Received: 18 June 2019 | Accepted: 4 November 2019 | Published online: 25 February 2020

DOI: 10.1002/csc2.20003

Crop Science

ORIGINAL RESEARCH ARTICLE

Genomics, Molecular Genetics & Biotechnology

Genomic prediction and quantitative trait locus discovery in a cassava training population constructed from multiple breeding stages

Mohamed Somo[1] | Heneriko Kulembeka[2] | Kiddo Mtunda[2] | Emmanuel Mrema[2] | Kasele Salum[2] | Marnin D. Wolfe[1] | Ismail Y. Rabbi[3] | Chiedozie Egesi[3] | Robert Kawuki[4] | Alfred Ozimati[4] | Roberto Lozano[1] | Jean-Luc Jannink[1,5]

[1] Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA

[2] Tanzania Agricultural Research Institute, Kibaha and Mwanza, Tanzania

[3] International Institute for Tropical Agriculture, Ibadan, Oyo, Nigeria

[4] National Crops Resources Research Institute, Namulonge, Uganda

[5] USDA–ARS, Ithaca, NY, USA

Correspondence
Mohamed Somo, Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA.
Email: msomowork@gmail.com

Funding information
Next Generation Cassava Breeding Project; Bill and Melinda. Gates Foundation; Department for International Development of the United Kingdom

Assigned to Associate Editor Timothy Beissinger

Abstract

Assembly of a training population (TP) is an important component of effective genomic selection-based breeding programs. In this study, we examined the power of diverse germplasm assembled from two cassava (*Manihot esculenta* Crantz) breeding programs in Tanzania at different breeding stages to predict traits and discover quantitative trait loci (QTL). This is the first genomic selection and genome-wide association study (GWAS) on Tanzanian cassava data. We detected QTL associated with cassava mosaic disease (CMD) resistance on chromosomes 12 and 16; QTL conferring resistance to cassava brown streak disease (CBSD) on chromosomes 9 and 11; and QTL on chromosomes 2, 3, 8, and 10 associated with resistance to CBSD for root necrosis. We detected a QTL on chromosome 4 and two QTL on chromosome 12 conferring dual resistance to CMD and CBSD. The use of clones in the same stage to construct TPs provided higher trait prediction accuracy than TPs with a mixture of clones from multiple breeding stages. Moreover, clones in the early

Abbreviations: AYTKIB, advanced yield trial at Kibaha; BLUPs, best linear unbiased predictors; CBSD, cassava brown streak disease; CBSDRS, cassava brown streak disease root necrosis severity; CETUKG, clonal evaluation trial at Ukiriguru; CMD, cassava mosaic disease; DM, dry matter; FYLD, fresh root yield; GEBVs, genomic estimated breeding values; GS, genomic selection; GWAS, genome-wide association study; HI, harvest index; MCBSDS, mean cassava brown streak disease severity; MCGMS, mean cassava green mite severity; MCMDS, mean cassava mosaic disease severity; PC, principal component; PVE, phenotypic variance explained; PYTKIB, preliminary yield trial at Kibaha; PYTUKG, preliminary yield trial at Ukiriguru; QTL, quantitative trait locus or loci; RTNO, root number; SHTWT, shoot weight; SNP, single nucleotide polymorphism; TARI, Tanzania Agriculture Research Institute; TP, training population.

breeding stage provided more reliable trait prediction accuracy and are better candidates for constructing a TP. Although larger TP sizes have been associated with improved accuracy, in this study, adding clones from Kibaha to those from Ukiriguru and vice versa did not improve the prediction accuracy of either population. Including the Ugandan TP in either population did not improve trait prediction accuracy. This study applied genomic prediction to understand the implications of constructing TP from clones at different breeding stages pooled from different locations on trait accuracy.

1 | INTRODUCTION

Cassava is an important source of dietary calories for millions of people in tropical regions (Howeler, Lutaladio, & Thomas, 2013; Salvador, Steenkamp, McCrindle, & Ethelwyn, 2014). Because of cassava's adaptation and resilience to dry and marginal environments, it is likely to continue to be a lead contributor to food security for many. Traditionally, cassava breeders have used phenotypes for selection but this strategy has not always generated adequate genetic gain for yield, especially in Africa (Ceballos, Kulakow, & Hershey, 2012). Similarly, the use of marker-assisted selection for complex traits in cassava has remained largely ineffective because of multiple minor loci, which are difficult to detect and deploy (Ceballos et al., 2012).

Genomic selection allows superior plants to be selected earlier in the seedling stage before phenotyping is started; these plants can then be used for crossing (Heffner, Sorrells, & Jannink, 2009; Jannink, Lorenz & Iwata, 2010; Lorenzana & Bernardo, 2009). In principle, the adoption of GS is expected to increase the rate of genetic gain and also reduce selection cycle time. For the case of clonal plants, the added advantage is that once elite clones are identified, they are genetically fixed and can thus immediately be advanced to downstream evaluation for faster variety replacement. Recent implementation of GS in three breeding programs in Africa have improved trait prediction accuracy. Wolfe et al. (2017) reported a prediction accuracy increase of 57% for CMD in a Nigerian cassava population. Similarly, Kayondo et al. (2018) reported improvements in prediction accuracy of 0.42 for both CBSD severity in leaves and roots in two Ugandan cassava populations.

In principle, successful implementation of a GS program hinges on several critical factors, one of which is how the TP is constituted. Mindful of breeding objectives, breeders can select individuals from the available diverse germplasm and improved breeding lines, within and across breeding stages to construct TPs. They could also construct a TP in clones from different breeding programs or even clones from different countries. Most early crop GS studies used genotypes from the same breeding stages to form TPs. These individuals were either tested in replicated trials in multiple environments or were part of larger historical phenotypic datasets (Dawson et al., 2013; Ly et al., 2013; Rutkoski et al., 2015; Storlie & Charmet, 2013).

Often, breeding programs transitioning to GS-based approaches may not have access to sufficient diverse genetic pools and adequate historical data to choose TP candidates from. For the case of clonal crops, the inadequacy of planting materials from potential candidates could further reduce the number of genotypes available for selection. In GS breeding, large TP size is generally preferred because it is associated with better trait prediction accuracy (Cericola et al., 2017; Zhong, Dekkers, Fernando, & Jannink, 2009).

Recently, The Tanzania Agriculture Research Institute (TARI) transitioned to a GS-based cassava breeding pipeline. Traditionally, TARI has maintained two distinct cassava breeding programs at Chambezi in Kibaha and Ukiriguru in Mwanza, serving the Coastal and Lake Zone parts of the country, respectively. The two breeding programs evolved independently because of restrictions on germplasm movement between regions to curb the spread of CBSD. Although the breeding objectives are quite similar between the two programs, cassava breeders at Kibaha largely selected for fresh root yield, whereas those at Ukiriguru focused mainly on dry matter content. This dichotomy of trait selection was largely driven by the consumption patterns of the local communities in the respective regions.

The TARI TPs were constructed from clones from different trial types. The Kibaha trials included clones in the preliminary yield trial (PYTKIB) and advanced yield trial (AYTKIB) selection stages, whereas those at Ukiriguru consisted of clones in the clonal evaluation trial (CETUKG) and the preliminary yield trial (PYTUKG). In a breeding pipeline, clonal evaluation trial genotypes are minimally selected in seedling nurseries before cloning, whereas those from preliminary and advanced yield trials would have undergone one

TABLE 1 The trial structure for the materials used for the training population, including the number of individual clones in each of the three replicates after sprouting

				Clones per replicate			Plot area (M^2)	
Trial	Stage	Location	Replicates	Replicate 1	Replicate 2	Replicate 3	Planted	Harvested
				n			m^2	
CET1	CET[a]	Ukiriguru	2	108	91	–	5	5
CET2	CET	Ukiriguru	2	213	214	–	5	5
PYT1	PYT	Ukiriguru	2	96	81	–	10	10
PYT1	PYT	Kibaha	2	64	65	–	14	14
PYT2	PYT	Kibaha	2	74	77	–	14	14
PYT3	PYT	Kibaha	2	45	48	–	14	14
PYT4	PYT	Kibaha	2	21	23	–	14	14
PYT5	PYT	Kibaha	2	24	23	–	14	14
PYT6	PYT	Kibaha	2	15	15	–	14	14
PYT7	PYT	Kibaha	2	20	19	–	14	14
PYT8	PYT	Kibaha	2	55	51	–	14	14
PYT9	PYT	Kibaha	2	58	63	–	14	14
AYT1	AYT	Kibaha	2	26	25	–	42	20
AYT2	AYT	Kibaha	3	12	12	12	42	20
AYT3	AYT	Kibaha	3	16	13	15	42	20
AYT4	AYT	Kibaha	2	23	23	–	42	20
AYT5	AYT	Kibaha	2	25	20	–	42	20
AYT6	AYT	Kibaha	2	21	21	–	42	20

[a]CET, clonal evaluation trial; PYT, preliminary yield trial; AYT, advanced yield trial.

and two specific selection rounds, respectively. In the present study, each breeding cycle represented several trials (Table 1). Each trial had different plot sizes, replications, and plants per plot. Additionally, there were no common checks for the trials across the two programs. Because there were too few clones in each trial to form a sizeable TP, clones from different trial types, breeding stages, and locations were pooled together to form TPs. The Ugandan TP was added to the clones from the two TARI programs to evaluate whether their inclusion could improve trait prediction accuracy. There is limited knowledge about the implications of using cassava clones from multiple trials, clones from multiple locations, and clones from multiple breeding stages to construct TPs regarding trait predictions. This study was therefore designed to: (a) examine which combination of breeding materials could provide the best training set; (b) evaluate trait prediction accuracy within and across breeding stages, locations, and combined TARI datasets; (c) assess the impact on trait predictions of adding Ugandan clones to clones from either program; and (d) identify the loci associated with disease resistance, particularly to CMD, CBSD, and CBSD root necrosis in Tanzanian cassava. The findings of this study will guide breeders in selecting the most suitable candidates as well as the best source of TP candidates when initiating GS-assisted breeding. We also expect the markers identified in this study will be a valuable addition for marker-assisted selection breeding in cassava.

2 | MATERIALS AND METHODS

2.1 | Germplasm and field design

A diverse panel of breeding lines from two independent breeding programs was used to construct the TP. The panel comprised 432 and 408 clones from the Kibaha and Ukiriguru breeding programs, respectively. The Kibaha clones were at the preliminary (PYTKIB) and advanced yield trial (AYTKIB) breeding stages, whereas those from Ukiriguru were at the clonal evaluation trial stage (CETUKG) and the preliminary yield trial stage (PYTUKG). The Kibaha clones were drawn from nine preliminary and six advanced yield trials, whereas Ukiriguru clones consisted of two clonal evaluation trial and one preliminary trial. Clones in each trial type arose from different progenitors and, as such, each trial in both programs had a unique set of individuals, a different number of replicates, and different plot sizes (Table 1). An additional 402 clones from Uganda previously described by Wolfe et al. (2016) were included in the study.

For each trial set, clones were evaluated in a randomized complete block design during the 2016–2017 cropping season at Kibaha (6.8138°S, 38.6949°E; 156 m a.s.l.) and Ukiriguru (2.71666667°S, 33.01666667°E; 1194 m a.s.l.). The Kibaha preliminary yield trials were laid out as a two-row plot with 14 plants, whereas the advanced yield trials were 42 plants planted in six-row plots. The clonal evaluation plots at Ukiriguru consisted of a single-row of five plants, whereas the preliminary yield trials were two-row plots with 10 plants. At both locations, the clones were spaced at 1 by 1 m. With exception of the two AYTKIB trials, which were replicated three times, the rest of the trials in both locations were replicated twice (Table 1). For all Kibaha trials, 'Kiroba', 'Mfaransa', and 'Mkuranga1' were used as common checks, whereas 'Mkombozi', 'Lwakitangaza', and 'Liongokwimba' were the checks in the Ukiriguru trials. For both locations, standard agronomic practices under rainfed cropping were applied. All the field management and phenotyping took place between March 2016 and May 2017.

2.2 | Phenotyping

Nine traits were evaluated for this study. Cassava mosaic disease severity, foliar CBSD severity, and cassava green mite [*Mononychellus tanajoa* (Bondar)] severity were scored at 3, 6, and 9 mo after planting on a scale of 1 (No symptoms) to 5 (severe symptoms). For analysis we used the season-wide mean severity [mean CMD severity (MCMDS), mean CBDS severity (MCBSDS), and mean cassava green mite severity (MCGMS)] for 3, 6, and 9 mo after planting. For CBSD root necrosis severity (CBSDRS), necrotic symptoms on a scale of 1 to 5 were scored 12 mo after planting in cross-sections of roots (Hillocks & Thresh, 2000), where 1 = no necrosis and 5 is >25% necrotic and severe root constriction. We then used average the disease scores for each experimental plot for analysis.

Root number (RTNO) is the number of fresh roots harvested per plot, which was used to calculate RTNO per hectare. Root weight and shoot weight (SHTWT) were both measured as kg per plot and then used to calculate below- and aboveground yield in Mg ha^{-1}. Harvest index (HI) was the ratio between root weight and total biomass (root weight +SHTWT) per unit of area. Dry matter (DM) was expressed as a percentage of fresh root weight (FYLD), calculated via the specific gravity method. The specific gravity of each sample (i.e., the whole root, including the peel) was determined from the weights in air and in water (Kawano, Fukuda, & Cenpukdee, 1987). The experimental and phenotypic data of the Ugandan population are as described previously (Kayondo et al., 2018; Ozimati et al., 2018; Wolfe et al., 2016). Eight variables were common to both the Ugandan and TARI populations, except MCGMS, which is specific to TARI.

2.3 | Genotyping

Single nucleotide polymorphism (SNP) marker genotypes were obtained via genotyping-by-sequencing (Elshire et al., 2011). The markers were called using the TASSEL version 5.0 genotyping-by-sequencing pipeline version 2 (Glaubitz et al., 2014) after aligning the resulting reads to the *M. esculenta* version 6 assembly available from Phytozome at http://phytozome.jgi.doe.gov/ (accessed 21 Jan. 2020). Genotype calls were accepted only when there was a minimum of two reads; otherwise, the genotype was set to missing and imputed downstream. The markers were initially converted to a matrix of allele dosages, with REF/REF, REF/ALT, ALT/ALT genotypes coded as −1, 0 and +1, respectively. The genotyping-by-sequencing data were filtered so that clones with >80% missing markers and markers with >60% missing genotype calls were removed. Beagle version 4.1 was used for imputation of the data (Browning & Browning, 2009). After imputation we retained 121,246 bi-allelic SNP markers with AR2 (Estimated Allelic r-squared) threshold higher than 0.3. Of the imputed markers, 116,837 markers with minor allele frequency higher than 0.01 were retained. These markers were then used for genomic prediction. Similarly, a total of 88,434 markers were retained after filtering with minor allele frequency threshold higher than at 0.05 and used for GWAS analysis. 37,776 markers common to both the Ugandan and TARI populations were filtered for minor allele frequency (0.01) and the resulting 36,847 markers were then used for downstream joint analysis.

3 | STATISTICAL ANALYSIS

3.1 | Estimation of observed values

To estimate these observed values, we fitted a mixed linear model across trial types and locations with the lme4 R package (Bates et al., 2015; Vazquez, Bates, Rosa, Gianola, & Weigel, 2010). For trials within locations, we fitted the model:

$$y = \mathbf{X\beta} + \mathrm{Z}_{\mathrm{clone}}\boldsymbol{u} + \mathbf{Z}_{\mathbf{Trial(Trial:Rep)}}r + \varepsilon, \qquad (1)$$

where y represents the raw phenotypic observations, β includes a fixed effect for the population mean and for plot-basis traits (FYLD, RTNO, and SHTWT), and the proportion of plants harvested per plot was included as a covariate with the design matrix **X**. The vector $\boldsymbol{u}$ and the corresponding incidence matrix $\mathbf{Z}_{\mathbf{clone}}$ represent the random effect for the clones, where $u \sim \mathrm{N}(0, \mathbf{I}\sigma_u^2)$ and **I** represents the identity matrix. The incidence of replication was nested in trial and was represented by the matrix $\mathbf{Z}_{\mathbf{Trial\ (Trial.Rep)}}$ and the random effects vector such that $r \sim \mathrm{N}(0, \mathbf{I}\sigma_r^2)$ and the residual ε was considered to be distributed as $\varepsilon \sim \mathrm{N}(0, \mathbf{I}\sigma_\varepsilon^2)$.

The model for the combined analysis across locations was:

$$y = \mathbf{X}\beta + \mathbf{Z}_{\mathbf{clone}}u + \mathbf{Z}_{\mathbf{rep(Loc.Trial.Rep)}}b + \boldsymbol{\varepsilon}, \qquad (2)$$

where the fixed and clone effects combined across locations were the same as those for each location described above. Trial and replicate effects were nested in locations and incorporated as random with the incidence matrix $\mathbf{Z}_{\mathbf{rep}}$ and the effects vector $\mathbf{b} \sim \mathrm{N}(0, \mathbf{I}\sigma_b^2)$. All the random effects in each of the models are independent and identically distributed. For both models described above, the best linear unbiased predictors (BLUPs) for the clones were extracted and used as the observed values that were correlated with the genomic estimated breeding values (GEBVs) to determine trait prediction accuracy.

In the combined location model, our emphasis was on predicting new genetic materials from another breeding program. Therefore, we considered the environmental effects to be nuisance parameters. We were not interested in location, trial, or replicate effects per se; instead, we were interested in GEBVs. We assumed that the three-way interaction of location, trial and replicate would be sufficient to correct for these nuisance terms. No main effects were required to be fitted, as they were simply linear combinations of the interaction predictors.

We also considered the issue of heterogeneity in the error variances for different trials across the two locations. Although error variance estimates differed across trials, those differences had little effect on the estimated BLUP values.The correlation between BLUPs with and without heterogeneous error variance were between 0.92 and 0.96, except for FYLD, which had reduced BLUP correlations (~0.48; data not shown). To overcome the challenges and complexity of modeling the heterogeneity of error variances, particularly for yield, we fitted the model for each trial type in each location separately and then extracted BLUPs from each experiment and used the combined BLUPs and correlated them with the GEBVs to determine the prediction accuracy. For the rest of the traits, because of the large number of trials (18), we felt that fitting the error variance for each trial would result in too many parameters to estimate, we used a simpler homogenous error variance model.

3.2 | Variance components, heritability, and trait correlation

The variance components were extracted from the model used for the trials within location. The model statement is as described above. Accordingly, broad sense heritability (H^2) was then computed for each trial as:

$$\boldsymbol{H}^2 = \frac{\boldsymbol{\sigma}_{\mathbf{c}}^2}{\boldsymbol{\sigma}_{\mathbf{c}}^2 + \boldsymbol{\sigma}_{\mathbf{e}}^2}, \qquad (3)$$

where $\boldsymbol{\sigma}_c^2$ is the clone variance and $\boldsymbol{\sigma}_{\mathbf{e}}^2$ is the residual variance.

The phenotypic and genotypic correlations between the nine traits were obtained from combined data from the two locations. We also assessed trait phenotypic correlations by using data from each breeding stage in each program, namely the Ukiriguru populations (CETUKG and PYTUKG) and the Kibaha populations (PYTKIB and AYTKIB), to determine the consistency of trait correlations across different breeding stages. For the combined locations, we used raw plot data to estimate phenotypic correlations, without accounting for the experimental design. For estimating the genetic correlations, we accounted for the trial, location, and replication effects, as described above. We then extracted and used BLUPs to estimate genetic correlations. The estimates were plotted (Figure 2) in R version 3.4.1 with the corrplot package (Wei, 2013). For all the cases, the correlation values were considered to be significantly different from zero at $P \leq 0.05$.

3.3 | Genomic prediction

To perform genomic prediction, we fitted separate genomic BLUP models as $y = \mathbf{X}\beta + Zg + \boldsymbol{\varepsilon}$ within breeding stages, across breeding stages, and across breeding programs. where y is a vector of the raw phenotype, β is the vector of fixed effects (which is different for different scenarios) with the design matrix **X**, the vector **g** is the random effect representing the GEBV for each individual, **Z** is a design matrix linking observations to genomic values, and ε is a vector of the residuals. For within-stage predictions, the fixed effect is the grand mean, whereas the fixed effects for across-stage predictions were the grand mean and trial. For cross-program predictions, the fixed effects included trial and location. The GEBV was obtained under the assumption that $\boldsymbol{g} \sim N(0, \boldsymbol{K}\sigma_g^2)$, where the additive genetic variance and σ_g^2 K is the square, symmetric genomic relationship matrix based on SNP markers. The genomic relationship matrix was constructed with the function A.mat in the R package rrBLUP (Endelman, 2011), which used VanRaden's (2008) Method 1. The genomic BLUP predictions were made with the function *emmreml* in the R version 3.1.0 package EMMREML (Akdemir & Godfrey, 2015).

3.4 | Assessment of prediction accuracy

In order to assess prediction accuracy, we used 30 replicates of fivefold cross-validation. For each replicate, we divided the population randomly into five equal and mutually exclusive subsets or folds. We then trained the prediction model on four of the five folds (training sets) to predict the fifth (validation set). For scenarios where we used one location to predict another location, our training and test sets were fixed and were therefore not replicated. The following prediction scenarios were explored in our analysis: (i) within breeding stages in each location: CETUKG, PYTUKG, PYTKIB, and AYTKIB; (ii) combined locations (Ukiriguru + Kibaha); (iii) within locations: prediction of the Ukiriguru sets alone (CETUKG + PYTUKG) and the Kibaha sets alone (PYTKIB + AYTKIB); and (iv) cross-location prediction: use of the Ukiriguru set to predict the Kibaha set, use of the Kibaha set to predict the Ukiriguru set, and use of the Ukiriguru + Kibaha set to predict the Ukiriguru set and use of the Ukiriguru + Kibaha set to predict the Kibaha set. We also used the Ukiriguru + Uganda set to predict the Ukiriguru set, the Kibaha + Uganda set to predict the Kibaha set, and the Ukiriguru + Kibaha + Uganda set to predict the Ukiriguru and Kibaha sets. This was done to determine if adding Ugandan clones to either TARI program would improve the trait prediction accuracy. For Scenarios 3 and 4, we maintained a fixed test set (20%) picked randomly from the population being predicted in each iteration. For each prediction, accuracy was computed as the Pearson's correlation coefficient between the GEBV predicted for the test set and the corresponding estimated observed breeding values or BLUPs. Correlation values were considered significantly different from zero at P value ≤ 0.05.

3.5 | Population structure

To visualize the population structure, principal component analysis of the common markers in the three breeding programs (Kibaha, Uganda, and Ukiriguru) was performed. The prcomp function in R was used to generate the principal components. The first two principal components were then used for plotting to visualize structure between the three populations (Kibaha, Ukiriguru and Uganda). Thereafter the loadings (eigenvector coefficients) for all the markers on 18 chromosomes on Principal Component (PC) 1 and PC2 were assessed to determine the markers contributing to the greatest variations in TARI alone and the TARI and Ugandan clones combined in the respective principal components.

3.6 | Genome-wide association study

A GWAS was performed on different subsets of the TARI accessions. The subsets included trial types (clonal evaluation, preliminary yield, and advanced yield), locations (Kibaha and Ukiriguru), and four clusters based on the marker kinship matrices of the individuals. We used these combined datasets because we suspected that certain chunks of data could provide better results than others. Clustering was considered because we thought it would be an objective way to find the population structure. By conducting a GWAS within each cluster, we were hoping to avoid population structure and find better allele frequencies.

In implementing the GWAS, we used BLUPs extracted from the linear mixed model described above as the phenotypes. For each trial, the GWAS was carried out with the genome-wide complex trait analysis tool (Yang, Lee, Goddard, & Visscher, 2011). This followed a leave-one-out approach, in which the chromosome with the tested candidate SNP markers was excluded from the genomic relationship calculation. The linear mixed model in Equation (4) was fitted for each case:

$$y = X\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon} \;\; \textbf{with var}\,(\mathbf{y}) = \mathbf{V} = \mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2, \tag{4}$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotype (BLUPs) ,with n being the sample size; $\boldsymbol{\beta}$ is a vector of fixed effects (genetic marker information); $\mathbf{g}$ is an $n \times 1$ vector of the total genetic effects of the individuals with $\mathbf{g} \sim N\,(\mathbf{0}, \mathbf{K}\boldsymbol{\sigma}_\mathbf{g}^\mathbf{2})$; $\mathbf{K}$ is the genetic relationship matrix among individuals, which is the same as the symmetric genomic realized relationship matrix based on SNP markers; $\mathbf{I}$ is an $n \times n$ identity matrix; and ε is a vector of the residual effects with $\boldsymbol{\varepsilon} \sim \mathbf{N}\,(\mathbf{0}, \mathbf{I}\boldsymbol{\sigma}_\boldsymbol{\varepsilon}^\mathbf{2})$.

A Bonferroni threshold was applied to correct for multiple testing for each dataset used for the GWAS. The cutoff was computed as $-\log_{10}(\alpha \div t \times n)$, where α is 0.05, which is the standard significance threshold, t is the number of different subsets of data, and n is the number of SNPs. In this study, the Bonferroni correction significance cutoff was $-\log_{10}$ (0.05 ÷ 15 × 88,434) = 7.42. We used the marker-wise *P*-value of 0.001 as the threshold to declare significant marker–trait associations and the *P*-value of 0.2 corrected for false discovery rate to select the highly significant marker–trait associations. The QTL boundaries were defined by a linkage disequilibrium (LD) pairwise correlation coefficient of $r^2 \geq 0.7$ coupled with the marker–trait association information. We chose the tagging marker for each QTL by the criteria of the smallest marker-wise *P*-value. Manhattan and quantile–quantile plots were generated by the R package *ggplots2* (Wickham, Chang, & Wickham, 2016), with customization for joint Manhattan plus quantile–quantile plot display through the R package *ggpubr (*Kassambara, 2018). We used the *M. esculenta*_305_v6.1 database available in Phytozome to report the presence of annotated genes that are related to plant defense systems within the QTL region.

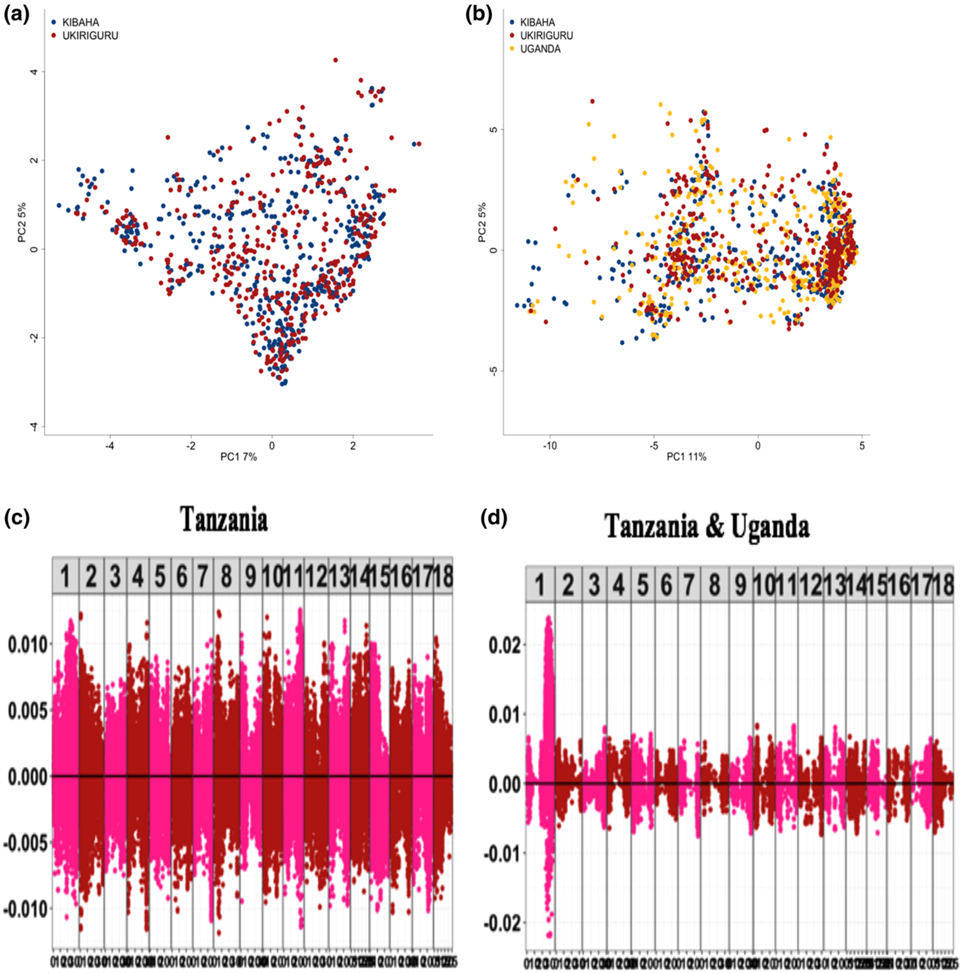

FIGURE 1 Principal component analysis (PCA) plot of the Tanzania Agriculture Research Institute (TARI) clones alone (a) and the combined TARI and Ugandan (UG) clones (b) based on a single nucleotide polymorphism (SNP(marker matrix and the loading values (eigenvector coefficients) for each marker on Principal Component (PC)1 against marker positions on all cassava chromosomes for TARI alone (c) and the combined TARI and Ugandan set (d). For the TARI PCA, 121K SNPs were used and for the combined PCA, we used 37K SNPs common to both the TARI and Ugandan sets. For both cases, the marker loadings were based on common markers (37,000)

4 | RESULTS AND DISCUSSION

4.1 | Population structure

We performed principal component analysis to assess structure within and among the Ukiriguru, Kibaha, and Ugandan populations (Figure 1). This is necessary to determine how individuals from these programs are related to each other in order to understand the effect of germplasm sharing between Uganda and Tanzania. It will also be useful to understand whether the existing restrictions on clone movement between the Kibaha and Ukiriguru populations might have facilitated population differentiation within Tanzanian germplasm. In the combined data (Ukiriguru + Kibaha + Uganda), the first and second principal components accounted for 11 and 5% of the genetic variation, respectively. In the TARI sets (Kibaha + Ukiriguru), PC1 and PC2 explained 7 and 5% of the variation, respectively. Some population stratification within each breeding program was observed (Figure 1).

Three clusters of clones were clearly arrayed along PC1 (Figure 1b). The loadings of markers on PC1 (Figure 1d) showed a strong effect of chromosome 1 on this component, suggesting that the clusters are caused by the known introgression from *Manihot glaziovii* Müll.Arg. on this chromosome (Bredeson et al., 2016). This clustering along PC1 has been observed previously in Ugandan cassava (Ozimati et al., 2019) and we suggest that it depends on the dosage of the chromosome 1 introgression, either no, one, or two copies. Interestingly, PC2 (Supplemental Figure S1) strongly separated the Ukiriguru from Kibaha sets, whereast the Ugandan accessions were intermediate between these two. When we analyzed the TARI sets alone, differences between Ukiriguru and Kibaha aligned along PC1 (Figure 1a), whereas the chromosome 1 introgression aligned along PC2 (Figure 1c, Supplemental Figure S1). This analysis may also provide valuable information on the role of introgression in driving population stratification. Despite Tanzania being the presumed origin of chromosome introgressions in the 1930s, (Fauquet, Fargette, & Munihor, 1990; Hahn et al., 1979, 1980), according to this analysis, this introgression may have been subjected to more recombination events in Tanzanian populations (Figure 1c,d).

The Ugandan population was constituted in 2011 by combining progenitors sourced from CIAT, IITA, and Tanzania, and a few clones from within Uganda (Kawuki, personal communication, 2019). Because of the recent construction of the Ugandan TP, it is possible that the Ugandan introgression is newer than the Tanzanian ones, resulting in differentiation between subpopulations with the introgression and the ones without it. However, further investigation is needed to validate these hypotheses.

4.2 | Trait correlations

Trait phenotypic and genetic correlations were quite similar (Figure 2; Supplemental Figure S2). Phenotypic and genotypic correlations of the yield traits were moderately positive, except for DM. The highest correlation was between FYLD and SHTWT, followed by RTNO and SHTWT, and RTNO and FYLD, ($r = 0.6$, $r = 0.5$, and $r = 0.4$, respectively). These values are only slightly lower than previous findings: Kundy, Mkamilo, and Misangu (2014) reported a strong positive correlation between RTNO and FYLD ($r = 0.7$) but fewer clones were tested by these authors, which could have skewed their findings. In addition, the effects associated with locations and trial types in the present study might have contributed to the observed deviations. Generally, the positive correlations of FYLD, RTNO, and SHTWT could be exploited for simultaneous trait improvements. We also observed a slight positive genetic correlation between MCMDS and MCBSDS ($r = 0.3$) but a very low association between MCBSDS and CBSDRS, which agrees with the findings in other studies (Ozimati et al., 2019; Rwegasira & Rey, 2012). The positive relationship between these two foliar diseases provides an opportunity to increase resistance against both concurrently. Mean cassava green mite severity showed weak negative correlations with MCMDS, SHTWT, and FYLD (r = –0.3, –0.3, and –0.4, respectively). The negative correlation between MCGMS and FYLD corroborates the result of Chipeta et al. (2013) ($r = –0.53$). Breeders should keep these negative associations in mind and incorporate sources of resistance when breeding for yield. Assessments of trait correlations in each breeding stage showed similar patterns observed in combined datasets to those across breeding stages. However, we also noticed a slight improvement in some traits, although we think this could just be noise in the data (Supplemental Figure S2).

4.3 | Variance components and broad-sense heritability estimates

Variance components and broad-sense heritability estimates for each of the nine traits are shown in Table 2. Generally, the Kibaha trials yielded the largest genetic variance estimates for MCMDS, CBSDRS, FYLD, and SHTWT, whereas the Ukiriguru trials had the largest estimates for MCBSDS and RTNO. Conversely, both FYLD and SHTWT had the largest residual estimates in the AYTKIB and Ukiriguru trials, respectively. Errors associated with CBSDRS, MCGMS, RTNO, HI, and DM in AYTKIB were lower than in other trials. However, both genetic and error estimates for RTNO in the Ukiriguru trials were higher than those in the Kibaha trials. Additionally, HI and MCBSDS had similar error estimates across all the trials.

Heritability estimates ranged between 0.06 and 0.91. On the basis of the cassava heritability classifications, only HI in AYTKIB, MCBSDS in PYTUKG, and MCMDS in both AYTKIB and PYTUKG could be classified as high (>0.50) (Bhateria, Sood, & Pathania, 2006). Other researchers reported higher heritability for MCMDS, MCBSDS, and HI in different cassava populations (Adeniji, Odo, & Ibrahim, 2011; Kanyondo et al., 2018; Wolfe et al., 2017). The high heritability estimates for MCMDS and MCBSDS could be associated with varying levels of resistance in Tanzanian accessions caused by specific responses to various strains of CBSD viruses and cassava green mite severity. In the Lake Zone, there is a high prevalence of *African cassava mosaic virus*, *Eastern African cassava mosaic virus*, the Ugandan variant of *Cassava mosaic virus*, CBSVD, and the Ugandan variant of the CBSD virus. However, only *Eastern African cassava mosaic virus* and the CBSD virus are known to occur in the Eastern Zone, which includes Kibaha (Jeremiah et al., 2015; Legg & Raya, 1998). Extensive pathogen variation within and between the two regions of Tanzania could have caused the variability detected in the germplasm. Significant differences in HI have been reported among cassava cultivars (Kawano, Daza, Amaya, Rios, & Goncalves, 1978). There is intensive selection against clones with heavy branching in early generations. Clones with fewer branches are known to have low HI. It appears from these data that the clones from AYTKIB trials were heavily selected for low branching, resulting in higher broad-sense heritability (0.80).

Cassava brown streak disease severity, SHTWT, FYLD, and DM in the CETUKG trials and MCBSDS and DM in the Kibaha trials all had medium heritability estimates. Similar estimates were reported in other cassava populations (Kayondo et al., 2018; Ozimati et al., 2019; Wolfe et al., 2017). We observed lower FYLD heritability in AYTKIB than in any other trials. We can only speculate that intensive selection for disease resistance and DM in early breeding stages could have

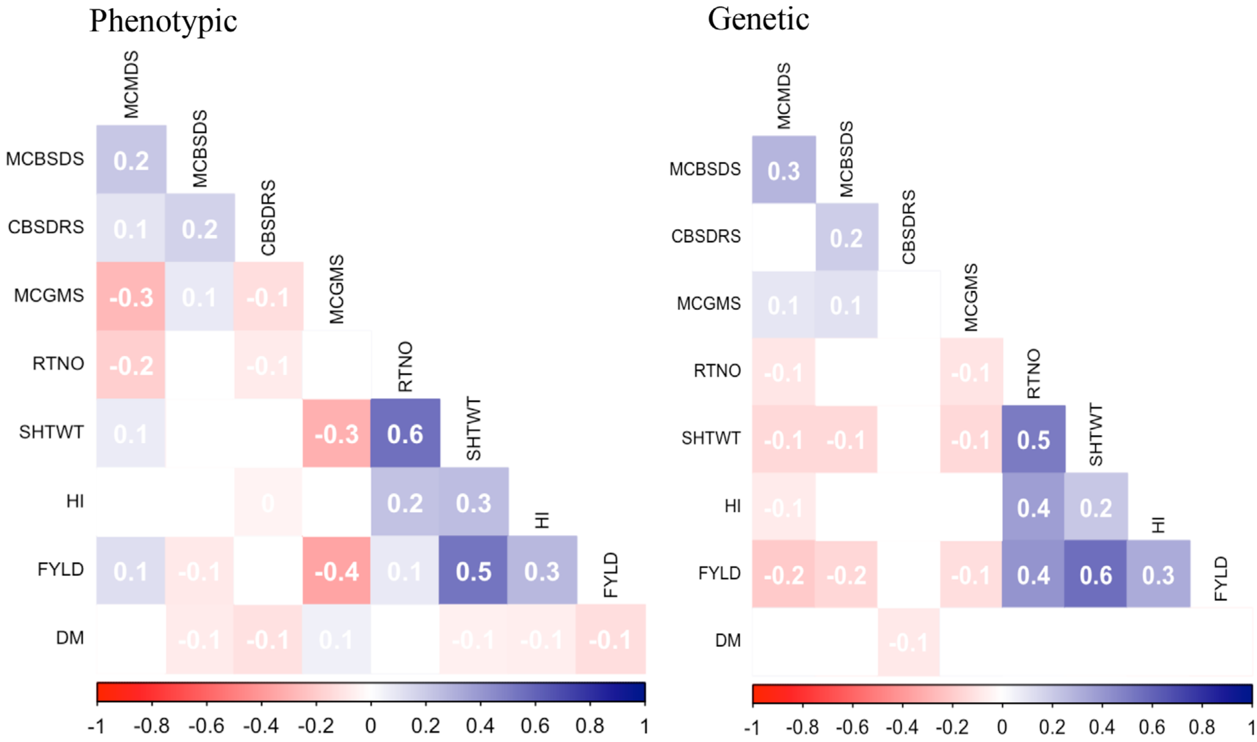

FIGURE 2 Phenotypic (left) and genetic (right) correlations among nine traits in cassava in combined Tanzania Agriculture Research Institute (TARI) data. Blue and red represent positive and negative correlations respectively. The strength of trait relationships is depicted by the intensity of the color. Cells with correlation values that are not significant at $P < 0.05$ have been left blank

TABLE 2 Heritability, genetic and residual variance components for four plot sizes

Plot size[a]	Variance component	MCMDS	MCBSDS	CBSDRS	MCGMS	RTNO	SHTWT	HI	FYLD	DM
5 m^2 (CETUKG)	σ^2_G	0.042	0.103	0.097	0.009	21950	12.71	0.005	11.01	7.11
	σ^2_{TR}	0.000	0.002	0.000	0.184	9516	9.10	0.005	0.46	1.69
	σ^2_e	0.049	0.121	0.215	0.142	71924	28.66	0.020	25.12	15.20
	H^2	0.46	0.46	0.31	0.06	0.23	0.31	0.20	0.30	0.32
10 m^2 (PYTUKG)	σ^2_G	0.050	0.114	0.008	0.037	22059	0.215	0.003	1.80	3.72
	σ^2_{TR}	0.000	0.000	0.005	0.000	15718	4.155	0.003	0.06	2.67
	σ^2_e	0.005	0.097	0.129	0.131	77142	10.605	0.017	5.80	16.84
	H^2	0.91	0.54	0.06	0.22	0.22	0.02	0.15	0.24	0.18
14 m^2 (PYTKIB)	σ^2_G	0.349	0.077	0.042	0.029	13763	14.89	0.004	50.09	5.23
	σ^2_{TR}	0.026	0.018	0.03	0.013	9203	28.08	0.002	24.69	1.74
	σ^2_e	0.432	0.100	0.21	0.118	43911	105.48	0.018	129.21	12.28
	H^2	0.43	0.39	0.14	0.18	0.24	0.10	0.17	0.25	0.27
20 m^2 (AYTKIB)	σ^2_G	0.28	0.083	0.031	0.051	4760	36.00	0.004	59.84	2.0885
	σ^2_{TR}	0.04	0.020	0.001	0.036	6688	70.08	0.003	66.79	0.9732
	σ^2_e	0.30	0.122	0.121	0.117	55303	197.90	0.001	211.45	10.2526
	H^2	0.44	0.38	0.20	0.25	0.08	0.12	0.80	0.18	0.16

[a] σ^2_G, genetic variance among clones; σ^2_{TR}, variance among replicates in trials; σ^2_e, residual error variance; H^2, plot-based broad-sense heritability estimates; CETUKG, clonal evaluation trial at Ukiriguru; PYTUKG, preliminary yield trial at Ukiriguru; PYTKIB, preliminary yield trial at Kibaha; AYTKIB, advanced yield trial at Kibaha; MCMDS, mean cassava mosaic disease severity; MCBSDS, mean cassava brown streak disease severity; CBSDRS, cassava brown streak disease root necrosis severity; MCGMS, mean cassava green mite severity; RTNO, root number; SHTWT, shoot weight; HI, harvest index; FYLD, fresh root yield; DM, dry matter.

limited and reduced the amount of FYLD genetic variation in advanced lines (AYTKIB). Therefore, we need to balance between trait selection preference and genetic variation in the breeding process.

For most traits, clonal evaluation trials with a smaller plot size (5 m^2) had lower error variance and better heritability than trials with medium and large plots (e.g. 14 and 20 m^2). High trait genetic variance has been reported in clonal evaluation trials of sugarcane (*Saccharum* spp.) at Louisiana State University Agricultural Center but the heritability estimates with different plot sizes were similar (Milligan, Balzarini, Gravois, & Bischoff, 2007). The substantial increase in genetic variance and trait heritability in smaller plots than larger plot sizes needs further investigations. This will help establish the optimal plot sizes for cassava breeding for better prediction accuracy. Generally, trait heritability estimates varied across trials, similar to the findings of Kayondo et al. (2018), Wolfe et al. (2017), and Ozimati et al. (2019). Confounding effects (locations and trials), different environmental conditions, trait selection priorities in each of the TARI breeding programs, and differences in data collection between the two programs could have affected the heritability estimates.

5 | GENOMIC PREDICTION

5.1 | Within-stage predictions

Within-stage prediction accuracy for all the traits in the CETUKG, PYTUKG, PYTKIB, and AYTKIB sets were 0.23, 0.20, 0.15, and 0.20, respectively (Table 3). The within-stage accuracy was lower than the accuracy from other cassava populations for most traits except DM and FYLD (Wolfe et al., 2017). In this study, the use of data from the CETUKG stage significantly improved prediction accuracy for all traits except MCMDS and CBSDRS compared with the PYTUKG, PYTKIB, and AYTKIB datasets. Overall prediction accuracy obtained from the cross-validations was higher for HI and FYLD when individuals in CETUKG and PYTUKG were used for model training. On the contrary, the two breeding stages evaluated at Kibaha produced markedly lower prediction accuracies for most traits except for MCGMS and DM.

Interestingly, there was consistent prediction of FYLD (~35%) in both CETUKG and PYTUKG. The stability in yield prediction in these breeding cycles at Ukiriguru suggests that breeders can opt to select candidates for the TP from either cycle, particularly when yield is the trait of interest. An added advantage of equal prediction in CETUKG and PYTUKG is that clones can directly be transferred from the clonal evaluation to advance yield trial stage for evaluation. Skipping the preliminary yield trial stage could help accelerate the varietal replacement process.

TABLE 3 Summary of cross-validated prediction accuracy by trait and breeding stage

Trait	CETUKG[a]	PYTUKG	PYTKIB	AYTKIB	Mean
MCMDS	0.11ns	0.06ns	0.17ns	0.32ns	0.14
MCBSDS	0.26*	0.33ns	0.08ns	0.24ns	0.23
CBSDRS	0.26ns	0.10ns	0.09ns	0.10ns	0.14
MCGMS	0.22*	0.09ns	0.21ns	0.43*	0.24
RTNO	0.28**	0.27ns	0.16ns	0.03ns	0.19
SHTWT	0.22ns	0.03ns	0.12ns	0.08ns	0.11
HI	0.22*	0.38*	0.13ns	0.19ns	0.23
FYLD	0.36*	0.35*	0.10ns	0.09ns	0.23
DM	0.14ns	0.18ns	0.28*	0.43*	0.26
Mean	0.23	0.20	0.15	0.20	0.19

[a] CETUKG, clonal evaluation trial at Ukiriguru; PYTUKG, preliminary yield trial at Ukiriguru; PYTKIB, preliminary yield trial at Kibaha; AYTKIB, advanced yield trial at Kibaha; MCMDS, mean cassava mosaic disease severity; MCBSDS, mean cassava brown streak disease severity; CBSDRS, cassava brown streak disease root necrosis severity; MCGMS, mean cassava green mite severity; RTNO, root number; SHTWT, shoot weight; HI, harvest index; FYLD, fresh root yield; DM, dry matter.

*Significant at the 0.05 probability level.

**significant at the 0.01 probability level; ns, nonsignificant prediction accuracy.

Dry matter prediction accuracy improved significantly by 53% when clones in the AYTKIB set were used for model training rather than those in PYTKIB. Comparison between breeding stages in the two locations showed high FYLD prediction accuracy in CETUKG and high DM in AYTKIB. One would have expected the opposite results because of the trait selection preferences between the two locations as well as the larger plot size associated with AYTs. We are not certain why clones in the early generation evaluated in smaller plots had higher yield than clones in advanced generations evaluated in larger plots. The difference could be attributed to higher genetic variation and higher heritability estimates for clonal evaluation clones than for advanced yield trial clones.

We also observed that using clones from the same breeding stage for TP gave higher prediction accuracy estimates than clones aggregated from multiple breeding stages. Our estimates agree with those found by Hofheinz, Borchardt, Weissleder, and Frisch (2012) for data from two consecutive breeding cycles of sugar beet (*Beta vulgaris* L.) and the study of Michel et al. (2018) of GS using multiple breeding cycles in bread wheat (*Triticum aestivum* L.). Ceballos et al. (2016) recommends the use of phenotypic information from clones at the advanced breeding stage during GS because of their “stable” genotypic performance. In this study, we observed that phenotypic records of clones in the early breeding stage predicted the test set better for most traits than advanced ones. We suggest that consideration should be given to within-stage clones, particularly clones from early generations when forming TPs. Additionally, CET clones are generally tested in smaller plots because of the limited number of planting stakes.

The use of smaller plots for trait evaluation can further help reduce phenotyping costs. Additional investigations of different cassava populations are needed to validate our results.

5.2 | Within- and cross-location predictions

One location's data can be used to predict performance in an independent location, which can help accelerate the breeding process. We evaluated cross-location predictions and compared the trait prediction accuracy estimates with the within-location prediction accuracy estimates. Except for MCGMS, the cross-location predictions were generally low, averaging between 0.10 and 0.14 when the Kibaha population was used to predict the Ukiriguru set and vice versa, respectively (Table 4). Mean cassava green mite severity was the highest and most consistently predicted trait both within and between locations. Adding the two populations together to increase the TP size did not improve trait prediction accuracy. We also noticed that the use of the Kibaha population to predict the Ukiriguru population reduced accuracy for CBSDRS, RTNO, SHTWT, HI, and FYLD compared with the within-Ukiriguru scenario. The unrelatedness of materials across cycles, confounding effects (trials and locations), and variable heritability across locations could have caused the low accuracy. Perhaps the existing quarantine between the two TARI programs associated with subtle population differentiation (Figure 1) as well as the absence of common checks between trials may have caused low trait prediction estimates between locations.

Reduced common ancestors over time results in reduced kinship, which reduces accuracy (de los Campos, Hickey, Pong-Wong, Daetwyler, & Calus, 2013). Lorenz and Smith (2015) reported a decrease in prediction accuracy when unrelated lines are added to the calibration set. Similarly, Song et al. (2017) reported a decrease in prediction accuracy when predicting yield across cycles compared with within cycles in wheat. Our results are also in agreement with the findings of Ly et al. (2013). They observed a decrease in prediction accuracy across locations in cassava. Therefore, according to this population, cross-location prediction may not be useful for programs implementing GS. Although, in general, it would be useful to implement GS across locations, the use of populations with structure and weak SNP–QTL linkage disequilibrium associations across populations could limit GS-assisted breeding in cassava and other species. This limitation has also been observed in livestock: Hayes, Bowman, Chamberlain, and Goddard (2009) reported that pooling animals from different populations did not improve trait predictions because of nonpersistent association between SNPs and QTL across breeds (populations).

In our study, we did not find a consistent relationship between heritability and prediction accuracy estimates across breeding cycles. In fact, we observed the same accuracies for FYLD in both UKGCET and UKGPYT. Other studies have reported a positive relationship between heritability and prediction accuracy across breeding cycles. For example, Sallam et al. (2015) reported higher prediction accuracy for *Fusarium head blight resistance* than for yield in cross-breeding cycles.

This is because highly heritable traits have a less complex genetic architecture and are therefore considered to be stable across multiple cycles. Combs and Bernardo (2013) and Daetwyler et al. (2010) have both attributed the positive relationship between heritability and accuracy to the preserved haplotype structures and relatedness across breeding cycles.

5.3 | Across-program prediction

The Ukiriguru population had slightly better prediction accuracy for most traits than the Kibaha set (Table 5). We added the Ukiriguru clones to the Kibaha clones and vice versa and predicted a fixed number (20%) of Kibaha or Ukiriguru accessions to determine whether including these clones could improve trait prediction. Adding Ukiriguru clones to the Kibaha training set did not improve prediction accuracy for most traits. On the other hand, adding the Kibaha clones to the Ukiriguru training set reduced the prediction accuracy for most traits. This effect was more severe on FYLD than any other trait (a change from $r = 0.30$ to $r = 0.08$). Similarly, adding the Ugandan clones to either the Kibaha or Ukiriguru clones did not improve prediction accuracy for either program. However, use of the Ukiriguru + Uganda set to predict the Ukiriguru test set slightly improved the results compared with use of the Kibaha + Uganda set to predict the Kibaha test set.

Furthermore, we used the Ukiriguru + Kibaha + Uganda set to predict the Ukiriguru and Kibaha sets to determine if adding Ugandan clones to TARI clones would improve trait prediction accuracy. There was no improvement in trait prediction. In fact, there were decreases for CBSD severity in leaves, CBSDRS, and SHTWT (from $r = 0.16$ to 0.07, from $r = 0.14$ to 0.07, and $r = 0.16$ to 0.08, respectively) when the Ukiriguru + Kibaha + Uganda set was used to predict the Kibaha set. We observed similar decreases for CBSDRS, SHTWT, and FYLD (from $r = 0.23$ to 0.15, from $r = 0.20$ to 0.14, and from $r = 0.30$ to 0.26, respectively) when the Ukiriguru + Kibaha + Uganda set was used to predict the Ukiriguru set. Decreases in accuracy as result of combining unrelated populations have been reported in other crops. For example, Lorenz and Smith (2015) reported low prediction accuracy estimates when they combined a population from a North Dakota state University barley (*Hordeum vulgare* L.) program and a second population from a University of Minnesota barley breeding program to form a TP. Other researchers have

TABLE 4 Average trait prediction accuracy for cross-location and combined Tanzania Agriculture Research Institute (TARI) populations

Trait	TARI	Ukiriguru to Ukiriguru[a]	Kibaha to Ukiriguru	Kibaha to Kibaha	Ukiriguru to Kibaha
MCMDS[b]	0.15*	0.09	0.02ns	0.18*	0.06ns
MCBSDS	0.23*	0.25	0.11*	0.08ns	0.09ns
CBSDRS	0.09*	0.21	0.16*	0.05ns	0.10*
MCGMS	0.25*	0.22	0.23*	0.23*	0.25*
RTNO	0.16*	0.24	0.08ns	0.09ns	0.11*
SHTWT	0.11ns	0.22	0.19ns	0.09ns	0.01ns
HI	0.16ns	0.18	0.15*	0.11*	0.12*
FYLD	0.18*	0.29	0.13*	0.08ns	0.05ns
DM	0.23*	0.13	0.16*	0.28*	0.24*
Mean	0.17	0.20	0.14	0.13	0.10

[a] Ukiriguru to Ukiriguru prediction accuracy indicates the accuracy of using the Ukiriguru training (sub)populations to predict the Ukiriguru test sets; similar notation is used for the other predictions.

[b] MCMDS, mean cassava mosaic disease severity; MCBSDS, mean cassava brown streak disease severity; CBSDRS, cassava brown streak disease root necrosis severity; MCGMS, mean cassava green mite severity; RTNO, root number; SHTWT, shoot weight; HI, harvest index; FYLD, fresh root yield; DM, dry matter.

* Significant at the 0.05 probability level; ns, nonsignificant prediction accuracy.

TABLE 5 Summary of cross-validated prediction accuracy by trait and across programs

Scenario	MCMDS[a]	MCBSDS	CBSDRS	RTNO	SHTWT	HI	FYLD	DM	Mean
Kibaha to Kibaha[b]	0.19ns	0.16ns	0.14ns	0.16ns	0.15ns	0.12ns	0.09ns	0.29*	0.16
Ukiriguru to Ukiriguru	0.01ns	0.28*	0.23*	0.26*	0.20*	0.17ns	0.30*	0.13ns	0.20
Kibaha + Ukiriguru to Kibaha	0.17ns	0.14ns	0.14ns	0.16ns	0.14ns	0.14ns	0.09ns	0.30*	0.16
Ukiriguru + Kibaha to Ukiriguru	0.01ns	0.28*	0.18ns	0.27*	0.14ns	0.19*	0.08ns	0.15ns	0.16
Kibaha + Uganda to Kibaha	0.20*	0.07ns	0.06ns	0.14ns	0.08ns	0.15ns	0.08ns	0.27*	0.13
Ukiriguru + Uganda to Ukiriguru	0.02ns	0.30*	0.15ns	0.24*	0.14ns	0.19*	0.26*	0.15ns	0.18
Kibaha + Ukiriguru + Uganda to Kibaha	0.17ns	0.11ns	0.06ns	0.18ns	0.10ns	0.15ns	0.08ns	0.28*	0.14
Ukiriguru + Kibaha + Uganda to Ukiriguru	−0.01ns	0.31*	0.17*	0.26*	0.12ns	0.20*	0.11ns	0.17ns	0.17
Mean	0.09	0.21	0.14	0.21	0.13	0.16	0.14	0.22	0.16

[a] MCMDS, mean cassava mosaic disease severity; MCBSDS, mean cassava brown streak disease severity; CBSDRS, cassava brown streak disease root necrosis severity; MCGMS, mean cassava green mite severity; RTNO, root number; SHTWT, shoot weight; HI, harvest index; FYLD, fresh root yield; DM, dry matter.

[b] Kibaha to Kibaha prediction accuracy indicates the accuracy of the Kibaha training (sub)populations to predict the Kibaha test sets; similar notation is used for the other predictions.

also reported low accuracies in cross-population predictions (Crossa et al., 2010; Endelman, 2011; Wolfe et al., 2017). Our results and evidence from other studies suggests that breeders can achieve better and more reliable prediction accuracy estimates with smaller populations with closely related genotypes than a large population with unrelated individuals. Population structure, different environmental conditions, different experimental designs, the direction of trait selection in each of the TARI breeding programs, and variations in heritability could have impacted the accuracy estimates. According to published results, the Ugandan TP had slightly higher accuracy for most traits than the TARI sets (Ozimati et al., 2018; Wolfe et al., 2017). More locations and replication of clones across environments could have improved their accuracies. The poor prediction results for cross-program prediction reported in our study will make it harder for breeders to use training data from different locations, breeding programs, and countries.

5.4 | Genetic architecture of disease resistance

Genetic associations with MCMDS were distributed across all chromosomes except 4, 5 6, 8, 11, 13, and 18 (Figure 3, Table 6, and Supplemental File S1). A previous GWAS for CMD resistance revealed the CMD2 locus on chromosome 12 in other cassava populations (Wolfe et al., 2016). The SNPs S12_5529819 and S12_11351823 delimit this QTL region (6.3–8.7 Mbp). S12_7270219 was the most significant marker tagging the QTL in all the TARI subpopulations and accounted for 14 to 37% of the phenotypic variance explained (PVE). This result validates the presence of the previously reported CMD2 locus. In addition to the CMD2 locus, we also detected another major CMD QTL (*QTL-cmds16-1)* on chromosome 16 tagged by the marker S16_421670 when we used the PYTUKG population. This region accounted for 87% of PVE (Figure 3, Table 6, and Supplemental File S1). There

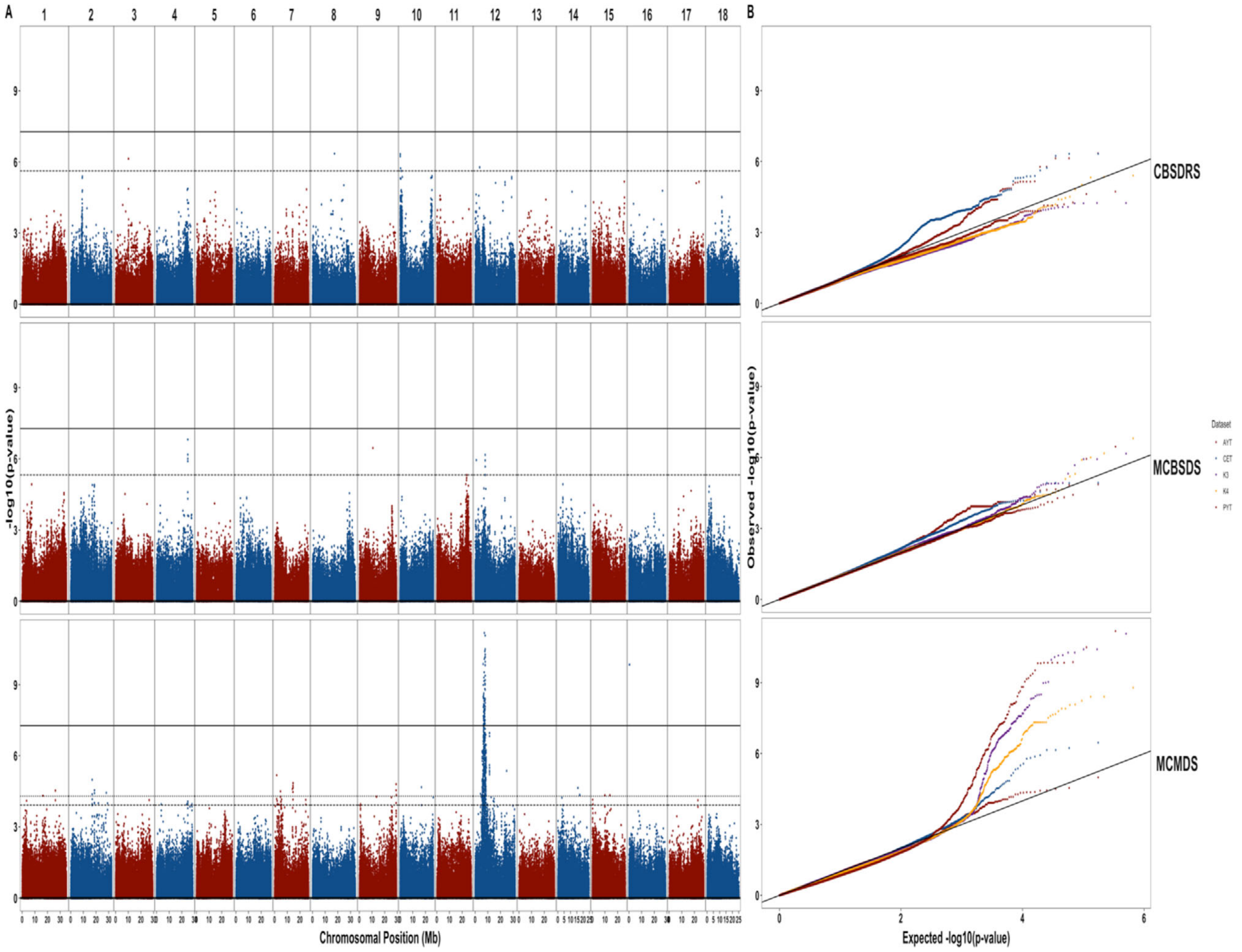

FIGURE 3 The Manhattan (a) and quantile–quantile (QQ) (b) plots from mixed linear models summarizing the genome-wide association results for three significant traits in all subsets of the combined Tanzania Agriculture Research Institute (TARI) population. The quantile–quantile plots demonstrate the differences among various population structure controls. CET, clonal evaluation trial; PYT, preliminary yield trial; AYT, advanced yield trial; K3 & K4, clusters 3 and 4 generated by the kinship relationship matrix. The horizontal solid line indicates the genome-wide Bonferroni significance threshold ($-\log_{10}$ (P-value) $= 7.27$), the dashed line indicates the false discovery rate (FDR) Part1 and the dotted line indicates FDR Part 2. Additional details of the significant markers, P-values, and positive FDR values are given in Supplemental File S1 (Sheet 1). Additional plots for all the other traits below the threshold are in Supplemental Figure S3

are no known CMD resistance genes in this region. However, the genes *Manes.16G036200.1*, *Manes.16G036300.1*, and *Manes.16G036400.1* have been annotated in this region and are known to encode pentatricopeptide repeats. These genes are positively expressed when plants are under attack from pathogens (Park et al., 2014).

A QTL on chromosome 4 (*QTL-cbsd4|cmd-1)* and two on chromosome 12 (*QTL-cbsd12|cmd-1* and *QTL-cbsd12|cmd-2)* showed significant associations with both MCBSDS and MCMDS resistance (Figure 3, Table 6, and Supplemental File S1). The marker S12_7929439 tagging the *cbsd12|cmd-2* QTL accounted for 8% of PVE for MCMDS and 33% of MCBSDS in several subsets of the TARI population. In the same region, a single marker (S12_929320) was found to be associated with a MCMDS and MCBSDS resistance QTL (*QTL-cbsd12|cmd-1)* in the PYTKIB dataset and accounted for 37 and 8% and PVE, respectively. *QTL-cbsd4|cmd-1*, tagged by marker S4_24670203, was detected in the K4_cluster1 dataset and explained 14% of PVE. The *Manes.04G113900.1* gene occurs in the marker region detected on chromosome 4. This gene is known to encode for phospholipased α and activates plant responses to pathogen attacks (De Torres et al., 2002). There are no known annotated genes in the region of *QTL-cbsd12|cmd-1* and *QTL-cbsd12|cmd-2.* Further studies need to be conducted to confirm the presence of QTL conferring resistance to both CMD and CBSD on chromosomes 4 and 12.

On chromosomes 9 and 11, *QTL-cbsd9-1* and *QTL-cbsd11-2*, tagged by SNP S9_10707044 and S11_22942418, were associated with responses to MCBSDS in the Ukiriguru

TABLE 6 Significant markers associated with mean cassava mosaic disease severity (MCMDS), MCBSDS (mean cassava brown streak disease severity), and cassava brown streak disease root necrosis severity (CBSDRS) resistance detected in the Tanzania Agriculture Research Institute (TARI) training population

			Region						Log_{10}	PVE	
QTL	Trait	Chr.	Mbs	Tag SNP	Position	Allele	Freq	b	*P*-value	%	Population
QTL-cbsdrs2-1	CBSDRS	2	9.15–9.16	S2_9258334	9,258,334	A/C	0.15	0.11	6.843	8	Ukiriguru
QTL-cbsdrs3-1	CBSDRS	3	10.16	S3_10165675	10,165,675	A/G	0.05	0.20	6.137	3	AYTKIB
QTL-cbsdrs8-1	CBSDRS	8	1.74	S8_17363721	17,363,721	A/G	0.05	0.21	6.35	32	AYTKIB
QTL-cbsdrs10-1	CBSDRS	10	0.16	S10_160775	160,775	G/A	0.08	0.16	5.24	11	CETUKG
QTL-cbsd9-1	MCBSDS	9	10.71	S9_10707044	10,707,044	G/A	0.5	0.57	6.457	9	PYTKIB
QTL-cbsd11-1	MCBSDS	11	22.88–22.94	S11_22942418	22,942,418	A/G	0.45	0.08	6.01	5	Ukiriguru
QTL-cmds16-1	MCMDS	16	0.42	S16_421670	421,670	T/C	0.49	1.86	9.844	87	PYTUKG
QTL-cmds12-1	MCMDS	12	6.3–8.7	S12_7270219	7,270,219	T/C	0.49	0.24	11.246	14 - 37	TARI
QTL-cbsd4/cmd-1	MCBSDS, MCMDS	4	24.67	S4_24670203	24,670,203	T/G	0.24	0.14	5.441	14	K4_Cluster1
QTL-cbsd12/cmd-1	MBSDS, MCMDS	12	0.93	S12_929320	929,320	T/G	0.06	0.13	5.941	37, 8	PYTKIB
QTL-cbsd12/cmd-2	MBSDS, MCMDS	12	7.93–7.95	S12_7929439	7,929,439	G/C	0.37	0.23	8.749	8, 33	Several

Note: Chr, chromosome; SNP, single nucleotide polymorphism; PVE, phenotypic variance explained; AYTKIB, advanced yield trial at Kibaha; CETUKG, clonal evaluation trial at Ukiriguru; PYTKIB, preliminary yield trial at Kibaha; PYTUKG, preliminary yield trial at Ukiriguru.

and PYTKIB subpopulations, respectively. *QTL-cbsd9-1* accounted for 9% of PVE, whereas *QTL-cbsd11-2* accounted for 5% of PVE. The annotated gene within the *QTL-cbsd11-2* is *Manes.11G120800.1*. This gene is known to encode for a protein kinase. Overexpression of a similar gene in tobacco (*Nicotiana tabacum* L.) stimulated plant defense responses (Li et al., 2018). Similarly, the region on chromosome 9, with a significant marker S9_10707044, contains a single gene (*Manes.09G074200.1*) that encodes a kinase family protein. Song et al. (1995) reported that a receptor kinase-like protein is encoded by the rice disease resistance gene *Xa21*. The two genes on chromosomes 9 and 11 could play a role in cassava's defense mechanism to confer CBSD resistance. However, further investigation to validate the presence of CBSD resistance is needed.

Nine markers representing four loci on chromosomes 2, 3, 8, and 10 were significantly associated with CBSDRS responses (Table 6, Supplemental File S1). Two loci on chromosomes 3 (*QTL-cbsdrs3-1)* and 8 (*QTL-cbsdrs8-1)*, which occurred in AYTKIB accessions, accounted for 30 and 32% of the phenotypic variation (chromosomes 3 and 8, respectively). Although the other two minor loci on chromosomes 2 (*QTL-cbsdrs2-1)* and 10 (*QTL-cbsdrs10-1)*, which were detected in the Ukiriguru accessions, explained 8 and 11% of PVE (chromosomes 2 and 10, respectively). The *Manes.08G079900.1* gene within the QTL region on chromosome 8 is known to express a wall-associated kinase-like receptor. Shi et al. (2016) cloned *Snn1*, which is a member of the wall-associated kinase class receptors and found that these receptors drive pathways for biotrophic pathogen resistance in wheat (*Triticum aestivum* L.). No annotated genes within the region were significant for resistance to CBSDRS on chromosome 3 are in the cassava reference genome.

In conclusion, discovery of new loci and associated markers will facilitate early selection during the season so breeders can have adequate information early enough to make decisions. Although CBSD resistance genes in some Ugandan accessions are thought to have originated from Tanzania during germplasm exchange, the MCBSDS genes discovered in this study are not localized on chromosomes 5, 11, and 18, as reported for Ugandan germplasm (Kayondo et al., 2018). In this study, we observed that the use of fewer but closely related individuals, particularly from the same clusters, improves the discovery of QTL compared with the use of a large number of individuals. These results are similar to the GWAS findings of Bradbury, Parker, Hamblin, and Jannink (2011) who reported that accounting for individual relatedness in barley improved the detection of true QTL.

6 | CONCLUSIONS

Genomic prediction and selection have been touted as tools that could greatly modernize plant breeding and accelerate genetic gain. In this study, we examined the power of diverse breeding lines assembled from two breeding programs, at different breeding stages, to predict traits and discover QTL. Differentiation of the TARI population could have resulted from existing restrictions on clonal movement between programs. This restriction was imposed to contain the spread of cassava foliar diseases between the Lake and Coastal Zones (Legg & Thresh, 2000). Although, there is a long tradition of germplasm sharing between Tanzania and Uganda, the introgression occurring in the Ugandan TP is large and is restricted to chromosome 1, whereas those in Tanzania are spread across all 18 chromosomes. This could suggest that the Ugandan introgression could be more recent than that in Tanzania. This may explain the population differentiation between subpopulations with the introgression versus those without.

There was no relationship between increased plot size and decreased error variance across all traits. For some traits, larger plots were preferable, though for other traits, smaller plots were. However, this conclusion needs more proven evidence to verify the results. If this finding is confirmed, then breeders need to reconsider whether utilizing larger plots is cost-effective.

An inverse relationship between heritability estimates and trait prediction accuracy were observed for some traits, contrary to other plant studies (Combs & Bernardo, 2013; Lian, Jacobson, Zhong, & Bernardo, 2014). We are not certain whether this inverse relationship is caused by noise in the data or is true; therefore, further assessments are needed to determine this relationship.

Prediction accuracies within and between locations (Kibaha and Ukiriguru) were generally lower than other cassava populations. Although larger TP sizes have been associated with improved accuracies, in this study, adding clones from Kibaha to those from Ukiriguru and vice versa did not improve the prediction accuracy of either population. Similarly, adding the Ugandan clones to either the Kibaha or Ukiriguru set did not improve the accuracy of either. The lack of relatedness between germplasm and population structure and the impact of the genotype × environment interaction negatively impacted accuracy estimates.

Generally, across breeding cycles, GS is more difficult in plants than in animals. This is because every year, plant breeders largely use new parents, some with an unknown background from other breeding programs or competitors, whereas animal breeders work in closed populations. This could make it challenging for breeders to keep materials adequately related in cross-cycle GS. The successful use of GS is dependent on a close relationship between individuals in the training and test sets. Clones at similar breeding stages were more valuable than a mixture of clones from different breeding stages when constructing TPs. This is because clones in the same generation are likely to share an ancestor a few generations back and therefore marker–QTL linkages are preserved because of the limited number of recombination events (Habier, Fernando, & Garrick, 2013). It is also possible that closely related population share more polymorphic loci and share large fraction of genetic background causing sufficient genetic variation (Lorenz & Cohen, 2012; Mohammadi, Tiede, & Smith, 2015).

Consistent with the findings of other researchers, we conclude that clones from the same breeding cycles are currently better option as candidates for GS TPs (Cericola et al., 2017; Michel et al., 2018; Song et al., 2017). In addition, it may not be useful to constitute TPs from programs with divergent populations or populations separated by barriers like the existing restriction on clone movement between the two Tanzanian programs. Cross-cycle GS, especially in more advanced breeding stages, needs to be investigated further because the prediction accuracy was very low. The impact of genotype × environment modeling on unreplicated clones across environments needs to be investigated further. Moreover, clones in the early breeding stage provided more reliable trait prediction accuracy because of their inherent genetic variation. Therefore, these clones are better candidates for TP construction.

We identified accessions carrying MCMDS, MCBSDS, and CBSDRS resistance. Some of the loci identified in these accessions have been reported previously. However, other loci are new. These results will be valuable for cassava breeding against CMD, CBSD, and CBSDRS. Although we have learned valuable lessons from this study, we still need to continue to improve our experimental designs, data capture, and construction of TPs so that genomic prediction and accuracy will be more reliable. We echo the lessons from Wolfe et al. (2017) to continue to improve on data quality and the selection of individuals to make TPs so that we can maximize genetic gains.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest

ACKNOWLEDGMENTS

This work was supported by the Next Generation Cassava Breeding Project through funds from the Bill and Melinda. Gates Foundation and the Department for International Development of the United Kingdom. We thank the technical field staff of the TARI and National Crops Resources Research Institute cassava breeding programs, who helped in trial management and/or data collection (Nsajigwa Mwakyusa, Karoline Leonard Sichalwe, Joseph Orone, Charles Majara, Gerald Adiga, and Vincent Kyaligonza). Special thanks to Guillaume Bauchet for initial curation of the sequence data and the entire Next Generation Cassava Breeding team (www.nextgencassava.org) as a whole.

DATA AVAILABILITY

The phenotypic and genotypic data generated and analyzed during this study are available in the CASSAVABASE repository (https://www.cassavabase.org/)

ORCID

Mohamed Somo https://orcid.org/0000-0003-1068-3266
Marnin D. Wolfe https://orcid.org/0000-0002-5929-5785
Alfred Ozimati https://orcid.org/0000-0003-0503-8107
Jean-Luc Jannink
https://orcid.org/0000-0003-4849-628X

REFERENCES

Akdemir, D., & Godfrey, O. U. (2015). EMMREML: Fitting mixed models with known covariance structures. R package version 3.1. Retrieved from https://cran.r-project.org/web/packages/EMMREML/

Adeniji, O. T., Odo, P. E., & Ibrahim, B. (2011). Genetic relationships and selection indices for cassava root yield in Adamawa State, Nigeria. *African Journal of Agricultural Research*, *6*(13), 2931–2934. https://doi.org/10.5897/AJAR10.143

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Grothendieck, G. (2015). Package 'lme4'. *Convergence*, *12*:1.

Bhateria, S., Sood, S. P., & Pathania, A. (2006). Genetic analysis of quantitative traits across environments in linseed (*Linum usitatissimum* L.). *Euphytica*, *150*(1–2), 185–194.

Bradbury, P., Parker, T., Hamblin, M. T., & Jannink, J. L. (2011). Assessment of power and false discovery rate in genome-wide association studies using the BarleyCAP germplasm. *Crop Science*, *51*(1), 52–59. https://doi.org/10.2135/cropsci2010.02.0064

Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Gonzales, E. E., & Rokhsar, D. S. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology*, *34*(5), 562–570.

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, *84*(2), 210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

Ceballos, H., Kulakow, P., & Hershey, C. (2012). Cassava breeding: Current status, bottlenecks and the potential of biotechnology tools. *Tropical Plant Biology*, *5*(1), 73–87. https://doi.org/10.1007/s12042-012-9094-9

Ceballos, H., Pérez, J. C., Joaqui Barandica, O., Lenis, J. I., Morante, N., Calle, F., … Hershey, C. H. (2016). Cassava breeding I: the value of breeding value. *Frontiers in Plant Science*, *7*, 1227. https://doi.org/10.3389/fpls.2016.01227

Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., & Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. A case of study in advanced wheat breeding lines. *PLoS One*, *12*(1), 0169606. https://doi.org/10.1371/journal.pone.0169606

Chipeta, M. M., Bokosi, J. M., Saka, V. W., & Benesi, I. R. (2013). Combining ability and mode of gene action in cassava for resistance to cassava green mite and cassava mealy bug in Malawi. *Journal of Plant Breeding and Crop Science*, *5*(9), 195–202. https://doi.org/10.5897/JPBCS2013.0388

Combs, E., & Bernardo, R. (2013). Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *The Plant Genome*, *6*, 1–7. https://doi.org/10.3835/plantgenome2012.11.0030

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., … Braun, H. J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, *186*, 713–724. https://doi.org/10.1534/genetics.110.118521

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*(3), 1021–1031. https://doi.org/10.1534/genetics.110.116855

Dawson, J. C., Endelman, J. B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., … Jannink, J. L. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research*, *154*, 12–22. https://doi.org/10.1016/j.fcr.2013.07.020

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*(2), 327–345. https://doi.org/10.1534/genetics.112.143313

de Torres Zabela, M., Fernandez-Delmond, I., Niittyla, T., Sanchez, P., & Grant, M. (2002). Differential expression of genes encoding *Arabidopsis* phospholipases after challenge with virulent or avirulent *Pseudomonas* isolates. *Molecular Plant—Microbe Interactions*, *15*(8), 808–816. https://doi.org/10.1094/MPMI.2002.15.8.808

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, *6*(5), 19379. https://doi.org/10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, *4*(3), 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Fauquet, C., Fargette, D., & Munihor, C. (1990). African cassava mosaic virus: Etiology, epidemiology, and control. *Plant Disease*, *74*, 404–411.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS One*, *9*(2), e90346. https://doi.org/10.1371/journal.pone.0090346

Habier, D., Fernando, R. L., & Garrick, D. J. (2013). Genomic BLUP decoded: A look into the black box of genomic prediction. *Genetics*, *194*(3), 597–607. https://doi.org/10.1534/genetics.113.152207

Hahn, S., Terry, E., & Leuschner, K. (1980). Breeding cassava for resistance to cassava mosaic disease. *Euphytica*, *29*, 673–683.

Hahn, S., Terry, E., Leuschner, K., Akobundu, I., Okali, C., & Lal, R. (1979). Cassava improvement in Africa. *Field Crops Research*, *2*, 193–226. https://doi.org/10.1016/0378-4290(79)90024-8

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *92*(2), 433–443. https://doi.org/10.3168/jds.2008-1646

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*(1), 1–12. https://doi.org/10.2135/cropsci2008.08.0512

Hillocks, R. J., & Thresh, J. M. (2000). Cassava mosaic and cassava brown streak virus diseases in Africa: A comparative guide to symptoms and aetiologies. *Roots*, *7*(1), 1–8.

Hofheinz, N., Borchardt, D., Weissleder, K., & Frisch, M. (2012). Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theoretical and Applied Genetics*, *125*(8), 1639–1645. https://doi.org/10.1007/s00122-012-1940-5

Howeler, R., Lutaladio, N., & Thomas, G. (2013). *Save and grow: Cassava. A guide to sustainable production intensification*. Rome: Food and Agriculture Organization of the United Nations.

Jannink, J. L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, *9*(2), 166–177. https://doi.org/10.1093/bfgp/elq001

Jeremiah, S. C., Ndyetabula, I. L., Mkamilo, G. S., Haji, S., Muhanna, M. M., Chuwa, C., … Legg, J. P. (2015). The dynamics and environmental influence on interactions between cassava brown streak disease and the whitefly, *Bemisia tabaci*. *Phytopathology*, *105*(5), 646–655. https://doi.org/10.1094/PHYTO-05-14-0146-R

Kassambara, A. (2018). ggpubr: 'Ggplot2' based publication ready plots. R package version 0.2. Retrieved from https://CRAN.R-project.org/package=ggpubr

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

Kawano, K., Daza, P., Amaya, A., Rios, M., & Goncalves, W. M. (1978). Evaluation of cassava germplasm for productivity 1. *Crop Science*, *18*(3), 377–380. https://doi.org/10.2135/cropsci1978.0011183X001800030006x

Kayondo, S. I., Del Carpio, D. P., Lozano, R., Ozimati, A., Wolfe, M., Baguma, Y., … Jannink, J. L. (2018). Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-19696-1

Kundy, A. C., Mkamilo, G. S., & Misangu, R. N. (2014). Correlation and path analysis between yield and yield components in Cassava (*Manihot esculenta* Crantz) in Southern Tanzania. *Journal of Natural Sciences Research*, *4*, 6–10.

Legg, J.,& Raya, M. D. (1998). Survey of cassava virus diseases in Tanzania. *International Journal of Pest Management 44*(1), 17–23. https://doi.org/10.1080/096708798228473.

Legg, J. P., & Thresh, J. M. (2000). Cassava mosaic virus disease in East Africa: A dynamic disease in a changing environment. *Virus Research*, *71*(1-2), 135–149. https://doi.org/10.1016/S0168-1702(00)00194-5

Li, W., Li, X., Chao, J., Zhang, Z., Wang, W., & Guo, Y. (2018). NAC family transcription factors in tobacco and their potential role in regulating leaf senescence. *Frontiers in Plant Science*, *21*(9), 1900. https://doi.org/10.3389/fpls.2018.01900

Lian, L., Jacobson, A., Zhong, S., & Bernardo, R. (2014). Genomewide prediction accuracy within 969 maize biparental populations. *Crop Science*, *54*(4), 1514–1522. https://doi.org/10.2135/cropsci2013.12.0856

Lorenzana, R. E., & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, *120*(1), 151–161. https://doi.org/10.1007/s00122-009-1166-3

Lorenz, A. J., & Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Science*, *55*(6), 2657–2667. https://doi.org/10.2135/cropsci2014.12.0827

Lorenz, K., & Cohen, B. A. (2012). Small-and large-effect quantitative trait locus interactions underlie variation in yeast sporulation efficiency. *Genetics*, *192*(3), 1123–1132. https://doi.org/10.1534/genetics.112.143107

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H. G., … Jannink, J. L. (2013). Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: A study in cassava. *Crop Science*, *53*(4), 1312–1325. https://doi.org/10.2135/cropsci2012.11.0653

Michel, S., Kummer, C., Gallee, M., Hellinger, J., Ametz, C., Akgol, B., … Buerstmayr, H. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theoretical and Applied Genetics*, *131*(2), 477–493. https://doi.org/10.1007/s00122-017-2998-x

Milligan, S. B., Balzarini, M., Gravois, K. A., & Bischoff, K. P. (2007). Early stage sugarcane selection using different plot sizes. *Crop Science*, *47*(5), 1859–1864. https://doi.org/10.2135/cropsci2006.12.0822

Mohammadi, M., Tiede, T., & Smith, K. P. (2015). PopVar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. *Crop Science*, *55*(5), 2068–2077. https://doi.org/10.2135/cropsci2015.01.0030

Ozimati, A., Kawuki, R., Esuma, W., Kayondo, S. I., Pariyo, A., Wolfe, M., … Jannink, J. L. (2018). Training population optimization for prediction of cassava brown streak disease resistance in West African clones. *G3: Genes, Genomes, Genetics*, *8*(12), 3903–3913. http://doi.org/10.1534/g3.118.200710

Ozimati, A., Kawuki, R., Esuma, W., Kayondo, S. I., Pariyo, A., Wolfe, M., & Jannink, J. L. (2019). Genetic variation and trait correlations in an East African cassava breeding population for genomic selection. *Crop Science*, *59*(2), 460–473. https://doi.org/10.2135/cropsci2018.01.0060

Park, Y. J., Lee, H. J., Kwak, K. J., Lee, K., Hong, S. W., & Kang, H. (2014). MicroRNA400-guided cleavage of pentatricopeptide repeat protein mRNAs renders *Arabidopsis thaliana* more susceptible to pathogenic bacteria and fungi. *Plant and Cell Physiology*, *55*(9), 1660–1668. https://doi.org/10.1093/pcp/pcu096

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., & Sorrells, M. E. (2015). Efficient use of historical data for genomic selection: A case study of stem rust resistance in wheat. *The Plant Genome*, *8*:1–10. https://doi.org/10.3835/plantgenome2014.09.0046

Rwegasira, G. M., & Rey, C. M. (2012). Response of selected cassava varieties to the incidence and severity of cassava brown streak disease in Tanzania. *The Journal of Agricultural Science*, *4*(7), 237–245. https://doi.org/10.5539/jas.v4n7p237

Sallam, A. H., Endelman, J. B., Jannink, J. L., & Smith, K. P. (2015). Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *The Plant Genome*, *8*(1), 1–15. https://doi.org/10.3835/plantgenome2014.05.0020

Salvador, E. M., Steenkamp, V., and McCrindle, C. M. E., & Ethelwyn, C.M. (2014). Production, consumption and nutritional value of cassava (*Manihot esculenta*, Crantz) in Mozambique: An overview. *Journal of Agricultural Biotechnology and Sustainable Development*, *6*(3), 29–38.

Shi, G., Zhang, Z., Friesen, T. L., Raats, D., Fahima, T., Brueggeman, R. S., … Frenkel, Z. (2016). The hijacking of a receptor kinase-driven pathway by a wheat fungal pathogen leads to disease. *Science Advances*, *2*(10), e1600822. https://doi.org/10.1126/sciadv.1600822

Song, J., Carver, B. F., Powers, C., Yan, L., Klápště, J., El-Kassaby, Y. A., & Chen, C. (2017). Practical application of genomic selection in a doubled-haploid winter wheat breeding program. *Molecular Breeding*, *37*(10), 117. https://doi.org/10.1007/s11032-017-0715-8

Song, W.-Y., Wang, G.-L., Chen, L.-L., Kim, H.-S., Pi, L.-Y., Holsten, T., … Roland, P. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science*, *270*(5243), 1804–1806. https://doi.org/10.1126/science.270.5243.1804

Storlie, E., & Charmet, G. (2013). Genomic selection accuracy using historical data generated in a wheat breeding program. *The Plant Genome*, *6*(1), 9. https://doi.org/10.3835/plantgenome2013.01.0001

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vazquez, A. I., Bates, D. M., Rosa, G. J. M., Gianola, D., & Weigel, K.A. (2010). An R package for fitting generalized linear mixed models in animal breeding 1. *Journal of Animal Science*, *88*(2), 497–504. https://doi.org/10.2527/jas.2009-1952

Wei, T. (2013). Corrplot: Visualization of correlation matrix. R package version 0.2. Retrieved from https://cran.r-project.org/web/packages/corrplot/

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics, version 2.1. Retrieved from https://cran.r-project.org/web/packages/ggplot2/index.html

Wolfe, M. D., Del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., … Jannink, J.-L. (2017). Prospects for genomic selection in cassava breeding. *Plant Genome*, *10*(3), 19. https://doi.org/10.3835/plantgenome2017.03.0015.

Wolfe, M. D., Rabbi, I. Y., Egesi, C., Hamblin, M., Kawuki, R., Kulakow, P., … Jannink, J. L. (2016). 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. https://doi.org/10.3835/plantgenome2015.11.0118

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zhong, S., Dekkers, J. C., Fernando, R. L., & Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics*, *182*(1), 355–364. https://doi.org/10.1534/genetics.108.098277

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Somo M, Kulembeka H, Mtunda K, et al. Genomic prediction and quantitative trait locus discovery in a cassava training population constructed from multiple breeding stages. *Crop Science*. 2020;60:896–913. https://doi.org/10.1002/csc2.20003