SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

# Improving root characterisation for genomic prediction in cassava

Bilan Omar Yonis[1,6], Dunia Pino del Carpio[2,5,6], Marnin Wolfe[2], Jean-Luc Jannink[2,3], Peter Kulakow[4] & Ismail Rabbi[4] ✉

**Cassava is cultivated due to its drought tolerance and high carbohydrate-containing storage roots. The lack of uniformity and irregular shape of storage roots poses constraints on harvesting and post-harvest processing. Here, we phenotyped the Genetic gain and offspring (C1) populations from the International Institute of Tropical Agriculture (IITA) breeding program using image analysis of storage root photographs taken in the field. In the genome-wide association analysis (GWAS), we detected for most shape and size-related traits, QTL on chromosomes 1 and 12. In a previous study, we found the QTL on chromosome 12 to be associated with cassava mosaic disease (CMD) resistance. Because the root uniformity is important for breeding, we calculated the standard deviation (SD) of individual root measurements per clone. With SD measurements we identified new significant QTL for Perimeter, Feret and Aspect Ratio on chromosomes 6, 9 and 16. Predictive accuracies of root size and shape image-extracted traits were mostly higher than yield trait prediction accuracies. This study aimed to evaluate the feasibility of the image phenotyping protocol and assess GWAS and genomic prediction for size and shape image-extracted traits. The methodology described and the results are promising and open up the opportunity to apply high-throughput methods in cassava.**

Cassava (*Manihot esculenta* Crantz), a tropical root crop with origins in Latin America, ranks as the 3rd most important crop in the tropics after rice and maize[1]. In Africa, more than 800 million people rely on cassava as a primary source of calories[2]. Cassava is widely cultivated due to its high drought tolerance and high carbohydrate-containing storage roots, and although most of the production is for human consumption, its use extends to animal feed and industrially processed products[2–4]. In addition to the edible, high-starch storage roots, cassava plants produce thin fibrous roots, which function to absorb water and nutrients from the soil[5]. The development and differentiation of fibrous roots, as well as the mechanism that triggers root storage formation in cassava, are poorly understood.

Cassava storage roots are morphologically diverse, the lack of uniformity and irregular shape between and within genotypes poses significant constraints on harvesting and post-harvest processing. The irregularity of root shape results in considerable losses of valuable root yield[3]. The waste of tuber flesh and the inefficiency of hand peeling could be avoided by peeling mechanization. However, breeding for root characteristics that facilitate that process requires a thorough understanding of the genetic basis of cassava root morphology. Several studies have attempted to characterize cassava root shape to support the development of peeling mechanization[6,7]. The root characteristics that were evaluated in those studies include root diameter, weight, length and peel thickness.

Routine assessment of storage root size and shape in breeding plots relies on visual scores (www.cassavabase.org/search/traits). The categorical scores for root size are 3, 5 and 7 for small, medium and large roots, respectively. A single categorical score is given to a harvested plot based on the most frequent size in that plot. The visual rating of shape is 1 (conical), 2 (conical-cylindrical), 3 (cylindrical), 4 (fusiform), 5 (Irregular), and 6 (Combination of shapes). Similar to root size, the shape scoring is based on the most common observation in a plot. These categorical scores suffer from person to person subjectivity and inability to describe the variation in size and shape within a plot. Thus, image analysis of roots offers a more objective means of obtaining unbiased quantitative data on important root traits.

The development of image software tools and analyses pipelines have gained increased relevance owing to advances in high-throughput phenotyping[8–10]. In Maize, imaging under controlled illumination followed by

[1]Montpellier SupAgro, 34060, Montpellier, Cedex, 02, France. [2]Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, 14850, USA. [3]US Department of Agriculture - Agricultural Research Service (USDA-ARS), Ithaca, NY, USA. [4]International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. [5]Department of Jobs, Precincts and Regions, AgriBio, Centre for AgriBioscience, Bundoora, Australia. [6]These authors contributed equally: Bilan Omar Yonis and Dunia Pino del Carpio. ✉e-mail: I.Rabbi@cgiar.org

| Trial | Design* | Location | Plots | Unique entries | Plot size |
|---|---|---|---|---|---|
| GG.C0.UBJ | CET, augmented | Ubiaja | 805 | 738 | 10 plants, single row |
| GS.C1.EC.IBA | CET, augmented | Ibadan | 293 | 265 | 20 plants (4 ×5) |
| GS.C1.EC.IKN | CET, augmented | Ikenne | 331 | 307 | 20 plants (4×5) |
| GS.C1.EC.MOK | CET, augmented | Mokwa | 329 | 278 | 20 plants (4×5) |
| Crossing block.C0_C1.UBJ | CET augmented | Ubiaja | 243 | 218 | |

**Table 1.** Summary of trials used in the present study including the trial names, design, locations, number of plots and number of unique clones in each trial. *CET = clonal evaluation trial.

automatic image-analysis has been successfully used to study root system architecture traits[11]. In cereals, grain shape is an important target for genetic improvement, because it is usually related to quality, consumer appeal or the intended end usage[12]. For rice grain shape description, SHAPE, a program based on Elliptical Fourier Descriptor (EFDs) has been used to derive shape-related phenotypes for genome-wide association and genomic prediction[13,14].

Genomic selection (GS) is a method first introduced in animal breeding to select candidates for crossing in the breeding program using only genomic information. GS is particularly relevant for the improvement of polygenic traits[15] because its implementation can lead to a reduction in cost and time compared to traditional plant breeding programs[16]. Because cassava is an outcrossing species mostly propagated by stem cuttings, conventional breeding methods can take more than five years to produce superior performing clones (www.nextgencassava.org). Genome-wide association studies (GWAS) are complementary to GS as they have proven effective for the identification of QTL regions associated with several traits that are critical for cassava breeding, including cassava mosaic disease resistance (CMD)[17], cassava brown streak disease resistance (CBSD)[18], and beta-carotene content and dry matter content[19].

To date, genomic prediction accuracies for root shape and size characteristics have not been reported in cassava. In this study, size and shape related traits describing cassava roots were obtained through automated image analysis. We first estimated their heritability and conducted a genome-wide association study to explore the genetic architecture of cassava roots shape characteristics; then we compared the genomic prediction accuracy of image size and shape traits to those of root yield. Our research contributes to a better understanding of the genetic architecture of cassava root size and shape traits and explores the possibility of in-field high-throughput phenotyping that would allow breeders to use GS to select varieties for quantitative root characteristics.

## Materials and methods

**Germplasm.** We processed and analyzed cassava roots images taken from several field trials conducted by the International Institute for tropical agriculture (IITA) as part of their genomic selection breeding program. The cassava germplasm collections that we analyzed are known as Genetic Gain (GG) and the progeny of the first genomic selection event (C1), which are thus progeny of a subset of the GG population. The GG constitutes a large collection of important landraces, breeding lines and released improved varieties of cassava developed by IITA over the last four decades. More detail about the origins and constituency of these populations is available in several published studies (Wolfe *et al*., 2016;Wolfe *et al*., 2017).

A summary of the trials used in the present study is presented in Table 1. The first set of trial was the GG trial which comprised 805 plots planted in the summer of 2014 in Ubiaja, Nigeria using an augmented design with two checks planted in each incomplete block. The trial comprised 758 unique clones. Each plot consisted of 10 stands in a single row with spacing of 1 m between rows and 0.8 m within rows. The second set of trials consisted of 86 C0 clones selected from the GG population and 158 C1 clones which were planted in Ubiaja in an augmented design, each plot consisted of 20 plants. It is important to note here that these clones were used as parents for the GS cycle 1 population. The third set of trials involved GS cycle 1 clones that were split into three sets and planted separately in three locations: Ibadan, Ikenne and Mokwa. Each set was planted as a clonal evaluation trial (CET) using an incomplete block design with common checks in each block. All trials had at least 10 clones in common. Plants were harvested after 12 months in all trials.

**Image acquisition.** The roots from four plants per plot were spread across a green board (160 cm by 120 cm). It was important that the roots were not touching each other and also not touching the board edges to get an individual root value (Supplementary Fig. 1). Five circles, each 7.5 cm in diameter were painted on the left and right sides of the board. Those circles were used as a reference to transform the final result from the pixel unit to cm. Labels were placed on the board for each image allowing images to be identified and renamed for further processing. Images are freely available at ftp://ftp.cassavabase.org/manuscripts/Yonis_et_al_2019/.

**Image processing and phenotype acquisition.** First, the images were coded to assign each photo to the plot from which the roots were taken. In some cases, several images were required per plotto capture all roots from all the plants. For the GG collection, after quality control we obtained 805 images of cassava roots for 738 clones of which 665 had genotypic information. For the C1 population, we had images originating from four

locations and a total of 1091 root images for 997 clones. All the image processing was performed with ImageJ Java version 1.8.0_11 (64-bit). The images were copied in two folders, one for processing and measuring the roots and the second for scaling the measurements. Thus, each image was processed and analysed twice.

**Image processing.** The first step of the image processing was to convert the RGB colour images into HSB stacks (hue, saturation and brightness images). We obtained three slices, but we only kept the first slice (the hue image). We then set a threshold from 0 to 255 for the roots and from 125 to 255 for the reference scaling circles before proceeding to run the "threshold" followed by the "make binary" commands. This threshold was determined by doing individual tests on some images. At the end of the processing, each image was binary, with our objects of interest (roots and scales) represented as white pixels and everything else as black. Most steps in the procedure were automated using customized ImageJ macros.

**Phenotypes acquisition and description.** The "analyze particles" command in ImageJ counts each contiguous area of white pixels within a binary image and gives some additional basic measurements. With the aim to get shape related traits, we used the "extended particle analyzer" function in the BioVoxxel Toolbox plugin (http://imagej.net/BioVoxxel_Toolbox#Extended_Particle_Analyzer). This function computes useful parameters of which we chose to keep seven for downstream analysis: Area, Perimeter, Feret, Circularity, Solidity, Roundness, and the Aspect Ratio (AR). The area and the perimeter describe the size of a root. The Feret, is the longest distance between any two points along the selection boundary, also known as maximum caliper. Circularity, Solidity, Roundness and aspect ratio (AR) describe shape.

The shape descriptors are ratio values that ranged from 0 to 1 except AR, which is not bounded. In addition, the shape descriptors do not have a unit, while area, perimeter, and feret are parameters expressed in pixels. The mean area value of the circles was used as a reference to convert pixels to centimetres (scaling coefficient). Since the exact diameter in centimetres of each circle was known, we used this value to calculate the mean number of pixels per $cm^2$ for each image.

$$Scaling\ coefficient = \sqrt{(Area\ (pixel^2)/Area(cm^2))}$$

**Genomic analyses.** We performed a two-step approach for the genomic analysis. In the first step, we used a linear mixed model to account for the variability in the field design and calculate the broad-sense heritability. The input data was: 1) the mean phenotype value for each plot (average phenotype of all imaged roots), 2) the same as (1) but adjusted to account for the potential effects of variation in cassava mosaic disease (CMD) severity among plots and 3) the standard deviation of the root shape and size measurements (across all imaged roots) per plot, also adjusted to remove the effect of CMD. We fit two different models, with CMD correction and without CMD correction, for each of the two focal populations (GG or C1).

For GG, the following models were fitted:

$$\boldsymbol{y} = \boldsymbol{X}m + \mathbf{Z}_{\mathbf{clone}}c + \mathbf{Z}_{\mathbf{range}}r + \varepsilon \quad (1)$$

$$\boldsymbol{y} = \boldsymbol{X}n + \mathbf{Z}_{\mathbf{clone}}c + \mathbf{Z}_{\mathbf{range}}r + \varepsilon \quad (2)$$

In both models $\boldsymbol{y}$ is a vector of phenotypes, $\mathbf{Z}_{\mathbf{clone}}$ and $\mathbf{Z}_{\mathbf{range}}$ are respectively the incidence matrices of the clones and range both fit as random with their effects vector $c \sim N(0, \mathbf{I}\sigma_c^2)$ for clones and $r \sim N(0, \mathbf{I}\sigma_r^2)$ for range. Ranges were equivalent to the row or column along which plots were arrayed. $\mathbf{X}$ is the incidence matrix for the fixed effects. In model 1, the number of harvested plants per plot (NOHAV) and CMD were accounted for as fixed and the vector $m$ contains the effect estimates. In model 2, we did not correct for CMD, $\mathbf{X}$ and $n$ therefore only reference NOHAV.

For C1, model (3) and (4) were fitted:

$$\boldsymbol{y} = \boldsymbol{X}m + \mathbf{Z}_{\mathbf{clone}}c + \mathbf{Z}_{\mathbf{loc:range}}r + \varepsilon \quad (3)$$

$$\boldsymbol{y} = \boldsymbol{X}n + \mathbf{Z}_{\mathbf{clone}}c + \mathbf{Z}_{\mathbf{loc:range}}r + \varepsilon \quad (4)$$

In model (3) and (4), we replace the range variable with the combination of the location and the range (Loc:range, i.e. range is nested in location) as the C1 population was planted in several locations unlike the GG. In all models, for all traits, $\boldsymbol{y}$ correspond to the log transformation of the original phenotypic values. Additional explanation of the models fitted for these populations can be found in Wolfe *et al.* (2017).

From these models, we extracted the clone-effect BLUP, which estimates the total genetic value (EGV) of each line and de-regressed the EGV by dividing them by their reliability to obtain the de-regressed BLUP[21]. Broad-sense heritability values were calculated using the variance components estimated using the mixed-models described above. EGV and de-regressed EGV are used in downstream analyses as described below.

In addition, the correlation across traits was estimated using using the de-regressed BLUP values obtained after fitting the aforementioned linear mixed model.

With the combined GG+C1 population, we calculated the phenotypic and genotypic coefficients of variation using the output values of the lmer model (3).

Genotypic and phenotypic coefficients of variation were calculated as:

$$GCV = \frac{\sqrt{\sigma_g^2}}{\mu} * 100$$

$$PCV = \frac{\sqrt{\sigma_p^2}}{\mu} * 100$$

Where $\mu$ is the grand mean value of the trait.

**Genotyping data.** Both populations (GG and C1) were genotyped using the genotyping-by-sequencing (GBS) method[22] and the SNP calling was performed with TASSEL 5.0 GBS pipeline v2[23]. Alignment of GBS reads was to the cassava reference genome v6.1 (http://phytozome.jgi.doe.gov; ICGMC, 2015). The condition for the genotype calls was the presence of a minimum of four reads. Extracted SNPs were filtered to remove clones with>80% missing and markers with >60% missing genotype calls. Markers were also removed when they had an extreme deviation from Hardy-Weinberg equilibrium ($\chi^2 > 20$).

A combination of custom scripts and common variant call file (VCF); manipulation tools were used to accomplish the above pipeline. The missing data were imputed using Beagle v4.0[24]. For the GG and C1 populations, we had 112,082 and 179,041 markers, respectively, with MAF > 0.01.

**Genomic prediction.** We estimated genomic prediction accuracy using 5-fold cross-validation repeated 25 times similar to what is described in Wolfe *et al.* (2017). Briefly, for each replicate of the process, the population was split into five approximately equal chunks (folds). Five genomic predictions were then made in which each fold (fifth of the population) served as the test set (no phenotypes) and were predicted by the remaining four-fifths (training set, with phenotypes). Prediction accuracy for each fold was defined as the correlation of the genome-estimated breeding values (GEBVs, which are BLUPs from the test-sets of each fold), with the de-regressed EGVs from the pre-adjustment stage of the analysis.

For genomic prediction, we used a mixed-model with a genotype (clone) random effect with covariance proportional to the genomic relationship matrix, also called GBLUP. The genomic relationship matrix was constructed using the function *A.mat* in the R package rrBLUP[25,26] De-regressed BLUPs were used as the response variable and the GBLUP models were fit with the function *emmreml* in the R package EMMREML[27].

**GWAS analyses.** Genome-wide association mapping (GWAS) analyses were performed using a linear mixed-model analysis (MLMA) implemented in GCTA (Version 1.90.0beta)[28]. Specifically, we followed a leave-one-chromosome-out approach and tested all markers with MAF > 0.05. The leave-one-chromosome-out approach involves excluding all markers on the chromosome of the current candidate SNP from the genomic relationship matrix (GRM) used to control population structure when estimating their marker effects. Manhattan plots were generated using the R package *qqman*[29]. SNP markers with a − log10(P-value) which exceeded the Bonferroni threshold >6.8 were considered to be statistically significant and were further annotated into coding regions (genes) of the cassava genome.

Candidate gene identification was performed using the significant GWAS results of the standard deviation + CMD correction GWAS results. Using the phytozome 12 portal link to biomart (https://phytozome.jgi.doe.gov/biomart/) we searched for genes located 10 kb around the top SNP hits.

**Multivariate GWAS analysis.** We used a multivariate linear mixed model as implemented in GEMMA (mvLMM)[30]. We tested marker associations with multiple phenotypes that are fitted jointly in the mvLMM while controlling for population stratification. Different combinations of phenotypes were fitted in six models, the phenotypes that were fitted together were selected based on their phenotypic correlation. Model 1: Circularity, Round, Solidity; Model 2: Area, Feret, Circularity, Solidity, AR; Model 3: Area, Perimeter, Round, Solidity, AR; Model 4: Area, Perimeter, Feret, Circularity, Round, Solidity, AR; Model 5: Area, Perimeter, Feret; Model 6: Circularity, Round, Solidity, AR.

P-values were corrected for multiple testing by computing a Bonferroni threshold similar to the univariate GWAS and Benjamini-Hochberg q-values with a q < 0.1 threshold.

## Results

**Phenotypes distribution.** Using the plugin BioVoxxel in ImageJ, we extracted quantitative measurements of the Area, Perimeter, Feret, Circularity, Solidity, Roundness, and the aspect ratio (AR) from root images collected in the field. The raw value datasets show similar ranges for root shape and size descriptors in GG and C1 populations (Supplementary Table 1). The individual root measurements with the maximum and minimum value of each trait in both populations are presented in Fig. 1.

The frequency distribution of the mean value per plot of the GG and C1 populations is presented in Supplementary Fig. 2 and the mean values per trait within a population are presented in Supplementary Table 1. Some genotypes exhibited large differences in their mean values for Area, Perimeter and Feret. For example, the maximum mean root Area in GG population was 339 $cm^2$ while the mean Area of the GG population was 121.5 $cm^2$. Similarly, those genotypes exhibited a maximum mean root Perimeter of 135 cm and a maximum mean Feret of 50 cm while the mean Perimeter and Feret value in the GG population were 66 cm and 26 cm, respectively.

In the C1 dataset, the maximum mean value for the root area in the C1 dataset was 372 $cm^{2,}$ while the mean area of that population was 128 $cm^2$. The maximum values for Perimeter and Feret were 132 and 49 cm respectively, while the C1 population mean value for the two traits was 68 cm and 28 cm respectively.

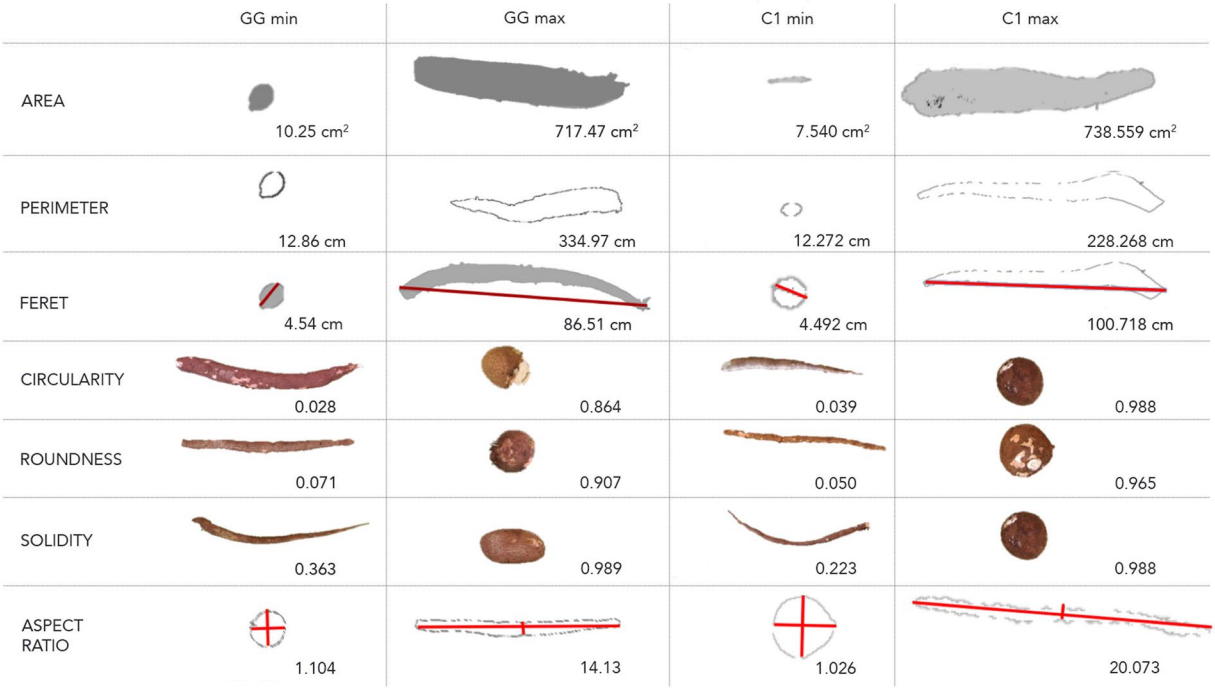


**Figure 1.** Phenotype description obtained using the extended particle analyzer plugin in ImageJ: Individual root measurements with the maximum and the minimum value of each trait in Genetic gain (GG) population and Cycle1 (C1) population, represented to highlight the range of values for each trait.

**Correlation plots.** Phenotypic correlations were calculated pairwise using de-regressed BLUPs of the mean values for each population separately (Fig. 2). In the GG dataset, the highest correlation within yield traits corresponded to root number and root weight ($r^2$ =0.79). Similarly, root number and root weight were highly correlated in C1 population ($r^2$=0.88). In both datasets, correlations between yield traits were significant and high ($r^2 > 0.5$) and these traits were also positively correlated with Area, Perimeter and Feret. However, a low correlation ($r^2 < 0.1$) was observed between yield traits and root shape descriptors such as Circularity, Roundness, Solidity and AR in both populations.

Size-related traits derived from root images (Area, Perimeter and Feret) showed the highest positive correlation (r > 0.7) with each other. In both datasets, the highest correlation between size-related traits corresponded to Perimeter and Feret (r=0.97). Additionally, Feret and Perimeter were negatively correlated with shape-related traits (Circularity, Roundness and Solidity) and positively correlated with AR. In the GG dataset, Area showed a negative correlation with Circularity (r= −0.26), Roundness (r= −0.21), Solidity (r= −0.19), and a positive correlation with AR (r= 0.19). While in the C1 population, a low correlation was observed between Area and shape descriptors.

Within the shape related traits, the highest correlation was found between Circularity and Roundness (GG r = 0.89, C1 r = 0.86) and Solidity (GG r = 0.87, C1 r = 0.84). AR showed a negative correlation with Circularity, Solidity and Roundness in both datasets.

**Phenotypic variability.** The analysis of variance per location is presented in Supplementary Table 2. The level of variation in variance components and coefficients across locations indicate the existence of variability among genotypes.

Using the combined GG+C1 populations, the extent of trait variability was assessed in terms of broad-sense heritability values (H2) phenotypic coefficient of variation (PCV) and genotypic coefficient of variation (GCV) (Table 2). Phenotypic coefficient of variation (PCV %) was found to be higher than the genotypic coefficient of variation (GCV %). Moderate genotypic variance of >10% was observed for most traits, except for Perimeter and solidity. The highest GCV was observed for the yield traits root weight (29.16%) and root number (26.36%). Moderate heritability values of >0.20 were found for most traits except for shoot weight (0.12). The top three heritability values corresponded to image extracted traits: AR (0.48), roundness (0.52) and circularity (0.46).

In addition, broad-sense heritability values ($H^2$) for root shape and yield-related traits were calculated for each population (Table 3). In the GG population, without adjusting the phenotypes for their CMD score, $H^2$ of root shape related traits ranged from 0.17 (Perimeter and Circularity) to 0.46 (aspect ratio) and for yield traits, $H^2$ ranged from 0.29 root weight (RTWT) to 0.44 shoot weight (SHTWT). In the GG dataset, Perimeter, Circularity and Solidity exhibited the lowest heritability values at 0.17, 0.17 and 0.12, respectively.

In the C1 population, the heritability of shape-related traits ranged from 0.36 (Perimeter) to 0.54 (Circularity) while for yield traits $H^2$ ranged from 0.36 (SHTWT) to 0.61 (RTWT). The heritability of most traits was higher in the C1 population than GG except for Area (0.39 to 0.38) and SHTWT (0.44 to 0.36). The inclusion of the CMD

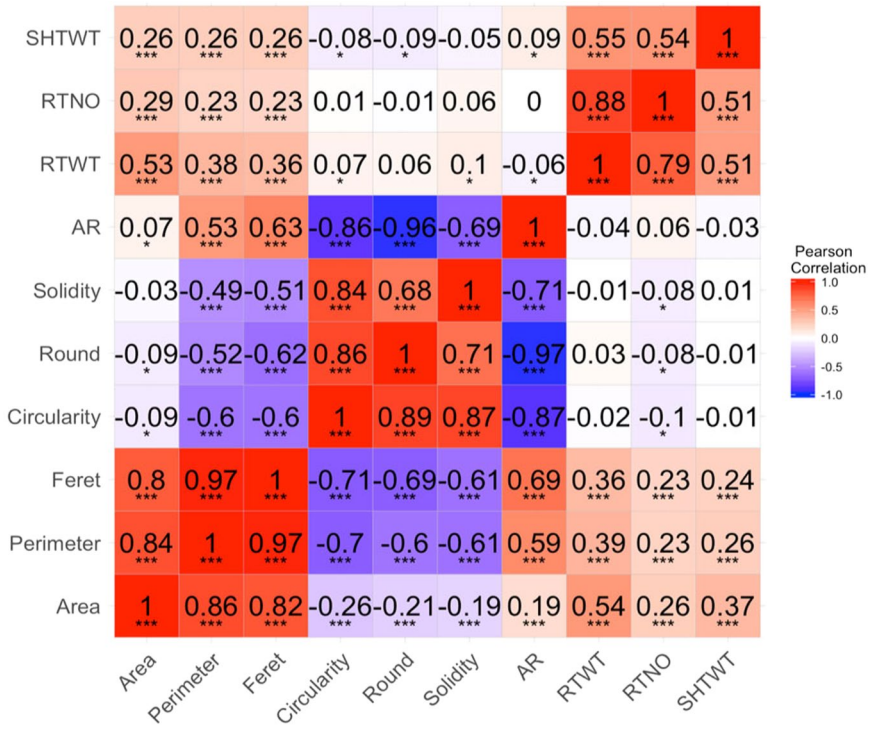


**Figure 2.** Heatmap with Pearson correlation coefficient: Trait correlation using the de-regressed BLUP value of GG dataset (lower triangle) and C1 dataset (upper triangle). The stars depict the significance according the p-value (***$P < 0.0001$, **$P < 0.001$, *$P < 0.05$).

| Trait | Range | Mean | VAR_G | VAR_P | GCV | PCV | H2 |
|---|---|---|---|---|---|---|---|
| AR | 2.05–8.99 | 4.78 | 0.433 | 0.929 | 13.78 | 19.94 | 0.48 |
| Area | 26.55–372.08 | 125.71 | 293.464 | 1497.949 | 13.63 | 30.02 | 0.21 |
| Perimeter | 31.11–135.42 | 67.66 | 39.161 | 180.040 | 9.25 | 19.44 | 0.23 |
| Circularity | 0.13–0.63 | 0.36 | 0.002 | 0.005 | 12.61 | 18.66 | 0.46 |
| Feret | 12.17–50.57 | 27.58 | 8.579 | 29.857 | 10.62 | 19.59 | 0.29 |
| Round | 0.11–0.53 | 0.25 | 0.002 | 0.003 | 16.03 | 22.27 | 0.52 |
| Solidity | 0.49–0.96 | 0.84 | 0.001 | 0.002 | 3.02 | 5.65 | 0.29 |
| RTWT | 0.5–121.5 | 28.92 | 71.103 | 226.537 | 29.16 | 50.68 | 0.33 |
| RTNO | 1–207 | 52.94 | 194.684 | 555.515 | 26.36 | 42.81 | 0.38 |
| SHTWT | 0.5–135 | 30.9 | 18.838 | 163.805 | 14.05 | 40.5 | 0.12 |

**Table 2.** Broad-sense heritability values and variance components of root size,shape and yield traits for the combined Genetic gain and cycle 1 breeding populations. (RTWT: root weight; RTNO: root number; SHTWT: Shoot weight, H2 broad sense heritability, genetic variance (VAR_G), phenotypic variance (VAR_P), genotypic coefficient of variation (GCV), phenotypic coefficient of variation (PCV)).

in the calculation of the variance components always reduced the heritability of all the traits in both populations by around 10%.

**Genome-wide association study of root traits.** Using a univariate genome-wide association approach for root image traits (root size and shape) and root yield traits we identified significant loci for all traits except for area (Fig. 3). We detected a total of 91 SNP markers exceeding the significance threshold ($-\log10\ P \geq 6.28$). The Manhattan plots of the univariate GWAS results for yield traits are shown in Supplementary Fig. 3 and detailed information on the significant markers is summarized in Supplementary Table 3.

We detected markers associated with Perimeter and Feret on chromosome 12, and with Solidity on chromosome 1, whereas for AR we identified significant loci on chromosome 1 and chromosome 12. Similarly, for Circularity and Roundness, we detected significant loci on chromosome 1 and chromosome 12.

For most shape-related traits several other regions on chromosomes 3, 4, 8, 9, 14, 15 and 18 did not reach the significance threshold but showed a $-\log10\ P \geq 5$ (Fig. 3). For root yield traits we detected a QTL on chromosome 12 associated to root number (RTNO) and RTWT (Supplementary Fig. 3, Supplementary Table 3). Notably, using the CMD adjusted phenotype removed the significance of the QTL on chromosome 12 but did not identify

| Trait | GG | | C1 | |
|---|---|---|---|---|
| | no correction | +MCMDS | no correction | +MCMDS |
| Area | 0.39 | 0.33 | 0.38 | 0.36 |
| Perimeter | 0.17 | 0.12 | 0.36 | 0.33 |
| Feret | 0.33 | 0.17 | 0.40 | 0.35 |
| Circularity | 0.17 | 0.15 | 0.54 | 0.53 |
| Round | 0.39 | 0.26 | 0.52 | 0.49 |
| Solidity | 0.12 | 0.12 | 0.48 | 0.48 |
| Aspect Ratio | 0.46 | 0.31 | 0.56 | 0.54 |
| RTWT | 0.29 | 0.23 | 0.60 | 0.50 |
| RTNO | 0.39 | 0.37 | 0.61 | 0.54 |
| SHTWT | 0.44 | 0.39 | 0.36 | 0.33 |

**Table 3.** Broad-sense heritabilities of root shape and root yield traits for the Genetic gain and cycle 1 breeding populations. (RTWT: root weight; RTNO: root number; SHTWT: Shoot weight).

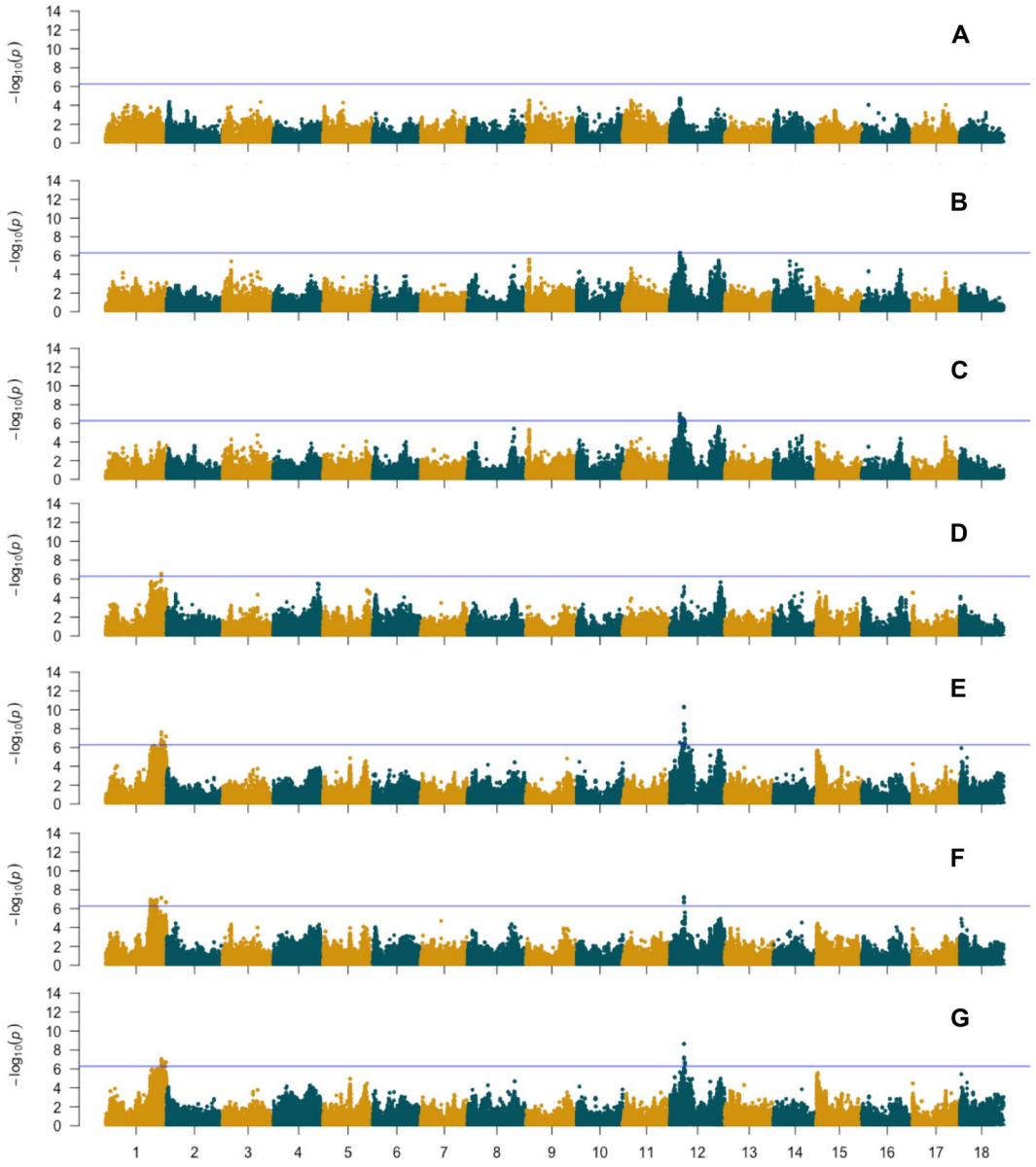


**Figure 3.** Genome-wide association results of size and shape-related traits using de-regressed BLUPs of mean values (not corrected for CMD). (**A**) Area. (**B**) Perimeter; (**C**). Feret; (**D**). Solidity; (**E**). Aspect ratio; (**F**). Circularity; (**G**). Roundness. Blue horizontal line indicates the Bonferroni statistical threshold.

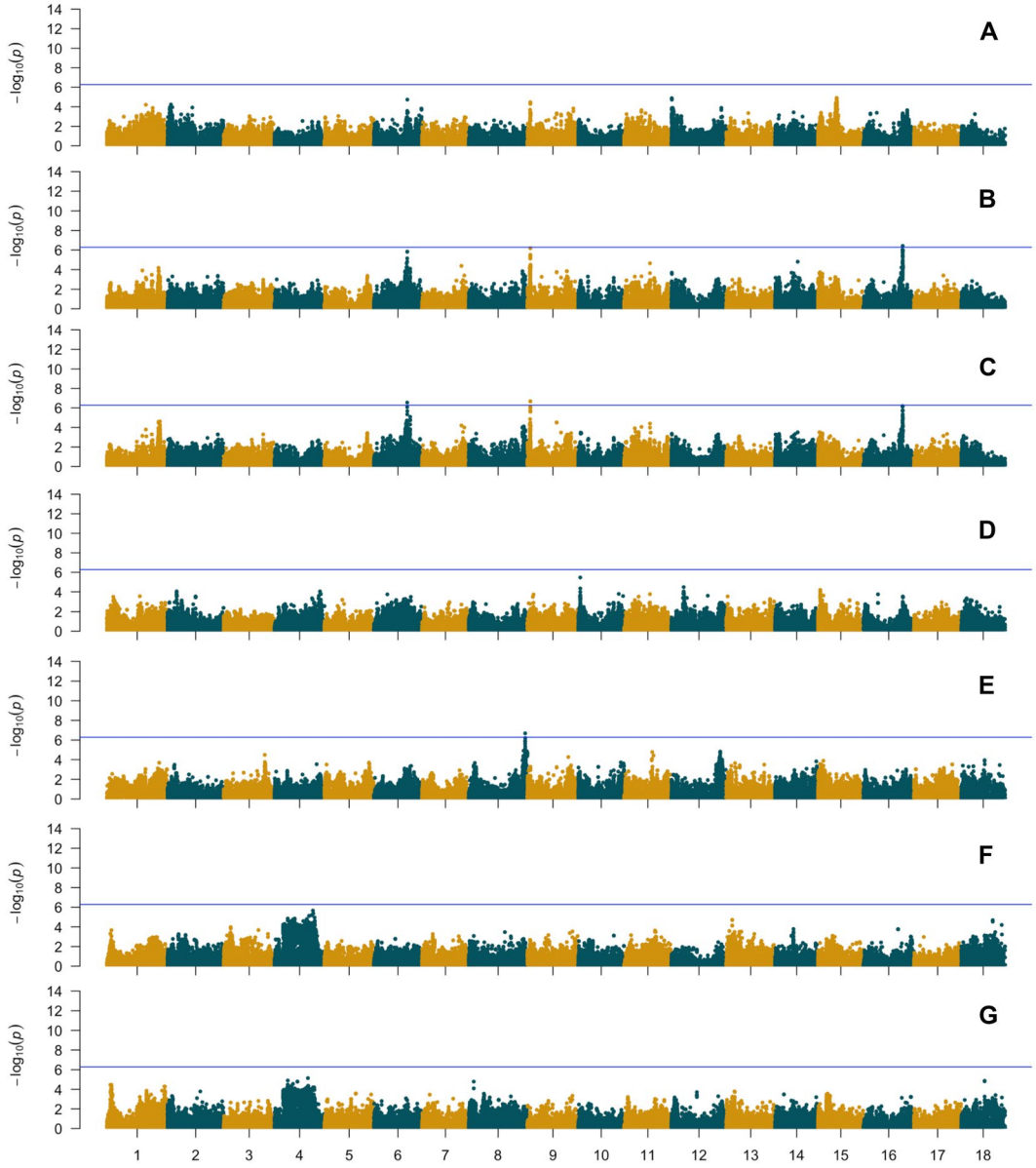


**Figure 4.** Genome-wide association results of standard deviation-derived size and shape-related traits using de-regressed BLUPs of mean CMD-corrected values. (**A**). Area; (**B**). Perimeter; (**C**). Feret; (**D**). Solidity; (**E**). Aspect ratio; (**F**). Circularity; (**G**). Roundness. Blue horizontal line indicates the Bonferroni statistical threshold.

new QTL for the image traits shape phenotypes (Supplementary Fig. 4). However, after the CMD adjustment we detected new loci associated with root number and shoot weight (Supplementary Table 4).

Significant SNP markers ($-\log10\ P \geq 6.28$) were detected for the standard deviation-derived traits of Perimeter (per-sd), Feret (feret-sd) and Aspect Ratio (AR-sd) (Fig. 4). For per-sd, a significant QTL was detected on chromosome 16, though it was not observed in the GWAS model with the mean values nor in the GWAS model with mean values with CMD adjusted phenotypes. For feret-sd, two significant QTL were identified, one on chromosome 9 and one on chromosome 6 and for AR-sd one significant QTL was found on chromosome 8 (Supplementary Table 5).

Different markers were significant in the multivariate GWAS model dependent on which phenotypes were included in the multivariate linear mixed model (mvLMM). Although the multivariate model can increase the power for detecting pleiotropic variants when using correlated traits, we identified few significant markers above the Bonferroni threshold (Supplementary Figs. 5–10). Nonetheless, when P-values were corrected for multiple testing by computing Benjamini-Hochberg q-values, four SNPs were identified as significant in the multivariate analysis. In the multivariate analysis using Area, Perimeter, Feret, Circularity, Round, Solidity and AR in the mvLMM (Model 4) we identified a significant marker at the same location on chromosome 4 (Supplementary Fig. 8). Similarly, using model 6 (Circularity, Round, Solidity, Aspect ratio) we identified one significant marker located on chromosome 4 (Supplementary Fig. 10). When Area, Perimeter and Feret were included in the

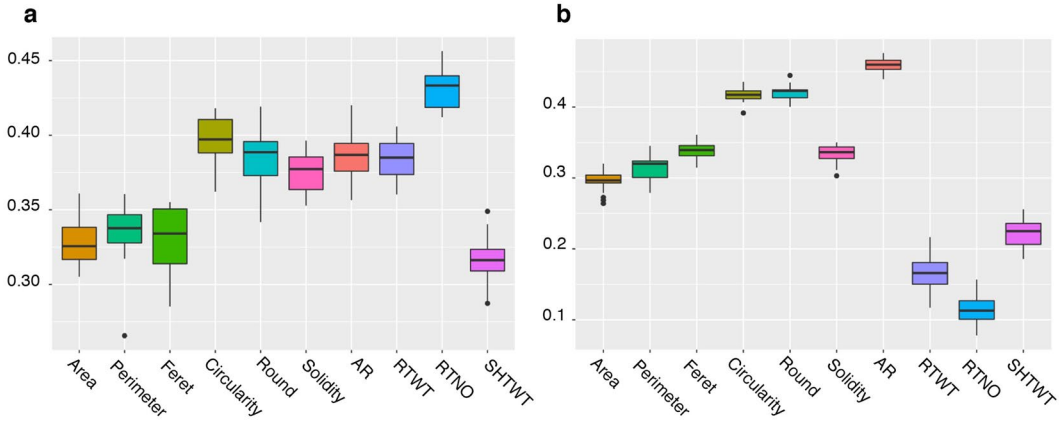


**Figure 5.** Prediction accuracy of root size and shape and yield traits. Predictive accuracies were obtained with 5 fold-cross-validation analysis using a mean plot value in the GBLUP model. (**a**) GG dataset and (**b**) C1 dataset.

mvLMM (model 5) we identified significant markers on chromosomes 6 and 9 using a q-value threshold of <0.1 (Supplementary Fig. 9).

**Genomic prediction.** Using the parental (GG) and offspring generation (C1) datasets independently, we calculated the prediction accuracies of size and shape image traits and compared those to root yield traits accuracies using de-regressed BLUPs of 1) the mean phenotype value (average phenotype of 4 plants) (Fig. 5, Supplementary Table 6), 2) the mean root size and shape phenotypes adjusted to account for the potential effect of cassava mosaic disease (CMD) on these traits (Fig. 6, Supplementary Table 6) and 3) the standard deviation of the root shape and size measurements adjusted to remove the effect of CMD (Fig. 6, Supplementary Table 6).

Prediction accuracy, calculated as the correlation between the genomic estimated breeding values (GEBVs) and the de-regressed BLUPs of the mean phenotype value, ranged from 0.32 (SHTWT) to 0.43 (RTNO) in the GG population and from 0.12 (RTNO) to 0.46 (AR) in the C1. For yield traits, accuracies in GG were higher than in C1 but were not different between populations for the shape and size related traits. In the GG population, the shape descriptors Circularity (mean = 0.40), Roundness (mean=0.39), Solidity (mean =0.37) and AR (mean =0.38) showed slightly higher accuracies than the size descriptors Area (mean =0.33), Perimeter (mean = 0.34) and Feret (mean = 0.33). In the C1 population, size and shape image traits exhibited a higher prediction accuracy than root yield traits. Among the size descriptors, Feret showed the highest accuracy (mean=0.34) and Area the lowest (mean=0.29). Among shape descriptors, AR showed the highest predictive value (mean=0.46) and Solidity the lowest (mean=0.33) (Supplementary Table 6). When the mean root size and shape phenotypes were adjusted to account for the effect of CMD, we observed a minimal decrease in predictive accuracy (Supplementary Table 6). A lower predictive accuracy was obtained for standard deviation of size and shape traits adjusted for CMD, in both populations. In the GG population, the decrease was pronounced with a maximum reduction of up to 55% for root perimeter (0.27 mean to 0.12 CMD adjusted) while in the C1 population the largest reduction was of 73% for circularity (0.41 mean to 0.11 CMD adjusted) (Supplementary Table 6).

## Discussion

Root number and root weight are among the most important targets for improvement in cassava breeding programs. Although cassava root characterisation has been the subject of several studies[31–33], the genetic architecture underlying cassava root shape remains unexplored. This study aimed to evaluate the feasibility of the image phenotyping protocol and to assess the use of genome-wide analyses for size and shape image-extracted traits.

Here, we phenotyped the GG and C1 populations from the International Institute of Tropical Agriculture (IITA) breeding program for root shape and size-related traits using image analysis of storage root photographs taken in the field. In both populations, the storage roots exhibited a wide range of shape variation. Roots with a large area were generally heavier and the circularity of storage roots was mostly inversely correlated to its area. Our results, suggest that rounded-shaped roots in cassava are generally smaller and hence lighter in weight. More importantly, the lowest correlation values found between shape traits (Circularity, Solidity, Roundness and aspect ratio (AR)) and yield suggest that there is no negative impact of differences in root storage shape on yield or that it is relatively minor.

In radish, rice and wheat, imaging-based studies of root shape and size traits have demonstrated first, that these have different genetic architectures[34] and second, that shape phenotyping can aid the identification of pleiotropic QTL. In our study, using univariate genome-wide association analysis, we detected for most shape and size related traits, significant QTL regions located on chromosomes 1 and 12. The QTL region on chromosome 1 has been previously shown to be segregating for an introgressed segment from *M. glaziovii*[35]. Furthermore, the QTL region on chromosome 1 has been associated, in the IITA genetic gain population, with other root traits such as dry matter and total carotenoid content[19].

For root weight and root number, we identified a significant QTL associated with those traits on chromosome 12. The QTL region on chromosome 12 has been previously associated, using IITA breeding populations, to cassava mosaic disease (CMD) resistance[17]. The effect of cassava mosaic disease (CMD) on root yield has been

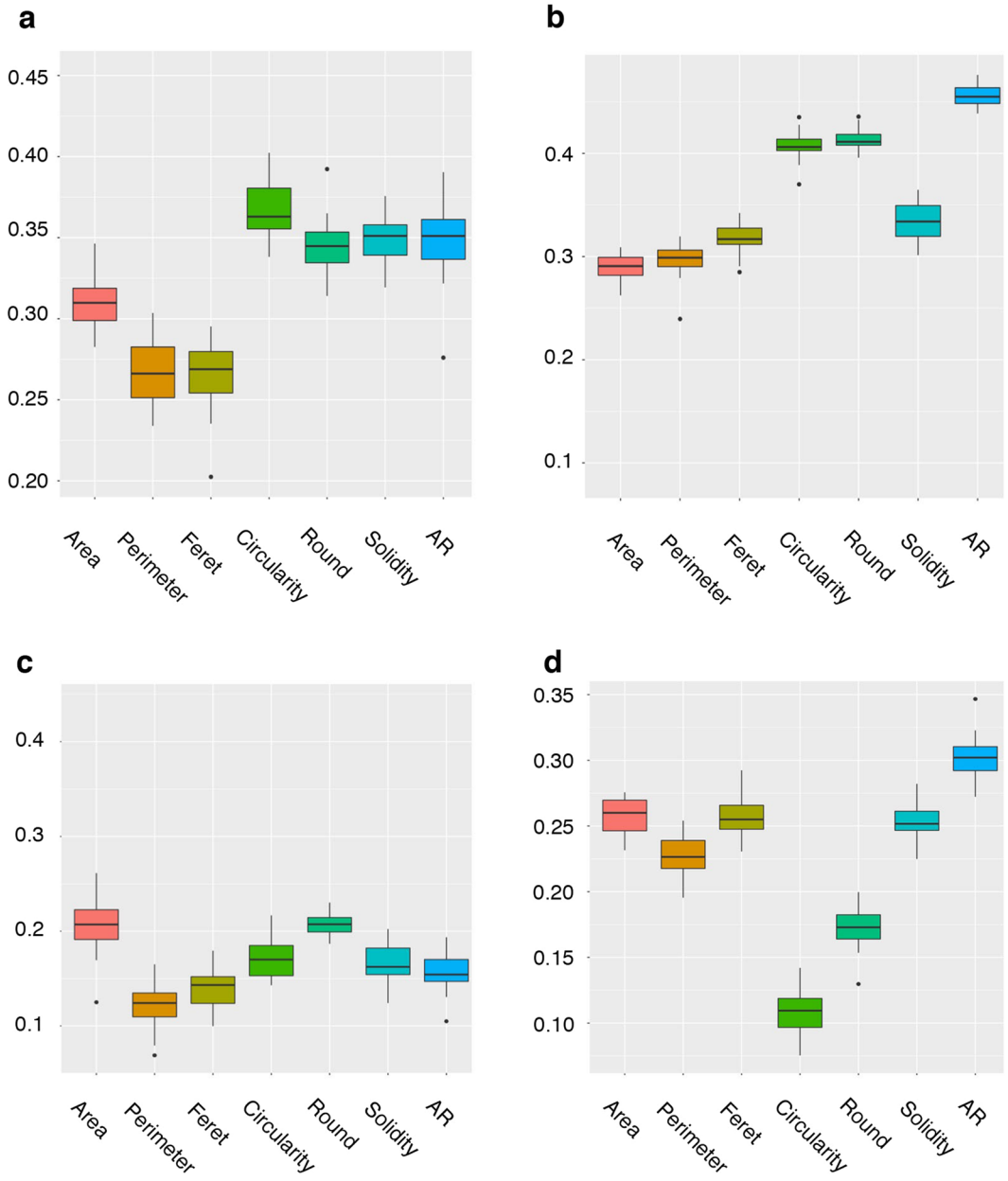


**Figure 6.** GBLUP model predictive accuracy of root size and shape traits. (**a**) GG population CMD adjusted phenotypes, (**b**) C1 population CMD adjusted phenotypes, (**c**) GG population standard deviation + CMD correction, (**d**) C1 population standard deviation + CMD correction.

previously investigated in fully and partly infected stands of cassava[36–38]. Nonetheless, the identification of new QTL for root number and shoot weight, when these traits were adjusted according to the CMD score, support the notion that CMD can have an effect on root yield traits.

Because the uniformity in size and shape of cassava roots is an important breeding goal we calculated the standard deviation of individual root measurements per clone. The use of standard deviation measurements allowed the identification of new significant QTL for Perimeter, Feret and Aspect Ratio on chromosomes 6, 9 and 16. For the new QTL regions located on chromosomes 9 and 16 we identified candidate genes related to the tocopherol and carotenoids pathways which are known regulators of plant development[39] (Supplementary Table 7). On chromosome 6, the most promising candidate is Manes.06G078700 a root meristem growth factor 1 related gene.

Together our GWAS results, 1) suggest that root-related traits are under genetic control by at least a few large effect loci, 2) reveal the impact of disease resistance loci on yield and 3) that introgressed genome regions contain QTL related to root yield, size and shape traits.

To increase the power of our study and to detect pleiotropic loci for size and shape traits[40,41], we used a multivariate linear mixed model approach which included groups of correlated root size and root size/shape traits. Considering multiple phenotypes in the mvLMM enabled us to identify new candidate loci on chromosomes 4, 6 and 9 that were not identified in the univariate analyses.

The potential of GS as a breeding tool to increase the rates of genetic gain was recently tested in three Next Generation Cassava Breeding programs[20]. The study showed promising results particularly for traits with consistent heritability values across programs and stable large-effect quantitative trait loci. Prediction accuracies for RTNO, RTWT and SHTWT were similar with those reported in previous cassava cross-validation analyses[20]. Root size and shape-related trait accuracies were lower than those reported for dry matter content (DM) and cassava mosaic virus resistance (CMD)[20].

Although the heritability of yield traits was higher in the offspring (Cycle 1, C1) than the parental generation (Genetic Gain, GG), the predictive accuracy of traits extracted from root images showed intermediate to high values in both populations. However, the C1 yield traits accuracies being lower than the GG, suggests that it is due to a reduction in variance caused by the strong selection on these yield traits.

Nonetheless, predictive accuracies of the mean values of root size and shape image-extracted traits were mostly higher than yield trait prediction accuracies in the C1 population. Adjusting the mean and standard deviation phenotypes for the effect of CMD reduced the predictive accuracy. However, that correction is necessary to unlink the effect of CMD from the causal loci that are responsible for the regulation size and shape root traits.

The successful identification of QTLs through GWAS analysis demonstrates that image size and shape traits from image analyses can capture genetic variation. The quantitative nature of these traits and the moderate to high predictive accuracy suggest that they can be included in a selection index to breed cassava with uniform storage root shape suitable for mechanical harvest.

Although these measurements were laborious in the field and not high-throughput, the analyses of the images are automated and quantitative, they avoid subjectivity in scoring and other human-errors and most importantly, they improve cassava root characterisation.

This study has shown that root phenotyping using image capture and analysis is feasible and can be included as part of routine yield data collection in field plots. We used fairly simple equipment including a green board platform, a cheap digital camera (can be substituted with a smartphone) and freely-available open source software for image processing. As expected, additional labour was required to lay and remove roots from the green board, a process that increase cost and time by about 15–20%. However, these costs could be offset through genomic prediction for shape-related traits in subsequent generations using training data from the parents and relatives.

The methodology described here and the results obtained in this study are promising and open up the opportunity to apply high-throughput methods in cassava. The image capture and analysis can now be performed using the OneKK (one thousand kernels) app (https://github.com/PhenoApps/OneKK), an inexpensive and user-friendly tool for automated measurement of seed size, shape, and weight using smart phones. The app is developed under the BREAD PhenoApps project and supported by the National Science Foundation. Still, further work is needed to automate all the steps of cassava root phenotyping to allow the characterization of more lines at different locations in multi-environmental trials that ultimately would allow the creation of larger training population sizes for genomic prediction of root size and shape traits.

Received: 12 July 2019; Accepted: 23 April 2020;
Published online: 14 May 2020

## References

1. Guira, F. *et al*. Origins, production, and utilization of cassava in Burkina Faso, a contribution of a neglected crop to household food security. *Food Sci Nutr* **5**, 415–423 (2017).
2. Howeler, R. H., Lutaladio, N., Nations, F. and A. O. of the U. & Thomas, G. *Save and Grow: Cassava: a Guide to Sustainable Production Intensification*. (Food & Agriculture Org, (2013).
3. Hahn, S. K., Reynolds, L., Egbunike, G. N. *Cassava as Livestock Feed in Africa: Proceedings of the IITA/ILCA/University of Ibadan Workshop on the Potential Utilization of Cassava as Livestock Feed in Africa: 14-18 November 1988, Ibadan, Nigeria*. (IITA (1992).
4. Lukuyu, B., Okike, I., Duncan, A. J., Beveridge, M. & Blummel, M. *Use of cassava in livestock and aquaculture feeding programs*. (ILRI (aka ILCA and ILRAD) (2014).
5. Alves, A. A. C. Cassava botany and physiology. In *Cassava: biology, production and utilization* 67–89 (CAB international. https://doi.org/10.1079/9780851995243.0067 (2002).
6. Ejovo, N. *et al*. Studies and Preliminary Design for a Cassava Tuber Peeling Machine. *Trans. ASAE* **31**, 380–385 (1988).
7. Onwueme, I. C. *The Tropical Tuber Crops: Yams, Cassava, Sweet Potato, and Cocoyams*. (John Wiley & Sons (1978).
8. Hartmann, A., Czauderna, T., Hoffmann, R., Stein, N. & Schreiber, F. HTPheno: an image analysis pipeline for high-throughput plant phenotyping. *BMC Bioinformatics* **12**, 148 (2011).
9. Furbank, R. T. & Tester, M. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* **16**, 635–644 (2011).
10. Fahlgren, N., Gehan, M. A. & Baxter, I. Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology* **24**, 93–99 (2015).
11. Colombi, T. *et al*. Next generation shovelomics: set up a tent and REST. *Plant Soil* **388**, 1–20 (2015).
12. Lestrel, P. E. *Biological Shape Analysis: Proceedings of the 1st International Symposium, Tsukuba, Japan, 3-6 June* 2009. (World Scientific (2011).
13. Iwata, H., Ebana, K., Uga, Y. & Hayashi, T. Genomic Prediction of Biological Shape: Elliptic Fourier Analysis and Kernel Partial Least Squares (PLS) Regression Applied to Grain Shape Prediction in Rice (Oryza sativa L.). *PLoS One* **10**, e0120610 (2015).
14. Iwata, H., Ebana, K., Uga, Y. & Hayashi, T. Genome-wide Association Study of Biological Shape Based on Elliptic Fourier Analysis: A Case Study in Rice Grain Shape Variation. *Biological Shape Analysis*. https://doi.org/10.1142/9789814704199_0007 (2015).
15. Heffner, E. L., Sorrells, M. E. & Jannink, J.-L. Genomic Selection for Crop Improvement. *Crop Sci.* **49**, 1 (2009).
16. Jannink, J.-L. L., Lorenz, A. J. & Iwata, H. Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* **9**, 166–177 (2010).
17. Wolfe, M. D. *et al*. Genome-wide association and prediction reveals the genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *Plant Genome* **9**, 1–13 (2016).
18. Kayondo, S. I. *et al*. Genome-wide association mapping and genomic prediction for CBSD resistance in Manihot esculenta. *Sci. Rep.* **8**, 1549 (2018).
19. Rabbi, I. Y. *et al*. Genome-Wide Association Mapping of Correlated Traits in Cassava: Dry Matter and Total Carotenoid Content. *Plant Genome* **10**, https://doi.org/10.3835/plantgenome2016.09.0094 (2017).

20. Wolfe, M. D. *et al*. Prospects for genomic selection in cassava breeding. *Plant Genome* **10**, https://doi.org/10.3835/plantgenome2017.03.0015 (2017).
21. Garrick, D. J., Taylor, J. F. & Fernando, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **41**, 55 (2009).
22. Elshire, R. J. *et al*. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, e19379 (2011).
23. Glaubitz, J. C. *et al*. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**, e90346 (2014).
24. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
25. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome J.* **4**, 250 (2011).
26. Endelman, J. B. & Jannink, J.-L. Shrinkage Estimation of the Realized Relationship Matrix. *G3 Genes|Genomes|Genetics* **2**, 1405–1413 (2012).
27. Akdemir, D. & Okeke, U. G. EMMREML: Fitting Mixed Models with Known Covariance Structures. https://cran.r-project.org/package=EMMREML. R package version 3.1 (2015).
28. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
29. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* https://doi.org/10.1101/005165 (2014).
30. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–9 (2014).
31. Adetan, D. A., Adekoya, L. O. & Aluko, O. B. Characterisation of some properties of cassava root tubers. *J. Food Eng.* **59**, 349–353 (2003).
32. Padonou, W., Mestres, C. & Nago, M. C. The quality of boiled cassava roots: instrumental characterization and relationship with physicochemical properties and sensorial properties. *Food Chem.* **89**, 261–270 (2005).
33. Anggraini, V., Sudarmonowati, E., Sri Hartati, N., Suurs, L. & Visser, R. G. F. Characterization of Cassava Starch Attributes of Different Genotypes. *Starch - Stärke* **61**, 472–481 (2009).
34. Iwata, H., Niikura, S., Matsuura, S., Takano, Y. & Ukai, Y. Diallel Analysis of Root Shape of Japanese Radish (Raphanus sativus L.) Based on Elliptic Fourier Descriptors. *Breed. Sci.* **50**, 73–80 (2000).
35. Bredeson, J. V. *et al*. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* **34**, 562–570 (2016).
36. Seif, A. A. Effect of Cassava Mosaic Virus on Yield of Cassava. *Plant Dis.* **66**, 661 (1982).
37. Otim-Nape, G. W., Thresh, J. M. & Shaw, M. W. The effects of cassava mosaic virus disease on yield and compensation in mixed stands of healthy and infected cassava. *Ann. Appl. Biol.* **130**, 503–521 (1997).
38. Owor, B., Legg, J. P., Okao-Okuja, G., Obonyo, R. & Ogenga-Latigo, M. W. The effect of cassava mosaic geminiviruses on symptom severity, growth and root yield of a cassava mosaic virus disease-susceptible cultivar in Uganda. *Ann. Appl. Biol.* **145**, 331–337 (2004).
39. Nisar, N., Li, L., Lu, S., Khin, N. C. & Pogson, B. J. Carotenoid metabolism in plants. *Molecular Plant*. https://doi.org/10.1016/j.molp.2014.12.007 (2015).
40. Korte, A. *et al*. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071 (2012).
41. Korol, A. B., Ronin, Y. I., Itskovich, A. M., Peng, J. & Nevo, E. Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics* **157**, 1789–1803 (2001).


## Acknowledgements

This research was supported by the Bill & Melinda Gates Foundation and the Department for International Development of the United Kingdom through the "Next Generation Cassava Breeding Project", and CGIAR-Research Program on Roots, Tubers and Bananas. The authors acknowledge the support of Andrew Smith Ikpan, Ogunpaimo Kayode and Cynthia Idhigu in acquiring the root images. We also acknowledge the PhenoApp project led by Jesse Poland, Kansas State University, for technical support in image-based phenotyping using their OneKK (one thousand kernel) application.

## Author contributions

I.R., P.K. and M.D.W. designed the experimental study. D.P.C. and B.O.Y. developed data analysis tools, interpreted the results and wrote the first version of the manuscript. J.L.J. and I.R. provided editorial advice and commented on the manuscript. D.P.C. and M.D.W. wrote the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41598-020-64963-9.

**Correspondence** and requests for materials should be addressed to I.R.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.