in the protoxylem/xylem parenchyma of petioles (Figure 5D) and green stems (Figure 5E), as well as slight staining in the pith tissue of auto- and heterotrophic stem tissue (Figures 5E, F). While storage roots displayed very little staining (Figure 5G), fibrous roots also displayed a slightly patchy staining (Figure 5H). Overall, the promoter sequence used, seemed mostly active in minor veins of source leaves but also seemed to convey some activity in non-phloem-related tissues in cassava. Despite the activity outside the leaf, the promoter could still be an interesting tool for biotechnological approaches centered on phloem loading.

In contrast to the *pCmGolS1::GUS* plant lines, which showed preferential activity in the loading phloem, the *pCoYMV::GUS* plant lines seemed to be more specific toward the transport- and unloading phloem. All lines studied, did not show any staining of source leaf vasculature, but rather displayed a staining pattern that seemed wound induced, due to the staining of the cutting site, as well as the punctual staining within the mesophyll (Figure 6A) or in fibrous roots (Figure 6H). In the sink leaves, the staining was observed just outside the vasculature, potentially representing the phloem parenchyma (Figures 6B, C). In addition to some staining in protoxylem and pith parenchyma (Figures 6D, F), pronounced staining was observed in the phloem tissues of petioles (Figure 6D), autotrophic stems (Figure 6E), heterotrophic stems (Figure 6F), and storage roots (Figure 6G). Interestingly, tissues with important functions in lateral transport, as indicated by the

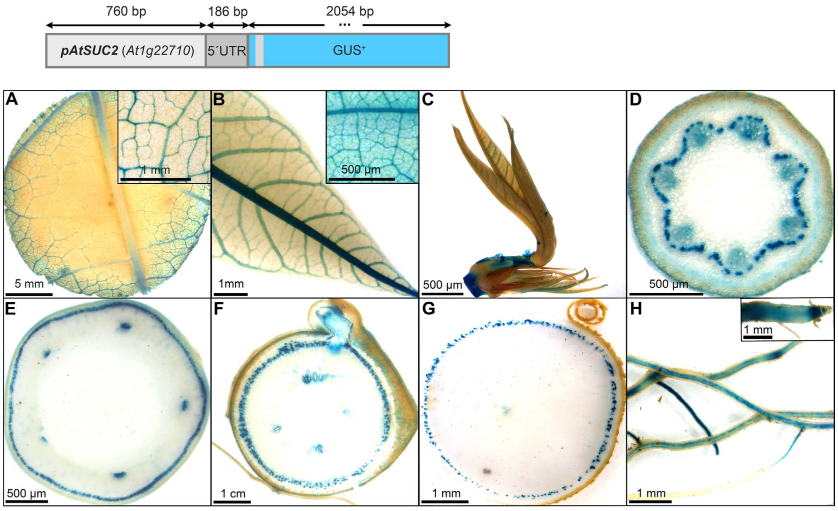


FIGURE 4
Representative GUS staining pattern of four *pAtSUC2::GUS* promoter-reporter lines. **(A)** Source leaf (Inlay = Close-up), **(B)** Sink leaf (Inlay = Close-up), **C** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, (**G**) Storage root cross-section, **(H)** Fibrous roots (Inlay = Root tip).

staining of vascular rays in the lower stems and storage roots, were also stained in these promoter-reporter plants (Figures 6F, G). Taken together, the analyzed sequence of *pCoYMV* seemed rather specific towards transport and unloading phloem tissues, which is in line with previous results, showing promoter activity in vascular and reproductive tissues (Medberry et al., 1992). Although not a quantitative measure, all *pCoYMV::GUS* lines stained within seconds of adding staining buffer, indicating a very strong activity for transport and unloading phloem tissues.

The promoter of *pMeSWEET1-like* (Figure 7) also displayed staining in the phloem areas, although less specific compared to p*AtSUC2* (Figure 4). The promoter element of *MeSWEET1-like* (2000 bp) was initially selected for testing because its transcript appeared highly abundant in storage roots in a RNA-seq dataset

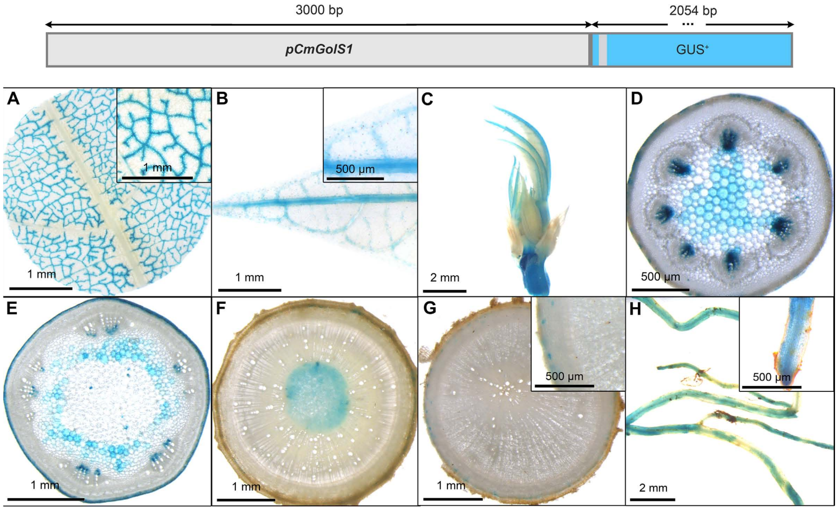


FIGURE 5
Representative GUS staining pattern of four *pCmGolS1* promoter-reporter lines. **(A)** Source leaf (Inlay = Close-up), **(B)** Sink leaf (Inlay = Close-up), **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section (Inlay = Close-up), **(H)** Fibrous roots (Inlay = Root tip).

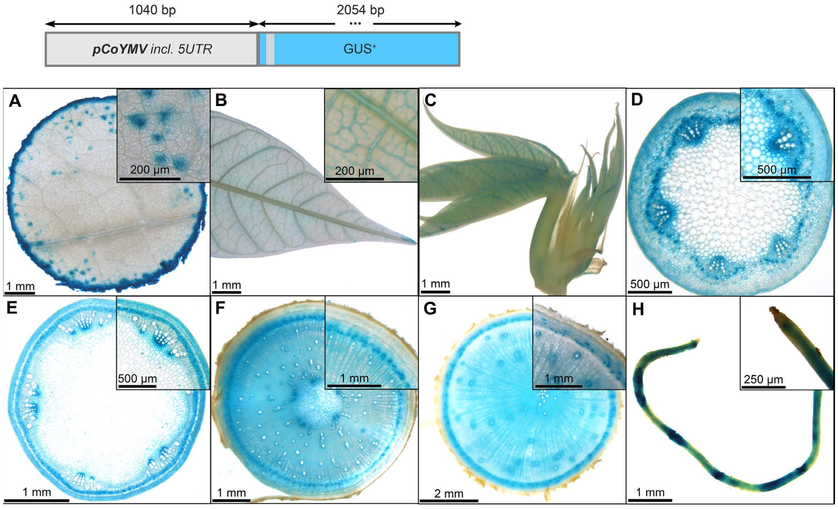


FIGURE 6
Representative GUS staining pattern of four *pCoYMV* promoter-reporter lines. **(A)** Source leaf (Inlay = Close-up), **(B)** Sink leaf (Inlay = Close-up), **(C)** Emerging leaves, **(D)** Petiole cross-section (Inlay = Close-up), **(E)** Upper stem cross-section (Inlay = Close-up), **(F)** Lower stem cross-section (Inlay = Close-up), **(G)** Storage root cross-section (Inlay = Close-up), **(H)** Fibrous roots (Inlay = Root tip).

(NCBI BioProject ID PRJNA784380). Promoter-GUS lines revealed staining in the vasculature of source- and sink leaves (Figures 7A, B), as well as staining in phloem and parenchyma tissues of petioles and stems (Figures 7D–G). The outer storage root region, containing phloem and phloem parenchyma, displayed pronounced GUS staining (Figure 7G). In addition, *pMeSWEET1-like* showed activity in the fibrous root vasculature and root tips (Figure 7H). These results indicate that *pMeSWEET1-like* has preferential activity in phloem and parenchyma cells in cassava.

The promoter of *MeSUS1* (2000 bp) was chosen for testing as a putative phloem promoter because *MeSUS1* transcripts were found to be highly abundant in the phloem fraction of cassava storage root tissues in a RNA-seq datasets (NCBI BioProject ID PRJNA784380). The *pSUS1::GUS* lines displayed an interesting staining pattern, resembling the *pCoYMV* promoter (Figure 6). *pSUS1* was active in the major veins of the leaf vasculature (Figures 8A, B), in the shoot apex (Figure 8C), in phloem and parenchyma cell types (Figures 8D–G), and in fibrous root vasculature (Figure 8H). It displayed pronounced staining in vascular rays of stems and storage roots (Figures 8F, G) and the staining pattern in the storage roots indicated preferential activity in the phloem unloading area, as well as in young xylem cells of the storage roots (Figure 8G). This staining pattern matches the previously described symplasmic unloading mode of cassava and the previously observed metabolic gradients within the storage root (Mehdi et al., 2019).

To determine the approximate expression strength of *pMeSUS1*, we tested the relative expression level of the different lines and compared them to the relative expression levels of *pCaMV35* as determined in three *pCaMV35S::GUS* lines. The promoter elements of *MeSUS1* showed approximately 5-20% activity compared to the promoter element of *CaMV35S* (Figure 8I). This is considerably weaker as the expression strength of the more parenchyma-dominated promoters shown below. However, the promoter is active in far less cells across the storage root, thinning out the specific signal. Overall, *pMeSUS1* is an interesting option as a promoter for applications focused on phloem transport and unloading.

Identification of promoter sequences with predominant activity in heterotrophic storage tissues

Storage root specific promoters are of special interest for cassava because they enable the modification of agronomically interesting storage root traits like starch content, starch quality, nutritional improvements, or shelf life. To our knowledge, there is only one storage root-specific promoter, which was been tested and confirmed in cassava by independent groups. A particular promoter sequence of the potato *Patatin Class I* promoter [*pStPat*; Bevan et al. (1986)], coding for the tuber storage protein patatin, mediates this specific expression pattern in cassava [e.g. Ihemere et al. (2006); Zidenga et al. (2012); Vanderschuren et al. (2014); Gaitan-Solis et al. (2015); Li et al. (2015); Zhou et al. (2017); Beyene et al. (2018); Wang et al. (2018); Narayanan et al. (2019)].

The promoter element of *pStPat* (999 bp) was included in this study to get confirmation of its tissue specificity and activity.

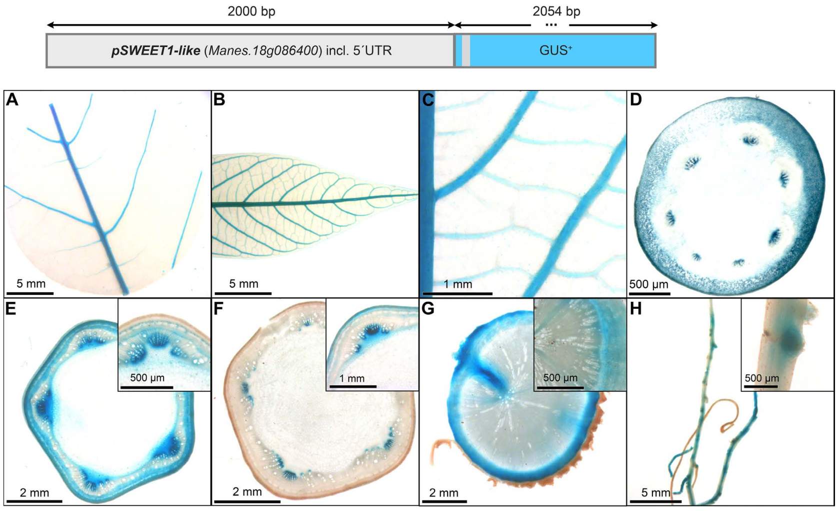


FIGURE 7
Representative GUS staining pattern of at least four *pMeSWEET1-like* promoter-reporter lines. **(A)** Source leaf, **(B)** Sink leaf, **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section (Inlay = Close-up), **(F)** Lower stem cross-section (Inlay = Close-up), **(G)** Storage root cross-section (Inlay = Close-up), **(H)** Fibrous roots (Inlay = Developing side root).

In addition, the promoter elements of *pStB33* (1529 bp), *pStGBSS1* (1061 bp), *pDjDIO3* (1925 bp), and *pMeGPT* (2000 bp) were selected for testing and assumed to be preferentially active in starch storage tissues. The promoters of *StB33*, *StGBSS1*, and *DjDIO3* were previously published with preferential storage organ activity in other plants (Rocha-Sosa et al., 1989; Van der Steege et al., 1992; Arango et al., 2010). The promoter of *MeGPT* was chosen, because *MeGPT* transcripts

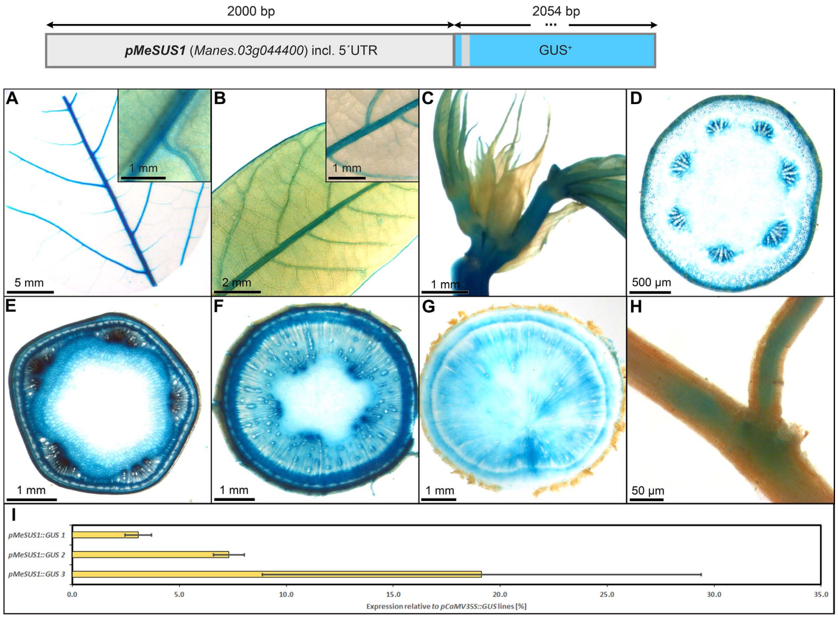


FIGURE 8
Representative GUS staining pattern of at least four *pMeSUS1* promoter-reporter lines. **(A)** Source leaf (Inlay = Close-up), **(B)** Sink leaf (Inlay = Close-up), **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section, **(H)** Fibrous roots, **(I)** GUS expression levels of three *pMeSUS1::GUS* lines relative to three *pCaMV35S::GUS* lines in %. Bars represent mean values with standard deviation (n=4).

were found highly abundant in storage root tissues in prior RNA-seq datasets (NCBI BioProject ID PRJNA784380) and where found to accumulate during storage root bulking (Rüscher et al., 2021).

As expected, *pStPat* displayed strong expression in storage roots, as well as the highest specificity for storage root expression among all promoters tested. The lines displayed no staining in leaves and petioles (Figures 9A–D), only faint staining in upper and lower stems (Figures 9E, F), as well as no staining in fibrous roots (Figure 9H). However, strong staining was observed in the xylem core area of the storage root (Figure 9G), consisting mostly of xylem parenchyma cells. The relative expression level of *pStPat*, compared to the relative expression level of *pCaMV35*, was approximately 40-160%, depending on the respective line (Figure 9I). These results underscore the storage root specificity of *pStPat* in cassava and confirm a high promoter activity in storage roots.

The promoter of *StB33*, also part of the class I family of patatin genes (Rocha-Sosa et al., 1989), appeared very suitable to drive strong expression in heterotrophic storage tissues in cassava as well. The p*StB33::GUS* lines, displayed staining of minor veins in source leaves and no staining in sink leaves and petioles (Figures 10A–D). Upper stem tissue showed staining of collenchyma and protoxylem (Figure 10E), while the heterotrophic lower stem section (Figure 10F) and storage roots displayed strong staining in xylem and phloem parenchyma (Figure 10G). In addition, the vasculature and root tips of fibrous roots were stained (Figure 10F). The relative expression level of *pStB33*, compared to the relative expression level of *pCaMV35*, was approximately 20-80%, depending on the respective line. Therefore, *pStB33* is rather specific for sink tissues and has a high activity in sink organs, although the activity might be slightly lower compared to the *StPat* promoter.

The p*StGBSS1::GUS* lines displayed a staining pattern with predominant activity in the phloem- and xylem parenchyma cells of storage roots (Figure 11G). They also displayed staining in the shoot apex (Figure 11C), in the collenchyma of petioles and stems (Figures 11D–F), the pith parenchyma (Figure 11E, F), and the vasculature of fibrous roots (Figure 11H). In contrast to the two patatin promoters *pStPat* and *pStB33* (Figures 9, 10), *pStGBSS1* showed activity in both source- and sink leaf vasculature (Figures 11A, B). The relative GUS expression level caused by *pStGBSS1*, compared to the relative GUS expression level caused by *pCaMV35*, was approximately 60-120%, depending on the respective line (Figure 11I). Therefore, *pStGBSS1* displays a similar sink activity as *pStPat*, but seems less specific due to its higher activity in some cell types of leaves, petioles and stems.

The promoter of *MeGPT* showed a similar staining pattern compared to *pStGBSS1*, with predominant activity in the phloem- and xylem parenchyma cells of storage roots (Figure 12G). It also displayed staining in the shoot apex (Figure 12C), in the collenchyma of petioles and stems

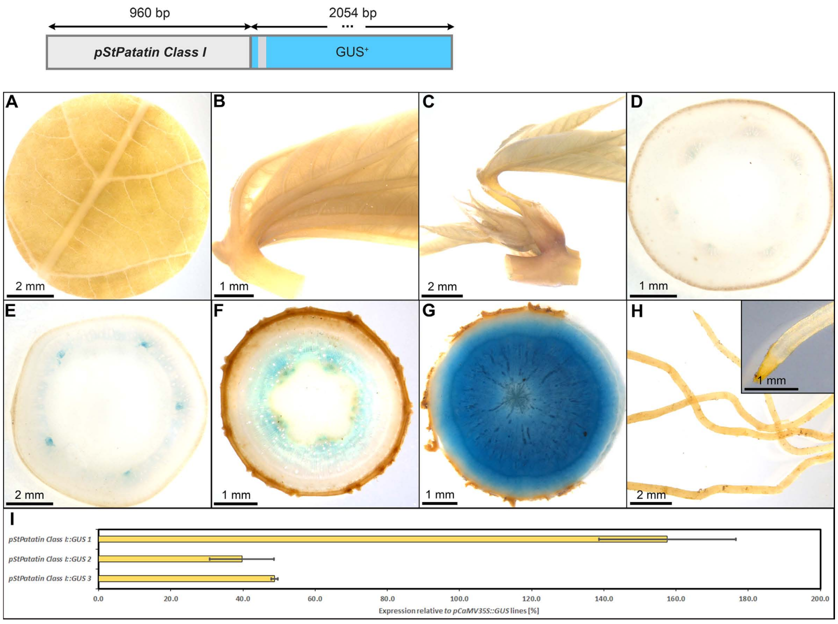


FIGURE 9
Representative GUS staining pattern of four *pStPatatin Class I* promoter-reporter lines. **(A)** Source leaf, **(B)** Sink leaf, **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section, **(H)** Fibrous roots, **(I)** GUS expression levels of three *pStPatatin : GUS* lines relative to three *pCaMV35S::GUS* lines in %. Bars represent mean values with standard deviation (n=4).

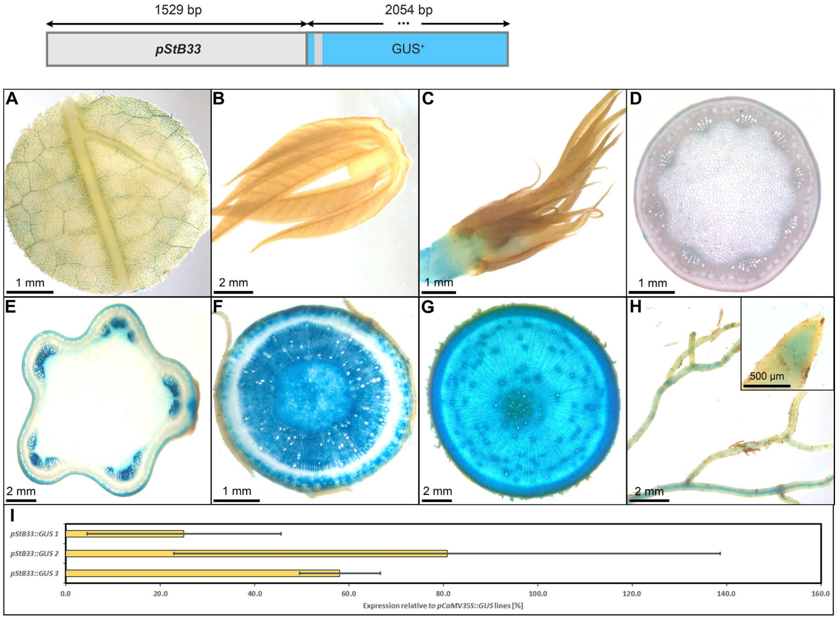


FIGURE 10
Representative GUS staining pattern of at least four *pStB33* promoter-reporter lines. **(A)** Source leaf, **(B)** Sink leaf, **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section, **(H)** Fibrous roots (Inlay = Root tip), **(I)** GUS expression levels of three *pStB33::GUS* lines relative to three *pCaMV35S::GUS* lines in %. Bars represent mean values with standard deviation (n=4).

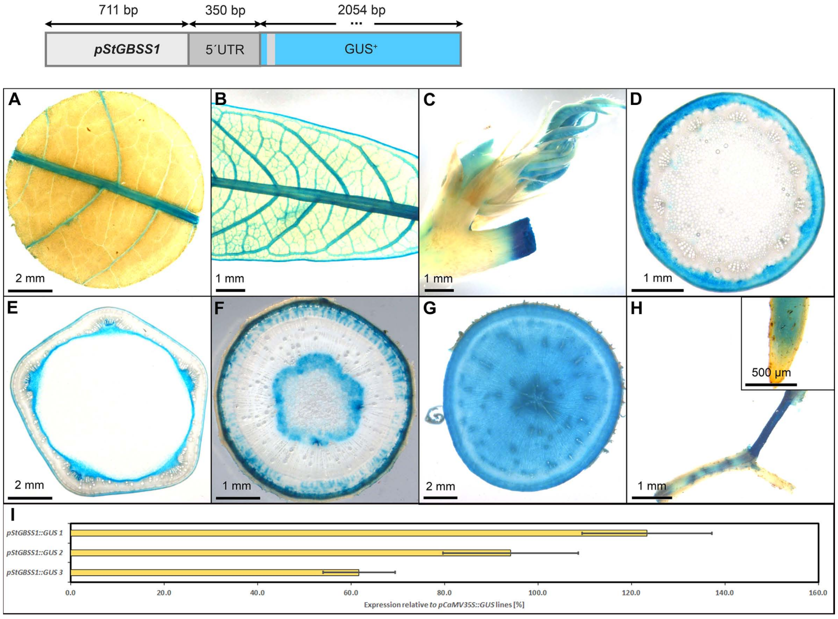


FIGURE 11
Representative GUS staining pattern of four *pStGBSS1* promoter-reporter lines. **(A)** Source leaf, **(B)** Sink leaf, **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section, **(H)** Fibrous roots (Inlay = Root tip), **(I)** GUS expression levels of three *pStGBSS1:GUS* lines relative to three *pCaMV35S::GUS* lines in %. Bars represent mean values with standard deviation (n=4).

(Figures 12D–F), the pith parenchyma (Figures 12E, F), and the vasculature of fibrous roots (Figure 12H). However, it had no staining in source leaves (Figure 12A) and only staining in sink leaf vasculature (Figure 12B). The relative GUS expression level caused by *pMeGPT*, compared to the relative GUS expression level caused by *pCaMV35*, was approximately 20-150%, depending on the respective line (Figure 12I). Taken together, the promoter of *MeGPT* appears rather specific for heterotrophic storage tissues and displays activity in the same range as the *StPat* promoter.

While *pStPat*, *pStB33*, *pStGBSS1*, and *pMeGPT* all show preferential activity in heterotrophic storage tissues, the promoter of the *dioscorin 3 small subunit* gene from *Discorea japonica* (*DjDIO3*) did not. In contrast to what was previously suggested by Arango et al. (2010), *pDjDIO3::GUS* lines displayed a rather ubiquitous staining pattern in cassava (Figure S4).

Identification of a promoter sequence with predominant activity in cambial tissues

To realize transgenic interventions targeting cassava secondary growth, promoters with distinct activity in the vascular cambium could be useful tools. We tested the tissue specificity of the sweet potato MADS-box transcription factor *pIbSRD1* (3011 bp) in cassava, a promoter that was previously characterized in thale cress, carrot, potato and sweet potato. In sweet potato, the *SRD1* expression was shown to be auxin-responsive and the transcript was localized in the primary cambium, secondary cambium, and primary phloem cells (Noh et al., 2010). The main promoter activity in thale cress could be demonstrated in the vasculature including pericycle and endodermis, while the promoter activity was strong in all cells of carrot taproots and potato tubers (Noh et al., 2012).

The promoter activity in cassava resembles the results obtained for sweet potato and thale cress. Pronounced staining was observed in the vasculature of source leaves, sink leaves, newly emerging leaves (Figures 13A–C), and the vasculature of fibrous roots (Figure 13H), as well as in the protoxylem and xylem vessels of petiols and stems (Figures 13D, E). In addition, strong staining was observed in the vascular cambium and cork cambium of stems and storage roots (Figures 13E-G). Together these results demonstrate that *pIbSRD1* has specific activity for cells with meristematic identity in cassava.

Summary of observed promoter specificities

Among the tested leaf promoters, *StFBPase*$_{cyt}$, *AtFBA2*, *AtGAPA*, *StLS1*, and *AtRBCS3B* displayed an either weak or

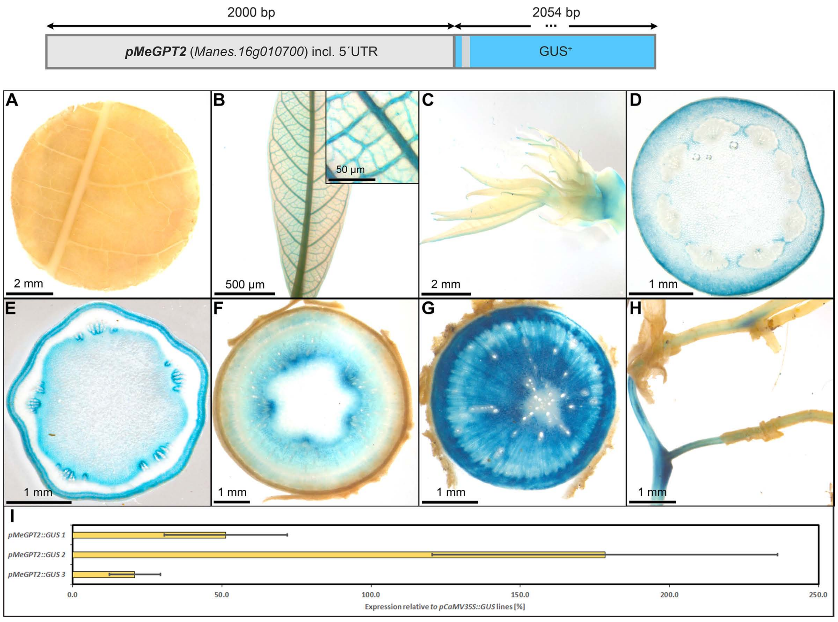


FIGURE 12
Representative GUS staining pattern of four *pMeGPT2* promoter-reporter lines. **(A)** Source leaf, **(B)** Sink leaf (Inlay = Close-up), **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section, **(H)** Fibrous roots, **(I)** GUS expression levels of three *pMeGPT2::GUS* lines relative to three *pCaMV35S::GUS* lines in %. Bars represent mean values with standard deviation (n=4).

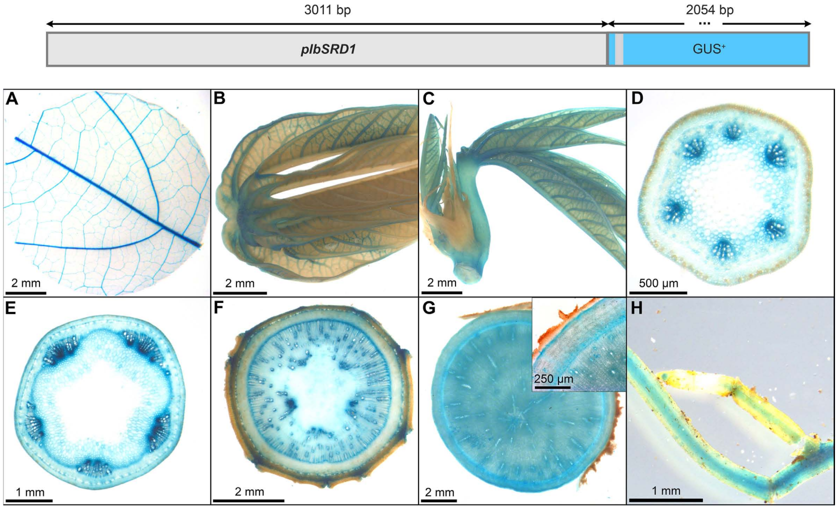


FIGURE 13
Representative GUS staining pattern of at least four *pIbSRD1* promoter-reporter lines. **(A)** Source leaf, **(B)** Sink leaf, **(C)** Emerging leaves, **(D)** Petiole cross-section, **(E)** Upper stem cross-section, **(F)** Lower stem cross-section, **(G)** Storage root cross-section (Inlay = Close-up), **(H)** Fibrous roots.

unspecific expression. However, the promoters of *AtCAB1* and *MePsbR* proved specific and reasonably strong, making them well-suited tools for transgene expression in photosynthetic tissues of cassava. Although no dedicated promoter-GUS lines were created for the promoters of *SlRBCS2* and *AtRBCS1A*, they appeared very active in source leaf tissues in transcript studies. In addition, *pSlRBCS2* also appeared to be specific for this tissue.

The tested promoters of *AtSUC2*, *CmGolS1*, *CoYMV*, *MeSWEET1-like*, and *StSTP1* can be used as expression tools for phloem tissues. While *pAtSUC2* has specific expression along the entire phloem, *pCmGolS1* or *pCoYMV* can target the loading or transport/unloading phloem, respectively. The promoter of *MeSWEET1-like and StSTP1* can be used to target phloem and especially phloem parenchyma tissues of cassava. The promoter of *MeSUS1* also has considerable phloem activity, as well as storage tissue activity, especially in the cells closer to the vascular cambium. This sequence could be an interesting tool for approaches centered on increased sink demand.

Among the promoters with predominant activity in heterotrophic storage tissue, *MeGBSS1* and *StSSS3* seemed less suitable promoters due to their low specificity, or in the case of *pStSSS3* weak activity. The promoter of *StPatatin Class I* proved to be very active and very storage root specific, as previously described. However, *pStB33*, *pMeGPT2* and *pStGBSS1* are also very good promoters for sink tissue expression, as they are predominantly active in starch-storing stem and storage root tissues. They also seem to have a comparable expression strength compared to *pStPatatin Class I*. These promoters will be useful to realize larger transgene stacks that try to avoid repetition of the same promoter sequence in order to avoid silencing or recombination effects.

While *pDjDIO3* is likely very strong (as it showed a strong GUS staining within seconds of staining buffer addition), the promoter is very unspecific and there is a large number of options for this expression pattern. In contrast, *pIbSRD1* showed a highly specific expression pattern with high activity in dividing cells. This promoter can be an interesting tool for more developmental focused approaches targeting stem cells.

Taken together, we have confirmed a number of tissue-specific promoter elements, allowing targeted transgene expression in a variety of cassava tissues. We summarize our recommendations for the most specific promoters per tissue in Table 3. We hope that these promoter sequences will support further transgenic studies in cassava and prove useful for the cassava community.

Discussion

Alongside cassava breeding, trait improvement for this important crop can be achieved through biotechnology by genome editing or transgene expression, introducing additional genetic variety or new functionalities. While traits like herbicide- or pathogen-resistance can sometimes be improved by transferring only a single gene, most traits, like significant nutritional improvements or even yield, often require the transfer and expression of multiple genes. In addition, it is desirable to combine different transgenic traits to aim for plants that are resistant to biotic and abiotic stress, high yielding and nutritious. Subsequent breeding in target genotypes is facilitated by linked transgenes, i.e. transgenes that have integrated into a particular genomic positon together.

TABLE 3 Promoter recommendations for target tissues.

Name	Code	Confirmed tissue specificity	Results in Fig.
CHLOROPHYLL A/B-BINDING PROTEIN	*AtCAB1*	Autotrophic tissues	3
PHOTOSYSTEM II SUBUNIT R	*MePsbR*	Autotrophic tissues	4
RIBULOSE BISPHOSPHATE CARBOXYYLASE SMALL SUBUNIT 2	*SlRBCS2*	Autotrophic tissues	1
BIDIRECTIONAL SUGAR TRANSPORTER SWEET1	*MeSWEET1*	Phloem and phloem parenchyma	8
SUCROSE-PROTON SYMPORTER 2	*AtSUC2*	Phloem companion cells	5
SUCROSE SYNTHASE 1	*MeSUS1*	Phloem and parenchyma cells	9
B33 GENE	*StB33*	Heterotrophic tissues	11
GLUCOSE-6-PHOSPHATE/PHOSPHATE TRANSLOCATOR	*MeGPT*	Heterotrophic tissues	13
GRANULE-BOUND STARCH SYNTHASE1	*StGBSS1*	Heterotrophic tissues	12
PATATIN CLASS 1	*StPat*	Heterotrophic tissues	10
MADS-BOX PROTEIN SRD1	*IbSRD1*	Cambium and metaxylem	14

Promoters observed to be specific for photosynthetic tissues are highlighted in light green, promoters observed to be specific for phloem tissues are highlighted in yellow, promoters observed to be specific for heterotrophic storage organs are highlighted in light red, and promoters observed to be specific for dividing tissues are highlighted in light blue.

However, expressing a variety of linked transgenes with particular strength and tissue specificity is a challenge and there are different ways of approaching it: Polycistronic- or polyprotein strategies have been developed, which can express multiple genes under the control of a single regulatory sequence by either combining all transgenes into a single transcript with subsequent individual translation, or by posttranslational cleavage of a long polypeptide chain, releasing the desired proteins [for review see Halpin (2005)]. However, both methods have limitations, especially if expression in different tissues or subcellular compartments is required. With recent advances in cloning strategies and falling prices for gene synthesis, as well as improvements to transformation protocols allowing for the transformation of larger pieces of DNA, multigene construct-based strategies have become favorable. In these constructs, the individual expression cassettes can be adjusted according to the desired subcellular localization and tissue specificity provided suitable promoters are available. There have been many reports, highlighting the potential of this strategy for i.e. nutritional or yield improvement in different plants (Ye et al., 2000; Paine et al., 2005; Jonik et al., 2012; Li et al., 2015; Kromdijk et al., 2016; Narayanan et al., 2019; South et al., 2019; Wu et al., 2019; Lopez-Calcagno et al., 2020; Narayanan et al., 2021).

This approach, however, requires the availability of a variety of promoters, especially if transgene expression in different tissues is desired. Reusing identical promoters to drive target gene expression in a particular tissue can work, but has the risk of causing recombination or transgene silencing effects, as reported already 30 years ago (Jorgensen, 1992). The existence of a variety of well-characterized promoters, with particular strength and specificity avoids these risks.

The promoter controlling the specificity and expression strength of a given transcript is always dependent on the particular sequence and sequence environment. For instance, combining multiple promoters in close proximity into multigene constructs might result in promoter crosstalk, altering promoter activity and/or specificity. Interestingly, we have observed a large overlap between the results for promoter specificity obtained from multigene constructs and the results obtained from individual promoter-gus plants, suggesting only limited crosstalk between the promoters in the multigene constructs. This observation supports the observed promoter specificities presented in this study and suggests that multiple transgenes can be simultaneously expressed in a tissue-specific manner through a multigene construct, provided suitable promoters are used.

However, our results underline that the described promoter activities and specificities from other plants are often not easily transferable to cassava, highlighting the current need for targeted promoter testing directly in cassava. Overall, we had more success isolating tissue-specific promoter sequences by relying on information from tissue-specific transcript datasets and testing endogenous promoters with a sequence length around 2000bp. Since tissue-specific expression is due to cell type-specific promoter activity, it is understandable that these endogenous promoters have a higher probability to contain the required cis elements for promoter activation in a particular cell type and/or have a higher probability to contain the necessary cis elements for suppression of promoter activity in other cell types. While the use of endogenous promoters seems to have a higher probability to achieve tissue-specific expression of the desired transgene, their activity also has a higher probability to be subject to endogenous regulation mechanisms. The use of the cassava´s own *GPT* promoter for instance would be beneficial to coordinate transgene expression with the onset of storage root formation, since the transcript greatly increases in expression during storage root bulking (Rüscher et al., 2021). At the same time, the likelihood of silencing at some point during cassava growth in response to certain environmental cues seems higher for the endogenous GPT promoter, compared to a potato-derived promoter like *PATATIN CLASS I* for instance.

Despite the higher likelihood of unexpected expression patterns while testing promoter sequences derived from other plants, sometimes exactly these unexpected findings are also the most interesting. Interestingly, almost all sink-specific promoters tested, including the potato *PATATIN B33* promoter, showed activity in tissues containing xylem parenchyma cells like the vasculature, the lower stem, and the storage root. By contrast, the potato *PATATIN CLASS I* promoter (approximately 70% sequence identity to the *PATATIN B33* promoter), which is also expected to be active in all xylem parenchyma cells, displayed a clearly higher specificity with almost exclusive activity in storage roots. Therefore, targeted testing of both endogenous and heterologous promoter sequences can yield highly useful expression tools for cassava research.

It would certainly be interesting for future studies to identify cell-type specific transcription factor regulatory elements for cassava promoters in an attempt to design artificial tissue-specific minimal promoters. However, such a study should contain a large amount of promoter sequences coupled with high-quality cell-type specific transcript data. If the recent progress made in single cell RNA sequencing in plants could also be adopted to different cassava tissues, this might be an interesting possibility. However, for the time being targeted testing of transgene expression tools will help to identify additional options for cassava.

In this study, we have carefully tested 24 individual promoter sequences for their specificity in stably transformed cassava plants. We find that approximately half of the tested promoters displayed an interesting tissue-specific expression pattern. We especially highlight *pAtCAB1*, *pMePsbR*, *pSlRBCS2* for their activity and specificity in autotrophic tissues, *pAtSUC2*, *pMeSWEET1*, *pMeSUS1* for their activity and specificity in different phloem parts, and *pStPat*, *pStB33*, *pStGBSS1*, and *pMeGPT* for their activity and specificity in heterotrophic storage tissues (starch-storing lower stems and storage roots). Furthermore, *pIbSRD1* represents an interesting option for targeting cambial tissues in cassava.

We hope that these promoter sequences will also facilitate the implementation of cassava biotechnology approaches in other research groups and that these approaches will contribute to positive impact on agriculture in the (sub-)tropics.

Material and methods

Plant material and growth conditions

Cassava plants cultivar 60444 were grown from tissue culture in a greenhouse in Erlangen, Germany, or in a confined field at NCHU Taichung, Taiwan. In the greenhouse, a light regime of 12 h light/12 h dark was employed, with a constant temperature of 30°C and 60% relative humidity.

Cloning

All plasmids were created using Golden Gate cloning. The promoters of *AtCAB1*, *pSlRBCS2*, *AtRBCS3B*, and *pStLS1* were taken from the "MoClo Plant Parts Kit" [Addgene Kit # 1000000047; *pICH45152*, *pICH71301*, *pICH45180*, *pICH41551*; Engler et al. (2014)]. All other promoter elements were created by either PCR amplification or DNA synthesis (All promoter sequences are provided in Supplementary Table 1 or the supplementary materials). The promoters of *AtCAB1*, *AtGAPA*, *AtFBA2*, *AtRBCS3B*, *MeGBSS1*, *StB33*, *StFBPase*$_{cyt}$, *StLS1*, *StSSS3*, and *StSTP1* were maintained in level 0 promoter modules (GGAT-TACT). The promoters of *AtSUC2*, *CmGolS1*, *CaMV35S*, *CoYMV*, *DjDIO3*, *IbSRD1*, *MeGPT*, *MePsbr*, *AtRBCS1A*, *MeSUS1*, *MeSWEET1-like*, *StGBSS1*, and *StPat* were maintained in level 0 promoter+5´UTR modules (GGAT-AATG). All level 0 promoter modules (GGAT-TACT) were fused with the *Tabacco mosaic virus* 5´UTR [*pICH41402*; Engler et al. (2014)], a modified *beta-glucuronidase* coding sequence ["GUSPlus"; Broothaerts et al. (2005)], the *E. coli* NOPALINE SYNTHASE 3´UTR+terminator [*pICH41421*; Engler et al. (2014)], and the level 1-1f acceptor [*pICH47732*; Engler et al. (2014)] to create the respective promoter-reporter cassette. All level 0 promoter+5´UTR modules (GGAT-AATG) were fused with a modified *beta-glucuronidase* coding sequence ["GUSPlus"; Broothaerts et al. (2005)], the *E. coli* NOPALINE SYNTHASE 3´UTR+terminator [*pICH41421*; Engler et al. (2014)], and the level 1-1f acceptor [*pICH47732*; Engler et al. (2014)] or level 1-3f acceptor [*pICH47751*; Engler et al. (2014)] to create the respective promoter-reporter cassette. The level 1 plasmids containing the respective promoter-reporter cassettes were transferred into the transformation vector *p134GG* (Mehdi et al., 2019) to create the final level 2 transformation plasmids. All promoter-GUS transformation plasmid maps are provided in supplementary material "Plasmid Maps".

Cassava transformation

Cassava genotype 60444 was transformed with promoter-reporter constructs as described previously (Bull et al., 2009). Hygromycin-resistant transformants were screened by ß-glucuronidase histological staining (see below). Plants with clear GUS staining were maintained in tissue culture and successively analyzed for their tissue specific expression patterns.

Histology and microscopy

Different cassava tissues (Figure S1) were sampled into ice-cold 90% acetone solution. Leaf-samples were taken with a leaf puncher and cross-sections were manually prepared with a razor blade. These sections were covered with GUS staining buffer (200mM NaP pH7, 100mM $K_3[Fe(CN_6)]$, 100mM $K_4[Fe(CN_6)]$, 500mM EDTA, 0.5% SILWET® gold) and thoroughly vacuum infiltrated for 10 minutes. The GUS staining buffer was removed and replaced with fresh GUS staining solution containing GUS staining buffer with 0.75mg/ml 5-bromo-4-chloro-3-indolyl-β-D-glucuronic acid (X-Gluc; pre-dissolved in a small amount of DMSO). The GUS staining solution was thoroughly vacuum infiltrated for 10 minutes. The infiltrated tissues were incubated in 37°C overnight or stopped shortly after incubation in case of very quick staining (e.g. *pCoYMV*, *pDjDIO3*). After removal of the GUS staining solution, 70% ethanol was added to the tissue sections and incubated in 37°C until the tissues were cleared. Light microscopic images were taken on a Zeiss Axioskop or a Zeiss STEMI SV11 Stereomicroscope (Zeiss, Wetzlar, Germany).

Quantification of GUS expression

RNA extraction of cassava source leaves and storage roots was performed using the Spectrum Plant Total RNA Kit (Sigma-Aldrich, St. Louis, MO, USA). cDNA was generated from 0.2-1µg of RNA using RevertAid H Minus Reverse Transcriptase as indicated by the manufacturer (Thermo Fisher Scientific, Waltham, MA, USA). The cDNA was diluted 1:10 and quantification of gene expression was examined using GoTaq® qPCR Master Mix (Promega, Madison, WI, USA). The assay was mixed in a 96-well plate and measured in an AriaMx Real-time PCR System (Agilent, Santa Clara, CA, USA).

The primer pairs "GCGGCCAAAGTCCATCTCCG/TGAAAGCCCGCAACGGTGTC" and "TCTTCGGCGTTAGGAACCCAG/GCAGCCTTATCCTTGTCGGTG" were used to determine *GUS* and *MeGAPDH* expression, respectively. Primer tests were performed and passed (Figures S5, 6). The normalized GUS expression of the promoter::GUS lines was determined by the $2^{-\Delta Ct}$ calculation method with *MeGAPDH* (*Manes.06g116400*) as a reference gene. The normalized GUS expression of the respective promoter::GUS lines was calculated in relation to the normalized expression of the *pCaMV35S::GUS lines* and displayed as relative expression *pCaMV35S::GUS* lines in percent to provide an approximate classification of expression strength.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

WZ performed the experiments and wrote the manuscript. RA transformed all constructs into cassava and provided all transgenic cassava plant lines. CL maintained cassava in tissue culture and assisted the experiments. S-HC managed the cassava field experiment. WG and US supervised the research. All authors contributed to the article and approved the submitted version.

Funding

We thank the Bill and Melinda Gates Foundation for funding this research through the grant INV-008053 'Metabolic Engineering of Carbon Pathways to Enhance Yield of Root and Tuber Crops' provided to U.S. We also acknowledge funding of this research by the Advanced Plant Biotechnology Center provided to W.G. as part of The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan. W.G. is a Yushan Scholar of the Ministry of Education in Taiwan.

Acknowledgments

We thank Michaela Reiser, Otilia Ciobotea, and Ingrid Schießl for excellent technical assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1042379/full#supplementary-material

References

Abel, G. J. W., Springer, F., Willmitzer, L., and Kossmann, J. (1996). Cloning and functional analysis of a cDNA encoding a novel 139 kDa starch synthase from potato (*Solanum tuberosum* l.). *Plant J.* 10, 981–991. doi: 10.1046/j.1365-313X.1996.10060981.x

Arango, J., Salazar, B., Welsch, R., Sarmiento, F., Beyer, P., and Al-Babili, S. (2010). Putative storage root specific promoters from cassava and yam: cloning and evaluation in transgenic carrots as a model system. *Plant Cell Rep.* 29, 651–659. doi: 10.1007/s00299-010-0851-7

Beltran, J., Prias, M., Al-Babili, S., Ladino, Y., Lopez, D., Beyer, P., et al. (2010). Expression pattern conferred by a glutamic acid-rich protein gene promoter in field-grown transgenic cassava (*Manihot esculenta* crantz). *Planta* 231, 1413–1424. doi: 10.1007/s00425-010-1144-7

Bevan, M., Barker, R., Goldsbrough, A., Jarvis, M., Kavanagh, T., and Iturriaga, G. (1986). The structure and transcription start site of major potato tuber protine gene. *Nucleic Acids Res.* 14, 4625–4638. doi: 10.1093/nar/14.11.4625

Beyene, G., Solomon, F. R., Chauhan, R. D., Gaitan-Solis, E., Narayanan, N., Gehan, J., et al. (2018). Provitamin a biofortification of cassava enhances shelf life but reduces dry matter content of storage roots due to altered carbon partitioning into starch. *Plant Biotechnol. J.* 16, 1186–1200. doi: 10.1111/pbi.12862

Broothaerts, W., Mitchell, H. J., Weir, B., Kaines, S., Smith, L. M. A., Yang, W., et al. (2005). Gene transfer to plants by diverse species of bacteria. *Nature* 433, 629–633. doi: 10.1038/nature03309

Bull, S. E., Owiti, J. A., Niklaus, M., Beeching, J. R., Gruissem, W., and Vanderschuren, H. (2009). Agrobacterium-mediated transformation of friable embryogenic calli and regeneration of transgenic cassava. *Nat. Protoc.* 4, 1845–1854. doi: 10.1038/nprot.2009.208

Chavarriaga-Aguirre, P., Brand, A., Medina, A., Prías, M., Escobar, R., Martinez, J., et al. (2016). The potential of using biotechnology to improve cassava: a review. *In Vitro Cell. Dev. Biol. - Plant* 52, 461–478. doi: 10.1007/s11627-016-9776-3

Dedonder, A., Rethy, R., Fredericq, H., Van Montagu, M., and Krebbers, E. (1993). Arabidopsis rbcS genes are differentially regulated by light. *Plant Physiol.* 101, 801–808. doi: 10.1104/pp.101.3.801

Ebneth, M. (1996). *Expressionsanalyse des promotors einer cytosolischen fruktose-1,6-bisphosphatase aus kartoffel in transgenen tabak- und kartoffelpflanzen (Dissertation).PhD thesis* (Deutschland: Frei Universität Berlin).

Engler, C., Youles, M., Gruetzner, R., Ehnert, T. M., Werner, S., Jones, J. D., et al. (2014). A golden gate modular cloning toolbox for plants. *ACS Synth. Biol.* 3, 839–843. doi: 10.1021/sb4001504

FAO, ECA and AUC (2020). Africa Regional overview of food security and nutrition 2019. *Accra* 1–104. doi: 10.4060/CA7343EN

FAO, IFAD, UNICEF, WFP and WHO (2022). "The state of food security and nutrition in the world 2022," in *Repurposing food and agricultural policies to make healthy diets more affordable* (Rome: FAO). doi: 10.4060/cc0639en

Gaitan-Solis, E., Taylor, N. J., Siritunga, D., Stevens, W., and Schachtman, D. P. (2015). Overexpression of the transporters AtZIP1 and AtMTP1 in cassava changes zinc accumulation and partitioning. *Front. Plant Sci.* 6, 492. doi: 10.3389/fpls.2015.00492

Halpin, C. (2005). Gene stacking in transgenic plants - the challenge for 21st century plant biotechnology. *Plant Biotechnol. J.* 3, 141–155. doi: 10.1111/j.1467-7652.2004.00113.x

Haritatos, E., Ayre, B. G., and Turgeon, R. (2000). Identification of phloem involved in assimilate loading in leaves by the activity of the *Galactinol synthase Promoter1*. *Plant Physiol.* 123, 929–938. doi: 10.1104/pp.123.3.929

Ihemere, U., Arias-Garzon, D., Lawrence, S., and Sayre, R. (2006). Genetic modification of cassava for enhanced starch production. *Plant Biotechnol. J.* 4, 453–465. doi: 10.1111/j.1467-7652.2006.00195.x

Jonik, C., Sonnewald, U., Hajirezaei, M. R., Flugge, U. I., and Ludewig, F. (2012). Simultaneous boosting of source and sink capacities doubles tuber starch yield of potato plants. *Plant Biotechnol. J.* 10, 1088–1098. doi: 10.1111/j.1467-7652.2012.00736.x

Jorgensen, R. A. (1992). Silencing of plant genes by homologous transgenes. *Agbiotech. News Inf.* 4, 265–273. doi: 10.1105/tpc.4.2.185

Khandal, H., Gupta, S. K., Dwivedi, V., Mandal, D., Sharma, N. K., Vishwakarma, N. K., et al. (2020). Root-specific expression of chickpea cytokinin oxidase/dehydrogenase 6 leads to enhanced root growth, drought tolerance and yield without compromising nodulation. *Plant Biotechnol. J.* 18, 2225–2240. doi: 10.1111/pbi.13378

Koehorst-van Putten, H. J., Wolters, A. M., Pereira-Bertram, I. M., Van Den Berg, H. H., van der Krol, A. R., and Visser, R. G. (2012). Cloning and characterization of a tuberous root-specific promoter from cassava (*Manihot esculenta* crantz). *Planta* 236, 1955–1965. doi: 10.1007/s00425-012-1796-6

Kromdijk, J., Glowacka, K., Leonelli, L., Gabilly, S. T., Iwai, M., Niyogi, K. K., et al. (2016). Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. *Science* 354, 857–861. doi: 10.1126/science.aai8878

Kuon, J. E., Qi, W., Schlapfer, P., Hirsch-Hoffmann, M., Von Bieberstein, P. R., Patrignani, A., et al. (2019). Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. *BMC Biol.* 17, 75. doi: 10.1186/s12915-019-0697-6

Kyozuka, J., Mcelroy, D., Hayakawa, T., Xie, Y., Wu, R., and Shimamoto, K. (1993). Light-regulated and cell-specific expression of tomato *rbcS-gusA* and rice *rbcS-gusA* fusion genes in transgenic rice. *Plant Physiol.* 102, 991–1000. doi: 10.1104/pp.102.3.991

Li, K. T., Moulin, M., Mangel, N., Albersen, M., Verhoeven-Duif, N. M., Ma, Q., et al. (2015). Increased bioavailable vitamin B6 in field-grown transgenic cassava for dietary sufficiency. *Nat. Biotechnol.* 33, 1029–1032. doi: 10.1038/nbt.3318

Lopez-Calcagno, P. E., Brown, K. L., Simkin, A. J., Fisk, S. J., Vialet-Chabrand, S., Lawson, T., et al. (2020). Stimulating photosynthetic processes increases productivity and water-use efficiency in the field. *Nat. Plants* 6, 1054–1063. doi: 10.1038/s41477-020-0740-1

Lu, W., Tang, X., Huo, Y., Xu, R., Qi, S., Huang, J., et al. (2012). Identification and characterization of fructose 1,6-bisphosphate aldolase genes in arabidopsis reveal a gene family with diverse responses to abiotic stresses. *Gene* 503, 65–74. doi: 10.1016/j.gene.2012.04.042

Medberry, S. L., Lockhart, B. E., and Olszewski, N. E. (1992). The commelina yellow mottle virus promoter is a strong promoter in vascular and reproductive tissues. *Plant Cell* 4, 185–192. doi: 10.1105/tpc.4.2.185

Mehdi, R., Lamm, C. E., Bodampalli, R., Müdsam, C., Saeed, M., Klima, J., et al. (2019). Symplasmic phloem unloading and radial post-phloem transport *via* vascular rays in tuberous roots of *Manihot esculenta*. *J. Exp. Bot.* 70, 5559. doi: 10.1093/jxb/erz297

Mitra, A., Han, J., Zhang, Z. J., and Mitra, A. (2009). The intergenic region of *Arabidopsis thaliana CAB1* and *CAB2* divergent genes functions as a bidirectional promoter. *Planta* 229, 1015–1022. doi: 10.1007/s00425-008-0859-1

Narayanan, N., Beyene, G., Chauhan, R. D., Gaitán-Solís, E., Gehan, J., Butts, P., et al. (2019). Biofortification of field-grown cassava by engineering expression of an iron transporter and ferritin. *Nat. Biotechnol.* 37, 144–151. doi: 10.1038/s41587-018-0002-1

Narayanan, N., Beyene, G., Chauhan, R. D., Grusak, M. A., and Taylor, N. J. (2021). Stacking disease resistance and mineral biofortification in cassava varieties to enhance yields and consumer health. *Plant Biotechnol. J.* 19, 844–854. doi: 10.1111/pbi.13511

Noh, S. A., Lee, H. S., Huh, E. J., Huh, G. H., Paek, K. H., Shin, J. S., et al. (2010). *SRD1* is involved in the auxin-mediated initial thickening growth of storage root by enhancing proliferation of metaxylem and cambium cells in sweetpotato (*Ipomoea batatas*). *J. Exp. Bot.* 61, 1337–1349. doi: 10.1093/jxb/erp399

Noh, S. A., Lee, H. S., Huh, G. H., Oh, M. J., Paek, K. H., Shin, J. S., et al. (2012). A sweetpotato *SRD1* promoter confers strong root-, taproot-, and tuber-specific expression in arabidopsis, carrot, and potato. *Transgenic Res.* 21, 265–278. doi: 10.1007/s11248-011-9528-4

Nuccio, M. L., Wu, J., Mowers, R., Zhou, H. P., Meghji, M., Primavesi, L. F., et al. (2015). Expression of trehalose-6-phosphate phosphatase in maize ears improves yield in well-watered and drought conditions. *Nat. Biotechnol.* 33, 862–869. doi: 10.1038/nbt.3277

Oyelakin, O. O., Opabode, J. T., Raji, A. A., and Ingelbrecht, I. L. (2015). A cassava vein mosaic virus promoter cassette induces high and stable gene expression in clonally propagated transgenic cassava (*Manihot esculenta* crantz). *S. Afr. J. Bot.* 97, 184–190. doi: 10.1016/j.sajb.2014.11.011

Paine, J. A., Shipton, C. A., Chaggar, S., Howells, R. M., Kennedy, M. J., Vernon, G., et al. (2005). Improving the nutritional value of golden rice through increased pro-vitamin a content. *Nat. Biotechnol.* 23, 482–487. doi: 10.1038/nbt1082

Ramireddy, E., Hosseini, S. A., Eggert, K., Gillandt, S., Gnad, H., Von Wiren, N., et al. (2018). Root engineering in barley: Increasing cytokinin degradation produces a larger root system, mineral enrichment in the shoot and improved drought tolerance. *Plant Physiol.* 177, 1078–1095. doi: 10.1104/pp.18.00199

Rocha-Sosa, M., Sonnewald, U., Frommer, W., Sratmann, M., Schell, J., and Willmitzer, L. (1989). Both developmantal and metabolic signals activate the promoter of a class I patatin gene. *EMBO J.* 8, 23–29. doi: 10.1002/j.1460-2075.1989.tb03344.x

Rüscher, D., Corral, J. M., Carluccio, A. V., Klemens, ,.P., Gisel, ,. A., Stavolone, ,. L., et al. (2021). Auxin signaling and vascular cambium formation enable storage metabolism in cassava tuberous roots. *J. Exp. Bot.* 72, 3688–3703. doi: 10.1093/jxb/erab106

Shih, M.-C., Heinrich, P., and Goodman, H. M. (1992). Cloning and chromosomal mapping of nuclear genes encoding chloroplast and cytosolic glyceraldehyde-3-phosphate-dehydrogenase from *Arabidopsis thaliana*. *Gene* 119, 317–319. doi: 10.1016/0378-1119(92)90290-6

Sonnewald, U., Basner, A., Greve, B., and Steup, M. (1995). A second l-type isozyme of potato glucan phosphorylase: cloning, antisense inhibition and expression analysis. *Plant Mol. Biol.* 27, 567–576. doi: 10.1007/BF00019322

Sonnewald, U., Fernie, A. R., Gruissem, W., Schläpfer, P., Anjanappa, R. B., Chang, S.-H., et al. (2020). The cassava source–sink project: opportunities and challenges for crop improvement by metabolic engineering. *Plant J.* 103, 1655–1665. doi: 10.1111/tpj.14865

South, P. F., Cavanagh, A. P., Liu, H. W., and Ort, D. R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. *Science* 363, 1–9. doi: 10.1126/science.aat9077

Stadler, R., and Sauer, N. (2019). *The AtSUC2 promoter: A powerful tool to study phloem physiology and development* (New York, NY:Humana). doi: 10.1007/978-1-4939-9562-2_22

Stockhaus, J., Eckes, P., Blau, A., Schell, J., and Willmitzer, L. (1987). Organ-specific and dosage-dependent expression of a leaf/stem specific gene from potato after tagging and transfer into potato and tabacco plants. *Nucleic Acids Res.* 15, 3479–3491. doi: 10.1093/nar/15.8.3479

Suhandono, S., Apriyanto, A., and Ihsani, N. (2014). Isolation and characterization of three cassava *Elongation factor 1 alpha* (*MeEF1A*) promoters. *PloS One* 9, e84692. doi: 10.1371/journal.pone.0084692

Truernit, E., and Sauer, N. (1995). The promoter of the *Arabidopsis thaliana SUC2* sucrose-h$^+$ symporter gene directs expression of β-glucuronidase to the phloem: evidence for phloem loading and unloading by SUC2. *Planta* 196, 564–570. doi: 10.1007/BF00203657

Vanderschuren, H., Nyaboga, E., Poon, J. S., Baerenfaller, K., Grossmann, J., Hirsch-Hoffmann, M., et al. (2014). Large-Scale proteomics of the cassava storage root and identification of a target gene to reduce postharvest deterioration. *Plant Cell* 26, 1913–1924. doi: 10.1105/tpc.114.123927

Van der Steege, G., Nieboer, M., Swaving, J., and Tempelaar, M. J. (1992). Potato granule-bound starch synthase promoter-controlled GUS expression: regulation of expression after transient and stable transformation. *Plant Mol. Biol.* 20, 19–30. doi: 10.1007/BF00029145

Wang, W., Hostettler, C. E., Damberger, F. F., Kossmann, J., Lloyd, J. R., and Zeeman, S. C. (2018). Modification of cassava root starch phosphorylation enhances starch functional properties. *Front. Plant Sci.* 9, 1562. doi: 10.3389/fpls.2018.01562

Werner, T., Nehnevajova, E., Kollmer, I., Novak, O., Strnad, M., Kramer, U., et al. (2010). Root-specific reduction of cytokinin causes enhanced root growth, drought tolerance, and leaf mineral enrichment in arabidopsis and tobacco. *Plant Cell* 22, 3905–3920. doi: 10.1105/tpc.109.072694

Wilson, M. C., Mutka, A. M., Hummel, A. W., Berry, J., Chauhan, R. D., Vijayaraghavan, A., et al. (2017). Gene expression atlas for the food security crop cassava. *New Phytol.* 213, 1632–1641. doi: 10.1111/nph.14443

Wu, T. Y., Gruissem, W., and Bhullar, N. K. (2019). Targeting intracellular transport combined with efficient uptake and storage significantly increases grain iron and zinc levels in rice. *Plant Biotechnol. J.* 17, 9–20. doi: 10.1111/pbi.12943

Ye, X., Al-Babili, S., Klöti, A., Zhang, J., Lucca, P., Beyer, P., et al. (2000). Engineering the provitamin a (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science* 287, 303–305. doi: 10.1126/science.287.5451.303

Zhang, P., Bohl-Zenger, S., Puonti-Kaerlas, J., Potrykus, I., and Gruissem, W. (2003). Two cassava promoters related to vascular expression and storage root formation. *Planta* 218, 192–203. doi: 10.1007/s00425-003-1098-0

Zhou, W., He, S., Naconsie, M., Ma, Q., Zeeman, S. C., Gruissem, W., et al. (2017). *Alpha-glucan, water dikinase 1* affects starch metabolism and storage root growth in cassava (*Manihot esculenta* crantz). *Sci. Rep.* 7, 9863. doi: 10.1038/s41598-017-10594-6

Zidenga, T., Leyva-Guerrero, E., Moon, H., Siritunga, D., and Sayre, R. (2012). Extending cassava root shelf life *via* reduction of reactive oxygen species production. *Plant Physiol.* 159, 1396–1407. doi: 10.1104/pp.112.200345

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fpls.2022.1041925

Utilizing evolutionary conservation to detect deleterious mutations and improve genomic prediction in cassava

Evan M. Long[1]*, M. Cinta Romay[2], Guillaume Ramstein[3], Edward S. Buckler[1,2,4] and Kelly R. Robbins[1]

[1]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States, [2]Institute for Genomic Diversity, Cornell University, Ithaca, NY, United States, [3]Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark, [4]United States Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, United States

OPEN ACCESS

EDITED BY
Xiaofei Zhang,
International Center for Tropical Agriculture (CIAT), Colombia

REVIEWED BY
Tong Wei,
Beijing Genomics Institute (BGI), China
Zhiqiang Xia,
Hainan University, China

*CORRESPONDENCE
Evan M. Long
eml255@cornell.edu

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 11 September 2022
ACCEPTED 06 December 2022
PUBLISHED 09 January 2023

CITATION
Long EM, Romay MC, Ramstein G, Buckler ES and Robbins KR (2023) Utilizing evolutionary conservation to detect deleterious mutations and improve genomic prediction in cassava. *Front. Plant Sci.* 13:1041925. doi: 10.3389/fpls.2022.1041925

Introduction: Cassava (Manihot esculenta) is an annual root crop which provides the major source of calories for over half a billion people around the world. Since its domestication ~10,000 years ago, cassava has been largely clonally propagated through stem cuttings. Minimal sexual recombination has led to an accumulation of deleterious mutations made evident by heavy inbreeding depression.

Methods: To locate and characterize these deleterious mutations, and to measure selection pressure across the cassava genome, we aligned 52 related Euphorbiaceae and other related species representing millions of years of evolution. With single base-pair resolution of genetic conservation, we used protein structure models, amino acid impact, and evolutionary conservation across the Euphorbiaceae to estimate evolutionary constraint. With known deleterious mutations, we aimed to improve genomic evaluations of plant performance through genomic prediction. We first tested this hypothesis through simulation utilizing multi-kernel GBLUP to predict simulated phenotypes across separate populations of cassava.

Results: Simulations showed a sizable increase of prediction accuracy when incorporating functional variants in the model when the trait was determined by<100 quantitative trait loci (QTL). Utilizing deleterious mutations and functional weights informed through evolutionary conservation, we saw improvements in genomic prediction accuracy that were dependent on trait and prediction.

Conclusion: We showed the potential for using evolutionary information to track functional variation across the genome, in order to improve whole genome trait prediction. We anticipate that continued work to improve genotype accuracy and deleterious mutation assessment will lead to improved genomic assessments of cassava clones.

KEYWORDS

genetic load, deleterious mutation, cassava (*Manihot esculenta*), genomic prediction, evolutionary conservation

1 Introduction

Cassava (Manihot esculenta) is a root crop that is clonally propagated and grown widely in the tropical regions of Africa, Asia, and South America. It is estimated that cassava is a major caloric source for almost half a billion people around the world (Parmar et al., 2017; Ferguson et al., 2019). Although it is naturally an outcrossing perennial, it has been clonally propagated and grown as an annual since its domestication between 5,000-10,000 years ago (Wang et al., 2014). During the colonial era it was also brought to Africa, where today it is valued for its ability to grow with minimal inputs in marginally fertile lands.

Many generations of clonal propagation have caused cassava to accumulate genetic load that inhibits its potential crop performance. This genetic load is most apparent in the heavy inbreeding depression exhibited in cassava, as observed through low performance of selfed offspring (Rojas et al., 2009; de Freitas et al., 2016). Studies have shown that this genetic load is present as deleterious recessive mutations that are masked by heterozygosity which can be maintained through the clonal propagation (Ramu et al., 2017). With minimal sexual reproduction these deleterious mutations are maintained (McKey et al., 2010) and inhibit current breeding efforts to improve cassava performance (de Freitas et al., 2016).

Plant breeders have worked on various methods to detect and manage genetic load throughout history. Many crop species exist as polyploids, which enables them to more easily mask recessive deleterious mutations responsible for genetic load (van de Peer et al., 2021). Hybrid crop breeding has been another common method of applying strong selection pressures by selecting on inbred lines (Labroo et al., 2021), eliminating the possibility of recessive deleterious mutations. Some crops with similar high inbreeding depression to cassava, like potato, have made recent efforts to breed with inbred diploids (Bachem et al., 2019), however the deleterious mutations targeted by this methodology reduce plant viability.

During the past decade, plant breeders have seen the emergence of methodical application of genotyping and genomic selection as a method to improve breeding selections and leverage understanding of genomic information. Genomic selection, which uses genome markers and a phenotyped training population to predict unobserved offspring performance, can decrease selection cycle time and improve selection accuracy. Efforts have been made to improve genomic selection by using causative knowledge, however understanding the true causative elements in the genome is not a trivial exercise. Many studies have shown that benefits from including genome-wide association (GWA) hits in genomic prediction can diminish when predicting unrelated material (Cheruiyot et al., 2022), indicating population specific quantitative trait locus (QTL) or a misinterpretation of a variant as causative, when it is only in high linkage disequilibrium (LD) with the causative variant (Cheruiyot et al., 2022). For cassava, an ideal genomic annotation would explain underlying causative elements, while being consistent across populations structures.

Regarding genetic load, evolutionary conservation has shown to be an effective method to assess deleterious mutations and explain functional variation (Xiang et al., 2019) in a population agnostic manner. Multiple studies in crops such as maize (Yang et al., 2016; Ramstein and Buckler, 2022), sorghum (Valluru et al., 2019; Lozano et al., 2021), and barley (Kono et al., 2019) have demonstrated potential benefits for detecting and using deleterious mutations in genomic prediction. The potential benefit of understanding these deleterious mutations in cassava will be limited by the absolute number of mutations and how much variation of agronomic traits they each explain.

The purpose of this study is first, to identify likely deleterious mutations in cassava, and second to evaluate their potential impact on genomic prediction for the goal of improving future breeding selections. We sequenced, assembled, and gathered 52 genomes from species that all shared ancestry within the last 50 million years in order to score conservation and detect deleterious mutations.

We designed an experiment that uses evolutionary information to augment genomic predictions within and across two different populations of 1048 cassava clones present in two different breeding programs in Sub-Saharan Africa, the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria, and the National Crops Resources Research Institute (NaCRRI), Namulonge, Uganda. By performing phenotype simulations using real genotypic data and generating genomic predictions with known, simulated QTL, we first evaluated the best possible benefit of including causative information in our genomic predictions under different scenarios. We then used genomic and phenotypic data from these cassava clones to test genomic predictions, while including various functional annotations based on deleterious mutations.

2 Results

2.1 Evolutionary conservation

Utilizing many germplasm resources, we sampled, sequenced and assembled 27 Euphorbiaceae species (Supplementary Table 1). These assemblies were combined with available genome from Euphorbiaceae and other related species to form a set of 53 species, including cassava. We obtained multiple sequence alignments from for each gene, requiring transcript alignment of ≥90% of length of the cassava gene. Only the best matching ortholog from each species was retained and, of the ~26k genes examined, 24565 genes had ≥4 orthologs, allowing them to be scored for evolutionary conservation using PAML's baseml tool. Over half of all base pairs across these genes have an alignment depth of ≥31 species (Figure 1). The large number of aligned orthologs from the many species to measure conservation is benefited from sampling species from within shorter evolutionary time, although it is limited by poorer gene reconstruction in assemblies from short-read sequence.

2.2 Deleterious mutations

We used evolutionary conservation and predicted protein mutation effects to classify the deleterious effects of 66k nonsynonymous SNPs segregating in the two target populations. Firstly, we used the intersection of baseml evolutionary rate and SIFT deleterious scores to classify 2,210 deleterious sites that are segregating in both cassava populations (Figure 2). While both methods rely on evolutionary information, the high coincidence of low evolutionary rate and low SIFT score support their signal for functionally important sites in the genome. Deleterious burden for each clone was then calculated as the number of derived alleles at these sites. We separated this deleterious burden into homozygous and heterozygous genetic load. Genome wide association for all nonsynonymous sites as well as the deleterious sites was performed on fresh root yield and dry matter percentage traits, and some loci passed Bonferroni significance testing for fresh root yield (Supplementary Figures 4, 5). Secondly, we leveraged a RandomForest prediction model to weight the functional importance of the nonsynonymous mutations. This prediction produces a score between 0-1, a quantitative weight for the functional importance of each amino acid residue altered by mutations at the nonsynonymous sites (Figure 3).

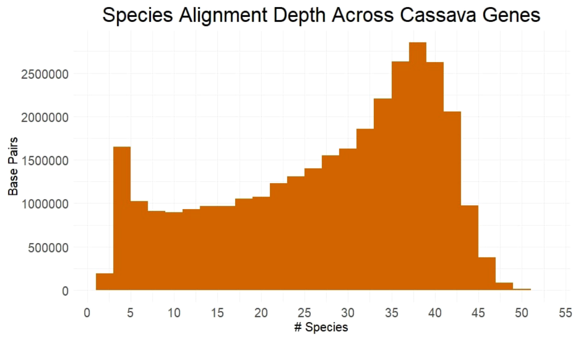


FIGURE 1
Species Alignment Depth Across Cassava Genes. Alignment depth represented by the number of species with homologous alleles in each multiple sequence alignment at any given protein coding base pair in the cassava genome.

2.3 Phenotype simulation

To validate our methodology and guide our expectations we performed genomic predictions using simulated phenotypes on 1048 cassava clones originating from IITA and NaCRRI breeding programs. These simulations represent some best-case scenarios for genomic prediction, where all QTL and their effect sizes are known.

The simulated QTL effects represent a suite of different genetic architectures ranging from highly complex genetic traits controlled by thousands of small effect QTL to oligogenic traits controlled by a handful of large effect QTL. These genetic architectures are represented by the proportion of the 66k variants simulated as causative QTL (Figure 4). These 66k variant sites were selected using nonsynonymous sites that showed high conservation (low evolution rate) from baseml. We modeled a range of dominance levels at each QTL in order to match our empirical scenario more closely in cassava (Supplementary Figure 1), where genetic load due to recessive deleterious alleles are expected to affect many agronomic, fitness related, traits (Bosse et al., 2019).

2.4 Genomic prediction with simulated phenotypes

Once QTL effects were modeled, we then calculated phenotypes for each of the 1048 clones (Supplementary Figure 2), where a positive effect is attributed to the ancestral allele. To Evaluate the effect of QTL structure, prediction model, and population, we performed genomic predictions. For all predictions in this study, we performed cross-population and within-population predictions designated as follows: IITA cross-validation (IITA_CV), NaCRRI cross-validation (NaCRRI_CV), Training with the IITA population and predicting in the NaCRRI population (IITA->NaCRRI), and "Training with the NaCRRI population and predicting in the IITA population (IITA->NaCRRI). Cross-population prediction accuracy is calculated by masking all phenotypes in one population and predicting using the other, then calculating the correlation between the true phenotype and the predicted phenotype. Within-population prediction accuracy is calculated similarly, using a 10-Fold prediction scheme where phenotypes in 10% of a population are masked and predicted by the other 90%.

We saw a marked increase in prediction accuracy when including the QTL information into the prediction model only when the trait was controlled by less than around 100 QTL (Figures 5C, D). Complex traits that are controlled by many small effect QTL across the genome show no increase in prediction accuracy with the inclusion of causative information (Figures 5A, B). For traits with an intermediate number of QTL (Figure 5C), the improvements in prediction accuracy are further increased by weighting the QTL information by their relative effect sizes. While the improvements are visible in both cross-population

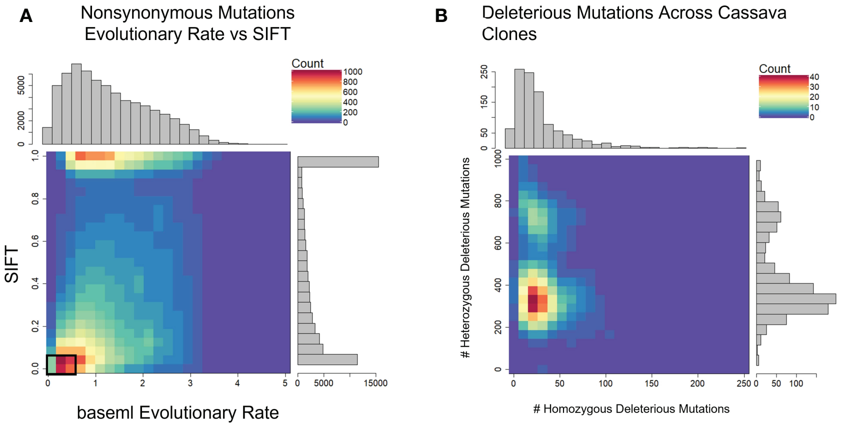


FIGURE 2
Defining Deleterious Mutations. **(A)** baseml evolutionary rate is plotted against SIFT scores. Deleterious mutations were classified as derived alleles at those sites with a baseml evolutionary rate < 0.5 and a SIFT score < 0.05 (Black box). **(B)** Distribution of homozygous and heterozygous deleterious mutations across 1048 cassava clones.

and within- population predictions, the improvements show some evidence of being more pronounced in cross-populations scenarios. These simulations show that even with perfect knowledge of QTL effects, improvements in prediction accuracy from using this information are limited by the relative abundance of those QTL.

2.5 Genomic prediction utilizing functional annotation

With deleterious mutations and functional weights for the segregating nonsynonymous sites, we mirrored the genomic predictions that we previously performed using simulated phenotypes, only this time using real data collected on the 1048 cassava clones.

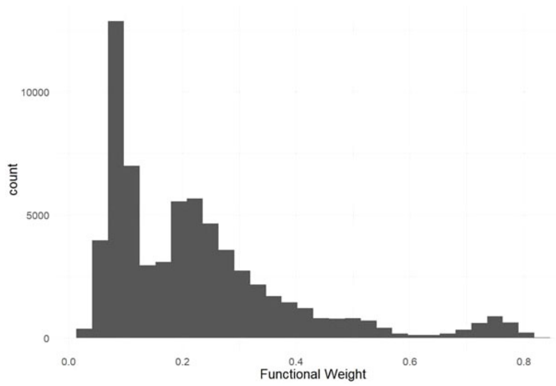


FIGURE 3
Predicted Functional Weights. Histogram of functional weights produced through RandomForest prediction of conservation for nonsynonymous variant sites. High functional weights correspond to highly conserved sites where nonsynonymous mutations are predicted to have large functional effects.

We predicted two different traits common in cassava breeding trials, fresh root yield and dry matter percentage, using the same cross-population and within-population scenarios previously shown. Multiple genomic prediction models were tested to evaluate the value of including the functional annotations.

Our two examples of a baseline prediction, where no functional information is present, are genomic prediction using the input marker data set and a genome-wide imputed dataset. In predicting fresh root yield, our results show that imputation alone does not improve cross-population prediction accuracy, however it does show some positive effect on within-population prediction (Figure 6). However, when including only imputed, segregating, non-synonymous variants, the prediction accuracy in cross-population predictions does increase over the two baseline models. Finally, we observed a further increase in prediction accuracy when weighting the non-synonymous variants and including derived genetic load from the deleterious mutations for both the cross-population predictions of fresh root yield and for within-population predictions in among the NaCRRI clones (Figure 6; Supplmentary Figure 6). For genomic prediction of cassava tuber dry matter percentage, we observed mostly negative or neutral effects of imputation and inclusion of deleterious annotations (Figure 7; Supplementary Figure 7). The improvements from functional information in predicting fresh root yield suggest it is correlated with fitness signals captured by the evolutionary information, while dry matter percentage may represent different, historical selection pressures.

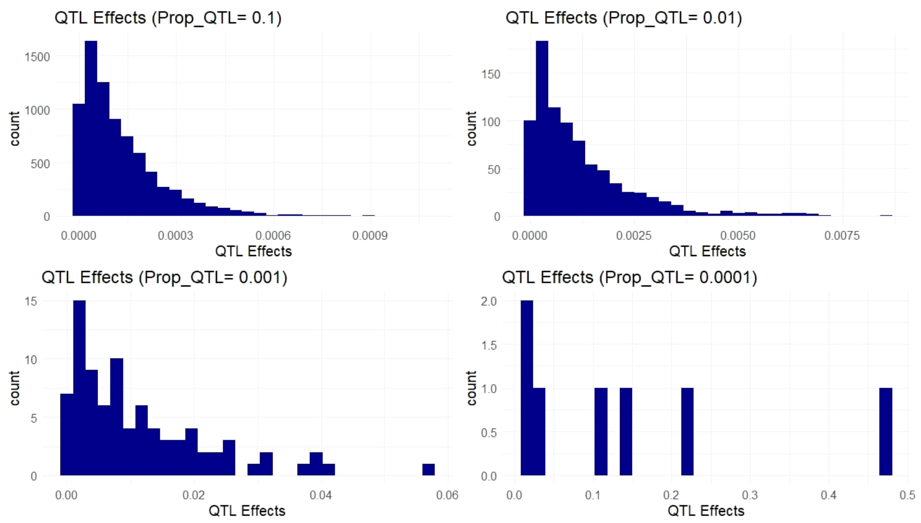


FIGURE 4
Simulated QTL Effects. Histograms show count of QTL effects in one example simulation. Each facet shows a genetic architecture with different proportions of the markers acting as QTL (resulting in ~ 6600, 660, 66, and 6 QTL on average). The x-axis represents the positive effect of carrying the ancestral allele at a given QTL.

3 Discussion

Genetic load, as defined as the accumulation of deleterious mutations through domestication, drift, mutation-selection balance and other means, has been identified as an impediment to the genetic value of a crop (Agrawal and Whitlock, 2012; Smýkal et al., 2018). Through simulation, we explored the possible scenarios in which knowing the exact deleterious mutations could improve breeding selections. In this study, we went on to use evolutionary conservation and genomic information to quantify deleterious mutations in cassava clones, as well as predict their potential effects.

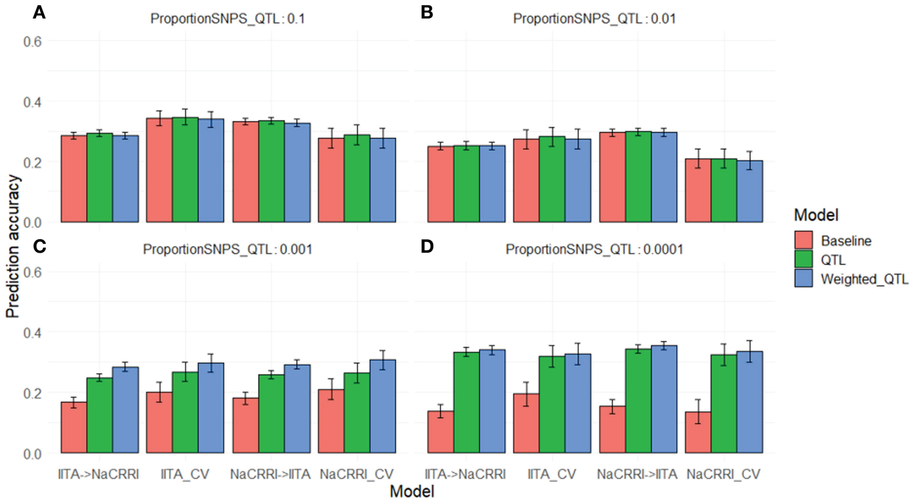


FIGURE 5
Genomic Prediction Accuracies with Simulated QTL. Prediction accuracies are shown on the y-axis as the correlation between predicted andtrue breeding values. The x-axis delineates the prediction scenario being tested. Barplot color corresponds to the genomic information used in the prediction model. Error bars represent a 95% confidence interval for simulations. Simulations were repeated with different proportions of the markers acting as causative QTL: 0.1 **(A)**, 0.01 **(B)**, 0.001 **(C)**, and 0.0001 **(D)**.

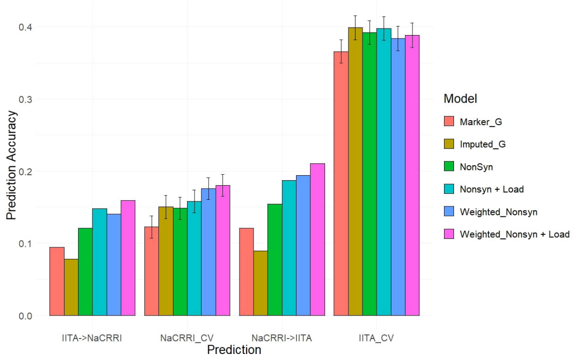


FIGURE 6
Fresh Root Yield Genomic Prediction Leveraging Deleterious Annotations. Prediction accuracy is measured in cross-population and within-population prediction scenarios. Genomic models are represented as bar graph colors where various genomic and deleterious data are used in the genomic prediction. Error bars represent a 95% confidence interval for within-population 10-fold prediction.

3.1 Simulation informs genomic prediction potential

The simulation of phenotypes under differing genetic architectures allowed us to manage expectations for the best possible scenarios in which understanding the causative variation of a trait could help inform genomic selection decisions. As we only observed benefits to genomic prediction under scenarios with<~100 QTL, it is clear that LD structure captured by genome wide markers is sufficient for genomic prediction under highly complex genetic architectures (Figure 5). The scenario with the fewest QTL (<~10) represents a more Mendelian or oligogenic architecture, which might benefit more from a marker assisted selection methodology, but it follows that traits with higher effect sizes of QTL will see more improvements from causative knowledge in genomic prediction. Interestingly, within-population predictions showed smaller, but still substantial benefits in genomic prediction accuracy. These results indicate that our empirical predictions have the potential to benefit from deleterious mutation annotations, only if there are a few or intermediate number of QTL (<~100) with substantial effects. Importantly, the expected benefits shown through simulations depend directly upon the population and LD structure in our tested clones and cannot necessarily be useful to interpret potential benefits in other scenarios.

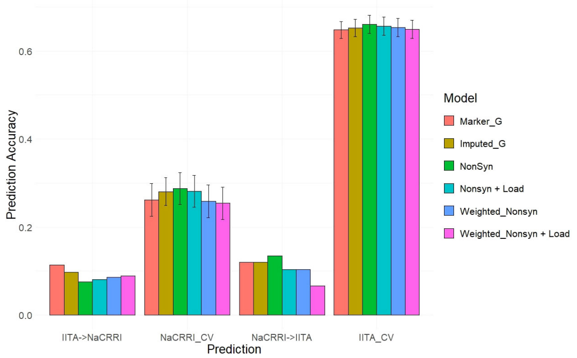


FIGURE 7
Dry Matter Percentage Genomic Prediction Leveraging Deleterious Annotations. Prediction accuracy is measured in cross-population and within-population prediction scenarios. Genomic models are represented as bar graph colors where various genomic and deleterious data are used in the genomic prediction. Error bars represent a 95% confidence interval for within-population 10-fold prediction.

3.2 Evolution conservation reveals deleterious mutations

We used evolutionary conservation and protein annotations to classify certain mutations as deleterious. By aligning over 50 species of relatively recent ancestry, we were able to assess the conservation status of a large majority of the cassava genome. We used separate neutral trees for each gene, rather than the entire chromosome or species, to address the difference between gene ancestry common in plants due to historical gene and genome duplication. Because of the millions of years of evolution, it is very difficult to predict the sizes of selection coefficients from evolutionary conservation alone (Huber et al., 2020). We then needed predicted protein effects of these mutations from SIFT to refine our set of putatively deleterious mutations. After defining our deleterious alleles, we separated the assessment of deleterious load into homozygous and heterozygous, because most deleterious mutations are assumed to be recessive (Bosse et al., 2019) and cassava has been shown to mask deleterious mutations through heterozygosity (Ramu et al., 2017). These assessments of genetic load are at least partially validated by a negative correlation (R=-0.18) between plant yield and homozygous deleterious mutations (Supplementary Figure 3).

As previously mentioned, evolutionary conservation alone cannot easily resolve effect sizes of mutations. For this reason, we used protein perturbation information from SIFT and UniRep to prioritize functional variants similar to work recently done in Maize (Ramstein and Buckler, 2022). Another advantage of this weighting method is that it does not imply a directional effect of the mutations, thereby allowing for potential positive or adaptive effects (Loewe and Hill, 2010) of derived mutations at conserved sites. While most derived alleles at conserved positions are predicted to be deleterious, these derived alleles could represent directed selection from domestication or adaptive evolution specific to cassava.

3.3 Leveraging functional data in genomic prediction

The inclusion of deleterious and functional mutations derived from evolutionary conservation showed promising value in informing the genetic value of cassava clones. Our results displayed improvements for cross-population predictions of fresh root yield as well as some of the within-population predictions in NaCRRI (Figure 6). This follows with the understanding that total plant growth, and even root yield, are correlated with total plant fitness (Pan and Price, 2001), while root dry matter percentage, which is primarily a quality trait, likely has little direct correlation with evolutionary fitness (Figure 7). We expect this trend would continue for other traits; however, few traits are measured identically across multiple populations.

In this study, we used multi-kernel GBLUP methods of genomic prediction to partition the additive and dominant genetic effects, while substituting unweighted and weighted genomic relationship matrices formed from subsets of the genomic data. These methodologies rely on the assumption that our selected functional variants, and the weights prescribed to them, are derived from a separate, and more functional, distribution of effects from a default, genome-wide relationship. Other methods, including Bayesian models, exist to prioritize functional information in genomic prediction, however multiple studies have found it to be difficult to prescribe consistent, significant differences in prediction accuracy results between them and GBLUP models, and the specific benefit of one method or the other are often situational (Moghaddar et al., 2019; Khansefid et al., 2020; Cheruiyot et al., 2022).

3.4 Reflections on load

In an effort to improve cassava's role as a reliable food source around the world, our results show the importance and potential of addressing the impact of genetic load. We used evolution and protein annotations to determine these deleterious mutations responsible for genetic load. It is important to note that, while the methods used in this study detected impactful deleterious variation across the genome, they ignore the many deleterious mutations likely found in regulatory regions of the genome.

The improvements made in genomic prediction validate the effects of these deleterious mutations and offer one possible avenue for their potential application. As observed in the within-population prediction of IITA, where prediction accuracy is higher and unaffected by our annotations, the application of this understanding of genetic load may not be beneficial in every breeding scenario, however cross-population prediction is not the only instance where deleterious information may prove informative. Rapid cycle recurrent selection, where generations of selection occur without phenotyping, could be another situation in which tracking functional information across the genome could improve genomic selection decisions. As generations of selection occur, linkage disequilibrium between causative mutations and genome-wide markers breaks down, making the functional tracking of causative effects more impactful in prediction.

In addition to genomic prediction scenarios, the understanding of the deleterious mutations responsible for genetic load in cassava could suggest alternative methods for crop improvement. Many crops today utilize hybrid breeding, where multiple groups of inbred parents are bred for use in creating a superior hybrid. Selecting on inbred individuals exposes recessive, or partially recessive, deleterious mutations, allowing them to be effectively purged in fewer generations. While difficulties due to severe inbreeding depression in cassava have hindered this genre of breeding, efforts being made in crops like potato show it's potential in a crop burdened by heavy genetic load (Bachem et al., 2019). Doubled haploidization has been a common tool in some inbred crops, while historically difficult to implement in some crops like cassava, however newer implementations such as those reported from ScreenSys (https://www.screensys.eu) offer a possible method of producing enough viable embryos for crops with heavy inbreeding depression like cassava. (Nasti and Voytas, 2021). With the understanding of the extent to which deleterious mutations account for missed potential in cassava performance, further consideration for how to effectively purge genetic load will be needed.

Historical evolution and population genetics continues to shed light on our understanding of genomic functions, as seen in our study in cassava. We showed the utility of using evolutionary derived deleterious mutations to improve genomic prediction across cassava populations. Additionally, the genetic load was identified from<~100 homozygous deleterious mutations per clone (Figure 2). This number of mutations could be the target of further improvement through gene editing or other means. In the future, as genome sequencing accelerates, coupled with our understanding of protein functions, we may be able to make targeted decisions to purge genetic load from cassava and advance genetic gains.

4 Methods

4.1 Euphorbiaceae sequencing & assembly

We gathered a total of 52 related species, 26 of which we sequenced and assembled, to evaluate evolutionary conservation across the cassava genome. In order to maximize the amount of evolutionary time sampled, while maintaining reliable alignments to cassava, we sampled 26 species across the Euphorbiaceae family, to which cassava belongs. We then sequenced these species using Illumina NovaSeq-6000. Genome sizes were estimated using k-mer spectra in order to estimate sequence input coverage for assembly (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/). Additional short-read sequences were downloaded from SRA corresponding to 11 unspecified Euphorbiaceae taxa (Liu et al., 2019). We then used a short-read sequence assembler MEGAHIT (Li et al.,), with modified parameters of "-m 0.2 -t 10 –no-mercy –min-count 3 –k-min 31 –k-step 20" to create contig assemblies. We additionally obtained long-read sequences using PacBio Sequel II for 7 species among our sampled Euphorbiaceae taxa. These sequences were assembled using Hifiasm (<xr rid="r6">Cheng et al., 2021</xr>) utilizing default settings. An additional 15 genome assemblies from other related species were downloaded from SRA and added to our assembled genomes resulting in a total of 52, excluding cassava (Supplementary Table 1).

4.2 Sequence alignment and evolutionary conservation

We used gene alignments from Cassava V7.1 gene annotations to the 52 species to extract homologous gene sequences for multiple sequence alignment. Gene transcripts were aligned using minimap2, and the best aligned region with >= 90% alignment length matching was retained as homologous coding sequences for each species were then extracted and aligned using MAFFT (Katoh et al., 2002) multiple sequence alignment. With a multiple sequence alignment for each gene, we then generated gene trees using RAxML (Stamatakis, 2014), and calculated evolutionary rates using baseml from the PAML (Yang, 2007) suite of tools. We then identified ancestral alleles at every site across the genic regions of the genome, using the ancestral node containing Manihot, Hevea, and Cnidoscolus genera. We used evolutionary conservation to select representative gene models for each gene, as well as only retaining genes with 5' and 3' untranslated regions annotated resulting in ~25k genes models.

4.3 Deleterious mutations

We used evolutionary conservations & protein structure conservation to identify deleterious mutations and produce weights for functional importance of sites across the cassava Genome. Deleterious mutations were categorized as sites with a baseml evolutionary rate of <0.5 and a "Sorting Intolerant From Tolerant" (SIFT) score of < 0.05 (Ng and Henikoff, 2003). Additionally, we required deleterious sites to have < 20% minor allele frequency in the cassava HapMap (Ramu et al., 2017) (Figure 2).

In addition to identifying a binary classification of deleterious, we used a RandomForest model to obtain a quantitative prediction of conservation similar to a previously reported method reported (Ramstein and Buckler, 2022). We used baseml evolutionary rates to classify nonsynonymous sites as either conserved (evolutionary rate< 0.3) or non-conserved (evolutionary rate > 2), while sites with values outside these ranges were excluded from model training. SIFT, UniRep, and 100bp windowed GC% totaling ~500 predictors in the RandomForest model implemented by the R package "ranger" (Wright and Ziegler, 2017). From the SIFT database, we used both the mutation type and SIFT score, which gives the predicted deleterious effect of a base-pair substitution. UniRep is a deep learning technique which characterizes protein structure (Alley et al., 2019), which we used to produce 256-unit representations of each protein and its associated mutated forms (https://github.com/churchlab/UniRep).

To increase the number of observations in the model, we used both the known HapMap mutations and *in silico* non-synonymous mutations at every possible site in our gene models. This resulted in over 1 million non-synonymous mutations whose genomic conservation could be modeled. We then used a leave-one-out prediction scheme where each of the 18 cassava chromosomes was left out of model training and predicted by the other 17. This method produced a predicted value between 0-1 for each of the ~66k nonsynonymous, segregating mutations used in this study (Figure 3).

4.4 Phenotypic & genotypic data

Phenotypic and genotypic data for 1048 cassava clones were downloaded from *cassavabase.org* representing two populations of breeding lines. The first population is from a breeding program at International Institute of Tropical Agriculture (IITA) in Nigeria, while the second is from a breeding program at National Crops Resources Research Institute National Crops Resources Research Institute (NaCRRI) in Uganda, representing breeding material for West and East Africa, respectively. Genotypes for the associated clones were downloaded from the "*East Africa Clones Dart-GBS 2020*" genotyping protocol on cassavabase.org containing 23,431 variants. Plant phenotypes for fresh root yield and dry matter percentage were downloaded from *cassavabase.org* and prepared according to previously described methods (https://wolfemd.github.io/GenomicSelectionManual/index.html).

We then performed genotype imputation using the cassava haplotype map using Beagle5 (Browning et al., 2018), with an **Ne=100,** resulting in ~26M variants. These variants were then filtered down to two genome-wide marker sets, one being a thinned sample of ~135k genome-wide SNPs, and the other being all non-synonymous sites segregating in both populations resulting in ~66k genome-wide variants. The input marker genotypes, the imputed sample, and the imputed non-synonymous sites will be used in genomic prediction analyses.

4.5 Causative variation simulation

We used quantitative trait loci (QTL) simulation, replicated 50 times, to model the potential benefits of knowing causative variants in genomic prediction. This simulation begins by sampling QTL across the 66K variant sites from a binomial distribution with the probability of being a QTL varied across possible values of 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}. The effect sizes for these QTL were then sampled from a gamma distribution using the *rgamma* function in R, with the shape parameter=1, with the ancestral allele set as having a positive effect. Lastly a dominance effect for each QTL was sampled from normal distribution "rnorm(mean = 2,sd=0.3)", restricting to dominance<=2 (Supplementary Figure S1). Phenotypes were then generated for the 1048 cassava clones. Residuals were then simulated such that the trait had a heritability of approximately 0.3.

We performed cross-population and 10-Fold within-population predictions using the simulated data, with and without QTL information incorporated into the prediction model. Genomic prediction was performed by using GBLUP methods fit using ASReml, with additive and dominance effects modeled as separate kernels. For all models described, residuals are represented by ε and modeled as random with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\ \sigma_\varepsilon^2)$.

For prediction using simulated phenotypes, we compared three different models. The first model represents our baseline prediction:

$$y = 1\mu + Z_A a + Z_D d + \varepsilon$$

Where y is the simulated phenotype, μ is the phenotype mean, **a** is the vector of additive genetic effects, $\mathbf{Z_A}$is the incidence matrix, and $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G_A}\ \sigma_a^2)$, G_A is an additive genomic relationship matrix produced using the VanRaden (VanRaden, 2008) method, and σ_a^2 is the additive genetic variance.

$$G_A = \frac{MM'}{\sum_i^n (2p_i * (1 - p_i))}$$

Where M is the centered genotype matrix (where genotypes are stored as dosages of 0,1, and 2 referring to being homozygous for reference allele, heterozygous, and homozygous for the alternate allele, respectively) and p_i is and allele frequency at the i^{th} locus. $\mathbf{Z_D}$ and **d** are analogous to the additive method, with the exception that a dominance genomic relationship matrix is produced using the Nishio and Satoh (Nishio and Satoh, 2014) method.

$$G_D = \frac{DD'}{\sum_i^n (2p_i * (1 - p_i))^2}$$

Where the entries of D are given as $-2p_i^2$ for the homozygous reference allele, $2p_i*(1-p_i)$ for the heterozygote, and $2(1-p_i)^2$ for the homozygous alternate allele.

The second model includes additive and dominance QTL relationship matrices formed in identical manner to the $\mathbf{G_A}$ & $\mathbf{G_d}$ matrices, but only utilizing the known QTL sites in the genomic relationship matrices:

$$y = 1\mu + Z_{AQTL} aQTL + Z_{DQTL} dQTL + \varepsilon$$

The final model includes weighted QTL matrices based on their effect size:

$$y = 1\mu + Z_{AW} aw + Z_{DW} dw + \varepsilon$$

Here the weighted matrices are formed using modified methods of the previously cited methods. The weighted additive matrix given by:

$$G_{AW} = \frac{MWM'}{\sum_i^n (2p_i * (1 - p_i) * w_i)}$$

Where M is the scaled genotype matrix. W is a diagonal matrix with w_i along the diagonal, w_i and p_i are the weight and frequency for the i^{th} locus, respectively.

The weighted dominance matrix is modified in a similar fashion to the additive matrix:

$$G_{DW} = \frac{DWD'}{\sum_i^n (2p_i * (1 - p_i) * w_i)^2}$$

Where the entries of D are given as $-2p_i^2$ for the homozygous reference allele, $2p_i*(1-p_i)$ for the heterozygote, and $2(1-p_i)^2$ for the homozygous alternate allele.

4.6 Genomic prediction models in empirical data

The genomic prediction models used for real breeding program phenotypes follow a similar pattern to our simulated scenario, with a few notable differences.

First, our ground truth for the phenotype of each clone was the best linear unbiased estimate (BLUE) using a model like those previously used in cassava plot level traits (Wolfe et al., 2017) and those suggested for use with African cassava breeding data (https://wolfemd.github.io/GenomicSelectionManual/index.html):

$$y = X\beta + Z_{block(rep)} b + Z_{rep(trial)} t + \varepsilon$$

where y is the vector of the phenotype, β included a vector of fixed effects for the population mean, the location–year combination, the number of plants harvested per plot, and germplasm ID with design matrix **X**. Replications were nested in trials, treated as random, and represented by the design matrix $\mathbf{Z}_{\mathbf{rep(trial)}}$ and the effects vector **t**~N(**0**,**I** σ_t^2). Blocks were nested in replications, treated as random, and represented by the design matrix $\mathbf{Z}_{\mathbf{block(rep)}}$ and the effects vector **b**~N(**0**,**I** σ_b^2).

Having a ground truth phenotype, we then compared multiple different genomic prediction models to measure the potential benefits to including the deleterious annotations. Each model followed a similar form:

$$y = X\beta + Z_{\text{block(rep)}}b + Z_{\text{rep(trial)}}t + Z_A a + Z_D d + \varepsilon$$

This generic model mirrors the previous one, with the exception that germplasm ID is no longer treated as fixed but is instead $\mathbf{Z_A}$ and $\mathbf{Z_D}$are design matrices indicating observations of germplasm IDs for the vectors of additive and dominance effects **a** and **d**, modeled as previously described in the simulated scenario. The six models we compared involve substituting different markers and methods of constructing genomic relationship matrices for $\mathbf{Z_A}$ and $\mathbf{Z_D}$, as well as adding fixed effects for derived homozygous and heterozygous load. The six models include:

- *Marker_G* where the 23,431 variants are used to produce the genomic relationship matrices.
- *Imputed_G* where ~135k imputed genome-wide segregating sites are used to produce the genomic relationship matrices.
- *Nonsyn* where 66k imputed, segregating, nonsynonymous mutation sites are used to produce the genomic relationship matrices.
- *Nonsyn + Load* which is identical to *Nonsyn* with the exception of including the derived load as fixed effects in the prediction
- *Weighted_Nonsyn* uses the same sites as *Nonsyn*, however the genomic relationship matrices are created using the weighted method described previously, with the deleterious weights for each SNP.
- *Weighted_Nonsyn + Load* which is identical to the *Weighted_Nonsyn* with the exception of including the derived load as fixed effects in the prediction

Each model was evaluated by performing the cross-population and within-population predictions as previously described and using the correlation between predicted phenotype and the BLUE as the prediction accuracy (Figures 6, 7). Prediction accuracy was also calculated as the number of the top 25 performing clones predicted as being among the top 25 performing clones (Supplementary Figures S6, S7).

For all simulated scenarios and for empirical within population cross-validations, 95% confidence intervals were calculated. 10-fold cross validation predictions were replicated 30 times, and confidence intervals (CI) were calculated using R:

$$CI = \frac{SD}{sqrt(n)} * \; qt(p = 0.05/2,\; df = (n-1), lower\,.\,tail = F)$$

Where n= # folds * # replications and SD=standard deviation. A true confidence interval assumes observations are independent, which is not true for replications of cross-fold validation, however this gives an estimate for variability in cross-validation prediction accuracies.

4.7 Data availability

Genotype and Phenotype data used in this study is available at cassavabase.org. Euphorbiaceae sequence reads and assemblies generated in this study will be available under bioprojects PRJNA608937 on the Sequence Read Archives and PRJEB55682 on the European Nucleotide Archive, respectively. Code used to process data and produce assemblies, simulations, genomic predictions as well as deleterious weights and mutation results are available at https://bitbucket.org/bucklerlab/cassava_load_and_gp.

Data availability statement

Euphorbiaceae sequence reads and assemblies generated in this study will be available under bioprojects PRJNA608937 on the Sequence Read Archives (SRA) and PRJEB55682 on the European Nucleotide Archive (ENA), respectively.

Author contributions

EL - Collected samples, performed analysis, and did majority of manuscript writing. MR - Managed and organized germplasm collection and genome sequencing. EB - Mentor and oversaw experiments for measuring evolutionary conservation and deleterious mutations, reviewed and edited manuscript. KR - Mentor and oversaw experiments for genomic prediction and deleterious mutations impact on traits, reviewed and edited manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work is supported by workforce development fellowship Project: NYC-149949, Award: 2021-67034-34970 from the USDA National Institute of Food and Agriculture as well as start-up funds from the Robbins lab at Cornell. Additionally, this study is made possible by the funding and support of the USDA-ARS and the NextGen Cassava project, through the Bill & Melinda Gates Foundation (Grant INV-007637 http://www.gatesfoundation.org) and Commonwealth & Development Office (FCDO).

Acknowledgments

We would like to acknowledge the many germplasm sources that contributed tissue for sequencing (Supplementary Table 1) including: the Denver Botanic Garden, Germplasm Resources Information Network, the Missouri Botanic Garden, the Montgomery Botanic Garden, the National Botanic Garden, the National Tropical Botanic Garden, The New York Botanic Garden, and the US Botanic Garden. Their support was essential in sampling the vast number of species used in this study. We would also like to thank IITA and NaCRRI for contribution of data that we used to cassavabase. In particular, we thank Peter Kulakow, Ismail Rabbi, and Prasad Peteti who were project leads at IITA and Robert Kawuki, a project leader at NaCRRI, and Chiedozie Egesi, overall project manager of the NextGen Cassava Project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1041925/full#supplementary-material

References

Agrawal, A. F., and Whitlock, M. C. (2012). Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. *Annu. Rev. Ecol. Evol. Syst.* 43, 115–135. doi: 10.1146/annurev-ecolsys-110411-160257

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 16 (12), 1315–1322. doi: 10.1038/s41592-019-0598-1

Bachem, C. W. B., van Eck, H. J., and de Vries, M. E. (2019). Understanding genetic load in potato for hybrid diploid breeding. *Mol. Plant* 12, 896–898. doi: 10.1016/J.MOLP.2019.05.015

Bosse, M., Megens, H. J., Derks, M. F. L., de Cara, Á. M. R., and Groenen, M. A. M. (2019). Deleterious alleles in the context of domestication, inbreeding, and selection. *Evol. Appl.* 12, 6. doi: 10.1111/EVA.12691

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18 (2), 170–175. doi: 10.1038/s41592-020-01056-5

Cheruiyot, E. K., Haile-Mariam, M., Cocks, B. G., MacLeod, I. M., Mrode, R., and Pryce, J. E. (2022). Functionally prioritised whole-genome sequence variants improve the accuracy of genomic prediction for heat tolerance. *Genet. Sel. Evol.* 54, 1–18. doi: 10.1186/S12711-022-00708-8/FIGURES/4

de Freitas, J. P. X., da Silva Santos, V., and de Oliveira, E. J. (2016). Inbreeding depression in cassava for productive traits. *Euphytica* 209, 137–145. doi: 10.1007/s10681-016-1649-7

Ferguson, M. E., Shah, T., Kulakow, P., and Ceballos, H. (2019). A global overview of cassava genetic diversity. *PLoS One* 14, 1–16. doi: 10.1371/journal.pone.0224763

Huber, C. D., Kim, B. Y., and Lohmueller, K. E. (2020). Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PloS Genet.* 16, e1008827. doi: 10.1371/JOURNAL.PGEN.1008827

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/NAR/GKF436

Khansefid, M., Goddard, M. E., Haile-Mariam, M., Konstantinov, K., Schrooten, C., de Jong, G., et al. (2020). Improving genomic prediction of crossbred and purebred dairy cattle. *Front. Genet.* 11. doi: 10.3389/FGENE.2020.598580

Kono, T. J. Y., Liu, C., Vonderharr, E. E., Koenig, D., Fay, J. C., Smith, K. P., et al. (2019). The fate of deleterious variants in a barley genomic prediction population. *Genetics* 213, 1531–1544. doi: 10.1101/442020

Labroo, M. R., Studer, A. J., and Rutkoski, J. E. (2021). Heterosis and hybrid crop breeding: A multidisciplinary review. *Front. Genet.* 12. doi: 10.3389/FGENE.2021.643761

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly *via* succinct de bruijn graph. *Bioinformatics* 31 (10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Liu, H., Wei, J., Yang, T., Mu, W., Song, B., Yang, T., et al. (2019). Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the ruili botanical garden. *Gigascience* 8, 1–9. doi: 10.1093/GIGASCIENCE/GIZ007

Loewe, L., and Hill, W. G. (2010). The population genetics of mutations: Good, bad and indifferent. *Philos. Trans. R. Soc. B.: Biol. Sci.* 365, 1153–1167. doi: 10.1098/rstb.2009.0317

Lozano, R., Gazave, E., dos Santos, J. P. R., Stetter, M. G., Valluru, R., Bandillo, N., et al. (2021). Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nat. Plants* 7 (1), 17–24. doi: 10.1038/s41477-020-00834-5

McKey, D., Elias, M., Pujol, M. E., and Duputié, A. (2010). The evolutionary ecology of clonally propagated domesticated plants. *New Phytol.* 186, 318–332. doi: 10.1111/J.1469-8137.2010.03210.X

Moghaddar, N., Khansefid, M., van der Werf, J. H. J., Bolormaa, S., Duijvesteijn, N., Clark, S. A., et al. (2019). Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. *Genet. Sel. Evol.* 51, 72. doi: 10.1186/S12711-019-0514-2

Nasti, R. A., and Voytas, D. F. (2021). Attaining the promise of plant gene editing at scale. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2004846117. doi: 10.1073/PNAS.2004846117/ASSET/F8F17C7C-565B-4915-A746-0A024AC2A114/ASSETS/IMAGES/LARGE/PNAS.2004846117FIG02.JPG

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Nishio, M., and Satoh, M. (2014). Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One* 9 (1), e85792. doi: 10.1371/JOURNAL.PONE.0085792

Pan, J. J., and Price, J. S. (2001). Fitness and evolution in clonal plants: The impact of clonal growth. *Evol. Ecol.* 15 (4), 583–600. doi: 10.1023/A:1016065705539

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

Ramstein, G. P., and Buckler, E. S. (2022). Prediction of evolutionary constraint by genomic annotations improves prioritization of causal variants in maize. *bioRxiv* 2021, 9.03.458856. doi: 10.1101/2021.09.03.458856

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi, C., Bredeson, J., et al. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* 49, 959–963. doi: 10.1038/ng.3845

Rojas, M. C., Pérez, J. C., Ceballos, H., Baena, D., Morante, N., and Calle, F. (2009). Analysis of inbreeding depression in eight s $_1$ cassava families. *Crop Sci.* 49, 543–548. doi: 10.2135/cropsci2008.07.0419

Smýkal, P., Nelson, M. N., Berger, J. D., and von Wettberg, E. J. B. (2018). The impact of genetic changes during crop domestication. *Agronomy* 8, 119. doi: 10.3390/AGRONOMY8070119

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312. doi: 10.1093/BIOINFORMATICS/BTU033

Valluru, R., Gazave, E. E., Fernandes, S. B., Ferguson, J. N., Lozano, R., Hirannaiah, P., et al. (2019). Deleterious mutation burden and its association with complex traits in sorghum (Sorghum bicolor). *Genetics* 211, 1075–1087. doi: 10.1534/GENETICS.118.301742

van de Peer, Y., Ashman, T. L., Soltis, P. S., and Soltis, D. E. (2021). Polyploidy: An evolutionary and ecological force in stressful times. *Plant Cell* 33, 11–26. doi: 10.1093/PLCELL/KOAA015

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/JDS.2007-0980

Wang, W., Feng, B., Xiao, J., Xia, Z., Zhou, X., Li, P., et al. (2014). Cassava genome from a wild ancestor to cultivated varieties. *Nat. Commun.* 5, 5110. doi: 10.1038/ncomms6110

Wolfe, M. D., del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., et al. (2017). Prospects for genomic selection in cassava breeding. *Plant Genome* 10, plantgenome2017.03.0015. doi: 10.3835/plantgenome2017.03.0015

Wright, M. N., and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *J. Stat. Softw.* 77, 1–17. doi: 10.18637/JSS.V077.I01

Xiang, R., van den Berg, I., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci.* 116, 19398–19408. doi: 10.1073/pnas.1904159116

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/MOLBEV/MSM088

Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E., Guill, K., McMullen, M., et al. (2016). Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLOS Genetics* 13 (9), e1007019. doi: 10.1101/086132

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 23 January 2023
DOI 10.3389/fpls.2023.1089759

OPEN ACCESS

EDITED BY
Jianfeng Zhou,
University of Missouri, United States

REVIEWED BY
Kusumiyati Kusumiyati,
Padjadjaran University, Indonesia
Karansher Singh Sandhu,
Bayer Crop Science, United States

*CORRESPONDENCE
Eder Jorge de Oliveira
eder.oliveira@embrapa.br

†PRESENT ADDRESS
Luciano Rogerio Braatz de Andrade,
Section on Plant Breeding and Genetics,
School of Integrative Plant Sciences,
Cornell University, Ithaca, NY, United States

SPECIALTY SECTION
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

RECEIVED 04 November 2022
ACCEPTED 09 January 2023
PUBLISHED 23 January 2023

CITATION
Sousa MB, Filho JSS, de Andrade LRB and de Oliveira EJ (2023) Near-infrared spectroscopy for early selection of waxy cassava clones via seed analysis.
Front. Plant Sci. 14:1089759.
doi: 10.3389/fpls.2023.1089759

Near-infrared spectroscopy for early selection of waxy cassava clones via seed analysis

Massaine Bandeira e Sousa[1], Juraci Souza Sampaio Filho[2], Luciano Rogerio Braatz de Andrade[1†] and Eder Jorge de Oliveira[1*]

[1]Embrapa Mandioca e Fruticultura, Cruz das Almas, Bahia, Brazil, [2]Universidade Federal do Recôncavo da Bahia, Cruz das Almas, Bahia, Brazil

Cassava (*Manihot esculenta* Crantz) starch consists of amylopectin and amylose, with its properties determined by the proportion of these two polymers. Waxy starches contain at least 95% amylopectin. In the food industry, waxy starches are advantageous, with pastes that are more stable towards retrogradation, while high-amylose starches are used as resistant starches. This study aimed to associate near-infrared spectrophotometry (NIRS) spectra with the waxy phenotype in cassava seeds and develop an accurate classification model for indirect selection of plants. A total of 1127 F_2 seeds were obtained from controlled crosses performed between 77 F_1 genotypes (wild-type, *Wx_*). Seeds were individually identified, and spectral data were obtained *via* NIRS using a benchtop NIRFlex N-500 and a portable SCiO device spectrometer. Four classification models were assessed for waxy cassava genotype identification: k-nearest neighbor algorithm (KNN), C5.0 decision tree (CDT), parallel random forest (parRF), and eXtreme Gradient Boosting (XGB). Spectral data were divided between a training set (80%) and a testing set (20%). The accuracy, based on NIRFlex N-500 spectral data, ranged from 0.86 (parRF) to 0.92 (XGB). The Kappa index displayed a similar trend as the accuracy, considering the lowest value for the parRF method (0.39) and the highest value for XGB (0.71). For the SCiO device, the accuracy (0.88–0.89) was similar among the four models evaluated. However, the Kappa index was lower than that of the NIRFlex N-500, and this index ranged from 0 (parRF) to 0.16 (KNN and CDT). Therefore, despite the high accuracy these last models are incapable of correctly classifying waxy and non-waxy clones based on the SCiO device spectra. A confusion matrix was performed to demonstrate the classification model results in the testing set. For both NIRS, the models were efficient in classifying non-waxy clones, with values ranging from 96–100%. However, the NIRS differed in the potential to predict waxy genotype class. For the NIRFlex N-500, the percentage ranged from 30% (parRF) to 70% (XGB). In general, the models tended to classify waxy genotypes as non-waxy, mainly SCiO. Therefore, the use of NIRS can perform early selection of cassava seeds with a waxy phenotype.

KEYWORDS

amylopectin, amylose, classification models, Manihot esculenta Crantz, portable NIRS

1 Introduction

Cassava (*Manihot esculenta* Crantz) is one of the most accessible and consumed sources of carbohydrates, being widely used as processed products and in its natural form as animal and human food. In Brazil, cassava has recently increased its value due to its different available applications, especially in the food industry. Starch is the main storage carbohydrate in plants, with its biosynthesis occurring in seeds, tubers, fruits, roots, and leaves. It is essential not only in the life cycle of plants but also in human nutrition as it provides large amounts of energy (Li et al., 2019). Along with corn, potato, wheat, and rice, cassava is one of the main commercial sources of starch globally (Agama-Acevedo et al., 2019).

Cassava starch comprises two types of glucose polymers, amylose and amylopectin, whose composition ranges from 15–27% amylose, with an average of 21% (Sánchez et al., 2009; Santos et al., 2021). Waxy starch comprises at least 95% amylopectin, and this is associated with certain advantages, including less starch retrogradation and syneresis from starch pastes during freeze/thaw cycles; this prevents the reduction of sensory quality and shelf life of processed foods (Demiate and Kotovicz, 2011; Wang et al., 2015; Morante et al., 2016). The waxy starches of roots and tubers, such as cassava and potato, compared to cereal waxy starches provide clearer gels, with a mild or neutral flavor (Koehorst-van Putten et al., 2012), and different, higher viscosity gel textures (Sánchez et al., 2010). Additionally, they are used in food products, such as nuggets, to provide crunchiness and prevent excessive oil penetration during preparation, and in the gummies industry, they provide 25–50% of the total starch used in the formulations (Cai et al., 2010; Li et al., 2019).

Developing cassava varieties with waxy starch has become an important goal for cassava breeders. However, the recessive nature of the trait and the long reproductive cycle of cassava make the selection of waxy genotypes relatively complex. The introgression of recessive traits requires multiple generations of recombination to reduce the linkage drag of unwanted alleles of the parental genotype that contain the waxy mutation(s), such as low dry matter content and root yield (Karlström et al., 2016). A crossing between an elite non-waxy and a waxy variety, which contains many undesirable genes besides the starch mutation, is expected to have a 100% frequency of non-waxy genotypes (wild-type, *Wx_*) in the F_1 generation and segregation of 3:1 (non-waxy:waxy) in the F_2 generation. Nonetheless, due to the high heterozygosity present in the population and the little variability between the waxy starch sources, the selected genotypes have lower or similar yield potential and lower starch content than the parental genotypes (Karlström et al., 2016; Rojanaridpiched et al., 2020; Ceballos et al., 2021). This result is a consequence of inbreeding depression caused by the increased frequency of homozygous genes, often deleterious, whose expressions are repressed in their heterozygous form. Currently, there are efforts to increase recombination cycles to maintain the waxy gene in homozygosity and break undesirable genetic linkage or even increase heterozygosity for loci associated with important agronomic attributes in cassava.

Genomic studies have enabled the identification of target genes that control amylose and amylopectin synthesis and enabled the selection of markers associated with these genes with potential use in marker-assisted selection (MAS) (Aiemnaka et al., 2012; Carmo et al., 2020). Starch biosynthesis is genetically controlled by target genes, including granule-bound starch synthases (GBSS), soluble starch synthases (SSS), starch branching enzyme (SBE or BE), debranching enzyme (DBE), and protein targeting to starch (PTST) (Zeeman et al., 2010; Bahaji et al., 2014; Seung et al., 2015; Seung et al., 2017). The SSS, BE, and DBE genes are involved in amylopectin synthesis, and GBSS and PTST are enzymes related to amylose biosynthesis in plants, including cassava (Zhao et al., 2011; Bull et al., 2018).

GBSSI-related SNP markers have not proven useful for MAS in populations with different genetic backgrounds (Aiemnaka et al., 2012; Carmo et al., 2020). Alternatively, the phenotypic identification between waxy and non-waxy genotypes is usually determined by staining the roots with iodine, which is a chemical method. Non-waxy, starchy roots stain dark blue due to the presence of amylose, and waxy phenotypes stain reddish brown (Ceballos et al., 2007). However, the screening of waxy clones by the iodine method requires the presence of tuberous roots, and for this reason, in most genetic breeding programs, the selection is conducted during or close to harvest, 10 months after planting. Thus, an evident disadvantage of this process is the difficulty of the early selection of waxy clones. Therefore, the development of rapid methodologies to identify the waxy phenotype, regardless of the genetic origin of the mutation, can help optimize the selection process.

Near-infrared spectroscopy (NIRS) technologies have been used with great accuracy as auxiliary tools in the phenotyping process, aiming to accelerate the selection steps. The performance of NIRS is comparable to other analytical chemistry methods with advantages including shorter analysis time, early evaluation, bulk sample analysis per day, and non-destruction of samples (Ikeogu et al., 2017). Near-infrared (NIR) electromagnetic region radiation (700–2500 nm) is absorbed by water and organic compounds, including carbohydrates, proteins, lipids, or alcohols (Agelet and Hurburgh, 2014). Therefore, NIRS can serve as an important predictor of these compounds in organic substances.

Carmo et al. (2019) evaluated Fourier-transform near-infrared spectroscopy (FT-NIRS) for indirect, early identification of waxy starch cassava genotypes by screening samples of dried, macerated leaves. In this study, the distribution between the classes of waxy and non-waxy genotypes was similar, and the results showed high accuracy, deeming it a potential technique for the classification of waxy genotypes. However, despite this analysis being earlier than the analysis of iodine in tuberous roots, it is still necessary to germinate a large batch of seeds in the greenhouse, collect and identify leaf samples, dry macerate, and perform screening *via* NIRS. Considering the typical segregation of genes with recessive inheritance, only 25% of the F_2 seeds will be classified as waxy and, therefore, most of the investments in germination and sampling for evaluation *via* NIRS were conducted in unwanted samples. Thus, the development of waxy and non-waxy seed classification models allows for an early, non-destructive seed selection that saves time and resources, ensuring only waxy seeds followed in the selection pipeline.

In fact, NIRS has been used as an efficient tool for classifying and predicting seed germination capacity, quality, and vigor (Al-Amery et al., 2018; Medeiros et al., 2020; Mortensen et al., 2021). This approach allows for the selection and classification of seeds according to specific traits without damaging or changing seed properties. Analyses in the endosperm of waxy, normal, and sweet corn varieties have demonstrated the ability to detect differences between amylopectin and amylose structures, shape, and size of starch granules as starch is synthesized within amyloplasts (Yu et al., 2015). This is useful for the selection of plants of interest in breeding programs.

With the interest in early selection of waxy genotypes, this study aimed to associate near-infrared spectrophotometry spectra with the waxy phenotype in cassava seeds and develop an accurate classification model for indirect selection of plants soon after the performance of the crossing's blocks.

2 Material and methods

2.1 Obtaining seeds and collecting spectra using NIRS

Two generations of recombination were performed to obtain segregating populations for the waxy gene. The genotypes were cultivated in a two-crossing blocks field located in the experimental area of Embrapa Cassava and Fruits in Cruz das Almas, Bahia, Brazil (12°39′25″ S, 39°07′27″W, 226 m altitude). The parent plants of the F_1 and F_2 populations were planted from 2016–2017 and 2018–2019, respectively. The weather conditions are hot, humid, and tropical (Aw/Am, according to the Koppen classification) with a photoperiod throughout the year of approximately 12 hours (Souza et al., 2020). Cuttings (16–20 cm long) with 5–7 buds were grown under rainfed conditions in plots containing two rows with eight plants each, spaced 1.20 m between rows and 0.80 m between plants. All cultivation practices were adopted by Souza et al. (2016).

The F_1 population was achieved through crossing a waxy (*wxwx*) genotype (Cassava-7909) with three non-waxy (wild-type, *Wx* _) genotypes (BGM-0131, BGM-0728, and BGM-0935). For the F_2 population, controlled crosses were randomly performed among 77 F_1 genotypes (wild-type, *Wx*_) to produce F_2 seeds. These parents were generated through crosses from three different F_1 families, with 13, 35, and 28 genotypes each. Overall, 39 genotypes were used as both male and female parents, while 69 and 46 were used only as female or male parents, respectively. In total, 197 F_2 families and 1127 F_2 seeds were obtained.

To prevent insect pollination, the female flowers were protected by a voile-type fabric bag 24 hours before anthesis, which is easily identifiable by experienced field workers. Male flowers, immediately following anthesis, were collected from 7–9 a.m., and the crosses were performed between 9 a.m. and 4 p.m. by distributing pollen grains on stigmas. One male flower was used to pollinate up to three female flowers, depending on the amount of pollen available. The female flowers were protected again, as previously described, shortly after pollination. One cross was defined as a single pollination event. After identifying female flowers ready for pollination, crosses were performed in one to four flowers per inflorescence, and the remaining flowers were removed. The protection bag covered the inflorescence until the seeds were released and collected, which occurred approximately 2–3 months post pollination. Each seed was labelled with the family information and the seed number, and they were individually stored in plastic bags in a refrigerator (10 ± 2°C) until further analysis.

Seed spectra were obtained in a laboratory at a room temperature of 22°C through ultraviolet-visible and near-infrared spectrophotometry using a benchtop NIRFlex N-500 spectrometer (Büchi, Flawil, Switzerland) and a portable SCiO (Consumer Physics, Tel-Aviv, Israel). The spectra were obtained by placing the samples (one whole seed at a time), directly at the output of the infrared source of the device. Four measurements were taken per seed using the NIRFlex N-500, with a wavelength ranging from 800–2500 nm (12500–4000 cm^{-1}). The NIRFlex N-500 was operated in diffuse reflectance mode at a spectral resolution of 8 cm^{-1}, interpolated at 4 cm^{-1}, resulting in 1501 data points per spectrum. For the SCiO portable device, three measurements were collected per seed (N=334) in diffuse reflectance mode with wavelengths ranging from 740–1070 nm (13.514–9.346 cm^{-1}). This device has a set of 12 photodiode detectors, each with a separate optical filter. The average spectral resolution of SCiO was 13 cm^{-1}, with the lowest resolution (18 cm^{-1}) found in the highest wavenumbers and the highest resolution (9 cm^{-1}) in the lowest wavenumbers. The SCiO™ Lab online app (Consumer Physics Inc., Tel-Aviv, Israel) was used for data collection, storage, and analysis.

2.2 Seedling trial and phenotypic data collection

After collecting the spectral data, the 1127 F_2 seeds were sown in 290 cm^3 plastic tubes and placed in trays in a greenhouse at 32 ± 3°C. The tube substrate comprised vermiculite and washed sand (1:1 ratio) in the upper quarter, and the lower three quarters was composed of vermiculite, sand, and coconut fiber (ratio 1:2:1) as well as 15 mg each of single superphosphate and ammonium sulfate. The seedlings were transplanted to the field when approximately 30 cm in height, around 45 days after germination. The cultural treatments were performed according to Souza et al. (2016).

The harvest was conducted at 10 months of age, and the evaluation was performed using the 2% iodine staining test (2 g Kl and 0.2 g I^2 in distilled water); stain was applied to the cross section of at least three roots of the seedlings for the identification of the type of starch (Karlström et al., 2016; Morante et al., 2016). A dark blue color in the treated root indicated the presence of amylose (non-waxy genotype), and a reddish-brown color indicated no or low amylose content (waxy genotype) (Denyer et al., 2001).

2.3 Discriminant analysis of principal components

The population structure of the genotypes was determined by principal component discriminant analysis (DAPC) (Ivandic et al., 2002), using the adegenet package (Jombart, 2008) of the R software version 4.1.3 (R Core Team, 2021). The find.clusters() function was used in detecting the number of clusters in the population. The function uses K-means clustering, which deconstructs the total variation of a variable into components between groups and within the group. The best number of subpopulations was chosen by the smallest Bayesian Information Criterion (BIC). The groups were plotted on a scatterplot of the first and second linear discriminant of the DAPC.

2.4 Pre-processing and adjustment of classification models

Several pre-processing techniques were evaluated to ensure spectral data reliability such as: first-order derivative (1st); detrend (DT); multiplicative scatter correction (MSC) and standard normal variation (SNV); Combined pretreatment methods, first-order derivative-detrend (1st-DT); first-order derivative-multiplicative scatter correction (1st-MSC); detrend-multiplicative scatter correction (DT-MSC); and first-order derivative with Savitzky–Golay-detrend (1st-SG-DT). The first-order derivative was used to substracted the influence of background and baseline drift, DT was used to eliminate the baseline drift in the spectra, and MSC and SNV methods were used to eliminate the scattering multiplicative interferences in the spectral signal.

The spectra were pre-processed for above tecniques and then smoothed with an N=11 filter at each end of the spectral set for noise reduction (Savitzky and Golay, 1964). The DT, MSC, SNV, and SG were implemented by the functions detrend(), msc(), standardNormalVariate(), and savitzkyGolay(), respectively, from the prospectr package (Stevens and Ramirez-Lopez, 2022) implemented in the R software version 4.1.3.

After pre-processing, the spectral data were arranged in an X matrix (predictors), and the starch type data (waxy and non-waxy) were allocated in a Y vector (response). Four classification models were assessed for waxy cassava genotype identification: k-nearest neighbor algorithm (KNN) (Cover and Hart, 1967), C5.0 decision tree (CDT) (Freund and Schapire, 1997), parallel random forest (parRF) (Breiman, 2001), and eXtreme Gradient Boosting (XGB) (Chen and Guestrin, 2016).

KNN is a commonly used non-parametric algorithm in Machine Learning. It is mathematically simple and based on the determination of distances, often Euclidean, between an unknown object and each of the objects in the training set. Thus, the smallest distance is selected for assigning the members of a given class. With k representing the number of neighbors, the k-nearest objects of the unknown sample are selected, and a majority rule is applied: the unknown sample is classified in the class to which most k objects belong. The choice of k is optimized by calculating the predictive power with different values of k.

C5.0 is an algorithm based on decision trees (Elsayad et al., 2020), which involve a set of decision nodes, among which the root and each internal node are labeled with a question (Pradhan, 2013). The arcs descend from each root node to leaf nodes, where a solution to the associated issue is offered. A split is created at each node by taking a binary decision, which separates a class or multiple classes from the global dataset.

The RF algorithm is a type of ensemble learning and is a method that generates several decision trees and combines the result of the classification from each of them. This combination of models makes it more powerful than Decision Tree. The algorithm works by growing a set of regression trees based on binary recursive partitioning, where the algorithm begins with a number of bootstrap samples from the predictor space (original data) (Cutler et al., 2012).

XGBoost is a machine learning algorithm based on a gradient boosting decision tree (GDBT) (Chen and Guestrin, 2016). XGBoost is an extension of RF (Svetnik et al., 2003), and, as a differential, it can use a regularization term to further reduce overfitting, improve prediction accuracy, and decrease the time needed to build decision trees (Luckner et al., 2017). All data analyses were performed with the R software version 4.1.3 using the caret package (Kuhn, 2008).

The selection of wavelengths with relative importance was conducted using the XGB model, as it automatically provides estimates of the importance of the variables. Variables with relative importance (≥30%) were selected. For this, the varImp() function from the caret package of the R software version 4.1.3 was used, which automatically scales importance scores between 0 and 100.

2.5 Cross-validation and external validation

Data were divided into a training set, for model development purpose, (80% of the data) and a testing set used as independent samples to test the classification models (used to obtain the confusion matrix), both with equitable distribution of genotypes according to the type of starch. The model performances were evaluated in the training set based on cross-validation, consisting of 10 repetitions with 5-folds each. Parameters that provide the best fit to the data were selected for each model evaluated (Table 1). The overall effectiveness of the classification

TABLE 1 Parameters used in the k-nearest neighbor algorithm (KNN), C5.0 decision tree (CDT), eXtreme Gradient Boosting (XGB), and parallel random forest (parRF) classification models using all variables and selected variables with relative importance (≥30%) using the XGB model.

Models	Parameters	NIRFlex N-500		SCiO	
		All variables	Selected variables	All variables	Selected variable
KNN	K	5	5	7	7
CDT	trials, model, and winnow	20, tree, and TRUE	20, tree, and FALSE	20, tree, and FALSE	20, tree, and FALSE
XGB	nrounds, lambda, alpha, and eta	150, $1e^{-4}$, 0, and 0.3	150, $1e^{-4}$, 0.1, and 0.3	50, 0, 0, and 0.3	50, 0.1, 0.1, and 0.3
ParRF	mtry*	1459	27	2	2

*number of predictors.

models was assessed based on mean values of accuracy and Cohen's Kappa statistic (unweighted) (Cohen, 1960), obtained in each repetition of the cross-validation. The accuracy was determined using the equation 1:

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn} \qquad (1)$$

where *tp* corresponds to the number of correctly recognized class examples (true positives), *tn* is the number of correctly recognized examples that do not belong to the class (true negatives), *fp* are examples that were incorrectly assigned to the class (false positives), and *fn* are examples that were not recognized as class examples (false negatives). The Kappa index is based on the number of concordant responses defined by equation 2:

$$Kappa = \frac{p_o + p_e}{1 - p_e} \qquad (2)$$

where p_o is the proportion of units that agreed, and p_e is the proportion of units for which agreement is expected by chance. This index indicates how well the models can correctly classify the two analyzed classes, and the closer to one, the greater the detection power.

The testing set (20% of the data) consisted of 225 and 67 genotypes for NIRFlex N-500 and SCiO, respectively. The prediction performance was evaluated with parameters generated from a confusion matrix. The parameters were accuracy, Kappa index, sensitivity, and specificity. Sensitivity measures the probability of the classifier hitting true positives ($\frac{tp}{tp+fn}$) , while specificity measures the probability of hitting true negatives ($\frac{tn}{tn+fn}$).

3 Results

3.1 Segregation and clustering of clones *via* multivariate analysis

Among the 1127 seedlings, 21.3% had waxy starch genotypes. Of the 197 families, 85 were used to assess the frequency of segregation for the mutant phenotype (Waxy – *wxwx*) because they had four or more individuals per family. As the population originated from the cross between waxy parents (*wxwx*) with a known genotype and non-waxy parents (wild-type, *Wx_*) with unknown genotypes, the expected frequencies of 3:1 and 1:1 were considered for the two possibilities of the non-waxy parent. As expected, the observed distribution of phenotypic classes in 86% of the evaluated families adjusted to a single-gene Mendelian inheritance (flex Table S1).

Both the spectral data collected by the NIRFlex N-500 (240 waxy and 887 non-waxy clones) and the SCiO portable NIR (291 waxy and 44 non-waxy clones) were used to assess the potential for classifying cassava genotypes based on a waxy phenotype. The density distributions of the waxy and non-waxy clones, were determined for each NIR equipment (Figure 1). It can be observed from the density curves that both equipments displayed overlapping curves, which represent areas of confusion, with the diferentiation between the groups not being clear by visual analysis.

3.2 Development of classification models

To evaluate the efficiency of the pre-processing techiniques were used the parameters Accuracy and the Kappa index from the KNN classification method (Figure S1). In general, according to cross-validation the results were similar between the pre-processing techniques, with lower performance when using the raw data without pre-processing. The 1st and MSC combination was selected to proceed with the analyses.

Accuracy and the Kappa index were used as parameters to evaluate the efficiency of the models with the best fit in the classification of waxy and non-waxy clones. Generally, the classification accuracy using NIRFlex N-500 spectral data varied among the different models analyzed. According to cross-validation, the accuracy ranged from 0.86 (parRF) to 0.92 (XGB) (Figure 2; Table 2). The Kappa index displayed a similar trend as the accuracy, considering the lowest value for the parRF method (0.39) and the highest value for XGB (0.69). Regarding the NIRFlex N-500 spectra collected, although the KNN classification method has presented similar accuracy (0.90) to the XGB model, the Kappa index was considerably lower (0.64) than the XGB.

Regarding NIRS SCiO, the classification accuracy was similar among the four models evaluated, with values ranging between 0.87 (CDT) and 0.89 (parRF and XGB). However, the Kappa index was lower than that of

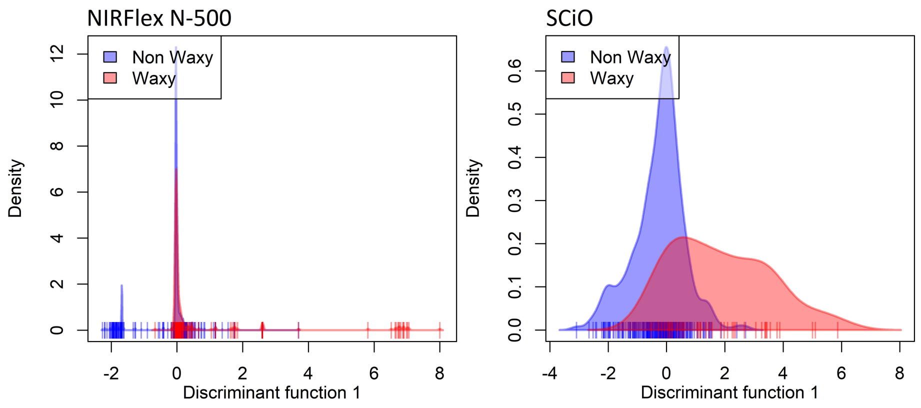


FIGURE 1
Density plot on the first discriminant function showing discriminant analysis of principal components (DAPC) based on near-infrared (NIR) spectral data obtained by NIRFlex N-500 and SCiO equipment, considering contrasting cassava genotypes for waxy and non-waxy starches.

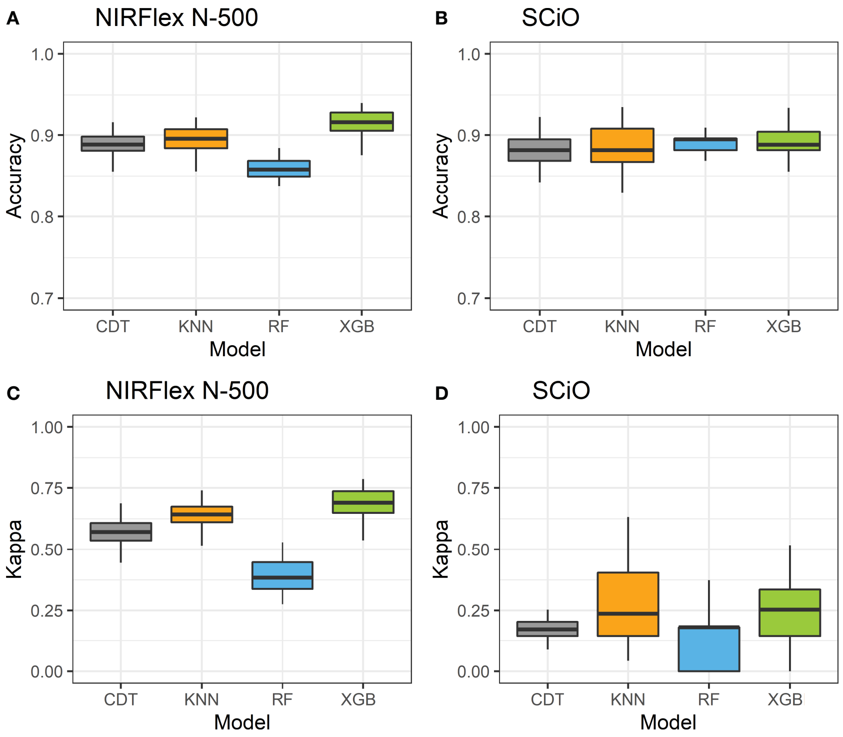


FIGURE 2
Accuracy **(A, C)** and kappa index **(B, D)** of cross-validation of classification models based on NIRFlex N-500 and SCiO near-infrared spectra evaluated in cassava seeds contrasting for waxy and non-waxy starch. KNN, k-nearest neighbor algorithm; CDT, C5.0 decision tree; XGB, eXtreme Gradient Boosting; parRF, parallel random forest.

the NIRFlex N-500, and this index ranged from 0.05 (CDT) to 0.22 (KNN). These results show that, despite high accuracy values, these models, especially CDT, are incapable of correctly classifying waxy and non-waxy clones based on the SCiO device spectra.

Despite high accuracy in classifying the waxy phenotype early during the seed stage, especially in the NIRFlex N-500 spectra, the possibility of improving classification accuracy was investigated further considering the selection of variables according to the importance scores of the spectra based on the XGB model. This was warranted because spectroscopic techniques tend to generate a high number of variables (wavelengths) with noise which are highly correlated, which reinforces the importance of removing non-informative variables. Thus, the construction of consistent classification and prediction models is possible, reducing the risk of

TABLE 2 Cross-validation parameters of the k-nearest neighbor algorithm (KNN), eXtreme Gradient Boosting (XGB), C5.0 decision tree (CDT) and parallel random forest (parRF) classification models obtained through spectral data analysis from the NIRFlex N-500 and SCiO in cassava seeds with waxy and non-waxy starch.

Models*		NIRFlex N-500		SCiO	
		Accuracy	Kappa	Accuracy	Kappa
All spectra	KNN	0.90 ± 0.01	0.64 ± 0.05	0.88 ± 0.02	0.22 ± 0.14
	CDT	0.89 ± 0.02	0.57 ± 0.08	0.88 ± 0.02	0.05 ± 0.15
	XGB	0.92 ± 0.01	0.69 ± 0.05	0.89 ± 0.02	0.20 ± 0.12
	parRF	0.86 ± 0.01	0.39 ± 0.06	0.89 ± 0.01	0.13 ± 0.10
Selected spectra	KNN_Sel	0.89 ± 0.02	0.61 ± 0.06	0.89 ± 0.02	0.26 ± 0.14
	CDT_Sel	0.92 ± 0.01	0.73 ± 0.06	0.89 ± 0.02	0.23 ± 0.17
	XGB_Sel	0.95 ± 0.01	0.82 ± 0.04	0.90 ± 0.02	0.37 ± 0.16
	parRF_Sel	0.92 ± 0.01	0.72 ± 0.05	0.89 ± 0.01	0.14 ± 0.13

* Sel: models using variables selected according to their relative importance by the xgbLinear model.

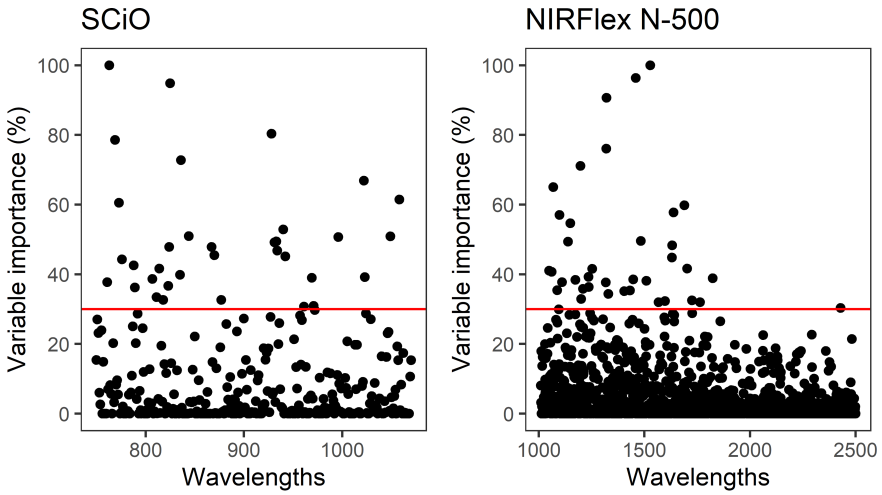


FIGURE 3
Relative importance of wavelengths collected by NIRFlex N-500 and SCiO equipment for classification of the waxy phenotype in cassava based on the eXtreme Gradient Boosting classification model.

inferences and the computational cost of the analyses. Thirty seven and 34 wavelengths were selected for the NIRFlex N-500 and the SCiO, respectively, with relative importance (≥30%) (Figure 3).

Overall, for the NIRFlex N-500, models built on the most important spectra only resulted in an increase in classification accuracy and Kappa index estimates compared to models built on all spectra, excluding the KNN model. The CDT and XGB models resulted in an average increase of 3.7% in accuracy, while the parRF model showed a 7% increase. Furthermore, the Kappa index significantly increased from 0.57, 0.69, and 0.39 to 0.73, 0.82, and 0.72 for the CDT, XGB and parRF models, respectively (Figure 4; Table 2). However, in relation to SCiO, the accuracy estimates remained practically unchanged after the selection of the most important spectra. Alternatively, the Kappa index increased

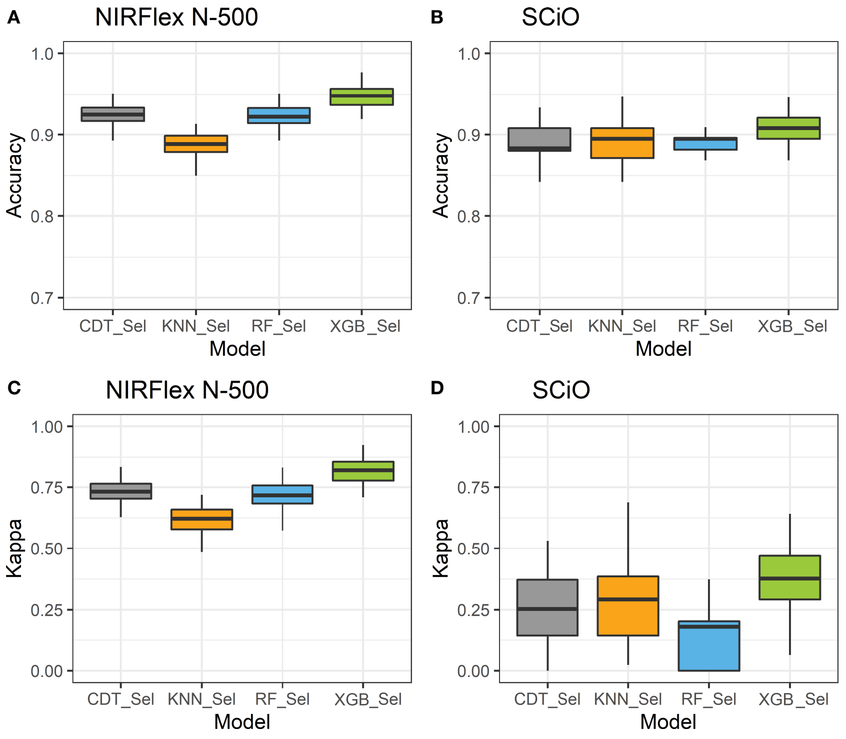


FIGURE 4
Accuracy **(A, B)** and Kappa index **(C, D)** of cross-validation of classification models based on NIRFlex N-500 and SCiO near-infrared spectra evaluated in cassava seeds contrasting for waxy and non-waxy starch. KNN, k-nearest neighbor algorithm; CDT, C5.0 decision tree; XGB, eXtreme Gradient Boosting; parRF, parallel random forest; Sel, models using variables selected according to relative importance by the XGB model.

significantly from 0.05 to 0.23 (CDT), and from 0.20 to 0.37 (XGB) (Figure 4). However, Kappa index estimates are considered very low (< 0.37) and highly biased in their estimates (Table 2).

Predictive capacity of classification models

The predictive capacity of the models was evaluated based on the accuracy, Kappa index, sensitivity, and specificity generated from the confusion matrix obtained by predicting the models in the testing set (Table 3; Figures 5, 6). Considering the testing population, high classification accuracy was identified for both NIRSs. The accuracies ranged from 0.85 (parRF) to 0.95 (XGB _Sel) for the NIRFlex N-500 (Table 3). The Kappa index was high (>0.60), except for the parRF model with a value of 0.37 (Table 3). Like cross-validation, the selection of the most important spectra for model calibration provided an increase in the accuracy values and, more importantly, in the Kappa index, excluding the KNN model.

Confusion matrix based on the spectra collected by SCiO resulted in similar values of accuracy and Kappa indices, regardless of whether the model uses all spectra or only the most important for the classification of waxy clones. Again, although the SCiO spectra resulted in high classification accuracies, the capability of reliable detection among the analyzed classes was null.

Overall, values equal to or close to one were obtained for sensitivity, indicating that the models were able to predict the true positives of each class. Specificity values ranged between 0.27–0.74, for NIRFlex N-500, and were close to zero for SCiO (Table 3). This result indicates that most models were not efficient in predicting the true negatives of the evaluated classes. The two classes evaluated present an imbalance in relation to the number of clones that comprise each class. Therefore, the differences in sensitivity and specificity estimates are attributed to this imbalance between the classes since the confusion matrix considers the non-waxy class as positive and waxy as negative.

The confusion matrix displays the results of classifying the different models in the external validation set (Figures 5, 6). For both NIRSs, the models were efficient in classifying non-waxy clones (considered the "positive" class) with hit percentages ranging between 95–100%. However, the NIRSs differ in the prediction potential of the waxy clone class. For the NIRFlex N-500, the hit percentage ranged from 27% (parRF) to 74% (KNN and XGB_Sel). In general, the models tended to classify waxy genotypes as non-waxy, especially for SCiO equipment.

4 Discussion

4.1 Evaluation of waxy phenotype classification efficiency

Several studies employ molecular markers to understand the genetic control of the waxy genotype, which guides the crossing planning of accessions, since the waxy phenotype is expressed in the recessive condition (Aiemnaka et al., 2012; Carmo et al., 2020). However, despite the development of protocols that allow the use of selection assisted by molecular markers related to the GBSSI (granule-bound starch synthase I) gene derived from the waxy starch source AM206-5, there remain obstacles when the population has a different genetic origin than the AM206-5 source (Carmo et al., 2020). Therefore, using technologies that allow a faster, earlier selection of waxy genotypes is desirable in the most diverse breeding programs.

In the present study, seeds from segregating populations of cassava for waxy starch were used as sample material for the identification/classification of waxy and non-waxy genotypes by near-infrared spectroscopy (NIRS). A previous study using spectral data collected from leaf tissue allowed the early and accurate identification of waxy genotypes (Carmo et al., 2019). The NIRS technique allows capturing differences in the chemical constitution of plants because of the expression of different genes. Further, leaves are complex assemblies of organic compounds and may be expected to exhibit different spectral responses. NIRS can be successfully used for the characterization of chemical components, like nitrogen, in different plant tissues (Li et al., 2022). In addition to leaf tissue, starch samples have been used to identify the waxy genotype based on

TABLE 3 Parameters from confusion matrix associated with grading efficiency of contrasting cassava seeds for waxy and non-waxy starch based on near-infrared (NIR) spectra collected by NIRFlex N-500 and SCiO equipment in test samples.

Models*		NIRFlex N-500				SCiO			
		Accuracy	Kappa	Sensitivity	Specificity	Accuracy	Kappa	Sensitivity	Specificity
All spectra	KNN	0.92	0.74	0.97	0.73	0.87	0.23	0.96	0.23
	CDT	0.90	0.60	1.00	0.49	0.89	0.19	0.99	0.14
	XGB	0.93	0.74	0.99	0.65	0.90	0.22	1.00	0.14
	parRF	0.85	0.37	1.00	0.27	0.88	0	1.00	0
Selected spectra	KNN	0.89	0.62	0.95	0.63	0.88	0.12	0.99	0.09
	CDT	0.94	0.79	1.00	0.71	0.88	0.26	0.97	0.23
	XGB	0.95	0.82	1.00	0.74	0.88	0.22	0.98	0.18
	parRF	0.93	0.73	1.00	0.63	0.89	0.08	1.00	0.05

* KNN, k-nearest neighbor algorithm; CDT, C5.0 decision tree; XGB, eXtreme Gradient Boosting; parRF, parallel random forest; Sel, models using variables selected according to their relative importance by the XGB model.

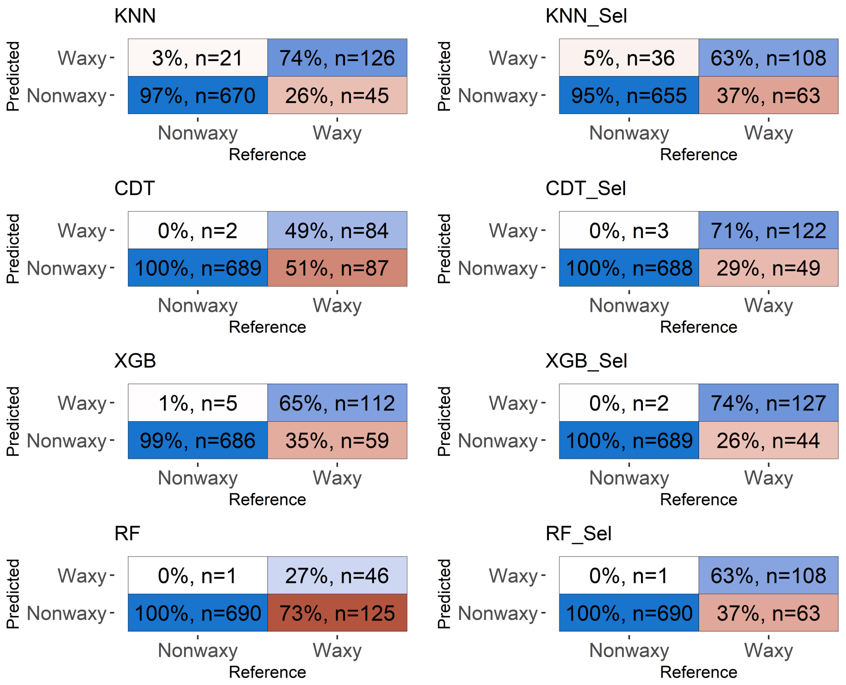


FIGURE 5
Confusion matrix of the testing set considering classification models based on near-infrared spectra by NIRFlex N-500 evaluated in cassava seeds contrasting for waxy and non-waxy starch. KNN, k-nearest neighbor algorithm; CDT, C5.0 decision tree; XGB, eXtreme Gradient Boosting; parRF, parallel random forest; Sel, models using variables selected according to relative importance by the XGB model.

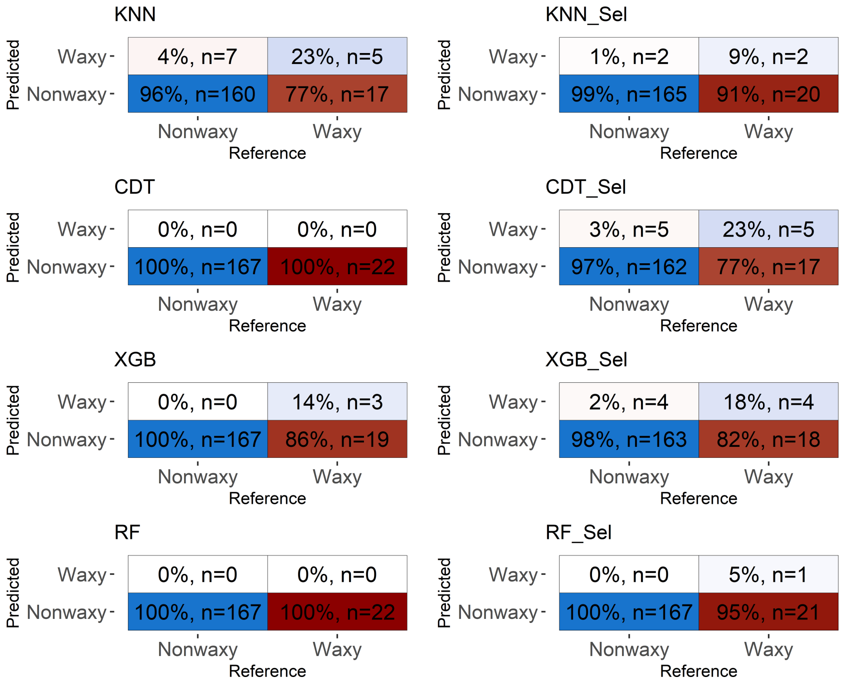


FIGURE 6
Confusion matrix performed in the testing set considering classification models based on near-infrared spectra by SCiO evaluated in cassava seeds contrasting for waxy and non-waxy starch. KNN, k-nearest neighbor algorithm; CDT, C5.0 decision tree; XGB, eXtreme Gradient Boosting; parRF, parallel random forest; Sel, models using variables selected according to relative importance by the XGB model.

NIR spectra in species such as wheat (Lavine et al., 2014; Delwiche and Graybosch, 2016; Delwiche et al., 2018).

The early analysis of greenhouse waxy cassava clones using NIR spectra in leaf tissues, before field planting, allows the exclusive selection of desired genotypes with a high probability to plant the waxy phenotype. Thus, a breeder can avoid planting large populations that do not contain the desired trait (~75% of individuals). However, the use of dried and macerated cassava leaves as sample material requires additional time and resources for the selection process, as it is necessary to sow seeds and grow plants in a greenhouse until the collection time of leaf tissues. The results of the present study indicate that it is possible to classify cassava seeds according to the type of starch with an accuracy close to 1 through classification models based on seed spectral data. Among the two evaluated NIRSs equipment, the NIRFlex N-500 proved to be more accurate, with Kappa values close to 0.80, compared to the portable NIR SCiO. This was possible as each device has different wavelength amplitudes, 740–1070 nm for SCiO and 800–2500 nm for NIRFlex N-500, in addition to the different sample sizes.

Although the NIRFlex N-500 has a higher cost, there is a better resolution in obtaining spectra that maximizes the chance of association with the phenotype of interest (Beć et al., 2022). Due to its numerous advantages, NIR spectra of 800–2500 nm have been used to predict several chemical components in plant seeds (Ferreira et al., 2013). Alternatively, although the SCiO equipment provided high classification accuracy (0.87–0.89), the Kappa indices were very low.

The accuracy values indicate that, in both NIRS equipment, there was a high proportion of correctly classified events in relation to the total number of samples. Accuracy is one of the most intuitive and widely used performance metrics for classification. The Kappa index is a widely used metric to measure classification performance, considering the probability of obtaining the classification by chance. Some authors warn that Kappa may be an inadequate estimate when an unbalanced distribution of classes is involved, where the marginal probability of a class is much (more or less) greater than the others (Donker et al., 1993; Forbes, 1995; Andrés and Marzo, 2004; Delgado and Tibau, 2019). In fact, the dataset evaluated by SCiO showed a greater imbalance between classes compared to the samples evaluated by the NIRFlex N-500.

Portable and smaller equipment, such as the SCiO, has a growing popularity in the agri-food industry. The NIR SCiO is a cost-effective device that stores data in a “cloud”, and it is affordable because it uses an LED light source and a simple 12-element Si photodiode detector, with a configuration matrix of 4 × 3, combined with optical filters on each pixel to form a 12-channel spectrometer (Beć et al., 2022). However, these characteristics give it lower optical performance due to the low number of wavelengths compared to benchtop equipment, such as the NIRFlex N-500 (Beć et al., 2022). Despite these limitations, the spectral region covered is sufficient for the prediction of important parameters related to food quality, such as total soluble solids, maturity, identification of fruits with a high concentration of dry matter (Li et al., 2018a), and sugar content and firmness in tomatoes (Goisser et al., 2018). Additionally, this equipment makes it possible to classify cultivars of barley, chickpeas, and sorghum seeds with 86–96% accuracy (Kosmowski and Worku, 2018).

The accuracies of cross-validation in training set and from confusion matrix in the testing set were high among the classification models analyzed, with emphasis on the XGB algorithm (>0.92). A recent study demonstrated the effectiveness of XGB in analyses with spectral data in food quality control (Li et al., 2018b), in comparison with the Back Propagation Neural Network and Support Vector Regression models, often used in analysis of products of vegetal origin. In addition to the high classification accuracy of waxy clones, the Kappa values obtained by this algorithm were high, at 0.69 and 0.82, respectively. Probably, because it is an extension of random forest and uses a regularization parameter to reduce overfitting, XGB was the algorithm with the highest detection power, allowing it to correctly classify the two classes analyzed (Luckner et al., 2017).

Due to the high number of variables (wavelengths) gathered, mainly by the NIRFlex N-500, the selection of variables makes it possible to remove noise, or highly correlated and non-informative variables, to improve computational performance. Therefore, the classification models were evaluated after selecting the most important spectra based on the XGB algorithm. Following this procedure, a slight increase in Kappa values was observed, and similar classification accuracies was revealed for the different models compared to the analyses performed with all spectra. Therefore, the selection of variables proved to be advantageous for increasing the power of the models to classify waxy cassava clones and in reducing the computational time for processing the analyses.

4.2 Prospects for the use of NIRS for early selection in cassava

NIR spectrometry has demonstrated a high potential in predicting key traits such as carotenoids, starch, and dry matter content in cassava (Ikeogu et al., 2017; Bantadjan et al., 2020; Maraphum et al., 2022). The correlation coefficient of prediction was 0.83 for starch content (Bantadjan et al., 2020), 0.88 for carotenoids, and 0.80 for dry matter content (Ikeogu et al., 2017), which ensures a sufficient predictive accuracy of new phenotypes to be generated and evaluated by the cassava breeding programs.

Furthermore, as it is a non-destructive technique, it can be incorporated as a new tool for cassava breeders, improving phenotyping efficiency. When compared to the conventional laboratory techniques for dry matter and carotenoid content in cassava breeding, the NIRS technique is rapid and cost-effective (Ikeogu et al., 2017). The current phenotyping techniques for key traits are laborious and time-consuming for large-scale screenings. Additionally, estimates could be influenced by sample preparation, including weight and number of roots used in the prevalent specific gravity method (Fukuda et al., 2010). For carotenoid quantification using color, the intensity could be subjective and inefficient in an advanced population of biofortified genetic materials (Sánchez et al., 2006). Moreover, laboratory processes using high-performance liquid chromatography (HPLC) or a UV-Visible spectrophotometer are low-throughput, processing less than 10 or 40 samples per day, respectively (Sánchez et al., 2014).

These results bring advances and new techniques for early identification of cassava genotypes with waxy starch at the seed stage, through non-destructive techniques. This allows cassava breeders to generate large F_2 segregating populations with thousands of individuals. From these populations, it is possible to select desirable genotypes with high classification accuracy before planting in the field.

Despite the initial investment to purchase the NIRS equipment, the economic return is readily apparent in the next seedling trials. After screening the seeds *via* NIRS, it is possible to reduce the planting area of seedlings by up to 75%. In terms of resource allocation, an estimated cost with phenotyping of a field plot, with a seedling per plot, in one environment is 2.20 U.S. dollars. This value was assumed for a single-plant field plot, including phenotyping with the iodine test. On average, 8000 seeds are obtained from segregating populations for waxy starch per year. Screening represents an average savings of $2.20 x 6000 = $13,200.00/year.

5 Conclusions

NIR spectroscopy in combination with the eXtreme Gradient Boosting algorithm (XGB) can be used to classify cassava seeds according to the type of waxy and non-waxy starch and select early genotypes with the desired phenotype. The methodology using NIRS techniques showed great potential for applicability, being a fast and efficient tool for the identification of waxy genotypes for practical use as an alternative to utilizing molecular markers in cassava breeding programs.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://figshare.com/, dx.doi.org/10.6084/m9.figshare.21071257.

Author contributions

MS, LA, and EO designed the experiments. JF Spectral data collection. LA and MS analyzed the spectral dataset. MS, LA, and EO were involved in the research design and improvement of the manuscript. MS, and EO wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

MS: Fundação Arthur Bernardes. Grant number: 4986. JF: FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia). Grant number: BOL0225/2018. LA: Empresa Brasileira de Pesquisa Agropecuária. Grant number: 20.18.01.012.00.00. EO: CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). Grant number: 409229/2018-0, 442050/2019-4 and 303912/2018-9. EO: FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia). Grant number: Pronem 15/2014. This work was partially funded by UK's Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation. Grant INV-007637. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. The funder provided support in the form of fellowship and funds for the research, but did not have any additional role in the study design, data collection and analysis, decision to publish, nor preparation of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1089759/full#supplementary-material

References

Agama-Acevedo, E., Flores-Silva, P. C., and Bello-Perez, L. A. (2019). "Cereal starch production for food applications," in *Starches for food application*. M.T.P. Silva Clerici and M. Schmiele Ed. (Academic Press: New York, NY, USA), 71–102.

Agelet, L. E., and Hurburgh, C. R. (2014). Limitations and current applications of near infrared spectroscopy for single seed analysis. *Talanta* 121, 288–299. doi: 10.1016/j.talanta.2013.12.038

Aiemnaka, P., Wongkaew, A., Chanthaworn, J., Nagashima, S. N., Boonma, S., Authapun, J., et al. (2012). Molecular characterization of a spontaneous waxy starch mutation in cassava. *Crop Sci.* 52, 2121–2130. doi: 10.2135/cropsci2012.01.0058

Al-Amery, M., Geneve, R. L., Sanches, M. F., Armstrong, P. R., Maghirang, E. B., Lee, C., et al. (2018). Nearinfrared spectroscopy used to predict soybean seed germination and vigour. *Seed Sci. Res.* 28, 245–252. doi: 10.1017/S0960258518000119

Andrés, A. M., and Marzo, P. F. (2004). Delta: A new measure of agreement between two raters. *Br. J. Math. Stat. Psychol.* 57 (1), 1–19. doi: 10.1348/000711004849268

Bahaji, A., Li, J., Sánchez-López, ÁM., Baroja-Fernández, E., Muñoz, FJ., Ovecka, M., et al. (2014). Starch biosynthesis, its regulation and biotechnological approaches to improve cropyields. *Biotechnol. Adv.* 32 (1), 87–106. doi: 10.1016/j.biotechadv.2013.06.006

Bantadjan, Y., Rittiron, R., Malithong, K., and Narongwongwattana, S. (2020). Rapid starch evaluation in fresh cassava root using a developed portable visible and near-infrared spectrometer. *ACS Omega* 5, 11210–11216. doi: 10.1021/acsomega.0c01346

Beć, K. B., Grabska, J., and Huck, C. W. (2022). Miniaturized NIR spectroscopy in food analysis and quality control: Promises, challenges, and perspectives. *Foods* 11 (10), 1465. doi: 10.3390/foods11101465

Breiman, L. (2001). Arcing classifier (with discussion and a rejoinder by the author). *Ann. Statist.* 26, 801–824. doi: 10.1214/aos/1024691079

Bull, S. E., Seung, D., Chanez, C., Mehta, D., Kuon, J. E., Truernit, E., et al. (2018). Accelerated ex situ breeding of GBSS-and PTST1-edited cassava for modified starch. *Sci. Adv.* 4 (9), eaat6086. doi: 10.1126/sciadv.aat6086

Cai, L., Shi, Y., Rong, L., and Hsiao, B. S. (2010). Debranching and crystallization of waxy maize starch in relation to enzyme digestibility. *Carbohydr. Polymers* 81 (2), 385–393. doi: 10.1016/j.carbpol.2010.02.036

Carmo, C. D., Sousa, M. B., and Silva, P. P. (2020). Identification and validation of mutation points associated with waxy phenotype in cassava. *BMC Plant Biol.* 20, 164. doi: 10.1186/s12870-020-02379-3

Carmo, C. D., Sousa, M. B., dos Santos Pereira, J. C. H., and de Oliveira, E. J. (2019). Identification of waxy cassava genotypes using Fourier-transform NearInfrared spectroscopy. *Crop Sci.* 60(2), 883–895. doi: 10.1002/csc2.20102

Ceballos, H., Sánchez, T., Morante, N., Fregene, M., Dufour, D., Smith, A. M., et al. (2007). Discovery of an amylose-free starch mutant in cassava. *J. Agric. Food Chem.* 55 (18), 7469–7476. doi: 10.1021/jf070633y

Ceballos, H., Hershey, C., Iglesias, C., and Zhang, X. (2021). Fifty years of a public cassava breeding program: evolution of breeding objectives, methods, and decision-making processes. *Theor. Appl. Genet.* 134, 2335–2353. doi: 10.1007/s00122-021-03852-9

Chen, T., and Guestrin, C. (2016). *XGBoost: A scalable tree boosting system* (New York, NY: Association for Computing Machinery), 785–794.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cover, T. M., and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 1, 21–27. doi: 10.1109/TIT.1967.1053964

Cutler, A., Cutler, D., and Stevens, J. (2012). "Random forests," in *Ensemble machine learning* (Boston, MA: Springer), 157–175.

Delgado, R., and Tibau, X. A. (2019). Why cohen's kappa should be avoided as performance measure in classification. *PloS One* 14(9), e0222916. doi: 10.1371/journal.pone.0222916

Delwiche, S., Jianwei, Q., Robert, A. G., Steven, R. R., and Moon, S. K. (2018). Near infrared hyperspectral imaging of blends of conventional and waxy hard wheats. *J. Spectral Imaging* 7 (1), 1. doi: 10.1255/jsi.2018.a2

Delwiche, S. R., and Graybosch, R. A. (2016). Binary mixtures of waxy wheat and conventional wheat as measured by NIR reflectance. *Talanta* 146, 496–506. doi: 10.1016/j.talanta.2015.08.063

Demiate, I. M., and Kotovicz, V. (2011). Cassava starch in the Brazilian food industry. *Ciec. e Tecnol. Alimentos* 31 (2), 388–397. doi: 10.1590/S0101-20612011000200017

Denyer, K., Johnson, P., Zeeman, S., and Smith, A. M. (2001). The control of amylose synthesis. *J. Plant Physiol.* 158, 479–487. doi: 10.1078/0176-1617-00360

Donker, D., Hasman, A., and Van Geijn, H. (1993). Interpretation of low kappa values. *Int. J. Bio-Med. Comput.* 33 (1), 55–64. doi: 10.1016/0020-7101(93)90059-F

Elsayad, A. M., Nassef, A. M., Al-Dhaifallah, M., and Elsayad, K. A. (2020). Classification of biodegradable substances using balanced random trees and boosted C5.0 decision trees. *Int. J. Environ. Res. Public Health* 14 (24), 9322. doi: 10.3390/ijerph17249322

Ferreira, D. S., Pallone, J. A. L., and Poppi, R. J. (2013). Fourier Transform near-infrared spectroscopy (FT-NIRS) application to estimate Brazilian soybean [Glycine max (L.) merril] composition. *Food Res. Int.* 51, 53–58. doi: 10.1016/j.foodres.2012.09.015

Forbes, A. D. (1995). Classification-algorithm evaluation: Five performance measures based onconfusion matrices. *J. Clin. Monit.* 11 (3), 189–206. doi: 10.1007/BF01617722

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55 (1), 119–139. doi: 10.1006/jcss.1997.1504

Fukuda, W. M. G., Guevara, C. L., Kawuki, R., and Ferguson, M. E. (2010). *Selected morphological and agronomic descriptors for the characterization of cassava*. Ibadan, Nigeria: International Institute of Tropical Agriculture (IITA), 19pp. doi: 10.25502/pfbm-9233/p

Goisser, S., Fernandes, M., Ulrichs, C., and Mempel, H. (2018). Non-destructive measurement method for a fast quality evaluation of fruit and vegetables by using food-scanner. *DGG-Proceedings* 8 (13), 1–5. doi: 10.5288/dgg-pr-sg-2018

Ikeogu, N. U., Davrieux, F., Dufour, D., Ceballos, H., Egesi, C.N., and 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, 1. doi: 10.1371/journal.pone.0188918

Ivandic, V., Hackett, C. A., Nevo, E., Keith, R., Thomas, W. T., and Forster, B. P. (2002). Analysis of simple sequence repeats (SSRs) in wild barley from the fertile crescent: associations with ecology, geography and flowering time. *Plant Mol. Biol.* 48, 511–527. doi: 10.1023/A:1014875800036

Jombart, T. (2008). Adegenet: A r package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Karlström, A., Calle, F., Salazar, S., Morante, N., Dufour, D., and Ceballos, H. (2016). Biological implications in cassava for the production of amylose-free starch: Impact on root yield and related traits. *Front. Plant Sci.* 7, 604. doi: 10.3389/fpls.2016.00604

Koehorst-van Putten, H. J., Sudarmonowati, E., Herman, M., Pereira-Bertram, I. J., Wolters, A. M., Meima, H., et al. (2012). Field testing and exploitation of genetically modified cassava with low-amylose or amylose-free starch in Indonesia. *Transgenic Res.* 21, 39–50. doi: 10.1007/s11248-011-9507-9

Kosmowski, F., and 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), 1. doi: 10.1371/journal.pone.0193620

Kuhn, M. (2008). Building predictive models in r using the caret package. *J. Stat. Softw.* 28 (5), 1–26. doi: 10.18637/jss.v028.i05

Lavine, B. K., Mirjankar, N., and Delwiche, S. (2014). Classification of the waxy condition of durum wheat by near infrared reflectance spectroscopy using wavelets and a genetic algorithm. *Microchem* 117, 178–182. doi: 10.1016/j.microc.2014.06.030

Li, C., Wu, Y., Yang, Y., Zhang, Y., and Zhang, H. (2018b). "Spectroscopy-based food internal quality evaluation with XGBoost algorithm," in *Web and big data* (New York: Springer International Publishing), 56–64.

Li, M., Qian, Z., Shi, B., Medlicott, J., and East, A. (2018a). Evaluating the performance of a consumer scale SCiO^{TM} molecular sensor to predict quality of horticultural products. *Postharvest Biol. Technol.* 145, 183–192. doi: 10.1016/j.postharvbio.2018.07.009

Li, Y., Sun, H., Tomasetto, F., Jiang, J., and Luan, Q. (2022). Spectrometric prediction of nitrogen content in different tissues of slash pine trees. *Plant Phenomics* 9892728. doi: 10.34133/2022/9892728

Li, H., Gidley, M. J., and Dhital, S. (2019). High-amylose starches to bridge the "Fiber gap": Development, structure, and nutritional functionality. *Compr. Rev. Food Sci. Food Saf.* 18, 1. doi: 10.1111/1541-4337.12416

Luckner, M., Topolski, B., and Mazurek, M. (2017). "Application of XGBoost algorithm in fingerprinting localisation task," in *Computer information systems and industrial management*. K. Saeed, W. Homenda and R. Chaki Eds. (New York: Springer). doi: 10.1007/978-3-319-59105-6_57

Maraphum, K., Saengprachatanarug, K., Wongpichet, S., Phuphuphud, A., and Posom, J. (2022). Achieving robustness across different ages and cultivars for an NIRS-PLSR model of fresh cassava root starch and dry matter content. *Comput. Electron. Agric.* 96, 106872. doi: 10.1016/j.compag.2022.106872

Medeiros, A. D., Silva, L. J., Ribeiro, J. P. O., Ferreira, K. C., Rosas, J. T. F., Santos, A. A., et al. (2020). Machine learning for seed quality classification: An advanced approach using merger data from FT-NIR spectroscopy and X-ray imaging. *Sensors* 20, 4319. doi: 10.3390/s20154319

Morante, N., Ceballos, H., Sánchez, T., Rolland-Sabaté, A., Calle, F., Hershey, C. , et al. (2016). Discovery of new spontaneous sources of amylose-free cassava starch and analysis of their structure and techno-functional properties. *Food Hydrocoll* 56, 383–395. doi: 10.1016/j.foodhyd.2015.12.025

Mortensen, A. K., Gislum, R., Jørgensen, J. R., and Boelt, B. (2021). The use of multispectral imaging and single seed and bulk near-infrared spectroscopy to characterize seed covering structures: Methods and applications in seed testing and research. *Agriculture* 11, 301. doi: 10.3390/agriculture11040301

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* 51, 350–365. doi: 10.1016/j.cageo.2012.08.023

R Core Team (2021) *R: A language and environment for statistical computing*. Available at: https://www.R-project.org/.

Rojanaridpiched, C., Vichukit, V., Ceballos, H., Aeimnaka, P., Pumichai, C., and Piyachomkwan, K. (2020). Development of waxy starch cassava varieties in Thailand. In *9thStarch World Asia Conference (CMT)*, Bangkok, Thailand.

Sánchez, T., Dufour, D., Moreno, I. X., and Ceballos, H. (2010). Comparison of pasting and gel stability of waxy and normal starches from cassava, potato, maize, and rice under thermal, chemical and mechanical stress. *J. Agric. Food Chem.* 58, 5093–5099. doi: 10.1021/jf1001606

Sánchez, T., Chávez, A.L., Ceballos, H., Rodriguez-Amaya, D. B., Nestel, P., Ishitani, M., et al. (2006). Reduction or delay of post-harvest physiological deterioration in cassava roots with higher carotenoid content. *J. Sci. Food Agric.* 86, 634–639. doi: 10.1002/jsfa.2371

Sánchez, T., Salcedo, E., Ceballos, H., Dufour, D., Mafla, G., Morante, N., et al. (2009). Screening of starch quality traits in cassava (Manihot esculenta crantz). *Starch/Stärke* 61, 12–19. doi: 10.1002/star.200800058

Sánchez, T., Ceballos, H., Dufour, D., Ortiz, D., Morante, N., Calle, F., et al. (2014). Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. *Food Chem.* 15, 444–51. doi: 10.1016/j.foodchem.2013.11.081

Santos, T. d., de Carvalho, C. W. P., de Oliveira, L. A., Oliveira, E. J., Villas-Boas, F., Franco, C. M. L., et al. (2021). Functionality of cassava genotypes for waxy starch. *Pesquisa Agropecuária Bras.* 56, 1. doi: 10.1590/s1678-3921.pab2021.v56.02414

Savitzky, A., and Golay, M. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36 (8), 1627–1639. doi: 10.1021/ac60214a047

Seung, D., Soyk, S., Coiro, M., Maier, B. A., Eicke, S., Zeeman, S. C., et al. (2015). Protein targeting to starch is required for localising granule-bound starch synthase to starch granules and for normal amylose synthesis in arabidopsis. *PloS Biol.* 13, 1–29. doi: 10.1371/journal.pbio.1002080

Seung, D., Boudet, J., Monroe, J., Schreier, TB., David, LC., Abt, M., et al. (2017). Homologs of protein targeting to starch control starch granule initiation in arabidopsis leaves. *Plant Cell* 29, 1657–1677. doi: 10.1105/tpc.17.00222

Souza, L. S., Alves, A. A. C., and Oliveira, E. J. (2020). Phenological diversity of flowering and fruiting in cassava germplasm. *Sci. Hortic.* 265, 109253. doi: 10.1016/j.scienta.2020.109253

Souza, L. S., Farias, A. R., Mattos, P. L. P., and Fukuda, W. M. G. (2006). *Aspectos socioeconômicos e agronômicos da mandioca*. Embrapa Mandioca e Fruticultura Tropical.

Stevens, A., and Ramirez-Lopez, L. (2022). *An introduction to the prospectr package*. R package Vignette R package version 0.2.6.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P., et al. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43 (6), 1947–1958. doi: 10.1021/ci034160g

Wang, S., Li, C., Copeland, L., Niu, Q., and Wang, S. (2015). Starch retrogradation: A comprehensive review. *Compr. Rev. Food Sci. Food Saf.* 14(5), 568–585. doi: 10.1111/1541-4337.12143

Yu, X., Yu, H., Zhang, J., Shao, S., Xiong, F., Wang, Z, et al. (2015). Endosperm structure and physicochemical properties of starches from normal, waxy, and super-sweet maize. *Int. J. Food Prop.* 18 (12), 2825–2839. doi: 10.1080/10942912.2015.1015732" 10.1080/10942912.2015.1015732

Zeeman, S. C., Kossmann, J., and Smith, A. M. (2010). Starch: its metabolism, evolution, and biotechnological modification in plants. *Annu. Rev. Plant Biol.* 61, 209–234. doi: 10.1146/annurev-arplant-042809-112301

Zhao, S. S., Dufour, D., Sánchez, T., Ceballos, H., and Zhang, P. (2011). Development of waxy cassava with different biological and physico-chemical characteristics of starches for industrial applications. *Biotechnol. Bioeng.* 108, 1925–1935. doi: 10.1002/bit.23120

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fpls.2023.1042701

Developing broad-spectrum resistance in cassava against viruses causing the cassava mosaic and the cassava brown streak diseases

Samar Sheat and Stephan Winter*

Plant Virus Department, Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

Check for updates

OPEN ACCESS

EDITED BY
Eder Jorge Oliveira,
Embrapa Mandioca e Fruticultura, Brazil

REVIEWED BY
Alfred Adebo Ozimati,
National Crops Resources Research Institute (NaCRRI), Uganda
Hamidun Bunawan,
National University of Malaysia, Malaysia
Mario Andrea Marchisio,
Tianjin University, China

*CORRESPONDENCE
Stephan Winter
✉ stephan.winter@dsmz.de

SPECIALTY SECTION
This article was submitted to Crop and Product Physiology, a section of the journal Frontiers in Plant Science

RECEIVED 12 September 2022
ACCEPTED 02 January 2023
PUBLISHED 26 January 2023

CITATION
Sheat S and Winter S (2023) Developing broad-spectrum resistance in cassava against viruses causing the cassava mosaic and the cassava brown streak diseases. *Front. Plant Sci.* 14:1042701.
doi: 10.3389/fpls.2023.1042701

Growing cassava in Africa requires resistance against the viruses causing cassava mosaic disease (CMD) and the viruses causing cassava brown streak disease (CBSD). A dominant CMD2 resistance gene from a West African cassava landrace provides strong resistance against the cassava mosaic viruses. However, resistance against cassava brown streak viruses is limited to cassava varieties that show tolerance to the disease. A recently identified cassava germplasm that cannot be infected with cassava brown streak viruses provides a new source of the resistance required to protect cassava from CBSD. We present a synopsis of the status of virus resistance in cassava and report on the research to combine resistance against CBSD and CMD. We improve the lengthy and erratic screening for CBSD resistance by proposing a virus infection and screening protocol for the viruses causing CBSD and CMD, which allows a rapid and precise assessment of cassava resistance under controlled conditions. Using this approach, we classified the virus responses of cassava lines from Africa and South America and identified truly virus-resistant clones that cannot be infected with any of the known viruses causing CBSD even under the most stringent virus infections. A modification of this protocol was used to test seedlings from cassava crosses for resistance against both diseases. A broad-spectrum resistance was identified in a workflow that lasted 9 months from seed germination to the identification of virus resistance. The workflow we propose dramatically reduces the evaluation and selection time required in a classical breeding workflow to reach the advanced field trial stage in only 9 months by conducting selections for virus resistance and plant multiplication in parallel. However, it does not bypass field evaluations; cassava resistance assessment prior to the field limits the evaluation to candidates with virus resistance defined as the absence of symptoms and the absence of the virus. The transfer of our virus screening workflow to cassava breeding programs enhances the efficiency by which resistance against viruses can be selected. It provides a precise definition of the plant's resistance response and can be used as a model system to tackle resistance in cassava against other diseases.

KEYWORDS

resistant germplasm, precise virus screening, disease tolerance, plant immunity, resistant cassava, dual virus resistance

1 Introduction

Viruses present a major threat to the cultivation of plants and, in particular, clonally propagated crops like potato, sweet potato, and cassava are menaced by a concoction of viruses from diverse genera. Host plant resistance is a key element of crop management but is limited by the availability of resistant sources. Breeding for virus resistance in a clonal crop is further complicated by the reproductive biology of the plant, the origin and inheritance of the genes conferring resistance, and the biology of the viruses threatening the crop.

In clonal plants, viruses are maintained and passed on to successive growing cycles through vegetative propagules. When plant propagation is not done *via* true seeds that clear viruses and effectively disrupt infection cycles, viruses become widely established within plant populations and evolve in uninterrupted plant infections. Thus, in vegetative crops, the challenge is to identify any resistance that interferes with the viral infection by blocking virus replication and efficiently preventing the establishment of an infection. As the virus does not replicate in a resistant plant, the infection is not carried over to the next growing cycle with vegetative propagules taken for planting. In disease-tolerant plants, viral infections are established but the diseases are associated with only a limited expression of symptoms and plant development is not critically impacted. Breeders and agronomists who assess losses from the disease set limits and thresholds for tolerance. However, since viral infections are maintained in successive cropping cycles, the tolerance assessment for a particular plant genotype may change over time because of the continuous use of virus-infected planting material that may lead to a higher incidence and severity of symptoms. Consequently, to be sustainable, disease-tolerant varieties require a strong seed system providing healthy planting materials.

Natural resistances against pathogens are mostly found in wild relatives of cultivated crops. Using such sources of resistance for breeding is often associated with major drawbacks concerning the genetic background of the progenitor carrying unwanted traits. Further impediments to rapid breeding progress are the inheritance of traits and the complex infection biology of the pathogen, complicating screening and selection of promising resistant candidates.

In this paper, we address the challenges associated with breeding cassava for resistance against the major viruses threatening the cultivation of the crop in Africa. We summarize the current knowledge of virus resistance in cassava and describe our approaches to accelerating resistance breeding by choosing defined sources of virus resistance and applying a precise virus infection workflow to shorten the virus screening processes in conventional breeding programs. The virus resistance we identified in cassava seedlings provides complete protection against the two most important cassava viral diseases in Africa, the cassava mosaic disease (CMD) and the cassava brown streak disease (CBSD).

2 Viruses infecting cassava in Africa

Two viral diseases, CMD and CBSD, caused by viruses from different families with distinct and diverse genomes and unique biological characteristics, threaten cassava in Africa. The major impact of these viral diseases is yield loss from severe symptoms on leaves leading to reductions in tuber sizes (CMD) (Thresh et al., 1994; Pita et al., 2001), necrosis of tuberous roots (CBSD) (Hillocks and Thresh, 2000; Kawuki et al., 2019), and plant decline (CMD & CBSD).

The viruses causing CMD are endemic in Africa (Fondong, 2017) and the disease is present wherever cassava is grown on the continent. The cassava mosaic begomoviruses comprise distinct virus species (Patil and Fauquet, 2009; Legg et al., 2015) causing a similar disease in cassava (Figures 1A–C), and all are readily transmitted by the whitefly *Bemisia tabaci*. This makes controlling the diseases and restricting their spread in open fields challenging; thus, host plant resistance is the most effective disease control (Figure 1D).

The viruses causing CBSD are the constituents of the current pandemic across East African countries, with epicenters in Uganda, Tanzania, Kenya, and Mozambique, and extending to neighboring countries (DR Congo and Zambia), where they present acute threats to cassava cultivation. Plant growth and development generally remain unaffected by the disease identifiable by characteristic leaf symptoms on older leaves (Figure 2B). However, the viruses cause root necrosis, and this destruction of the tubers renders them inedible (Figure 2A) (Nichols, 1950; Hillocks et al., 2001). The two distinct ipomovirus species causing CBSD, the cassava brown streak virus (CBSV) and the Ugandan cassava brown streak virus (UCBSV) (Winter et al., 2010), have complex infection strategies (Sheat et al., 2021) which complicate virus diagnosis and assessment of the disease. The viruses are inefficiently transmitted by *B. tabaci* (Maruthi et al., 2005) (Figure 2C) and their spread is bound to seasons with high whitefly populations. Human-assisted spread is the main pathway for their distribution. Although phytosanitary options exist, genetic

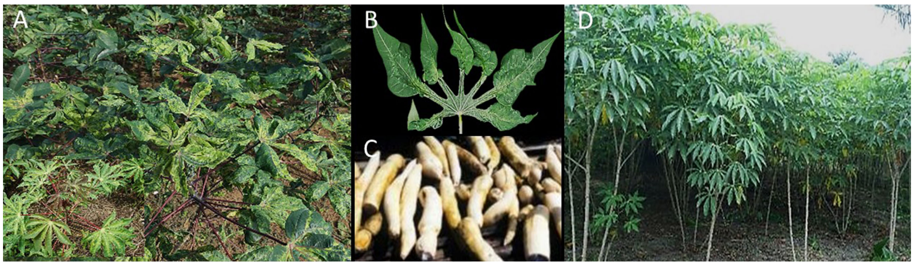


FIGURE 1
Severe mosaic symptoms of CMD **(A)**, leaf deformation **(B)**, and plant stunting in sensitive varieties result in small-sized root tubers **(C)**. The disease does not affect the highly resistant cassava variety TME 419 growing in an open field near Yangambi, DR Congo **(D)**.

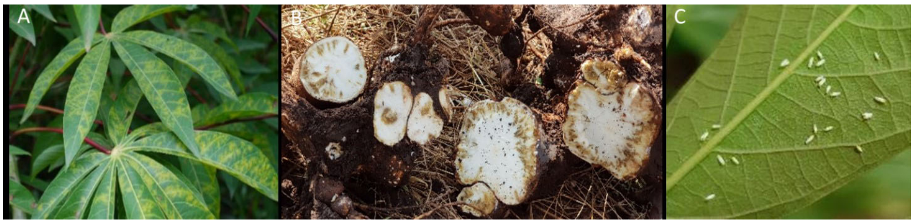


FIGURE 2
Vein chlorosis, and yellow blotches on leaves **(A)**, severe necrosis symptoms on tuberous roots **(B)** from CBSD on a sensitive cassava variety. Symptoms are mostly visible on older leaves and on tuberous roots reaching maturity. Semi-persistent virus transmission by the vector **(B)** *tabaci* **(C)** depends on high numbers of adult insects.

resistance against the viruses causing CBSD is key to the cultivation of healthy cassava and preventing further spread and transboundary movement of the disease.

2.1 Recovery resistance and immunity in cassava against begomoviruses causing cassava mosaic disease

Breeding for virus resistance in cassava goes back to the Amani breeding program in Tanzania. It started in 1937 (Jennings, 1957), when CMD resistance from the wild relative *Manihot glaziovii* was introgressed in African cultivars. Seeds (clone 58308) from this interspecific hybrid backcrossed against *M. esculenta* were used extensively in the IITA breeding program, resulting in TMS 3001, TMS 30395, and TMS 30572 (Hahn et al., 1980). The improved cassava varieties had resistance against CMD, showed good breeding values, and, consequently, were widely distributed throughout the cassava regions of Africa. Their inherent CMD1 resistance, originating from *M. glaziovii*, is polygenic and recessive (Akano et al., 2002). CMD1 cassava lines can become infected with the virus but respond with milder symptoms and some eventually recover from symptoms and appear healthy while the infection persists. However, CMD1 resistance does not sufficiently protect against the species and strains of East African cassava mosaic virus (EACMV) now prevalent in East and Central Africa. CMD1 varieties respond with more severe symptoms to EACMV species and strains and do not recover from the disease.

Intensive search efforts led to the discovery of virus resistance in the west African landrace TME 3 (Akano et al., 2002), which provides a high level of resistance against many African and East African cassava mosaic viruses that can completely protect cassava against begomovirus infections. This CMD2 resistance has a dominant inheritance and is found in TME 204 (TME 419, Obama), Albert, Nsansi (TME 96/0160), Tz130 (NaroCASS1), and many other cassava lines and varieties. Today this is the basis of begomoviruses resistance in cassava.

In our laboratory, we infected several cassava lines carrying CMD2 with a broad range of begomovirus isolates, including African cassava mosaic virus, East African cassava mosaic Cameroon virus, East African Cassava virus-Uganda, and Sri Lankan cassava mosaic virus, and found that begomoviruses could not establish themselves in CMD2 cassava lines. This immunity was effective against all known begomoviruses and, although leaf symptoms on a few leaves initially developed on some cassava lines, virus replication was not further supported and the plants remained symptom-free and free of the virus. Despite its monogenetic nature and its wide use in modern varieties, the resistance provided is robust; during more than 20 years of its use, resistance breaking has never been observed. Recent evidence suggests that the outstanding characteristics of CMD2 resistance in cassava are associated with mutations in the DNA polymerase δ subunit 1 (MePOLD1) located within the CMD2 locus on chromosome 12 (Lim et al., 2022).

Introducing CMD2 to confer begomovirus resistance in cassava is an ideal breeding target because the resistance is clearly defined. High-performing African varieties with CMD2 resistance are available and breeding tools are on hand to support a controllable and reproducible screening process (Okogbenin et al., 2012; Rabbi et al., 2014; Thuy et al., 2021).

2.2 Tolerance in cassava against ipomoviruses causing cassava brown streak disease

Cassava brown streak disease has been known for a long time, and an early report on CBSD in Tanzania (Storey, 1936) was swiftly followed by a resistance breeding program at the Amani research station (Jennings, 1957) despite the causal agent(s) of the disease not being known (Hillocks and Thresh, 2000). Similar to CMD, interspecific hybrids with wild relatives of *M. esculenta* were generated and one offspring of the program, clone 46106/27, also known as Namikonga (syn. Kaleso), became an important source of CBSD resistance. Namikonga was less affected by the disease and developed only moderate leaf symptoms and limited necrosis on tuberous roots. Namikonga was considered tolerant, and when CBSD-resistance breeding intensified in the early 2000s, Namikonga was incorporated into many crosses like NASE 1 and NASE 14 (Kawuki et al., 2016). In recent years, cassava lines with resistance against U/CBSV have been developed that show only mild symptoms on leaves and stems when infected and much less root necrosis (Jennings, 1957; Kaweesi et al., 2014; Kawuki et al., 2016; Masinde et al., 2018; Mukiibi et al., 2019). Nevertheless, despite progress to enhance the level of tolerance, cassava genotypes with high levels of resistance have not yet been found (Bart and Taylor, 2017).

We infected cassava germplasm from South America and cassava varieties from Africa with well-characterized virus isolates (Sheat et al., 2019) and followed the virus infections over many months. We confirmed earlier reports (Ogwok et al., 2014) on the high sensitivity of NASE 14 and NASE 3 to CBSV and showed that KBH 2006/18 and KBH 2006/26 (Mkuranga), two varieties that were considered immune (Anjanappa et al., 2016), can in fact be infected with such viruses (Sheat et al., 2019). Finally, we concluded that all African cassava varieties were susceptible to the viruses and responded to the disease with mild to severe leaf symptoms (Table 1).

In our infection studies, we recorded pronounced differences in plant responses against the two viruses, the most striking being disease progress. While CBSV symptoms developed within weeks after grafting, it could take many months (6-8 months), even under stringent virus infection conditions, before UCBSV symptoms became evident. Moreover, it could even take much longer before root necrosis symptoms became visible. Secondary plant infections, from infected cuttings, showed root necrosis earlier because the higher virus loads in persisting virus infections led to early tissue necrosis in developing tubers that increased along with secondary growth. Thus, the extent of root necrosis was correlated with the length of infection and the species of the virus.

Furthermore, virus species-specific responses were recorded for several cassava genotypes. The popular variety, "Mkuranga" responded with mild symptoms when infected with UCBSV but showed severe leaf symptoms and wilting when infected with CBSV. TMS 30572 was highly sensitive to CBSV but this genotype could not be infected with UCBSV. In contrast, the breeding line KBH 2016B/504 could not be infected with CBSV (Table 1) but was sensitive to UCBSV. Both cassava genotypes could be highly interesting sources of resistance; however, their potential was not evident when the diseases were evaluated without resolving the causal viruses. Thus, knowledge about the virus species present in a particular genotype is a prerequisite for reproducible and comparable resistance/tolerance evaluations.

Kaweesi et al. (2014) evaluated the response of NASE 14 to CBSD infection in the field and recorded in some plants (15) a complete absence of symptoms, while others (4) showed mild leaf symptoms only, and two plants had a high incidence and severity of necrosis symptoms on tuberous roots. The latter observation indicated that this variety was highly susceptible to CBSD but may not become readily infected; thus, the absence of symptoms may be due to a lack of infection. In any case, such variations cause uncertainties concerning the assessment of CBSD disease/tolerance due to the lack of control over virus infection processes. In field situations, the transmission of U/CBSV was highly erratic and the time points of the virus infections could vary dramatically between each individual plant, which then led to highly variable root necrosis symptoms and severity scores.

The virus infection and screening protocol we adopted in our laboratory reduced the uncertainties associated with virus species and time point of infection. A highly effective plant infection with a known virus isolate ensured that almost 100% of the plants become infected at a given time point. This assured that plant responses against the viruses are reproducible.

In all our experimental infections, and complemented by several years of field trials in DR Congo, the advanced breeding lines KBH 2016B/504 and KBH 2016B/521 have shown extraordinary resilience against U/CBSV (Figure 3). Although susceptible to the viruses, it was very difficult, even under our stringent conditions, to establish infections. Only limited CBSD symptoms were found on leaves, and

TABLE 1 Response of cassava lines and varieties upon infection with CBSV and UCBS.

Name/accession	CBSV	UCBSV
KBH 2016B/504	S0	S+
KBH 2016B/185	S++	S+
KBH 2016B/521	S+	S++
KBH 2016B/020	S+++	S+
KBH 2016B/087	S+++	S+
Yizaso	S+++	not tested
Eyope	S+++	not tested
NaroCASS1 (TZ 130)	S++	S+
Orera	S++	not tested
Mkuranga (KBH 2006/26)	S+++	S+
Kipusa (KBH 2002/066)	S+++	S+
Mkumbozi (MM96/4684)	S+	S+
Pwani (B2c20-65)	S++	S+
Mkumba (3C20-10)	S+	S+
Kizimbani	S+++	S+
Kiroba	S+	S+
Nase 19 (72-TME14)	S+++	S+
Nase 1 (TMS 60142)	S+	not tested
Nase 3 (TMS 30572)	S++	S0
Nase 14 (MM192/0248, MM96/4271)	S+++	not tested
TME 419 (TME204, Obama)	S++	S++
MM2006/0123	S+	S+
MM2006/0128	S+	S+
NaroCASS2 (MM2006/0130)	S+	S+
UG120198	S+	S+
UG120001	S+	not tested
UG120024	S+	S+
UG120156	S+	S+
Game changer	S+	not tested
Poundable (TME 693)	S+	not tested

S, sensitivity status; +++, very severe leaf symptoms, deformation, wilting, plant death; ++, severe leaf symptoms; + mild to moderate leaf symptoms; 0, plant cannot be infected.

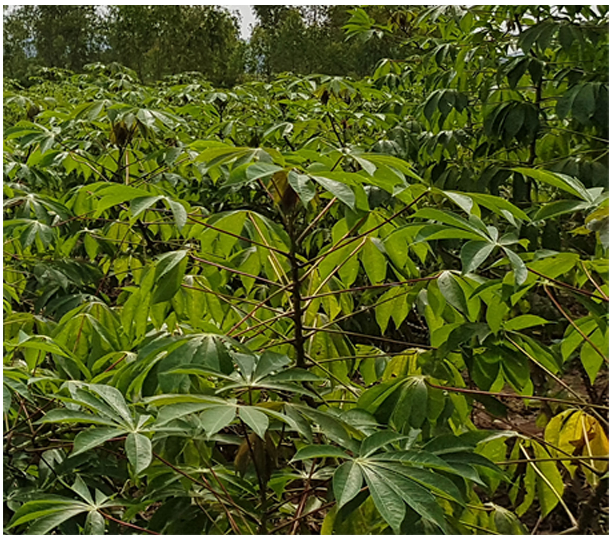


FIGURE 3
The advanced breeding line KBH 2016B/521, 6 months after planting in an epidemic zone for CBSD and CMD near Uvira, DR Congo. The line is free of symptoms after 3 consecutive growing seasons.

root necrosis eventually developed but was limited to small areas of the roots only. The plants are highly resistant to CMD and show good field performance, which emphasizes their potential as parents for further cassava improvement.

2.3 Immunity, differential resistance, and tissue-specific resistance in cassava against ipomoviruses causing cassava brown streak disease

Our search for new sources of U/CBSV resistance in South American cassava germplasm (Sheat et al., 2019) was motivated to find cassava lines that could not be infected by any U/CBSV isolate, thus expressing an immunity status similar to the begomovirus response provided by CMD2. This would create complete protection against CBSD and resolve ambiguities associated with categories of tolerance, breaking of tolerance, and concerns about persisting virus infections in clonal crops.

In a stringent virus screening workflow applied to approximately 300 cassava germplasm lines of the CIAT collection, we infected cassava plants with the most severe CBSV isolates (DSMZ PV949; FN434436), and those with virus symptoms and testing positive by qRT-PCR were eliminated. Plants that stayed healthy were further subjected to infections with UCBSV.

Only three lines passed this stringent virus screening, and high resistance against U/CBSV viruses was identified in COL 40, COL 2182, and PER 556 (Table 2), the first two varieties originating from Columbia and the latter from Peru. Even under high virus pressure from a grafted virus-infected branch, these cassava lines did not become infected. There were no symptoms expressed and no virus was detected in any tissue. U/CBSV remained in the phloem

TABLE 2 Cassava germplasm from South America with resistance against cassava brown streak viruses.

DSMZ acronym.	CIAT accession	CBSV status	UCBSV status
DSC 118	COL 40	resistant	resistant
DSC 167	COL 2182	resistant	resistant/susceptible*
DSC 196	ECU 41	resistant	susceptible
DSC 250	PER 221	resistant	susceptible
DSC 269	PER 556	resistant	resistant
DSC 120	COL 144	resistant	susceptible
DSC 258	PER 333	resistant	root restricted
DSC 199	ECU 159	root restricted	susceptible
DSC 257	PER 315	root restricted	susceptible
DSC 260	PER 353	root restricted	root restricted
DSC 261	PER 368	root restricted	not tested
DSC 272	PER 597	root restricted	susceptible
DSC 122	COL 262	root restricted	susceptible
DSC 248	PER 206	root restricted	susceptible
DSC 251	PER 226	root restricted	susceptible
DSC 186	CUB 40	susceptible	susceptible
DSC 142	COL 1107	susceptible	not tested

*Under specific experimental conditions, the line can be infected with UCBSV only (Sheat et al., 2021).

companion cells and there was no virus replication (Sheat et al., 2019; Sheat et al., 2021). We identified further cassava lines (e.g. PER 333 and PER 353) that restricted U/CBSV to the roots associated with necrosis symptoms, while leaves remained free of symptoms and virus infection. In this case, the virus replicated in phloem companion cells but was not able to translocate to adjacent parenchymatic tissues of stems for replication (Sheat et al., 2019; Sheat et al., 2021).

The cassava line COL 40 has been subjected to field infections for several growing cycles and, so far, U/CBSV have never been detected, emphasizing the outstanding resistance performance of this line. Similarly, the resistance against U/CBSV identified in the South American cassava germplasm accessions PER 556 and COL 2182 (Table 3) is considered plant immunity: the lines do not support virus replication, and the cassava brown streak disease does not establish.

3 Breeding for resistance against cassava mosaic viruses and cassava brown streak viruses

The South American cassava lines selected for U/CBSV resistance (Table 3) were very sensitive to CMD and instantly became infected with severe disease symptoms. Thus, our NextGen Cassava partners, CIAT (Columbia), IITA (Uganda), and TARI (Tanzania) used the new sources of CBSD resistance to generate crosses between South American and African lines (Figure 4) to also include the most promising CBSD resistant lines COL 2182, PER 556, and COL 40. The latter is currently the most widely used CBSD parent because of its readiness to flower and its resilience to CMD in Africa.

3.1 Immunity against U/CBSV is expressed in F1seedling populations

Seedlings from crosses comprising U/CBSV resistant parents (Table 3) were subjected to a stringent virus-resistance screening (Sheat et al., 2022). Since the resistance status of both seedling parents was known, we modified our resistance discovery workflow (Sheat et al., 2019) and adopted a more rapid virus screening process that would identify/confirm virus resistance/susceptibility in seedlings within a few weeks if this resistance phenotype was evident in F1 seedlings.

The high throughput virus screening workflow consisted of two cycles (identification and confirmation). In the first cycle, we grafted scions from CBSV-infected plants to infect each seedling. Seedlings that showed virus symptoms and tested positive with qRT-PCR (Sheat et al., 2019) were eliminated from further testing. Seedlings that tested negative were grafted with scions of cassava plants that were mixed-infected with UCBSV and EACMV-UG (GenBank accessions UCBSV, MW961202; EACMV-UG OL44492, OL444943), as described in (Sheat et al., 2022). All plants that passed cycle two and tested negative by qRT-PCR and qPCR for U/CBSV and EACMV entered the confirmation cycle, in which all steps of the workflow were repeated with a higher number of plants. Along with the confirmation cycle, the resistant cassava candidates were transferred and established in African fields to evaluate virus resistance under natural conditions. All experiments, including molecular testing, are described in detail (Sheat et al., 2019). A graphical overview of this workflow and further descriptions can be found in (Sheat et al., 2022).

TABLE 3 Cassava crossing populations generated at CIAT and IITA with U/CBSV-resistant parents.

Population Nr.	Mother	Father
1 CIAT	PER 353	GM 7673-3
	GM10054B-1	PER 221
	GM10054B-1	PER 353
	GM10054B-2	PER 353
	GM10055B-2	PER 353
	GM10062-1	PER 353
	C 33	PER 221
	C 33	PER 353
	C 39	PER 353
	C 243	PER 353
	C 413	PER 353
	COL 40	**C 33**
	COL 144	GM 7673-3
	COL 144	GM10055B-1
	COL 144	GM10055B-2
	COL 144	C 19
	COL 144	**C 33**
	COL 144	C 39
2 IITA	KBH 2016B/504	TME 14
	COL 40	KBH 2016B/185
	COL 40	KBH 2016B/087
	COL 40	TME 14
3 CIAT	COL 144	C 19
	COL 144	C 39
	GM6127-15	PER 221
	GM7672-8	PER 221
	COL 40	GM6127-13
	COL 144	TME 3
	COL 40	NN
	ECU 41	NN
4 IITA	COL 40	KBH 2016B/504
	COL 40	TME 14
	COL 40	MM2016/1487

COL 40 crosses provide broad-spectrum resistance against all U/CBSV viruses. C33, TME3, and TME14 have proven resistances against cassava mosaic viruses.

The seeds obtained from the breeding programs (Table 3) included crosses with: COL 40, providing complete immunity against U/CBSV; PER 221, which has a differential resistance

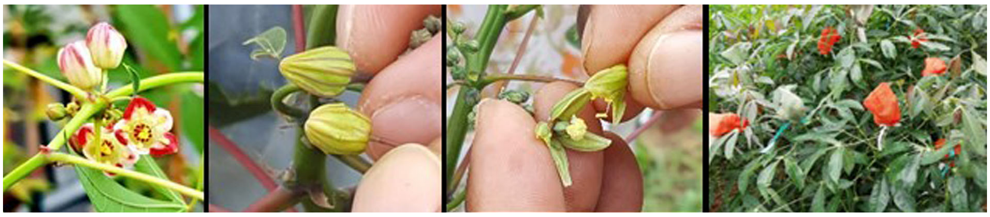


FIGURE 4
Cassava flowers: unusual colors mark the flowers of some South American x African cassava crosses (left), COL 2182 parents setting seeds at the TARI crossing block in Maruku, Tanzania.

against CBSV; and PER 353, in which U/CBSV remains restricted to the roots. Complete control over CBSD can only be reached when COL 40 is used as a CBSD parent. However, crosses with other parents can provide insights into the resistance mechanisms.

Infecting seedlings from population 1 crosses (Table 3) with CBSV resulted in virus infections that became readily evident with symptoms developing within six weeks after grafting. Several seedlings in population 1 families did not become infected with CBSV; there were no symptoms indicating for virus infection and the virus was not detected. As the resistance phenotype was visible in the F1 population, we can assume that the CBSD resistance we identified in South American germplasm is a dominant trait.

Virus infection assays comprising further populations (Table 3) and higher numbers of seedlings are still ongoing to assess inheritance from infecting a large number of seedlings from different crossing families (Table 3). However, it is already clear that a resistance phenotype expressed in F1 as a binary response radically facilitates selection processes and speeds up resistance breeding.

3.2 Broad spectrum immunity against viruses causing CMD and CBSD

We screened for resistance against both diseases by subjecting the CBSV-resistant seedlings to further infections with UCBSV and a severe isolate of East African cassava mosaic virus (EACMV-UG). As COL 40 is immune to all U/CBSV isolates, UCBSV testing was taken as further confirmation of the broad spectrum of the resistance.

Graft transmission of the cassava mosaic virus resulted in infections in susceptible seedlings within 3 to 4 weeks. In unclear symptomatology, excision of the apical parts of a graft-infected plant (comprising the first three leaves) provoked new leaf flushes, along with the expression of pronounced symptoms. A persistent symptomatic phase verified a cassava mosaic virus infection. Longer observation times were needed to identify true CMD resistance in seedlings because resistant plants can initially develop symptoms on a few leaves but thereafter recover, with new leaves free of both symptoms and the virus.

In this first screening for dual resistance against viruses causing CBSD and CMD, we identified five seedlings from the 18 families of population 1 (Table 4) having complete immunity. The plants stayed healthy and virus-free even under high virus pressures from grafted virus-infected scions.

From our predictions, only seedling 12-1 (Table 4) carries the broad-spectrum U/CBSV resistance from its COL 40 parent. Even after 18 months of infection, the four seedlings with predicted sensitivities to UCBSV did not show UCBSV symptoms and the virus was never detected.

While this warrants further explanation, it also discloses a weakness of the glasshouse-based virus testing. While this virus screening is very powerful for rapidly identifying virus-susceptible plants, proof of virus resistance/immunity can only be comprehensively provided when tuberous roots are tested. This is very difficult to achieve under screen/glasshouse conditions; hence, a confined field trial with infected plants at early screening stages is needed to provide further clarifications on the fate of the virus in an infected plant and on the immunity status of the genotype.

Phenotyping for virus resistance becomes more complex and lengthier when the chosen parents have a level of tolerance against U/CBSV. This is because the infection processes are dramatically delayed and thus viral infection and phenotyping of the best-predicted crossing combinations, e.g., COL 40 x KBH 2016B/504 (Table 3, 4 IITA), may require controlled infections in the field followed by prolonged observation times. However, because seedlings from virus-sensitive crosses have already been eliminated, the efforts can focus on fewer final candidates.

The 12-1 seedling (DSC 493) (Figure 5) and the other four candidates selected from population 1 (Table 4) are currently being subjected to confirmation-round testing under greenhouse conditions. At the same time, the lines are being grown at three

TABLE 4 Cassava seedlings (population 1) with resistance against viruses causing CMD x CBSD.

Mother	Father	Nr. of resistant seedlings	Name
PER 353	GM 7673-3	1	1-1 (DSC 673)
C 33	PER 221	2	8-1, 8-10 (DSC 516, DSC 525)
COL 40	C 33	1	12-1 (DSC 493)
COL 144	C 39	1	18-8 (DSC 510)

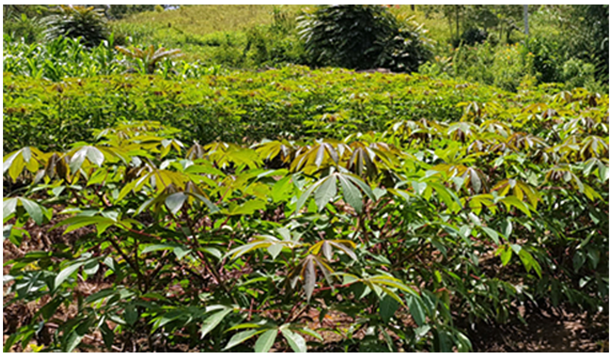


FIGURE 5
Cassava seedling DSC 493, 6 months after planting in an epidemic zone for CBSD and CMD near Uvira, DR Congo. The line is free of symptoms and shows vigorous growth.

cassava stations in Africa (DR Congo AVPD, DR Congo IITA, and Tanzania IITA) to check for their resistance status and agronomic performance under natural conditions to ultimately prove their potential to mitigate the impact of CMD and CBSD in Africa.

4 Screening for resistance against viruses causing CBSD X CMD in the field

Several strategies can be followed to accelerate virus-resistance screening under field conditions. The high throughput screening protocol (Sheat et al., 2022) we developed solved the main uncertainties associated with U/CBSV resistance assessment in cassava from uncontrolled and erratic infection processes in the field. When CBSD x CMD crosses were tested (3.2), the screening started with CBSV infections of seedlings because we assumed that the parents, including COL 40, PER 221, and PER 353, would either be highly resistant or highly sensitive to CBSV. When such crosses are tested under field conditions in Africa, it can be assumed that the seedlings are either highly resistant to CMD begomoviruses from C33 and TME14 parents, or highly sensitive because South American cassava varieties lack this resistance. As whitefly transmission of CMD begomoviruses is very efficient in the field and sensitive seedlings from CBSD x CMD crosses react rapidly and with pronounced symptoms, the first step of field screening comprises monitoring of CMD symptoms to eliminate susceptible seedlings. All seedlings that did not become naturally infected are then infected by grafting with U/CBSV. The U/CBSV used for infections are sequence-characterized viruses representing the isolates predominant in the region. As such, a set of resistant candidates is created that can be subjected to further infection experiments with other virus combinations for confirmation. Resistance testing of CBSD x CMD crosses under field conditions is feasible when appropriate conditions for U/CBSV infections are established. It requires a seedling nursery, a propagation plot to maintain virus-infected cassava source plants, and a screenhouse for virus infections and to protect sensitive seedlings and rootstocks during the first weeks after grafting. When complemented with a limited laboratory infrastructure for virus testing, cassava virus resistance screening in the field converts to a precise and reproducible process to accelerate breeding.

5 Future perspectives

Resistance against the two most widely distributed and severe virus diseases, cassava mosaic disease and cassava brown streak disease, is a prerequisite for growing a healthy and productive crop in Africa. Considering the geographic extension of CBSD on the continent and the current invasion of the viruses causing CMD to spread into new regions in South East Asia, Cambodia, Vietnam, and Thailand, the incorporation of resistance to viruses should be a global requirement for cassava just like it is for the potato (*Solanum tuberosum* L.). Therefore, developing cassava with resistance against these two viral diseases is a response to the acute threat of CBSD and a preemptive measure for regions at risk. The CMD2 resistance from African cassava that protects from CMD also presents the cure against the Sri Lankan Cassava mosaic virus causing CMD in South East Asia.

CMD2 provides complete resistance against all currently known begomoviruses infecting cassava. This high resistance is considered immunity because an infecting virus cannot establish itself and the infection is aborted. Furthermore, there are no reports of resistance-breaking viruses; hence, this resistance appears to be broad-spectrum and durable. A likely explanation is that a vital interaction and critical element for geminivirus replication is disrupted.

The resistance to viruses causing CBSD in South American germplasm lines blocked cassava brown streak viruses from replication and confined the pathogens to the phloem companion cells. There is no evidence so far that the viruses can establish themselves in a plant when grafted with scions of infected plants. The resistant plants were grown in the field under disease pressure without developing symptoms or viral infections. However, because CBSD resistance was only recently found (Sheat et al., 2019) and characterized (Sheat et al., 2021), nothing is known about its mechanism and, more critically, no field data have been collected over several seasons, including assessments of tuberous roots. COL 40 provides strong resistance against a broad range of U/CBSV isolates; however, we need to consider the limited repertoire of isolates tested and the rather short time of observation. Long-term data do not exist and we cannot rule out that viruses, variants, and/or strains may not appear that escape the resistance response, accumulate in vegetative cycles, and cause disease. This can only be elucidated in virus studies accompanying field trials of CBSD x CMD crosses.

Cassava brown streak disease is a very complicated disease because of its infection biology and its impact on tuberous roots. The rapid disease phenotyping approach we have developed solves major impediments in resistance screening by providing a defined workflow for virus infection and testing. By subjecting seedlings from resistance crosses to this workflow, precise phenotyping and identification of CBSV-resistant genotypes have been made possible that provide the fundament for a genome-wide association study (GWAS) to further our understanding of the resistance and to guide future breeding.

The first generation of prototypes with CBSD x CMD immunity was selected from crosses with South American and African germplasm using our high-precision virus screening (Sheat et al., 2022). Indeed, the limited number of resistant plants does not represent a diversity sufficient for breeding populations, but does provide the resistances for further breeding. As this method is highly efficient and precise in identifying resistant candidates, it will even be more useful when advanced crosses between highly CBSV-tolerant parents (e.g., KBH 2016/504) and CBSV-immune lines (e.g., COL 40) and their progenies are to be tested.

There is no doubt that the effective and precise workflow developed for CBSD-resistance evaluation will accelerate resistance breeding. Its future lies in the transfer of the concept and methods to breeding sites. This requires a change of perspective regarding how screening and selection for virus resistance in cassava is done, but success is one step closer when the field is considered an open laboratory space.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

Conceptualization, methodology, validation, writing, review, and editing: SS, SW. All authors contributed to the article and approved the submitted version

Funding

This research was funded by the Bill & Melinda Gates Foundation through the Next Generation Cassava Breeding Project, agreement no. 84941-11220 (under prime agreement no. OPP1175661).

Acknowledgments

The authors wish to thank the Plant Virus team at DSMZ for their continuous support, motivation, and competent assistance in this demanding project. Our sincere thanks go to the NextGen partners for constructive cooperation and we are indebted to our partners in South and North Kivu, DR Congo, for supporting the resistance evaluation in the field.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akano, A. O., Dixon, A. G. O., Mba, C., Barrera, E., and Fregene, M. (2002). Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *Theor. Appl. Genet.* 105 (4), 521–525. doi: 10.1007/s00122-002-0891-7

Anjanappa, R. B., Mehta, D., Maruthi, M. N., Kanju, E., Gruissem, W., and Vanderschuren, H. (2016). Characterization of brown streak virus-resistant cassava. *Mol. Plant Microbe Interact.* 29 (7), 527–534. doi: 10.1094/MPMI-01-16-0027-R

Bart, R. S., and Taylor, N. J. (2017). New opportunities and challenges to engineer disease resistance in cassava, a staple food of African small-holder farmers. *PloS Pathog.* 13 (5). doi: 10.1371/journal.ppat.1006287

Fondong, V. N. (2017). The search for resistance to cassava mosaic geminiviruses: How much we have accomplished, and what lies ahead. *Front. Plant Sci.* 8, 408. doi: 10.3389/fpls.2017.00408

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

Hillocks, R. J., Raya, M. D., Mtunda, K., and Kiozia, H. (2001). Effects of brown streak virus disease on yield and quality of cassava in Tanzania. *J. Phytopathology-Phytopathologische Z.* 149 (7-8), 389–394. doi: 10.1046/j.1439-0434.2001.00641.x

Hillocks, R. J., and Thresh, J. M. (2000). Cassava mosaic and cassava brown streak virus diseases in Africa: A comparative guide to symptoms and aetiologies. *Root* 7 (1), 1–8.

Jennings, D. L. (1957). Further studies in breeding cassava for virus resistance. *East Afr. Agric. J.* 22 (4), 213–219. doi: 10.1080/03670074.1957.11665107

Kaweesi, T., Kawuki, R., Kyaligonza, V., Baguma, Y., Tusiime, G., and Ferguson, M. E. (2014). Field evaluation of selected cassava genotypes for cassava brown streak disease based on symptom expression and virus load. *Virol. J.* 11, 216. doi: 10.1186/s12985-014-0216-x

Kawuki, R. S., Esuma, W., Ozimati, A., Kayondo, I. S., Nandudu, L., and Wolfe, M. (2016). Eleven years of breeding efforts to combat cassava brown streak disease. *Breed. Sci.* 66 (4), 560–571. doi: 10.1270/jsbbs.16005

Kawuki, R. S., Kaweesi, T., Esuma, W., Pariyo, A., Kayondo, I. S., Ozimati, A., et al. (2019). Alternative approaches for assessing cassava brown streak root necrosis to guide resistance breeding and selection. *Front. Plant Sci.* 10, 1461. doi: 10.3389/fpls.2019.01461

Legg, J. P., Lava Kumar, P., Makeshkumar, T., Tripathi, L., Ferguson, M., Kanju, E., et al. (2015). Cassava virus diseases: Biology, epidemiology, and management. *Adv. Virus Res.* 91, 85–142. doi: 10.1016/bs.aivir.2014.10.001

Lim, Y. W., Mansfeld, B. N., Schlapfer, P., Gilbert, K. B., Narayanan, N. N., Qi, W., et al. (2022). Mutations in DNA polymerase delta subunit 1 co-segregate with CMD2-type resistance to cassava mosaic geminiviruses. *Nat. Commun.* 13 (1), 3933. doi: 10.1038/s41467-022-31414-0

Maruthi, M. N., Hillocks, R. J., Mtunda, K., Raya, M. D., Muhanna, M., Kiozia, H., et al. (2005). Transmission of cassava brown streak virus by bemisia tabaci (Gennadius). *J. Phytopathol.* 153 (5), 307–312. doi: 10.1111/j.1439-0434.2005.00974.x

Masinde, E. A., Mkamillo, G., Ogendo, J. O., Hillocks, R., Mulwa, R. M. S., Kimata, B., et al. (2018). Genotype by environment interactions in identifying cassava (Manihot esculenta crantz) resistant to cassava brown streak disease. *Field Crops Res.* 215, 39–48. doi: 10.1016/j.fcr.2017.10.001

Mukiibi, D. R., Alicai, T., Kawuki, R., Okao-Okuja, G., Tairo, F., Sseruwagi, P., et al. (2019). Resistance of advanced cassava breeding clones to infection by major viruses in Uganda. *Crop Prot.* 115, 104–112. doi: 10.1016/j.cropro.2018.09.015

Nichols, R. F. W. (1950). The brown streak disease of cassava. *East Afr. Agric. J.* 15 (3), 154–160. doi: 10.1080/03670074.1950.11664727

Ogwok, E., Alicai, T., Rey, C., Beyene, G., and Taiylor, N. J. (2014). Distribution and accumulation of cassava brown streak viruses within infected cassava (Manihot esculenta) plants. *Plant Pathol.* p, 1–12.

Okogbenin, E., Egesi, C. N., Olasanmi, B., Ogundapo, O., Kahya, S., Hurtado, P., et al. (2012). Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. *Crop Sci.* 52 (6), 2576–2586. doi: 10.2135/cropsci2011.11.0586

Patil, B. L., and Fauquet, C. M. (2009). Cassava mosaic geminiviruses: Actual knowledge and perspectives. *Mol. Plant Pathol.* 10 (5), 685–701. doi: 10.1111/j.1364-3703.2009.00559.x

Pita, J. S., Fondong, V. N., Sangare, A., Kokora, R. N. N., and Fauquet, C. M. (2001). Genomic and biological diversity of the African cassava geminiviruses. *Euphytica* 120 (1), 115–125. doi: 10.1023/A:1017536512488

Rabbi, I. Y., Hamblin, M. T., Kumar, P. L., Gedil, M. A., Ikpan, A. S., Jannink, J. L., et al. (2014). High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. *Virus Res.* 186, 87–96. doi: 10.1016/j.virusres.2013.12.028

Sheat, S., Fuerholzner, B., Stein, B., and Winter, S. (2019). Resistance against cassava brown streak viruses from Africa in cassava germplasm from south America. *Front. Plant Sci.* 10, 567. doi: 10.3389/fpls.2019.00567

Sheat, S., Margaria, P., and Winter, S. (2021). Differential tropism in roots and shoots of resistant and susceptible cassava (Manihot esculenta crantz) infected by cassava brown streak viruses. *Cells* 10 (5). doi: 10.3390/cells10051221

Sheat, S., Zhang, X. F., and Winter, S. (2022). High-throughput virus screening in crosses of south American and African cassava germplasm reveals broad-spectrum resistance against viruses causing cassava brown streak disease and cassava mosaic virus disease. *Agronomy-Basel* 12 (5). doi: 10.3390/agronomy12051055

Storey, H. H. (1936). Virus diseases of East African plants: VI. - a progress report on studies of the disease of cassava. *East Afr. Agric. J.* 34–39.

Thresh, J. M., Fargette, D., and Otim-Nape, G. W. (1994). The viruses and virus diseases of cassava in Africa. *Afr. Crop Sci. J.* 2 (4), 459–478.

Thuy, C. T. L., Lopez-Lavalle, L. A. B., Vu, N. A., Hy, N. H., Nhan, P. T., Ceballos, H., et al. (2021). Identifying new resistance to cassava mosaic disease and validating markers for the CMD2 locus. *Agriculture* 11 (9). doi: 10.3390/agriculture11090829

Winter, S., Koerbler, M., Stein, B., Pietruszka, A., Paape, M., and Butgereitt, A. (2010). Analysis of cassava brown streak viruses reveals the presence of distinct virus species causing cassava brown streak disease in East Africa. *J. Gen. Virol.* 91, 1365–1372. doi: 10.1099/vir.0.014688-0

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 17 March 2023
DOI 10.3389/fpls.2022.1079254

CRISPR-Cas9-mediated knockout of *CYP79D1* and *CYP79D2* in cassava attenuates toxic cyanogen production

Michael A. Gomez[1*], Kodiak C. Berkoff[1,2], Baljeet K. Gill[1], Anthony T. Iavarone[3], Samantha E. Lieberman[1,4], Jessica M. Ma[1,4], Alex Schultink[4†], Nicholas G. Karavolias[1,4], Stacia K. Wyman[1], Raj Deepika Chauhan[5†], Nigel J. Taylor[5], Brian J. Staskawicz[1,4], Myeong-Je Cho[1], Daniel S. Rokhsar[1,2,6,7,8] and Jessica B. Lyons[1,2*]

[1]Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA, United States, [2]Department of Molecular & Cell Biology, University of California, Berkeley, Berkeley, CA, United States, [3]California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, Berkeley, CA, United States, [4]Department of Plant & Microbial Biology, University of California, Berkeley, Berkeley, CA, United States, [5]Donald Danforth Plant Science Center, St. Louis, MO, United States, [6]US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, United States, [7]Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, Japan, [8]Chan-Zuckerberg BioHub, San Francisco, CA, United States

OPEN ACCESS

EDITED BY
Ralf Alexander Wilhelm,
Julius Kühn-Institute, Germany

REVIEWED BY
Evans N. Nyaboga,
University of Nairobi, Kenya
Rahul Mahadev Shelake,
Gyeongsang National University,
Republic of Korea

*CORRESPONDENCE
Michael A. Gomez
michaelgr@berkeley.edu
Jessica B. Lyons
jblyons@berkeley.edu

†PRESENT ADDRESS
Alex Schultink,
Fortiphyte, Berkeley, CA, United States
Raj Deepika Chauhan,
Pairwise, Durham, NC, United States

SPECIALTY SECTION
This article was submitted to Plant Biotechnology, a section of the journal Frontiers in Plant Science

RECEIVED 25 October 2022
ACCEPTED 21 December 2022
PUBLISHED 17 March 2023

CITATION
Gomez MA, Berkoff KC, Gill BK, Iavarone AT, Lieberman SE, Ma JM, Schultink A, Karavolias NG, Wyman SK, Chauhan RD, Taylor NJ, Staskawicz BJ, Cho M-J, Rokhsar DS and Lyons JB (2023) CRISPR-Cas9-mediated knockout of *CYP79D1* and *CYP79D2* in cassava attenuates toxic cyanogen production.
Front. Plant Sci. 13:1079254.
doi: 10.3389/fpls.2022.1079254

Cassava (*Manihot esculenta*) is a starchy root crop that supports over a billion people in tropical and subtropical regions of the world. This staple, however, produces the neurotoxin cyanide and requires processing for safe consumption. Excessive consumption of insufficiently processed cassava, in combination with protein-poor diets, can have neurodegenerative impacts. This problem is further exacerbated by drought conditions which increase this toxin in the plant. To reduce cyanide levels in cassava, we used CRISPR-mediated mutagenesis to disrupt the cytochrome P450 genes *CYP79D1* and *CYP79D2* whose protein products catalyze the first step in cyanogenic glucoside biosynthesis. Knockout of both genes eliminated cyanide in leaves and storage roots of cassava accession 60444; the West African, farmer-preferred cultivar TME 419; and the improved variety TMS 91/02324. Although knockout of *CYP79D2* alone resulted in significant reduction of cyanide, mutagenesis of *CYP79D1* did not, indicating these paralogs have diverged in their function. The congruence of results across accessions indicates that our approach could readily be extended to other preferred or improved cultivars. This work demonstrates cassava genome editing for enhanced food safety and reduced processing burden, against the backdrop of a changing climate.

KEYWORDS

cassava (*Manihot esculenta* Crantz), cyanide, cyanogenesis, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas9 (CRISPR-associated protein 9), genome editing, CYP79D, climate resilience, food safety

1 Introduction

The starchy root crop cassava (*Manihot esculenta* Crantz, also known as tapioca, yuca, or manioc) is an important staple for over a billion people in tropical and subtropical regions of the world, including roughly 40% of Africans (Nweke, 2004; Lebot, 2019). It is an excellent food security crop due to its tolerance for drought and marginal soils, and because its tuberous roots can remain in the ground until needed (Howeler et al., 2013). A major challenge, however, is the presence of toxic cyanogenic compounds (e.g., cyanogenic glucosides) in cassava, which must be removed by post-harvest processing to prevent cyanide exposure and illness. Cassava root processing can be laborious, results in nutrient loss, and in Africa falls disproportionately on women and girls (Chiwona-Karltun et al., 1998; Curran et al., 2009; Maziya-Dixon et al., 2009; Montagnac et al., 2009; Boakye Peprah et al., 2020). Troublingly, cyanogen levels in cassava increase under drought stress (El-Sharkawy, 1993; Okogbenin et al., 2003; Vandegeer et al., 2013; Brown et al., 2016). As drought frequency, duration, and severity are projected to increase due to climate change (Ayugi et al., 2022), cassava consumers' risk of cyanide exposure may increase as well.

Following cellular disruption (e.g., during ingestion), cyanogenic glucosides are broken down to release the toxin cyanide. Distributed throughout the body *via* the bloodstream, cyanide halts mitochondrial electron transport, thereby preventing cells from using oxygen to produce energy and causing cell death (Dobbs, 2009). The central nervous system is particularly impacted by this toxin due to its substantial oxygen demand. The risks of insufficient cassava processing include acute cyanide poisoning which can be fatal. Chronic cyanide exposure from dietary intake induces the paralytic disease konzo, is associated with neurodevelopmental deficits, and exacerbates tropical ataxic neuropathy and goiter (Nhassico et al., 2008; Nzwalo and Cliff, 2011; Tshala-Katumbay et al., 2016; Kashala-Abotnes et al., 2018). Sulfur-containing amino acids are required to detoxify cyanide in the body; thus, those with a protein-poor diet heavily reliant on cassava are particularly at risk for adverse effects from cyanide exposure (Nzwalo and Cliff, 2011). Konzo is more likely to occur in women of childbearing age and children (Baguma et al., 2021).

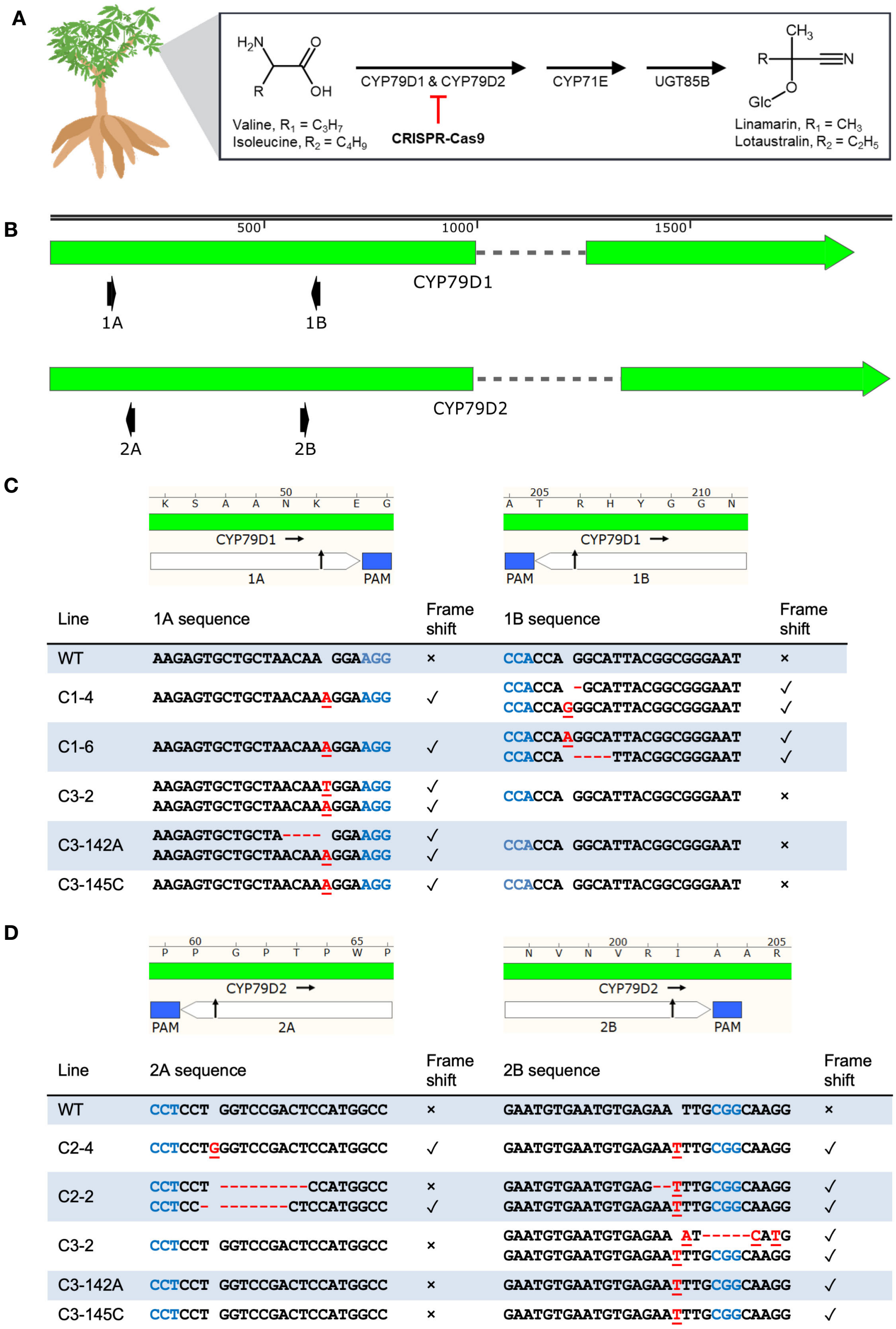
Processing to remove cyanogenic content from tuberous roots can be achieved by chipping and air drying, grinding, mashing and steeping, and/or fermentation. All require 24 hours to several days to complete. Though premature consumption exposes consumers to risk, shortcuts are sometimes taken during processing, especially when food is in short supply (Banea et al., 1992; Essers et al., 1992; McKey et al., 2010; Fitzpatrick et al., 2021). Processing approaches vary by region and specific cultivated variety (cultivar) used. There are cultural preferences for growing high cyanogenic (known as "bitter") cultivars in some contexts, for example to deter theft (Chiwona-Karltun et al., 1998). A mismatch between expected and actual cyanide levels (due to use of a different cultivar or environmental factors) may render the usual processing insufficient. Industrial scale processing of cassava poses risks to the environment and to workers through cyanide release into wastewater and the air, respectively (Adewoye et al., 2005; Ehiagbonare et al., 2009; Dhas et al., 2011). Cyanide levels above WHO recommendations have been found in commercial cassava products as well as household flour (Burns et al., 2012; Kashala-Abotnes et al., 2018).

The biosynthetic pathway for cyanogenic glucosides requires cytochrome P450 (CYP) enzymes of the CYP79 family (Luck et al., 2016). In cassava, the enzymes CYP79D1 and CYP79D2 catalyze the first, limiting step of cyanogen biosynthesis (Andersen et al., 2000) (Figure 1A). The genes *CYP79D1* and *CYP79D2* are paralogous, having arisen through the whole-genome duplication found in this lineage (Bredeson et al., 2016). Cassava's principal cyanogens, linamarin and lotaustralin, derived from valine and isoleucine, respectively, are synthesized in the canopy and transported to the storage roots (Nartey, 1968; Jørgensen et al., 2005). Linamarin accounts for greater than 90% of cassava cyanogens (Nartey, 1968).

Cyanogens play multiple roles in plants including defense and metabolism. Although cyanogens can deter herbivores (Bernays et al., 1977; Rajamma and Premkumar, 1994; Gleadow and Møller, 2014), this is not the case for cassava in all contexts or against all herbivores, possibly due to coevolution (Riis et al., 2003; Pinto-Zevallos et al., 2016). For example, the whitefly *Bemisia tabaci* detoxifies cyanogenic glucosides by enzymatic conversion to inert derivatives (Easson et al., 2021). It has been proposed that cyanogens shuttle reduced nitrogen to cassava roots for protein synthesis; increased nitrate reductase activity in roots, however, may compensate for reduced cyanogen availability (Siritunga and Sayre, 2004; Jørgensen et al., 2005; Narayanan et al., 2011; Zidenga et al., 2017). Cyanogens are also hypothesized to play a role in initiating postharvest physiological deterioration of the roots by triggering reactive oxygen species production (Zidenga et al., 2012). Modulation of cyanide levels may, therefore, bolster the longevity of harvested roots. Generation of acyanogenic cassava will facilitate further investigation of cyanogenic potential in these roles.

Cyanogen production varies naturally among cultivars (Whankaew et al., 2011; Ogbonna et al., 2021; Ospina et al., 2021). RNAi knockdown of the *CYP79D* genes reduced cyanogen levels in cassava; knockdown plants displayed wildtype morphology in soil (Siritunga and Sayre, 2003; Jørgensen et al., 2005; Piero, 2015). These observations indicate that cyanogen levels can be modulated without disrupting other desirable plant properties.

Here, we show that cassava cyanogenesis can be prevented *via* genome editing. We used CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas9 (CRISPR-associated protein 9) mutagenesis to knock out the *CYP79D* genes in the model variety 60444; the popular West African



Line	1A sequence	Frame shift	1B sequence	Frame shift
WT	AAGAGTGCTGCTAACAA GGAAGG	×	CCACCA GGCATTACGGCGGGAAT	×
C1-4	AAGAGTGCTGCTAACAAAGGAAGG	✓	CCACCA -GCATTACGGCGGGAAT CCACCAGGGCATTACGGCGGGAAT	✓ ✓
C1-6	AAGAGTGCTGCTAACAAAGGAAGG	✓	CCACCAAGGCATTACGGCGGGAAT CCACCA ----TTACGGCGGGAAT	✓ ✓
C3-2	AAGAGTGCTGCTAACAATGGAAGG AAGAGTGCTGCTAACAAAGGAAGG	✓ ✓	CCACCA GGCATTACGGCGGGAAT	×
C3-142A	AAGAGTGCTGCTA---- GGAAGG AAGAGTGCTGCTAACAAAGGAAGG	✓ ✓	CCACCA GGCATTACGGCGGGAAT	×
C3-145C	AAGAGTGCTGCTAACAAAGGAAGG	✓	CCACCA GGCATTACGGCGGGAAT	×

Line	2A sequence	Frame shift	2B sequence	Frame shift
WT	CCTCCT GGTCCGACTCCATGGCC	×	GAATGTGAATGTGAGAA TTGCGGCAAGG	×
C2-4	CCTCCTGGGTCCGACTCCATGGCC	✓	GAATGTGAATGTGAGAATTTGCGGCAAGG	✓
C2-2	CCTCCT ---------CCATGGCC CCTCC- -------CTCCATGGCC	× ✓	GAATGTGAATGTGAG--TTTGCGGCAAGG GAATGTGAATGTGAGAATTTGCGGCAAGG	✓ ✓
C3-2	CCTCCT GGTCCGACTCCATGGCC	×	GAATGTGAATGTGAGAA AT-----CATG GAATGTGAATGTGAGAATTTGCGGCAAGG	✓ ✓
C3-142A	CCTCCT GGTCCGACTCCATGGCC	×	GAATGTGAATGTGAGAATTTGCGGCAAGG	✓
C3-145C	CCTCCT GGTCCGACTCCATGGCC	×	GAATGTGAATGTGAGAATTTGCGGCAAGG	✓

FIGURE 1

CRISPR-Cas9 induces indels at *CYP79D1* and *CYP79D2* gRNA target sites in transgenic 60444 lines. **(A)** Cassava biosynthetic pathway for the cyanogenic glucosides linamarin and lotaustralin. This process primarily occurs in the leaves. The step catalyzed by CYP79D1 and CYP79D2 enzymes was selected for disruption by the CRISPR-Cas9 system. Respective side chains are labeled as R1 and R2. Glc, glucose. Chemical structures created using ACD/Chemsketch ("ACD/Chemsketch, Version 2021.1.1" 2021). **(B)** Lengths of *CYP79D1* and *CYP79D2* genes are to nucleotide scale (top bar). Exons are denoted by solid blocks and introns are represented as dashed lines. Arrowheads indicate the 3′ terminus. Diagrams of the protospacers (white) and protospacer adjacent motifs (PAMs, blue) of *CYP79D1* **(C)** and *CYP79D2* **(D)** gRNA targets are aligned to edited line genotypes. Edited lines are identified by the CRISPR construct with which they were modified (C1, C2, C3), followed by an index number (e.g., 142A). Black arrow indicates predicted CRISPR-Cas9 cut site. Lengths are to amino acid (top bar) and nucleotide (bottom table) scale. Homozygous genotypes are shown as a single sequence per line. Bi-allelic genotypes are shown as two sequences per line. Mutations on the same haplotype at 1A and 1B sites in a given mutant line are shown in the same row. Insertions are denoted by red, underlined nucleotides. Deletions are denoted by red dashes. Presence of a frameshift mutation at the corresponding target site is denoted by ✓; absence of a frameshift mutation is denoted by ×. WT, wildtype. Maps created with SnapGene.

landrace TME 419; and the improved variety TMS 91/02324, which retains robust resistance to cassava mosaic disease following regeneration through somatic embryogenesis (Chauhan et al., 2018). *Agrobacterium*-mediated CRISPR-Cas9 editing is efficient in cassava (Odipio et al., 2017; Bull et al., 2018; Hummel et al., 2018; Gomez et al., 2019; Veley et al., 2021). In contrast to RNAi knockdown, our targeted genome editing approach provides a precise, complete, and permanent loss of function, not requiring the ongoing expression of a transgene. The efficiency and precision of CRISPR-Cas9 editing are advantageous in this vegetatively propagated crop for which conventional breeding is laborious. We find that dual knockouts eliminate cyanogenic potential in all three cassava accessions. Single gene knockout lines reveal differential contribution of the two *CYP79D* genes to cassava cyanogenesis. The knockout lines described here facilitate further research into the role of cyanogens in cassava, and chart a course toward the development of acyanogenic planting materials.

2 Results

We disabled *CYP79D1* and *CYP79D2* using CRISPR-Cas9 constructs with guide RNAs (gRNAs) targeting the two genes, both singly and in combination (Table 1; Figure 1B; Supplementary Figure 1, Methods). For each gene, we selected two gRNAs with minimal off-target potential that were ~500 bp apart. We assembled these gRNAs into CRISPR constructs, and confirmed construct functionality *in planta* by adapting a geminivirus system in the surrogate model *Nicotiana benthamiana* (Baltes et al., 2014) (Supplementary Note 1; Supplementary Figure 2, Methods). These constructs were transformed using *Agrobacterium* into friable embryogenic calli (FEC) from the three cassava accessions. For each construct-accession pair, we recovered multiple independent T0 transgenic plant lines and characterized the induced mutations using Sanger sequencing (Figures 1C, D; Supplementary Figures 3, 4; Supplementary Data File 1). We found mutagenesis of targets 1A and 2B occurred at a higher frequency than of 1B and 2A, which may be due to differences in the gRNA sequences and their respective binding efficiencies. Rarely was the region between the two target sites within a gene deleted. This result was unexpected since this excision, by all CRISPR constructs, was easily detectable in the *N. benthamiana* surrogate assay. Simultaneous cleavage of the target sites and excision may have occurred more frequently in this assay due to the great number of target DNA copies delivered into the surrogate plant.

TABLE 1 CRISPR constructs used in this work.

CRISPR Construct	Target Gene(s)	gRNAs
C1	*CYP79D1*	1A, 1B
C2	*CYP79D2*	2A, 2B
C3	*CYP79D1*, *CYP79D2*	1A, 1B, 2A, 2B

We obtained four classes of CRISPR-induced mutations for each of the target loci: bi-allelic (carrying two different mutations, one for each copy of the targeted gene); homozygous (having two identical mutations of their alleles); heterozygous (carrying one mutagenized allele and one wildtype allele); and complex (carrying more than two sequence patterns, indicating genetic mosaicism or chimerism; Frank and Chitwood, 2016). For further analysis we selected mutant lines showing bi-allelic or homozygous frameshift mutations leading to premature stop codons (Supplementary Data File 2), and confirmed their genotypes using Illumina amplicon sequencing (Methods). This amplicon sequencing revealed one putative dual knockout mutant as complex, bearing some wildtype alleles as well (Supplementary Figure 5; Supplementary Note 2). This result highlights the importance of thorough sequence analysis. This line was excluded from further analysis. We found no evidence of off-target mutagenesis in 60444-derived edited lines based on the sequencing of candidate off-target sites for our gRNAs (Supplementary Tables 1, 2; Methods). Furthermore, cDNA of *CYP79D1* and *CYP79D2* transcripts in 60444-derived lines confirmed the expected sequences.

To test the impact of *CYP79D* edits on cyanogen levels, we measured linamarin and lotaustralin in leaves of edited 60444 and TME 419 *in vitro* plantlets using liquid chromatography-mass spectrometry (LC-MS). Linamarin was not detected in dual knockout lines (Supplementary Note 2; Supplementary Figures 6, 7). We also measured cyanide levels in leaves and tuberous roots of adult wildtype and mutant 60444, TME 419, and TMS 91/02324 plants, using a picrate assay (Bradbury et al., 1999). Assays were performed on greenhouse grown synchronous cohorts of plants 6–11 months after transfer to soil. Edited plants were morphologically indistinguishable from wildtypes (Supplementary Figure 8). Up to nine root samples from at least three plants per line were analyzed. As observed in *in vitro* plantlets, dual knockout lines showed no cyanogenic potential, and *CYP79D2* knockouts showed a more drastic reduction in cyanogenic potential relative to wildtype than did *CYP79D1* knockouts (Figure 2; Supplementary Figures 9, 10). As cyanogens have been implicated in nitrogen storage and transport in cassava, we tested the ability of our acyanogenic plants to grow in nitrogen limited media. In this context, dual knockout plantlets displayed a morphology typical of, and indistinguishable from, wildtypes (Supplementary Figure 11; Supplementary Table 3).

We found significant differences in cyanide content between edited and unedited lines despite the well-known variability of cyanide levels between roots of the same plant and plants of the same cultivar (Cooke et al., 1978). To account for observed

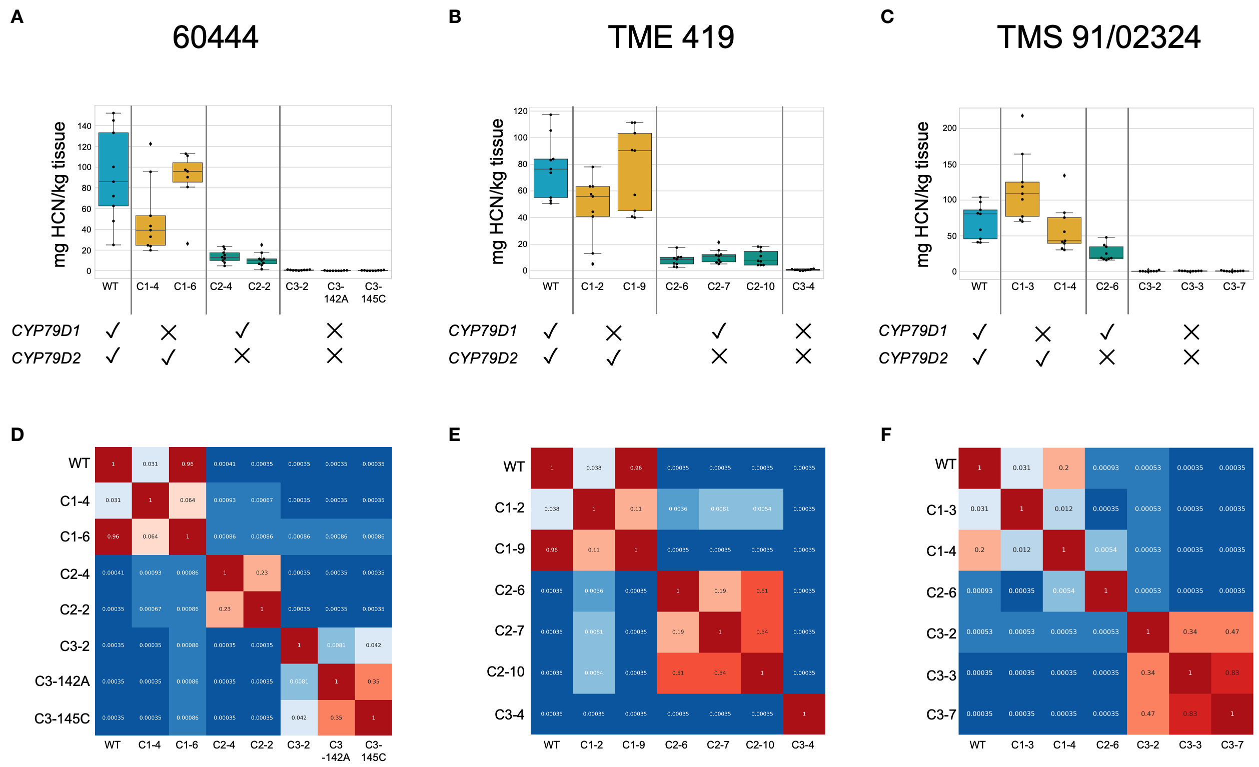


FIGURE 2

Dual knockout of *CYP79D1* and *CYP79D2* eliminates cyanide production in accessions 60444, TME 419, and TMS 91/02324. Assays were conducted 6.5, 8, and 7.5 months, respectively, after plants were transferred to soil. **(A–C)** Box and whisker plots of cyanide values in storage roots in mg HCN per kg tissue, as detected by picrate assay. Black dots are biological replicates. The median, and lower (25th percentile) and upper (75th percentile) quartiles are indicated. Whiskers define the minimum and maximum regions of the data; data points outside of these are outliers. **(D–F)** Wilcoxon Rank Sum p-values calculated from pairwise comparisons between lines of root cyanide content. Values less than 0.05, indicating the distributions of the values are statistically different between the two lines, are colored in shades of blue. Values greater than 0.05 are colored in shades of red. WT, wildtype. Edited lines are identified by the CRISPR construct with which they were modified (C1, C2, C3), followed by an index number (e.g., 142A).

variability we conducted pairwise comparisons of cyanide levels using Wilcoxon Rank Sum tests (Figure 2; Supplementary Figure 9). In dual knockout lines generated from the three accessions, the zero (or very near zero) assayed cyanide levels were distinct from those of wildtype and single knockout lines. In each of the three accessions, one *CYP79D1* knockout line had cyanide levels distinct from wildtype and the other did not. This is consistent with our observation that knocking out *CYP79D1* alone does not reliably reduce cyanogenic potential below wildtype levels. All *CYP79D2* knockout lines had cyanide levels distinct from corresponding wildtype, *CYP79D1* knockouts, and dual knockouts. This is consistent with our assessment that knocking out *CYP79D2* alone provides a near complete, but not total, reduction in cyanogenic potential.

To identify cultivars that could be prospective targets for *CYP79D* gene modification, we measured leaf and storage root cyanide content in accessions that have established transformation protocols: 60444, TME 419, TMS 91/02324, TMS 98/0505, TME 3, MCol 22, and MCol 2215 (Figure 3) (Li et al., 1996; Siritunga and Sayre, 2003; Taylor et al., 2012; Zainuddin et al., 2012; Chauhan et al., 2018). Genotype and environment both impact cassava root cyanide levels (Ogbonna et al., 2021). In our controlled environment, the ranges of root cyanide values largely overlapped; 60444 and TME 419 were significantly lower, but all others were not distinguishable from each other. Relative cyanide content in leaves was not predictive of relative cyanide content in roots across accessions. This result is consistent with previous work that showed weak correlation between leaf and storage root cyanide content, using a different cyanide assay method and field-grown landraces (Ospina et al., 2021).

3 Discussion

This study marks the first report of engineering acyanogenic plants *via* the CRISPR-Cas system. Genome editing is a powerful and heritable method to disable genes of interest for functional assessment and crop improvement. Here, we targeted cyanogenesis genes *CYP79D1* and *CYP79D2* to achieve reduction in cassava's cyanogenic potential. We demonstrated

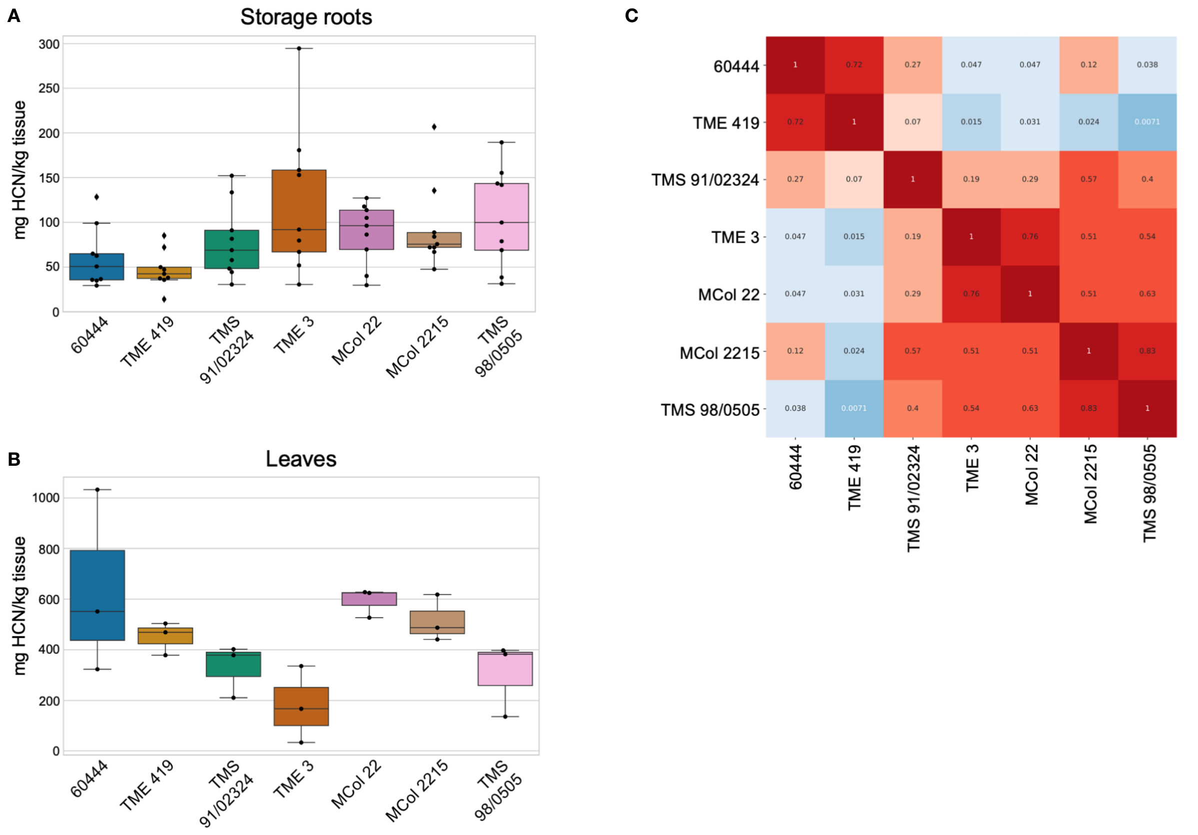


FIGURE 3
Cyanide levels from seven transformable cassava accessions. Assays were conducted seven months after plants were transferred to soil. **(A, B)** Cyanide levels in **(A)** storage roots and **(B)** leaves of the indicated cassava accessions in mg HCN per kg tissue, as detected by picrate assay. Black dots are assayed values. The median, upper, and lower quartiles are indicated. Whiskers define the minimum and maximum regions, and dots outside of these are outliers. **(A)** Root measurements. For each accession, samples were taken from a total of nine tuberous roots, from four to six plants. **(B)** Cyanide measured from leaf samples. For each accession, a total of three leaf samples were taken, from three plants. **(C)** Wilcoxon Rank Sum p-values calculated from pairwise comparisons between accessions of root cyanide content. Values less than 0.05, indicating the distributions of the values are statistically different between the two lines, are colored in shades of blue. Values greater than 0.05 are colored in shades of red.

the elimination of cyanogenic potential in eight dual knockout cassava lines. We discovered differing contributions to cyanogenesis by the two *CYP79D* genes, and observed that loss of *CYP79D2* alone is sufficient for dramatic and stable reduction in cyanide levels. These results were consistent across three different cassava accessions, in *in vitro* plantlets as well as adult plant leaves and storage roots, and *via* measurement of cyanogenic glucosides or evolved cyanide, respectively. We also assayed cyanide levels among a group of cassava lines that have elucidated transformation protocols, and are thus poised for genome engineering. The observed incongruity between relative root and leaf cyanide content indicates that these lines differ in terms of cyanogen biosynthesis, transport and/or metabolism. Understanding the mechanism and regulation of these differences may be useful for future modulation of this pathway.

The paralogous *CYP79D* genes were duplicated during the ancient paleotetraploidy in the cassava lineage (Bredeson et al., 2016). Gene duplication provides an evolutionary substrate for functional diversification (Otto and Yong, 2002), since initially redundant genes can then accumulate novel and/or complementary mutations, including variation in substrate specificity and gene expression, and may also alter the regulation of biochemical networks. Our data indicate that *CYP79D2* is likely responsible for a greater proportion of CYP79D enzymatic function than is *CYP79D1*. Differences in gene expression and/or protein sequence may explain this disparity: *CYP79D2* shows higher transcriptional activity than *CYP79D1* (Wilson et al., 2017) (Supplementary Figure 12), and lies in the cyanide biosynthetic gene cluster, whereas *CYP79D1* does not (Takos et al., 2011; Bredeson et al., 2021; Ogbonna et al., 2021). The 1000-bp regions upstream of the *CYP79D1* and *CYP79D2* transcript sequences have 50.17% identity; the amino acid sequences of CYP79D1 and CYP79D2 have 85.74% identity (Bredeson et al., 2021), with multiple mismatches in the transmembrane and P450 domains. There may be differences in spatial and/or temporal expression between these genes. It is also possible that genetic and metabolic feedback loops are influencing cyanogenic output of the intact *CYP79D* gene in single knockout lines. Thus, regulation and expression of the *CYP79D* and associated genes merit further study. *CYP79D2* knockouts showed a three- to ten-fold reduction of cyanide content, when comparing the mean values of these edited lines to mean values from the corresponding wildtypes. Hence, knockout of *CYP79D2* alone provides a straightforward mechanism for generating stably low-cyanide plants, if so desired.

A recent article reported a significant reduction of cyanide in the leaves of cassava plants *via* CRISPR-Cas9 targeting of the *CYP79D1* gene alone (Juma et al., 2022). The target sequence used, however, also matches a site in the *CYP79D2* gene (Bredeson et al., 2021). Thus, perhaps the degree of cyanide reduction observed in the edited lines resulted from mutations in *CYP79D2*, in addition to the reported mutations in *CYP79D1*.

Field testing of our edited lines with their corresponding wildtypes will allow well-controlled interrogation of the roles of cyanogens in cassava, including herbivory defense, stress response, nitrogen metabolism, and postharvest physiological deterioration. In addition to the *CYP79D* genes, the CRISPR-Cas9 system can be applied to modulate cyanogenic potential through modification of other genes of interest. A recent genome-wide association study, for example, identified two proteins that regulate cassava cyanide levels in storage roots (Ogbonna et al., 2021).

For all of our mutant lines, both copies of the target gene(s) are mutated, and we demonstrated that these changes to the genomic DNA are stably inherited *via* clonal propagation, which is typically used in cassava cultivation. We thus anticipate that the mutant alleles would be stably inherited *via* sexual propagation, and if crossed with other cassava, would segregate in a typical Mendelian fashion. The potential for changing a particular trait in outbred cassava varieties, without disrupting the complement of other traits for which they are preferred, is an alluring aspect of this precision breeding method, and can play a role in the maintenance of genetic diversity across the global population of cassava cultivars. For example, the editing approach demonstrated here can be applied to cassava varieties popular in regions of Africa projected to experience increased drought as a result of climate change, and hence higher cyanide risk (Nhassico et al., 2016).

With the increasing severity and frequency of drought, the ability to modulate cyanide levels in preferred cassava cultivars will increase in importance. TME 419, popular in Nigeria, is known as a low cyanide "boil and eat" cultivar. Under environmental conditions that would increase cyanogenesis, farmers and consumers using acyanogenic or reduced cyanide TME 419 would not have to alter farming and preparation practices intended for low cyanide roots. Previous work indicates that the disruption of cyanogenesis can impede cassava growth in the absence of sufficient reduced nitrogen (Siritunga and Sayre, 2004; Jørgensen et al., 2005). We observed no impact on morphology on plants in soil in glasshouse, potentially due to the provision of fertilizer containing ammoniacal nitrogen (Supplementary Figure 8, Methods). We also did not observe any differences between wildtype and acyanogenic plantlets grown on media with limited nitrogen (Supplementary Figure 11). If the complete absence of cyanogens proves deleterious to the crop in nitrogen-limited soils in the field, releasing low-cyanide *CYP79D2* knockout versions of farmer preferred cultivars may serve as a suitable approach for the reduction of cyanide risk.

Reduction of cyanogen content in cassava has the potential for broad socioeconomic benefits for cassava producers and consumers as well as positive effects on the environment. As detoxification of cassava can take days (Tewe, 1992), acyanogenic cassava can reduce processing time and labor. Women and girls, who disproportionately bear the burden of this labor, may then be at greater liberty to pursue other forms of work and education. If reduction of cyanide levels delays postharvest physiological deterioration of storage roots, the increased shelf life could economically benefit farmers and other stakeholders in the value chain (Zainuddin et al., 2018). At the industrial scale, processing of acyanogenic cassava would not release cyanide into wastewater, thereby reducing the labor and cost of wastewater treatment and/or the toxicity to local terrestrial and aquatic life (Adewoye et al., 2005; Silva et al., 2017). Moreover, acyanogenic cassava cultivars would be a boon for food safety and consumer wellness. As described above, excessive consumption of cyanide with a protein-poor diet can lead to neurological harm including decline in motor proficiency and cognitive performance, and, in severe cases, paralysis (Kashala-Abotnes et al., 2019). Acyanogenic cassava could preclude these debilitating conditions and open at-risk consumers and their would-be caretakers to other pursuits.

4 Materials and methods

4.1 gRNA and CRISPR construct design

Candidate target sequences were identified in *CYP79D1* and *CYP79D2* genes (Manes.13G094200 and Manes.12G133500, respectively, in cassava AM560-2 reference assembly v8.1, https://phytozome-next.jgi.doe.gov/info/Mesculenta_v8_1) of cassava using the online CRISPR-P 2.0 software (Liu et al., 2017). The software CasOT was used with default settings to search for potential off-targets in a 60444 genome assembly (Xiao et al., 2014; Gomez et al., 2019). Candidate gRNAs with minimal off-target potential and targeting sites approximately 500 bp apart were selected for assembly. Matching CRISPR targets in 60444 were verified by PCR amplification of targeted regions from genomic DNA extracts and Sanger sequencing (Supplementary Table 4).

The CRISPR-Cas9 expression entry plasmid (Thomazella et al., 2021) was re-engineered to carry the optimized gRNA scaffold with stem loop extension and A-U flip for improved Cas9 binding and gRNA transcription, respectively (Chen et al., 2014). The *BsaI* site in the backbone of the binary destination vector pCAMBIA2300 was removed *via* the QuikChange Site-Directed Mutagenesis Kit (Agilent) (Hajdukiewicz et al., 1994). The cassette carrying the CRISPR expression system was Gateway cloned into the *BsaI*-removed pCAMBIA2300 vector. The assembled binary vector with the CRISPR expression system, pCAMBIA2300 CR3-EF, requires a single cloning step for insertion of desired gRNA with white colony screen and kanamycin selection in *Escherichia coli*.

The selected CRISPR spacers were assembled into a polycistronic tRNA-gRNA (PTG) gene for multiplex targeting (Xie et al., 2015). The protocol was modified to incorporate the aforementioned stem loop extension and A-U flip. The Golden Gate cloning method was used to BsaI digest the pCAMBIA2300 CR3-EF vector and PTG ends, and then ligate the PTG into the vector. Sequences of assembled CRISPR constructs were verified *via* Sanger sequencing.

CRISPR construct activity was verified *via in planta* gemini-vector mutagenesis assay. Briefly, targeted regions were cloned into a derivative of the pLSL.D.R gemini-vector and pEAQ-HT vector maintaining replication elements for the generation of replicons bearing the target sites (Sainsbury et al., 2009; Baltes et al., 2014). Geminivirus constructs and CRISPR constructs were separately transformed into *Agrobacterium tumefaciens* strain GV3101 cultures *via* heat shock and rifampicin, gentamicin, and kanamycin selection. Transformants were grown overnight and diluted to OD600 = 0.3 each in infiltration medium (10 mM MES pH 5.6, 10 mM $MgCl_2$, 150 µM acetosyringone). After incubation at room temperature for 3 h, *A. tumefaciens* cultures bearing the CRISPR construct and corresponding geminivirus-targets construct were mixed 1:1 and infiltrated into *N. benthamiana* leaves. After five days, DNA was extracted from infiltrated leaf material *via* a modified CTAB procedure (Murray and Thompson, 1980). Frozen leaf tissue was ground by 3-mm glass beads in Minibeadbeater (Biospec Products, Inc.) and resuspended in extraction buffer (1.4 M NaCl, 100 mM Tris-HCl pH 8.0, 20 mM EDTA, 2% CTAB). Following incubation at 65°C for at least 10 min, the extract was emulsified with chloroform and centrifuged at 16,000 *g* for 5 min. DNA was precipitated from the aqueous phase with an equal volume of isopropanol and centrifuged for 10 min at 4°C. The supernatant was decanted and the DNA pellet was washed with 70% ethanol. After re-centrifugation for 2 min, ethanol was removed by pipette and air drying for 5–10 min. The DNA pellet was resuspended in 1X TE Buffer. Dissolution was advanced by incubation at 60–65°C for 5 min, or overnight incubation at room temperature. Target regions were PCR amplified and run on 1.5% agarose gel (Supplementary Table 4). CRISPR-mediated excision of 500 bp between CRISPR targets resulted in an amplified band that was visibly smaller on the gel.

4.2 Genetic transformation of cassava

Agrobacterium-mediated transformation was utilized to deliver CRISPR-Cas9 genome editing tools into friable embryogenic calli (FEC) of cassava accessions 60444, TME 419, and TMS 91/02324, with subsequent plant regeneration, following the protocol described by Taylor et al. (2012) and Chauhan et al. (2015). Somatic embryos were induced from leaf explants of *in vitro* micro-propagated plants by culture on Murashige and Skoog basal medium (MS) (Murashige and Skoog, 1962) supplemented with 20 g/L sucrose (MS2) plus 50 µM picloram. Pre-cotyledon stage embryos were subcultured onto Gresshoff and Doy basal medium (GD) (Gresshoff and Doy, 1974) supplemented with 20 g/L sucrose and 50 µM picloram (GD2 50P) in order to induce production of FEC. Homogenous four-month-old FEC were selected and used as target tissue for transformation with *A. tumefaciens* strain LBA4404 (Taylor et al., 2012) or AGL1, carrying CRISPR constructs targeting *CYP79D1* (C1), *CYP79D2* (C2), and both genes (C3). In some cases, infection was performed using sonication with a Branson 3510-DTH Ultrasonic Cleaner for three seconds. Transgenic tissues were selected and proliferated on GD2 50P containing paromomycin, prior to regeneration of embryos on MS2 medium supplemented with naphthalene acetic acid (NAA). Somatic embryos were germinated on MS2 medium containing 6-benzylaminopurine (BAP). Regenerated plants were maintained on MS2 in Phytatrays II (Sigma-Aldrich, St. Louis, MO), incubated at 28°C in high light (90–150 µmol m^{-2} s^{-1} for 16 h light/ 8 h dark conditions) and subcultured every 3 weeks.

4.3 Plant growth and maintenance

Plantlets were maintained in MS2 agar medium for stem elongation and stable growth. Well-developed growing shoots were maintained in growth chambers in Phytatrays, one to two shoots per tray, following the conditions described by Taylor et al. (2012). Regenerated plants were micro-propagated and rooted in MS2 medium containing 2.2 g/L phytagel at two to three plantlets per petri dish. For the nitrogen limitation experiment shown in Supplementary Figure 11, plantlets were grown in standard MS2 medium; MS2 medium at ½ X concentration; or MS2 medium with ½ the amount of ammonium nitrate (see Supplementary Table 3). After three weeks rooted plantlets were synchronously transferred into soil (BM7 45% bark mix, Berger) in 3" square (0.37 L^3) Kord pots and grown in a glasshouse. During soil transfer, pots were subirrigated with an aqueous solution containing, per gallon, Gnatrol WDG larvicide (Valent) at the label rate, ½ tsp Jack's Professional LX 15-5-15 Ca-Mg fertilizer (JR Peters, Inc.), ¼ tsp Jack's Professional M.O.S.T. mix of soluble traces (JR Peters, Inc.), and ½ tsp Sprint 330 (Becker Underwood). Plants were transferred to soil and watered, and the pots placed in trays with drainage holes. The trays were covered with a low (2”) dome and kept on a heating pad set to 80°F in a misting bench for 100% humidity, under 40% white shade cloth. After 8–16 days, the low domes were replaced with 6” high vented domes and the trays moved into a room without shade, with misting three times per day. The domes were removed after 9–11 days and plants placed approximately six per tray in 28-pocket spacing trays. Some plants that were lagging were kept under non-vented high domes for longer periods. For the first four weeks after transfer to soil, plants were watered with Jack’s Peat Lite 15-16-17 (JR Peters, Inc.) at 200 ppm approximately three times per week, and thereafter with Jack’s Blossom Booster 10-30-20 (JR Peters, Inc.) at 100 ppm two times per week (Taylor et al., 2012). Plants were watered with tap water on days fertilizer was not administered.

4.4 Sequence analysis

Putative transgenic lines (based on growth on antibiotic) were genotyped at the target loci. Genomic DNA extraction, PCR amplification, and Sanger sequencing were conducted as described in Gomez et al. (2019). PCR amplification and Sanger sequencing were performed using the gemini-CYP79D primers (Supplementary Table 4). The genotypes of all plants sampled in the Figure 2 picrate assay were confirmed by DNA sequence analysis of the target loci.

Percent identity of the *CYP79D1* and *CYP79D2* promoter sequences and protein amino acid sequences was evaluated using sequences derived from the cassava v6.1 genome assembly (Bredeson et al., 2016). Sequences were aligned for analysis using Clustal Omega under default settings (Madeira et al., 2019). Protein domains were predicted using SMART (Letunic et al., 2021).

4.4.1 Amplicon sequencing of selected CRISPR-Cas9 edited lines

For each line, leaf samples from two parts of an individual plant were collected for DNA extraction. PCR reactions for amplicon sequencing were performed using Phusion HF Polymerase and amplicon sequencing primers (Supplementary Table 4), and amplified for 25 cycles. Samples were deep sequenced on an Illumina MiSeq using 300 bp paired-end reads to a depth of at least 10,000 reads per sample. Cortado (https://github.com/staciawyman/cortado) was used to analyze editing outcomes. Briefly, reads were adapter trimmed and then merged using overlap into single reads. These joined reads were then aligned to the target sequence using NEEDLE (Li et al., 2015) to identify any insertions or deletions (indels) overlapping the targeted cut site. Genotypes found in less than 1% of reads were considered to be PCR or sequencing errors.

4.4.2 Off-target analysis

Identification of potential off-target loci was performed using CasOT software and a reference-based genome assembly of accession 60444, as described previously (Xiao et al., 2014; Gomez et al., 2019). Sites that contained the protospacer adjacent motif (PAM) region of the gRNA were selected and ranked according to sequence similarity to the target site, and the 2–3 highest ranking potential off-targets (those with the fewest mismatches to the gRNA spacer) for each gRNA were selected for sequence analysis (Supplementary Table 1). Genomic DNA was extracted from cassava leaves using the modified CTAB protocol described above. The selected potential off-target regions were amplified using Phusion polymerase (New England Biolabs [NEB]) and touchdown PCR (TD-PCR) (Korbie and Mattick, 2008), Phusion polymerase and 30 cycles of PCR with an annealing temperature of 63°C, or OneTaq Quick-Load 2X Master Mix with Standard Buffer (NEB) with 35 PCR cycles and an annealing temperature of 47°C. PCR primer sequences are listed in Supplementary Table 4. The TD-PCR protocol began with an annealing temperature of Tm + 10°C for the first cycling phase (Tm calculated by NEB Tm Calculator tool). The annealing temperature was then decreased by 1°C per cycle until the primers’ Tm was reached, followed by 20 or 25 cycles using the primer Tm as the annealing temperature. PCR amplicons were visualized using gel electrophoresis, then the remaining reaction was purified for sequencing *via* the AccuPrep PCR/Gel Purification Kit (Bioneer), the Monarch PCR & DNA Cleanup Kit (NEB), or SPRI magnetic nucleic acid purification beads (UC Berkeley DNA Sequencing Facility). We sequenced purified amplicons containing potential off-targets using Sanger sequencing. Putative off-target loci were then examined in SnapGene for potential sequence discrepancy with the 60444 reference sequence.

4.4.3 RT-PCR

Cassava cDNA was generated using the Spectrum Plant Total RNA Kit (Sigma-Aldrich) following Protocol A and performing On-Column DNase Digestion. Concentrations of RNA extracts were measured by NanoDrop One (Thermo Fisher Scientific). Quality of RNA was examined by first denaturing aliquots at 70°C for 5 min (followed by 4°C on ice), then electrophoresing 200 ng of RNA on 1.5% UltraPure Agarose (Invitrogen). 450–1000 ng of RNA was added to SuperScript III Reverse Transcriptase (Invitrogen) reaction mix with Oligo(dT)$_{20}$. Reaction was run for 60 min at 50°C followed by RNase H treatment. Primers were designed to amplify the *CYP79D* transcripts from the 5’ UTR to the 3’ UTR (Supplementary Table 4). 2 μL of cDNA mix was added to 50 μL Phusion High-Fidelity DNA Polymerase (NEB) reaction mix. cDNA was amplified for 35 cycles. PCR reactions were run on 1.5% agarose and desired bands were extracted. Amplicons were cloned into the Zero Blunt PCR Cloning Kit (Thermo Fisher Scientific) and 10–12 colonies subsequently sequenced *via* the UC Berkeley DNA Sequencing Facility.

4.5 Measurement of cyanogenic potential

4.5.1 Measurement of cyanogens from *in vitro* plantlets

We used LC-MS to measure linamarin and lotaustralin in *in vitro* plantlets. Regenerated transgenic plants were micropropagated and established in MS2 medium at two plantlets per Phytatray in a growth chamber at 28°C +/- 1°C, 41% relative humidity, 120–150 µmol/m^2/s light for 16 h light/8 h dark conditions. After four to five weeks plants were ready for tissue sampling. One leaf was harvested from each plantlet and stored in a plastic bag on ice until extraction, approximately 1–3 h later. For biological replicates, we harvested tissue from three plantlets per line.

Approximately 20 or 30 mg of leaf tissue was excised from fresh leaves and placed in a 1.5-mL tube (Safe-Lock, Eppendorf) with 600 or 900 µL of 85% MeOH warmed to approximately 68°C. Sample weight was recorded. Negative controls contained no tissue. A cap lock was added and the tube was floated in boiling water for 3 min, then returned to ice. One to three tubes were boiled at a time. Cooled tubes were spun down briefly. A 1:10 or 1:20 dilution was prepared from each extract, pipetted up and down to mix, and spun through a 0.45-µM spin filter (Ultrafree MC HV Durapore PVDF, EMD Millipore) for 2 min, 10,000 x *g*, 4°C. 20 µL of filtered extract was placed in a glass autosampler vial with insert (Fisher Scientific), and the LC-MS run begun the same day. Three samples were submitted from each extract, for technical replicates.

To facilitate absolute quantitation, standard stocks were prepared from solid linamarin (Cayman Chemical, purity ≥98%) and lotaustralin (Millipore Sigma, purity ≥95%) resuspended in 85% MeOH to 3 or 4 mM, aliquoted into dark glass vials, and stored at –20°C. On the day of assay, these standards were further diluted in 85% MeOH and submitted for LC-MS analysis. Lotaustralin standards ranged in concentration from 0.01 to 1 µM, and linamarin from 0.05 to 5 µM. Submitted standard samples contained both linamarin and lotaustralin in known quantities. As with extracts, three technical replicates were performed for each standard. To buffer against any potential position/timing effects, samples were analyzed by LC-MS in three consecutive cohorts, where each cohort had one technical replicate from each sample.

Samples of cassava extracts were analyzed using a liquid chromatography (LC) system (1200 series, Agilent Technologies, Santa Clara, CA) that was connected in line with an LTQ-Orbitrap-XL mass spectrometer equipped with an electrospray ionization (ESI) source (Thermo Fisher Scientific, San Jose, CA). The LC system contained the following modules: G1322A solvent degasser, G1311A quaternary pump, G1316A thermostatted column compartment, and G1329A autosampler (Agilent). The LC column compartment was equipped with a reversed-phase analytical column (length: 150 mm, inner diameter: 1.0 mm, particle size: 5 µm, Viva C18, Restek, Bellefonte, PA). Acetonitrile, formic acid (Optima LC-MS grade, 99.5+%, Fisher, Pittsburgh, PA), and water purified to a resistivity of 18.2 MΩ·cm (at 25°C) using a Milli-Q Gradient ultrapure water purification system (Millipore, Billerica, MA) were used to prepare LC mobile phase solvents. Solvent A was 99.9% water/0.1% formic acid and solvent B was 99.9% acetonitrile/0.1% formic acid (volume/volume). The elution program consisted of isocratic flow at 2% B for 2 min, a linear gradient to 6% B over 1 min, a linear gradient to 90% B over 0.5 min, isocratic flow at 90% B for 4.5 min, a linear gradient to 2% B over 0.5 min, and isocratic flow at 2% B for 16.5 min, at a flow rate of 200 µL/min. The column compartment was maintained at 40°C and the sample injection volume was 10 µL. Full-scan mass spectra were acquired in the positive ion mode over the range of mass-to-charge ratio (*m*/*z*) = 200 to 800 using the Orbitrap mass analyzer, in profile format, with a mass resolution setting of 100,000 (at *m*/*z* = 400, measured at full width at half-maximum peak height, FWHM). For tandem mass spectrometry (MS/MS or MS^2) analysis, selected precursor ions were fragmented using collision-induced dissociation (CID) under the following conditions: MS/MS spectra acquired using the linear ion trap, in centroid format, normalized collision energy: 35%, activation time: 30 ms, and activation Q: 0.25. Mass spectrometry data acquisition and analysis were performed using Xcalibur software (version 2.0.7, Thermo Fisher Scientific).

To convert LC-MS concentration values in µM to grams per kg fresh weight (fw), the following formula was used:

$$\begin{aligned} &\mu\mathrm{M} \times \text{dilution factor} \times \text{extraction vol (L)} \\ &\quad \times \text{molecular weight (g/mol)/mg fw} \\ &\quad = \text{g/kg fresh weight} \end{aligned}$$

where the molecular weight of linamarin is 247.248 g/mol and of lotaustralin is 261.272 g/mol. Concentration values reported as <LLOQ (below the lower limit of quantification) were treated as 0 µM.

4.5.2 Measurement of cyanide from adult cassava plants

Each assay cohort was synchronously grown from *in vitro* plantlets transferred to soil on the same day. Plants were collected from the glasshouse and assayed on the same day, six to 11 months after transfer to soil. The height of each plant's stem(s) was measured from the topsoil to the apical meristem in a direct line, without forcing the stems to bend, to the nearest 0.5 cm. Leaf and tuberous root samples were collected for cyanide content analysis *via* the picrate paper assay using Konzo Kit A from Australia National University Konzo Prevention Group (Bradbury, Egan, and Bradbury 1999) (https://biology.anu.edu.au/research/resources-tools/konzo-kits). Leaf samples were collected from three plants of each line. The third, fourth, and fifth expanded leaves from the top were cut perpendicular to the midribs into 0.5 cm wide pieces with clean scissors. Leaf cuts were immediately ground with a mortar and pestle. One hundred milligrams of ground leaf was loaded into a vial containing buffer paper, and 1 mL of water and the cyanide indicator paper were added immediately and the vial capped. For negative controls, no tissue was added to a vial. Positive controls were conducted using the standard provided with the kit. Sample vials were incubated overnight (minimum 12 h) at room temperature. Up to nine tuberous roots were collected from each line, with no more than three roots coming from a single plant. To buffer against any cyanide variation over the course of processing samples, root collection was staggered among groups to three to five roots per line at a time. Roots of minimum 1 cm in diameter were collected, washed, and photographed by a ruler for scale. Each root was cut at its widest section, and a 1.5 mm slice was taken crosswise *via* a kitchen mandoline. The peel (rind) was removed and 100 mg of tuberous root loaded into a vial and sealed as described above.

Indicator papers were removed from vials and compared to a cyanide color chart for an approximate cyanide content reference. These papers were then placed in 15 mL culture tubes and completely immersed in 5 mL of water. Solutions were incubated at room temperature for 30–60 min with occasional gentle stirring. The absorbance of each pipette-mixed solution was measured at 510 nm using an Ultrospec 3000 UV/Visible Spectrophotometer (Pharmacia Biotech). Absorbance was normalized to the value of the negative control (no plant sample). Absorbance was multiplied by 396 to acquire the total cyanide content in ppm (equivalent to mg HCN per kg of tissue).

4.5.3 Statistical analyses

Box and whisker plots were generated for cyanide measurements from picrate assays. The box of each plot represents the interquartile range (IQR) which is bounded by a lower quartile (Q1, 25th percentile) and an upper quartile (Q3, 75th percentile) of the data. The whiskers of each plot are defined as the approximate minima and maxima of the data, with the minimum data value defined as Q1 – 1.5 × IQR, and the maximum defined as Q3 + 1.5 × IQR. Data outside of the maximum and minimum values were considered outliers.

The Wilcoxon rank-sum statistic (also known as the Mann–Whitney *U* test) was performed on the picrate data to test whether there were statistically significant differences between the various lines, using pairwise comparisons. This method was used primarily due to the nonparametric and continuous nature of our picrate data.

For data shown in Figures 2, 3 and Supplementary Figure 9, two-group Wilcoxon rank-sum statistical comparisons between all lines were performed *via* SciPy's stats.ranksum function (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ranksums.html; Scipy v1.4.1), which uses a normal approximation of the rank sum. The p-values of each two-group test were compiled to form heatmaps. In these heatmaps, p-values less than 0.05 were considered significant and shaded blue, indicating that for two compared groups, the data appeared to come from two separate distributions. P-values greater than 0.05 were shaded red and indicated that for two compared groups, the data appeared as if drawn from the same distribution, demonstrating statistical insignificance from each other. Software versions used in these statistical analyses: Python v3.7.13; Numpy v1.21.6; Pandas v1.3.5; Matplotlib v3.2.2. Plots in Figures 2, 3 and Supplementary Figure 9 were generated in Google Colab notebooks with the Seaborn data visualization library (v0.11.2).

For data shown in Supplementary Figure 10, Wilcoxon rank-sum statistical comparisons were calculated as described above using R studio v1.1.456. Figures were generated using the ggplot2 package in R studio (Wickham, 2016).

Data availability statement

The data produced in this study are available in the article and Supplementary Material. The raw Illumina sequence data generated for this study were deposited in the NCBI Sequence Read Archive (SRA), under BioProject PRJNA906627. Vectors are available upon request to BS stask@berkeley.edu. Plants are available upon request to M-JC mjcho1223@berkeley.edu.

Author contributions

Designed the project: MG, NT, BS, M-JC, DR, JL. Molecular biology lead: MG. Gemini-vector transient assay: MG, AS. Cassava transformation: BG, RC. Genotyping of on- and off-target sites: MG, KB, SL, JM. Extractions for LC-MS: MG, BG, JL. LC-MS: AI. LC-MS plots: JL. Picrate assays: MG, KB, SL, NK, JL. Picrate assay plots and statistical analysis: KB, NK. Illumina amplicon data analysis: SW. Project leadership: JL, with co-PIs BS, M-JC, DR; Danforth Center lead NT. Wrote the paper: MG and JL; with contributions by KB, BG, AI, SL, SW, NT, M-JC; and edited by NT, M-JC, and DR. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Innovative Genomics Institute. The QB3/Chemistry Mass Spectrometry Facility received support from the National Institutes of Health (grant 1S10OD020062-01). Publication made possible in part by support from the Berkeley Research Impact Initiative (BRII) sponsored by the UC Berkeley Library.

Acknowledgments

The IGI is grateful to the Danforth Center for support in initiating its cassava transformation and regeneration platform, including Tira Jones and Danielle Stretch for technical assistance, and Claire Albin for sharing plant care protocols. The authors thank Tina Wistrom and UCB Oxford Tract greenhouse staff for plant care; Jonathan Vu and Netravathi Krishnappa, IGI Center for Translational Genomics, for Illumina Amplicon sequencing; UC Berkeley DNA Sequencing Facility; Elaine Zhang, Dominick Tucker, Xiuli Shen, and Ankita Singh for technical assistance; and Benton Cheung for cassava graphic. We are grateful to Ros Gleadow for advice on cyanide measurements, Kirsten Jørgensen for sharing the extraction protocol for LC-MS, John Young and Nikki Kong for advice on statistical analyses, and Susan Abrahamson for consultation on intellectual property. The authors would like to thank our two reviewers for helpful feedback.

Conflict of interest

The University of California Regents have filed for protection of intellectual property related to CRISPR-Cas around the world.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1079254/full#supplementary-material

A **Supplementary Materials** file including Supplementary Notes 1 and 2, Supplementary Figures 1–12, Supplementary Tables 1–4, and legends for Supplementary Data Files 1–4.

SUPPLEMENTARY DATA FILE 1
Genotypes of CRISPR-Cas9-edited 60444, TME 419, and TMS 91/02324 lines.

SUPPLEMENTARY DATA FILE 2
Predicted CYP79D amino acid sequences from selected 60444, TME 419, and TMS 91/02324 lines.

SUPPLEMENTARY DATA FILE 3
LC-MS, picrate, and height readings.

SUPPLEMENTARY DATA FILE 4
Python notebook for Figures 2, 3 and Supplementary Figure 9.

References

ACD/Chemsketch (2021) (Toronto, ON, Canada: Advanced Chemistry Development, Inc). Available at: www.acdlabs.com.

Adewoye, S. O., Fawole, O. O., Owolabi, O. D., and Omotosho, J. S. (2005). Toxicity of cassava wastewater effluents to African catfish: Clarias gariepinus (Burchell 1822). *SINET: Ethiop. J. Sci.* 28 (2), 189–194. doi: 10.4314/sinet.v28i2.18254

Andersen, M. D., Busk, P. K., Svendsen, I., and Møller, B. L. (2000). Cytochromes P-450 from cassava (*Manihot esculenta* Crantz) catalyzing the first steps in the biosynthesis of the cyanogenic glucosides linamarin and lotaustralin: CLONING, FUNCTIONAL EXPRESSION IN *PICHIA PASTORIS*, AND SUBSTRATE SPECIFICITY OF THE ISOLATED RECOMBINANT ENZYMES. *J. Biol. Chem.* 275 (3), 1966–1755. doi: 10.1074/jbc.275.3.1966

Ayugi, B., Eresanya, E. O., Onyango, A. O., Ogou, F. K., Okoro, E. C., Okoye, C. O., et al. (2022). Review of meteorological drought in Africa: Historical trends, impacts, mitigation measures, and prospects. *Pure Appl. Geophys.* 179 (4), 1365–1386. doi: 10.1007/s00024-022-02988-z

Baguma, M., Nzabara, F., Balemba, G. M., Malembaka, E. B., Migabo, C., Mudumbi, G., et al. (2021). Konzo risk factors, determinants and etiopathogenesis: What is new? A systematic review. *Neurotoxicology* 85 (July), 54–67. doi: 10.1016/j.neuro.2021.05.001

Baltes, N. J., Gil-Humanes, J., Cermak, T., Atkins, P. A., and Voytas, D. F. (2014). DNA Replicons for plant genome engineering. *Plant Cell* 26 (1), 151–635. doi: 10.1105/tpc.113.119792

Banea, M., Poulter, N. H., and Rosling, H. (1992). Shortcuts in cassava processing and risk of dietary cyanide exposure in Zaire. *Food Nutr. Bull.* 14 (2), 1–75. doi: 10.1177/156482659201400201

Bernays, E. A., Chapman, R. F., Leather, E. M., McCaffery, A. R., and Modder, W. W. D. (1977). The relationship of *Zonocerus variegatus* (L.) (Acridoidea: Pyrgomorphidae) with cassava (*Manihot esculenta*). *Bull. Entomol. Res.* 67 (3), 391–404. doi: 10.1017/S0007485300011202

Boakye Peprah, B., Parkes, E. Y., Harrison, O. A., van Biljon, A., Steiner-Asiedu, M., and Labuschagne, M. T. (2020). Proximate composition, cyanide content, and carotenoid retention after boiling of provitamin A-rich cassava grown in Ghana. *Foods* 9 (12), 1800. doi: 10.3390/foods9121800

Bradbury, M. G., Egan, S. V., and Bradbury, J.H. (1999). Picrate paper kits for determination of total cyanogens in cassava roots and all forms of cyanogens in cassava products. *J. Sci. Food Agric.* 79 (4), 593–6015. doi: 10.1002/(SICI)1097-0010(19990315)79:4<593::AID-JSFA222>3.0.CO;2-2

Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G.A., Ha, C. M., Edsinger-Gonzales, E., et al. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* 34 (5), 562–570. doi: 10.1038/nbt.3535

Bredeson, J. V., Shu, S., Berkoff, K., Lyons, J. B., Caccamo, M., Santos, B., et al. (2021). *An improved reference assembly for cassava (Manihot esculenta* Crantz*).* Available at: https://phytozome-next.jgi.doe.gov/info/Mesculenta_v8_1.

Brown, A. L., Cavagnaro, T. R., Gleadow, R., and Miller, R. E. (2016). Interactive effects of temperature and drought on cassava growth and toxicity: Implications for food security? *Global Change Biol.* 22 (10), 3461–3735. doi: 10.1111/gcb.13380

Bull, S. E., Seung, D., Chanez, C., Mehta, D., Kuon, J.-E., Truernit, E., et al. (2018). Accelerated ex situ breeding of *GBSS*- and *PTST1*-edited cassava for modified starch. *Sci. Adv.* 4 (9), eaat6086. doi: 10.1126/sciadv.aat6086

Burns, A. E., Bradbury, J.H., Cavagnaro, T. R., and Gleadow, R. M. (2012). Total cyanide content of cassava food products in Australia. *J. Food Composit. Anal.* 25 (1), 79–82. doi: 10.1016/j.jfca.2011.06.005

Chauhan, R. D., Beyene, G., Kalyaeva, M., Fauquet, C. M., and Taylor, N. (2015). Improvements in *Agrobacterium*-mediated transformation of cassava (*Manihot esculenta* Crantz) for large-scale production of transgenic plants. *Plant Cell Tissue Organ Cult.* 121 (3), 591–6035. doi: 10.1007/s11240-015-0729-z

Chauhan, R. D., Beyene, G., and Taylor, N. J. (2018). Multiple morphogenic culture systems cause loss of resistance to cassava mosaic disease. *BMC Plant Biol.* 18 (1), 1325. doi: 10.1186/s12870-018-1354-x

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., et al. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell.* 155(7), 1479–1491. doi: 10.1016/j.cell.2013.12.001

Chiwona-Karltun, L., Mkumbira, J., Saka, J., Bovin, M., Mahungu, N. M., and Rosling, H. (1998). The importance of being bitter–a qualitative study on cassava cultivar preference in Malawi. *Ecol. Food Nutr.* 37 (3), 219–455. doi: 10.1080/03670244.1998.9991546

Cooke, R. D., Howland, A. K., and Hahn, S. K. (1978). Screening cassava for low cyanide using an enzymatic assay. *Exp. Agric.* 14 (4), 367–372. doi: 10.1017/S0014479700009017

Curran, S., Anderson, C.L., Gugerty, M. K., and Cook, J. (2009). "Gender and cropping: Cassava in Sub-Saharan Africa," in *Evans School policy analysis and research (EPAR)* (Seattle, WA, USA: Evans School of Public Affairs, University of Washington).

Dhas, P. K., Chitra, P., Jayakumar, S., and Mary, A. R. (2011). Study of the effects of hydrogen cyanide exposure in cassava workers. *Indian J. Occup. Environ. Med.* 15 (3), 133–365. doi: 10.4103/0019-5278.93204

Dobbs, M. R. (2009). "Cyanide," in *Clinical Neurotoxicology: Syndromes, Substances, Environments* (Philadelphia, PA: Elsevier), 515–522.

Easson, M. L.A.E., Malka, O., Paetz, C., Hojná, A., Reichelt, M., Stein, B., et al. (2021). Activation and detoxification of cassava cyanogenic glucosides by the whitefly *Bemisia tabaci*. *Sci. Rep.* 11 (1), 13244. doi: 10.1038/s41598-021-92553-w

Ehiagbonare, J. E., Adjarhore, R. Y., and Enabulele, S. A. (2009). Effect of cassava effluent on Okada natural water. *Afr. J. Biotechnol.* 8 (12), 2816–2818. doi: 10.4314/ajb.v8i12.60890

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

Essers, A. J. A., Alsen, P., and Rosling, H. (1992). Insufficient processing of cassava induced acute intoxications and the paralytic disease Konzo in a rural area of Mozambique. *Ecol. Food Nutr.* 27 (1), 17–27. doi: 10.1080/03670244.1992.9991222

Fitzpatrick, M. C., Kurpad, A. V., Duggan, C. P., Ghosh, S., and Maxwell, D. G. (2021). Dietary intake of sulfur amino acids and risk of kwashiorkor malnutrition in Eastern Democratic Republic of the Congo. *Am. J. Clin. Nutr.* 114 (3), 925–335. doi: 10.1093/ajcn/nqab136

Frank, M. H., and Chitwood, D. H. (2016). Plant chimeras: The good, the bad, and the 'Bizzaria'. *Dev. Biol.* 419 (1), 41–535. doi: 10.1016/j.ydbio.2016.07.003

Gleadow, R. M., and Møller, B. L. (2014). Cyanogenic glycosides: Synthesis, physiology, and phenotypic plasticity. *Annu. Rev. Plant Biol.* 65 (February), 155–185. doi: 10.1146/annurev-arplant-050213-040027

Gomez, M. A., Lin, Z.D., Moll, T., Chauhan, R. D., Hayden, L., Renninger, K., et al. (2019). Simultaneous CRISPR/Cas9-mediated editing of cassava *eIF4E* isoforms *nCBP-1* and *nCBP-2* reduces cassava brown streak disease symptom severity and incidence. *Plant Biotechnol. J.* 17 (2), 421–434. doi: 10.1111/pbi.12987

Gresshoff, P. M., and Doy, C. H. (1974). Derivation of a haploid cell line from *Vitis vinifera* and the importance of the stage of meiotic development of anthers for haploid culture of this and other genera. *Z. Für Pflanzenphysiol.* 73 (2), 132–141. doi: 10.1016/S0044-328X(74)80084-X

Hajdukiewicz, P., Svab, Z., and Maliga, P. (1994). The small, versatile *pPZP* family of *Agrobacterium* binary vectors for plant transformation. *Plant Mol. Biol.* 25 (6), 989–994. doi: 10.1007/BF00014672

Howeler, R., Lutaladio, N., and Thomas, G. (2013). *Save and grow: Cassava: A guide to sustainable production intensification* (Rome: Food and Agriculture Organization of the United Nations).

Hummel, A. W., Chauhan, R. D., Cermak, T., Mutka, A. M., Vijayaraghavan, A., Boyher, A., et al. (2018). Allele exchange at the EPSPS locus confers glyphosate tolerance in cassava. *Plant Biotechnol. J.* 16 (7), 1275–1825. doi: 10.1111/pbi.12868

Jørgensen, K., Bak, S., Busk, P. K., Sørensen, C., Olsen, C. E., Puonti-Kaerlas, J., et al. (2005). Cassava plants with a depleted cyanogenic glucoside content in leaves and tubers. Distribution of cyanogenic glucosides, their site of synthesis and transport, and blockage of the biosynthesis by RNA interference technology. *Plant Physiol.* 139 (1), 363–745. doi: 10.1104/pp.105.065904

Juma, B. S., Mukami, A., Mweu, C., Ngugi, M. P., and Mbinda, W. (2022). Targeted mutagenesis of the *CYP79D1* gene *via* CRISPR/Cas9-mediated genome editing results in lower levels of cyanide in cassava. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1009860

Kashala-Abotnes, E., Okitundu, D., Mumba, D., Boivin, M. J., Tylleskär, T., and Tshala-Katumbay, D. (2019). Konzo: A distinct neurological disease associated with food (cassava) cyanogenic poisoning. *Brain Res. Bull.* 145 (February), 87–91. doi: 10.1016/j.brainresbull.2018.07.001

Kashala-Abotnes, E., Sombo, M. T., Okitundu, D. L., Kunyu, M., Makila-Mabe, G. B., Tylleskär, T., et al. (2018). Dietary cyanogen exposure and early child neurodevelopment: An observational study from the Democratic Republic of Congo. *PloS One* 13 (4), e0193261. doi: 10.1371/journal.pone.0193261

Korbie, D. J., and Mattick, J. S. (2008). Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat. Protoc.* 3 (9), 1452–1565. doi: 10.1038/nprot.2008.133

Lebot, V. (2019). *Tropical root and tuber crops. 2nd Edition* (Wallingford, Oxfordshire, UK; and Boston, MA, USA: CABI).

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: Recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49 (D1), D458–D460. doi: 10.1093/nar/gkaa937

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., et al. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* 43 (W1), W580–W584. doi: 10.1093/nar/gkv279

Li, H. Q., Sautter, C., Potrykus, I., and Puonti-Kaerlas, J. (1996). Genetic transformation of cassava (*Manihot esculenta* Crantz). *Nat. Biotechnol.* 14 (6), 736–740. doi: 10.1038/nbt0696-736

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., and Chen, L.-L. (2017). CRISPR-p 2.0: An improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant* 10 (3), 530–325. doi: 10.1016/j.molp.2017.01.003

Luck, K., Jirschitzka, J., Irmisch, S., Huber, M., Gershenzon, J., and Köllner, T. G. (2016). CYP79D enzymes contribute to jasmonic acid-induced formation of aldoximes and other nitrogenous volatiles in two erythroxylum species. *BMC Plant Biol.* 16 (1), 2155. doi: 10.1186/s12870-016-0910-5

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47 (W1), W636–W641. doi: 10.1093/nar/gkz268

Maziya-Dixon, B., Dixon, A. G.O., and Ssemakula, G. (2009). Changes in total carotenoid content at different stages of traditional processing of yellow-fleshed cassava genotypes. *Int. J. Food Sci. Technol.* 44 (12), 2350–2575. doi: 10.1111/j.1365-2621.2007.01638.x

McKey, D., Cavagnaro, T. R., Cliff, J., and Gleadow, R. (2010). Chemical ecology in coupled human and natural systems: People, manioc, multitrophic interactions and global change. *Chemoecology* 20 (2), 109–335. doi: 10.1007/s00049-010-0047-1

Montagnac, J. A., Davis, C. R., and Tanumihardjo, S. A. (2009). Nutritional value of cassava for use as a staple food and recent advances for improvement. *Compr. Rev. Food Sci. Food Saf.* 8 (3), 181–945. doi: 10.1111/j.1541-4337.2009.00077.x

Murashige, T., and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant.* 15 (3), 473–975. doi: 10.1111/j.1399-3054.1962.tb08052.x

Murray, M. G., and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8 (19), 4321–4325. doi: 10.1093/nar/8.19.4321

Narayanan, N. N., Ihemere, U., Ellery, C., and Sayre, R. T. (2011). Overexpression of hydroxynitrile lyase in cassava roots elevates protein and free amino acids while reducing residual cyanogen levels. *PloS One* 6 (7), e219965. doi: 10.1371/journal.pone.0021996

Nartey, F. (1968). Studies on cassava, *Manihot utilissima* Pohl–I. Cyanogenesis: The biosynthesis of linamarin and lotaustralin in etiolated seedlings. *Phytochemistry* 7 (8), 1307–1312. doi: 10.1016/S0031-9422(00)85629-0

Nhassico, D., Bradbury, J. H., Cliff, J., Majonda, R., Cuambe, C., Denton, I. C., et al. (2016). Use of the wetting method on cassava flour in three konzo villages in Mozambique reduces cyanide intake and may prevent konzo in future droughts. *Food Sci. Nutr.* 4 (4), 555–561. doi: 10.1002/fsn3.317

Nhassico, D., Muquingue, H., Cliff, J., Cumbana, A., and Bradbury, J.H. (2008). Rising African cassava production, diseases due to high cyanide intake and control measures. *J. Sci. Food Agric.* 88 (12), 2043–2495. doi: 10.1002/jsfa.3337

Nweke, F. I. (2004). *New challenges in the cassava transformation in Nigeria and Ghana* (Washington, D.C., USA: Intl Food Policy Res Inst).

Nzwalo, H., and Cliff, J. (2011). Konzo: From poverty, cassava, and cyanogen intake to toxico-nutritional neurological disease. *PloS Negl. Trop. Dis.* 5 (6), e10515. doi: 10.1371/journal.pntd.0001051

Odipio, J., Alicai, T., Ingelbrecht, I., Nusinow, D. A., Bart, R., and Taylor, N. J. (2017). Efficient CRISPR/Cas9 genome editing of phytoene desaturase in cassava. *Front. Plant Sci.* 8 (October), 1780. doi: 10.3389/fpls.2017.01780

Ogbonna, A. C., de Andrade, L. R. B., Rabbi, I. Y., Mueller, L. A., de Oliveira, E. J., and Bauchet, G. J. (2021). Large-scale genome-wide association study, using historical data, identifies conserved genetic architecture of cyanogenic glucoside content in cassava (*Manihot esculenta* Crantz) root. *Plant J.* 105 (3), 754–705. doi: 10.1111/tpj.15071

Okogbenin, E., Ekanayake, I. J., and Porto, M. C. M. (2003). Genotypic variability in adaptation responses of selected clones of cassava to drought stress in the Sudan savanna zone of Nigeria. *J. Agron. Crop Sci.* 189 (6), 376–389. doi: 10.1046/j.1439-037X.2003.00050.x

Ospina, M. A., Pizarro, M., Tran, T., Ricci, J., Belalcazar, J., Luna, J. L., et al. (2021). Cyanogenic, carotenoids and protein composition in leaves and roots across seven diverse population found in the world cassava germplasm collection at CIAT, Colombia. *Int. J. Food Sci. Technol.* 56 (3), 1343–1353. doi: 10.1111/ijfs.14888

Otto, S. P., and Yong, P. (2002). The evolution of gene duplicates. *Adv. Genet.* 46, 451–483. doi: 10.1016/S0065-2660(02)46017-8

Piero, N. M., Murugi, N. J., Richard, O. O., Jalemba, M. A., Omwoyo, O. R., and Cheruiyot, R. C. (2015). Determination of cyanogenic compounds content in transgenic acyanogenic Kenyan cassava (*Manihot esculenta* Crantz) genotypes: Linking molecular analysis to biochemical analysis. *J. Anal. Bioanal. Tech*. 6 (5), 264. doi: 10.4172/2155-9872.1000264

Pinto-Zevallos, D. M., Pareja, M., and Ambrogi, B. G. (2016). Current knowledge and future research perspectives on cassava (*Manihot esculenta* Crantz) chemical defenses: An agroecological view. *Phytochemistry* 130, 10–21. doi: 10.1016/j.phytochem.2016.05.013

Rajamma, P., and Premkumar, T. (1994). Influence of moisture content/equilibrium relative humidity of cassava chips on the infestation by *Araecerus fasciculatus* DeGeer (Coleoptera: Anthribidae) and *Rhyzopertha dominica* (Fabricius) (Coleoptera: Bostrichidae). *Int. J. Pest Manage*. 40(3), 261–265. doi: 10.1080/09670879409371894

Riis, L., Bellotti, A. C., Bonierbale, M., and O'Brien, G. M. (2003). Cyanogenic potential in cassava and its influence on a generalist insect herbivore *Cyrtomenus bergi* (Hemiptera: Cydnidae). *J. Econ. Entomol.* 96 (6), 1905–1145. doi: 10.1603/0022-0493-96.6.1905

Sainsbury, F., Thuenemann, E. C., and Lomonossoff, G. P. (2009). pEAQ: Versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* 7 (7), 682–935. doi: 10.1111/j.1467-7652.2009.00434.x

Silva, V. C., de Oliveira, L. A., Lacerda, M. S. C., Pimentel, L. A., Santos, W. S., Macêdo, J. T. S. A., et al. (2017). Experimental poisoning by cassava wastewater in sheep. *Pesquisa Vet. Brasileira = Braz. J. Vet. Res.* 37 (11), 1241–1246. doi: 10.1590/s0100-736x2017001100008

Siritunga, D., and Sayre, R. T. (2003). Generation of cyanogen-free transgenic cassava. *Planta* 217 (3), 367–735. doi: 10.1007/s00425-003-1005-8

Siritunga, D., and Sayre, R. (2004). Engineering cyanogen synthesis and turnover in cassava (*Manihot esculenta*)". *Plant Mol. Biol.* 56 (4), 661–695. doi: 10.1007/s11103-004-3415-9

Takos, A. M., Knudsen, C., Lai, D., Kannangara, R., Mikkelsen, L., Motawia, M. S., et al. (2011). Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in lotus japonicus and suggests the repeated evolution of this chemical defence pathway. *Plant J.* 68 (2), 273–286. doi: 10.1111/j.1365-313X.2011.04685.x

Taylor, N., Gaitán-Solís, E., Moll, T., Trauterman, B., Jones, T., Pranjal, A., et al. (2012). A high-throughput platform for the production and analysis of transgenic cassava (*Manihot esculenta*) plants. *Trop. Plant Biol.* 5 (1), 127–395. doi: 10.1007/s12042-012-9099-4

Tewe, O. O. (1992). "Detoxification of cassava products and effects of residual toxins on consuming animals," in Roots, tubers, plantains and bananas in animal feeding. (D. Machin and S. Nyvold, editors) *FAO Animal Production and Health Paper*, (Rome, Italy: FAO) vol. 95, 81–98.

Thomazella, D. P. de T., Seong, K., Mackelprang, R., Dahlbeck, D., Geng, Y., Gill, U. S., et al. (2021). Loss of function of a DMR6 ortholog in tomato confers broad-spectrum disease resistance. *Proc. Natl. Acad. Sci. U.S.A.* 118 (27), e2026152118. doi: 10.1073/pnas.2026152118

Tshala-Katumbay, D. D., Ngombe, N. N., Okitundu, D., David, L., Westaway, S. K., Boivin, M. J., et al. (2016). Cyanide and the human brain: Perspectives from a model of food (cassava) poisoning. *Ann. New York Acad. Sci.* 1378(1), 50–57. doi: 10.1111/nyas.13159

Vandegeer, R., Miller, R. E., Bain, M., Gleadow, R. M., and Cavagnaro, T. R. (2013). Drought adversely affects tuber development and nutritional quality of the staple crop cassava (*Manihot esculenta* Crantz). *Funct. Plant Biol. FPB* 40 (2), 195–2005. doi: 10.1071/FP12179

Veley, K. M., Okwuonu, I., Jensen, G., Yoder, M., Taylor, N. J., Meyers, B. C., et al. (2021). Gene tagging *via* CRISPR-mediated homology-directed repair in cassava. *G3* 11 (4), jkab028. doi: 10.1093/g3journal/jkab028

Whankaew, S., Poopear, S., Kanjanawattanawong, S., Tangphatsornruang, S., Boonseng, O., Lightfoot, D. A., et al. (2011). A genome scan for quantitative trait loci affecting cyanogenic potential of cassava root in an outbred population. *BMC Genomics*. 12, 266. doi: 10.1186/1471-2164-12-266

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (New York: Springer-Verlag).

Wilson, M. C., Mutka, A. M., Hummel, A. W., Berry, J., Chauhan, R. D., Vijayaraghavan, A., et al. (2017). Gene expression atlas for the food security crop cassava. *New Phytol.* 213 (4), 1632–1415. doi: 10.1111/nph.14443

Xiao, A., Cheng, Z., Kong, L., Zhu, Z., Lin, S., Gao, G., et al. (2014). CasOT: A genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics* 30 (8), 1180–1825. doi: 10.1093/bioinformatics/btt764

Xie, K., Minkenberg, B., and Yang, Y. (2015). Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc. Natl. Acad. Sci. U.S.A.* 112 (11), 3570–3755. doi: 10.1073/pnas.1420294112

Zainuddin, I. M., Fathoni, A., Sudarmonowati, E., Beeching, J. R., Gruissem, W., and Vanderschuren, H. (2018). Cassava post-harvest physiological deterioration: From triggers to symptoms. *Postharvest Biol. Technol.* 142 (August), 115–123. doi: 10.1016/j.postharvbio.2017.09.004

Zainuddin, I. M., Schlegel, K., Gruissem, W., and Vanderschuren, H. (2012). Robust transformation procedure for the production of transgenic farmer-preferred cassava landraces. *Plant Methods* 8 (1), 245. doi: 10.1186/1746-4811-8-24

Zidenga, T., Leyva-Guerrero, E., Moon, H., Siritunga, D., and Sayre, R. (2012). Extending cassava root shelf life *via* reduction of reactive oxygen species production. *Plant Physiol.* 159(4), 1396–1407. doi: 10.1104/pp.112.200345

Zidenga, T., Siritunga, D., and Sayre, R. T. (2017). Cyanogen metabolism in cassava roots: Impact on protein synthesis and root development. *Front. Plant Sci.* 8 (February), 220. doi: 10.3389/fpls.2017.00220

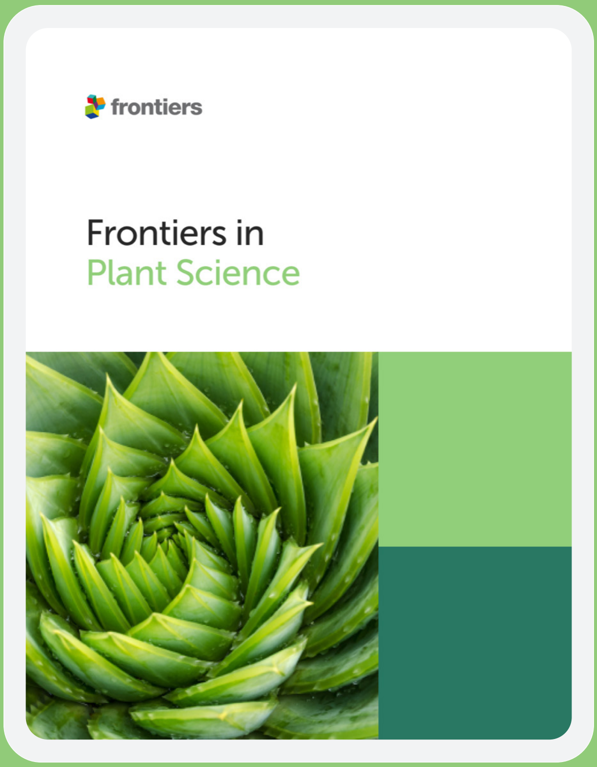
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact