Running Title

Candidates for training populations and implication on accuracy

Title

Genomic prediction and QTL discovery in a cassava training population constructed from multiple breeding stages

Authors

Mohamed Somo1*
Heneriko Kulembeka2
Kiddo Mtunda2
Emmanuel Mrema2
Kasele Salum2
Marnin D. Wolfe1
Ismail Y. Rabbi3
Chiedozie Egesi3
Robert Kawuki4

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/csc2.20003.

This article is protected by copyright. All rights reserved.
Affiliations

(1) Dept. 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

Abbreviations

Please list abbreviations in alphabetical order with the abbreviation first, separated from its definition by a comma. Please use semicolons to distinguish separate abbreviations.

BLUPs, Best Linear Unbiased Predictors

GBS, Genotyping-by-Sequenceing

GEBVs, Genomic Estimated Breeding Values

GS, Genomic Selection
MAF, Minor Allele Frequency

MAP, Months After Planting

PCA, Principal Component Analysis

SNP, Single Nucleotide Polymorphism

TP, Training Population
Abstract

Assembly of training population (TP) is an important component of effective genomic selection (GS) based breeding program. In this study, we examined the power of diverse germplasm assembled from two breeding programs in Tanzania, at different breeding stages to predict traits and discover QTL. This is the first GS and Genome-wide association study (GWAS) analysis on Tanzanian cassava data. We detected QTL associated with resistance to Cassava mosaic disease (CMD) 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 (CBSDR). We also detected a QTL on chromosome 4 and two QTL on chromosome 12 conferring dual resistance to CMD and CBSD. Using clones in the same breeding stage to construct TP provided higher trait prediction accuracies than using TP with mixture of clones from multiple breeding stages. Moreover, clones in the early breeding stage provided more reliable trait prediction accuracies and therefore are better candidates for constructing TP. Although larger training population sizes have been associated with improved accuracies, in this study, adding clones from Kibaha location to Ukiriguru and vice versa did not improve prediction accuracy of either population. Similarly including Uganda TP in either population did not improve trait prediction accuracies. This study applied genomic prediction to understand the implication of constituting TP using clones at different breeding stages and pooled from different locations on trait accuracies in cassava.
Introduction

Cassava is an important source of dietary calories for millions of people in tropical regions (Howeler et al., 2013; Salvador et al., 2014). Because of cassava’s adaptation and resiliency 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 et al., 2012). Similarly, the use of marker assisted selection (MAS) on complex traits in cassava has remained largely ineffective because of the multiple minor loci which are difficult to detect and deploy (Ceballos et al., 2012).

Genomic selection (GS) allows superior plants to be selected earlier on at seedling stage before phenotyping is started, and then used for crossing (Heffner et al., 2009; Lorenzana and Bernardo, 2009; Jannink et al., 2010). In principle, 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 thus can immediately be advanced to downstream evaluation for faster variety replacement. Recent implementation of GS in three breeding programs in Africa have improved trait prediction accuracies. Wolfe et al. (2017) reported a prediction accuracy increase of 57% for CMD in Nigerian cassava population. Similarly, Kayondo et al. (2017) reported improved prediction accuracy of 0.42 for both CBSD foliar severity and root severity in two Ugandan cassava populations.
In principle, successful implementation of GS program hinges on several critical factors, one of which is how TP is constituted. Mindful of breeding objectives, breeders can select individuals from available diverse germplasm, improved breeding lines, within and across breeding stages to construct TPs. They could also compose them using 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 data (Dawson et al., 2013; Storlie and Charmet. 2013; Ly et al., 2013; Rutkoski et al., 2015).

Often breeding programs transitioning to GS-based approach may not have access to sufficient and diverse genetic pool and adequate historical data to choose TP candidates from. For the case of clonal crops, the inadequacy of planting materials of 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 accuracies (Zhong et al., 2009; Cericola et al., 2017).

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

The TARI TPs were constructed using clones from different trial types. The KIB trials included clones at preliminary yield trial (PYT) and advanced yield trial (AYT) selection stages while those at UKG consisted of clones at clonal evaluation trial (CET) and PYT. In a breeding pipeline, CET genotypes are minimally selected in seedling nurseries before cloning, whereas PYT and AYT would have undergone one 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 different number of plants per plot. Additionally, there were no common checks for the trials across the two programs. Because there were few clones in each trial to form sizeable TP, clones from different trial types, breeding stages, and locations were pooled together to form TPs. The Uganda (UG) training population was added to the clones from the TARI two programs to evaluate whether their inclusion could improve trait prediction accuracies. There is limited knowledge on the implication of using cassava clones from multiple trials, multiple locations, and clones from multiple breeding stages to construct TP on trait predictions. This study was therefore designed to: (1) examine which combination of breeding materials could provide the best training set, (2) evaluate trait prediction accuracies for within and across breeding stages, locations, and combined TARI data sets, (3) assess the impact of adding Ugandan clones to clones of either program on trait predictions, and (4) identify loci associated with disease resistance particularly for CMD, CBSD, and CBSDRS 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 valuable addition for MAS breeding in cassava.
Materials And Methods

Germplasm and field design

A diverse panel of breeding lines from two independent breeding programs were used to construct the training population. The panel comprised of 432 and 408 clones from Kibaha and Ukiriguru breeding programs, respectively. Kibaha clones were at PYT (PYTKIB) and AYT (AYTKIB) breeding stages while those from Ukiriguru were at CET (CETUKG) and PYT (PYTUKG). The KIB clones were drawn from nine PYTs and six AYTs while UKG clones consisted of two CETs and a PYT. Clones in each trial type arose from different progenitors and as such each trial in both programs had a unique set of individuals, varied number of replicates, and different plot sizes (Table 1). An additional 402 clones from Uganda (UG) training population 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 (lat/Ing: 6.8138° S, 38.6949° E, Alt: 156 m above sea level) and Ukiriguru (lat/Ing: 2° 43' 0" S, 33° 1' 0" E, Alt: 1194 m above sea level). Kibaha PYT trials were laid as a 2-row plot with 14 plants while AYT trials were 42 plants planted in 6-row plots. The CET plots at UKG consisted of a single-row of 5 plants while the PYT were 2 row plots with 10 plants. At both locations, the clones were spaced at 1 m x 1m. With exception of two AYT trials which were replicated three times, the rest of the trials in both locations were replicated twice (Table 1). For all KIB trials, Kiroba, Mfaransa, and Mkuranga1 were used as common checks while Mkombozi, Lwakitangaza, and Liongokwimba were the checks in UKG 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.
**Phenotyping**

Nine traits were evaluated for this study. Cassava mosaic disease severity (CMDS), foliar cassava brown streak disease severity (CBSDS), and cassava green mite severity (CGMS) were scored at 3, 6, and 9 months after planting (MAP) on a scale of 1 (No symptoms) to 5 (severe symptoms). For analysis we used the season wide mean severity (MCMDS, MCBSDS, and MCGMS) for 3, 6, and 9 MAP. For root necrosis severity (CBSDRS), necrotic symptoms on a scale of 1-5 were scored 12 MAP using cross sections of root (Hillocks and Thresh, 2000), where 1 = no necrosis and 5 is > 25% necrotic and severe root constriction. We then used average disease scores for each experimental plot for analysis.

The root number (RTNO) are the number of fresh roots harvested per plot and used to calculate RTNO per hectare. Root weight (RTWT) and shoot weight (SHTWT) were both measured in kg per plot and then used to calculate below and above ground yield in tons per hectare, respectively. Harvest index (HI) was the ratio between RTWT and total biomass (RTWT+SHTWT) per unit area.

Dry matter (DM) was expressed as a percentage of fresh root (FYLD) weight using the specific gravity method. The specific gravity of each sample (of the whole root, including the peel) was determined from the weights in air and in water (Kawano et al., 1987). The experimentation and phenotypic data capture of UG population is as described previously (Wolfe et al., 2016; Ozimati et al. 2018; Kayondo et al., 2018). Eight variables were common to both UG and TARI populations except MCGMS which is specific to TARI.

**Genotyping**

SNP marker genotypes were obtained using genotyping-by-sequencing (GBS; Elshire et al., 2011). The markers were called using the TASSEL 5.0 GBS pipeline version 2 (Glaubitz et al., 2014) after aligning the resulting reads to the *Manihot esculenta* Version 6 assembly available at

This article is protected by copyright. All rights reserved.
http://phytozome.jgi.doe.gov/. 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 GBS 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 and 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 (MAF) 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 MAF threshold higher than at 0.05 and used for GWAS analysis. 37,776 markers common to both UG and TARI populations were filtered for minor allele frequency (0.01) and the resulting 36,847 markers were then used for downstream joint analysis.

Statistical analysis

Estimation of observed values

To estimate these observed values, we fitted a mixed linear model across trial types and locations using lme4, R package (R Development Core Team 2010; Vazquez et al., 2010; Bates et al., 2015). For trials within locations, we fitted the model: \( y = X\beta + Z_{\text{clone}}u + Z_{\text{Trial (Trial: Rep)}}r + \epsilon \), where \( y \) is raw phenotypic observations, \( \beta \) included a fixed effect for the population mean and for plot-basis traits (FYLD, RTNO, and SHTWT), the proportion of plants harvested per plot was included as a covariate; vector \( u \) and corresponding incidence matrix \( Z_{\text{clone}} \) represented a random effect for clone
where $u \sim N(0, \sigma_u^2)$ and $I$ represented the identity matrix. The incidence of replication was nested in trial and was represented by matrix $Z_{\text{Trial}}(\text{Trial.Rep})$ and random effects vector such that $r \sim N(0, \sigma_r^2)$ and residuals $\epsilon$ were considered to be distributed as $\epsilon \sim N(0, \sigma_\epsilon^2)$.

The model for combined analysis across locations was: $y = X\beta + Z_{\text{clone}}u + Z_{\text{rep}}(\text{Loc.Trial.Rep})b + \epsilon$. The fixed and clone effects for combined across location were the same as those for each location described above. Trial and rep effects were nested in locations and incorporated as random with incidence matrix $Z_{\text{rep}}$ and the effects vector $b \sim N(0, \sigma_b^2)$. All the random effects in each of the models are independent and identically distributed (iid). For both models described above, the best linear unbiased predictors (BLUPs) for clones were extracted and used as the “observed values” that were correlated with genomic estimated breeding values (GEBVs) to determine trait prediction accuracies.

Under combined location model, our emphasis was on prediction of new genetic materials from another breeding program, and therefore we considered the environment effects to be nuisance parameters. Since we were not interested in location, trial or rep effects per se, but instead interested in genomic estimated breeding values (GEBVs). We assumed the three-way interaction of location, trial and rep is sufficient to correct for these nuisance terms, and no main effects are required to be fit as they are simply linear combinations of the interaction predictors.

We also considered the issue of heterogeneity of error variances for different trials across the two locations. Although error variance estimates did differ 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 fresh yield (FYLD) which had reduced BLUP correlations (~0.48 data not shown). To overcome the challenges and complexity of modelling for heterogeneity of error variances particularly for yield, we did fit model for each trial type in each location separately and then extracted BLUPs from each experiment and then used combined BLUPs to correlate with genomic estimated breeding values to determine prediction accuracies. For the rest of the trait, because of the large number of trials (18), we felt that fitting an error variance for each trial
would result in too many parameters to estimate, and therefore we used a simpler homogenous error variance model.

**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 under “Estimation of observed values.” Accordingly, broad sense heritability ($H^2$) was then computed for each trial using

$$H^2 = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}$$

where $\sigma_c^2 =$ clone variance, and $\sigma_e^2 =$ residual variance.

The phenotypic and genotypic correlations between the nine traits were obtained using combined data from the two locations. We also assessed trait phenotypic correlations using data from each breeding stage in each program; UKG (CETUKG & PYTUKG), and KIB (PYTKIB & 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 designs. 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 (Fig. 2) in R 3.4.1 (R Core Team 2017) using the ‘corrplot’ package (Wei, 2013). For all the cases, correlation values were considered significantly different from zero at P value ≤ 0.05.

**Genomic Prediction**

To perform genomic prediction, we fitted separate genomic best linear unbiased prediction (GBLUP) model as $y = X\beta + Zg + \epsilon$ for within breeding stages, across breeding stages, and across breeding programs. where $y$ is a vector of the raw phenotype $\beta$ is a vector of fixed effect (different for different
scenarios) with design matrix $X$, the vector $g$ is the random effect representing GEBV for each individual, $Z$ is a design matrix linking observations to genomic values, and $\varepsilon$ is a vector of residuals. For within breeding stages, the fixed is the grand mean, while the fixed effects for across breeding stages were grand mean and trial. For across breeding programs the fixed effect included trial and location. The GEBV was obtained assuming $g \sim N(0, K \sigma_g^2)$, where $\sigma_g^2$ the additive genetic variance and $K$ 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) that used VanRaden’s (2008) method 1. The GBLUP predictions were made with the function $emmreml$ in the R 3.1.0 package EMMREML (Akdemir and Okeke, 2015).

**Assessment of Prediction Accuracy**

In order to assess prediction accuracy, we used 30 replicates of 5-fold 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 (training sets) of the five folds 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 therefore were not replicated. The following prediction scenarios were explored in our analysis: (1) within breeding stages in each location: CETUKG, PYTUKG, PYTKIB, and AYTIB, (2) combined locations (UKG+KIB), (3) within locations: prediction of UKG alone (CETUKG + PYTUKG) and KIB alone (PYTKIB + AYTIB), (4) inter-location prediction: UKG to predict KIB and KIB to predict UKG, and UKG+KIB to predict UKG and vice versa. We further used UKG+UG to predict UKG, KIB+UG to predict KIB and UKG+KIB+UG to predict UKG and KIB. This was done to determine if adding UG clones to either TARI program would improve trait prediction accuracies. For scenarios 3 and 4, we maintained a fixed test set (20%) picked randomly from the population under prediction in each iteration. For each prediction, accuracies were computed as the Pearson correlation coefficient between the GEBV predicted for the test set and the corresponding estimated observed breeding values; BLUPs. Correlation values were considered significantly different from zero at $P \leq 0.05$. 

This article is protected by copyright. All rights reserved.
Population structure

To visualize population structure, principal component analysis (PCA) using common markers in the three breeding programs; KIB, UG, and UKG was performed. The prcomp function in R was used to generate PCs. The first two PCs were then used for plotting to visualize structure between the three populations (KIB, UKG and UG). Thereafter the loadings (eigenvector coefficients) for all the markers on 18 chromosomes on PC1 and PC2 were assessed to determine the markers contributing to the greatest variations in TARI alone and combined TARI and UG in the respective PCs.

Genome-Wide Association Study

GWAS was performed on different subsets of TARI accessions. The subsets included trial types (CET, PYT, AYT), locations (KIB and UKG), and four clusters generated based on marker kinship matrices of the individuals. We used these sets of data combination because we suspected that certain chunks of data could provide better results than others. Clustering was considered because we thought it is an objective way to find population structure. By doing GWAS within each cluster we were hoping to avoid population structure and find better allele frequencies.

In implementing GWAS, we used BLUPs extracted from the linear mixed model described under estimated observed breeding values section as phenotypes. For each trial, the GWAS was carried out using genome-wide complex trait analysis tool (GTCA, Yang et al., 2011). This followed a leave one chromosome out approach in which the chromosome with the tested candidate SNP markers is excluded from the genomic relationship calculation. The linear mixed model: $y = X\beta + g + \varepsilon$ with $\text{var}(y) = V = K\sigma_g^2 + I\sigma_e^2$, were fitted for each case. Here $y$ is an $n \times 1$ vector of phenotype (BLUPs) with $n$ being the sample size, $\beta$ is a vector of fixed effect (genetic marker information), $g$ is an $n \times 1$ vector of the total genetic effects of the individuals with $g \sim N(0, K\sigma_g^2)$, $K$ is the genetic relationship matrix between individuals, same as symmetric genomic realized relationship matrix based on SNP markers, $I$ is an $n \times n$ identity matrix, and $\varepsilon$ is a vector of residual effects with $\varepsilon \sim N(0, I\sigma_e^2)$.
Bonferroni threshold was applied to correct for multiple testing for each data set used for GWAS. The cut off was computed as −log(α/t*n) where α is 0.05, which is the standard significance threshold, t is the number of different subset of data, and n is the number of SNPs. In this study, the Bonferroni correction significance cut off was −log10 (0.05/15*88,434) = 7.42. We used the marker-wised \( P \) value of 0.001 as the threshold to declare significant market trait association (MTA) and the false discovery rate (FDR) corrected \( P \) value of 0.2 to select the highly significant MTA. QTL boundaries were defined with linkage disequilibrium (LD) pairwise correlation coefficient \( r^2 \geq 0.7 \) coupled with the MTA information. We chose the tagging marker for each QTL with the criteria of the smallest marker-wised \( P \) value. Manhattan and QQplots plots were generated using R package ggplots2 (Wickham et al., 2016) with customization for joint Manhattan-QQplot display using R package ggpubr (Kassambara, 2018). We used the \( M. esculenta \_305 \_v6.1 \) available in Phytozome to report presence of annotated genes that are related to plant defense system within the QTL region.

Results and Discussion

Population Structure

We performed principal component analysis to assess structure within and among the UKG, KIB, and UG populations (Fig.1). This is necessarily 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 KIB and UKG might have facilitated population differentiation within Tanzanian germplasm. In the combined data (UKG+KIB+UG), the first and second PC accounted for 11% and 5% of the genetic variation, respectively. In TARI (KIB+UKG), PC1 and PC2 explained 7% and 5% of the variation, respectively. Some population stratification within each breeding program was observed (Fig. 1).

This article is protected by copyright. All rights reserved.
Three clusters of clones were evident arrayed along PC1 (Fig. 1b). The loadings of markers on PC1 (Fig. 1d) showed a strong effect of chromosome 1 on this component, suggesting that the clusters are due to the known introgression from *M. glaziovii* 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 0, 1, or 2 copies. Interestingly, PC2 (Supplementary Fig. 1) strongly separated UKG from KIB while the UG accessions were intermediate between these two. When analyzing TARI alone, differences between UKG and KIB aligned along PC1 (Fig. 1a) whereas the chromosome 1 introgression aligned along PC2 (Fig. 1c, Supplementary Fig. 1). 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, (Hahn et al., 1979, 1980; Fauquet et al., 1990), based on this analysis, this introgression may have been subjected to more recombination events in Tanzanian population (Fig 1:c &d).

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

**Trait Correlations**

This article is protected by copyright. All rights reserved.
Trait phenotypic and genetic correlations were quite similar (Fig. 2; Supplementary Fig. 2). 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, r=0.4, respectively). These values are only somewhat lower than previous findings; Kundy et al. (2014) reported a strong positive correlation between RTNO and FYLD (r=0.7), but fewer clones were tested which could have skewed their findings. In addition, 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 CBSD root necrosis which agrees with findings in other studies (Rwegasira and Rey, 2012; Ozimati et al., 2019). The positive relationship between these two foliar diseases provides an opportunity to increase resistance against both concurrently. MCGMS showed weak negative correlations with MCMDS, SHTWT, and FYLD (r=-0.3, -0.3, -0.4, respectively). The negative correlation between MCGMS and FYLD corroborates Chipeta et al., (2008) result (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 data sets from across breeding stages. However, we also noticed slight improvement in some trait although we think it could just be noise in the data (Supplementary Fig. 2).

Variance Components and Broad-Sense Heritability Estimates
Variance components and broad-sense heritability estimates for each of the nine traits are in Table 2. Generally, KIB trials yielded the largest genetic variance estimates for MCMDS, CBSDRS, FYLD, and SHTWT while UKG trials had the largest estimates for MCBSDS and RTNO. Conversely, both FYLD and SHTWT had the largest residual estimates in the AYTKIB and UKG trials, respectively. Errors associated with CBSDRS, MCGMS, RTNO, HI, and DM in AYTKIB were lower than other trials. Meanwhile both genetic and error estimates for RTNO in the UKG trials were higher than in the KIB trials. Additionally, HI and MCBSDS had similar error estimates across all the trials.

Heritability estimates ranged between 0.06 and 0.91. Based on cassava heritability classification, only HI in AYTKIB, MCBSDS in PYTUKG, and MCMDS in both AYTKIB and PYTUKG could be classified as high (>0.50) (Bhateria et al., 2006). Other researchers reported higher heritability for MCMDS, MCBSDS, and HI in different cassava populations (Wolfe et al., 2017; Kanyondo et al., 2018; Adeniji et al., 2011). High heritability estimates for MCMDS and MCBSDS could be associated with varying levels of resistance in Tanzanian accessions owing to specific responses to various strains of CBSVs and CGMs. In the Lake Zone, there is a high prevalence of Africa cassava mosaic virus, Eastern African cassava mosaic virus (EACMV), the Ugandan variant of cassava mosaic virus, cassava brown streak virus (CBSV), and the Ugandan variant cassava brown streak virus. However, only EACMV and CBSV are known to occur in the Eastern Zone; includes Kibaha (Raya et al., 1993; Jeremiah et al., 2015). Extensive pathogen variation within and between the two regions of Tanzania could have caused the variability detected in the germplasm. Significant differences in the HI were reported among cassava cultivars (Kawano et al., 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 this data that the clones from AYTKIB trials were heavily selected for low branching and consequently resulting in higher heritability (h²=0.80).

The CBSDRS, SHTWT, FYLD, and DM in the CETUKG trials and MCBSDS and DM in the KIB trials, all had medium heritability estimates. Similar estimates were reported in other cassava
populations (Wolfe et al., 2017; Kayondo et al., 2018; Ozimati et al., 2019). We observed low 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 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 breeding stages.

For most traits, CET trials with a smaller plot size (5m²) had lower error variance and a better heritability than trials with medium and large plots (e.g. 14m² and 20m²). High trait genetic variances were reported in sugarcane clonal evaluation trials at Louisiana State University Agricultural Center, but the heritability estimates in different plot sizes were similar (Milligan et al., 2007). The substantial increase in genetic variance and trait heritability in smaller plots than larger plot sizes needs further investigations. This will help establish optimal plot sizes for cassava breeding for better prediction accuracies. Generally, trait heritability estimates varied across trials similar to the findings of Wolfe et al. (2017), Kayondo et al. (2018), and Ozimati et al. (2019). Confounding effects (locations and trials), environmental condition differences, trait selection priority in each of the TARI breeding programs, and differences in data collection between the two programs could have affected heritability estimates.

Genomic prediction

Within Breeding Stage Prediction

Within breeding stage prediction accuracies for all the traits in CETUKG, PYTUKG, PYTKIB and AYTKIB were 0.23, 0.20, 0.15, and 0.20, respectively (Table 3). The accuracies within the breeding stages were lower than accuracies from other cassava populations for most traits except DM and FYLD (Wolfe et al., 2017). In this study, the use of data from CETUKG stage significantly improved prediction accuracies for all the traits except MCMDS and CBSDRS compared to PYTUKG, PYTKIB, and AYTKIB data sets. Overall prediction accuracies obtained from cross-validations were
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 training population from either cycle particularly when the interest is yield. An added advantage of equal prediction in CETUKG and PYTUKG is that clones can directly be transferred CET to AYT stage for evaluation. Skipping of PYT stage could help accelerate the varietal replacement process.

Dry matter prediction accuracy improved significantly by 53% when clones in advanced Kibaha trials (AYTKIB) were used to model training than those in PYTKIB. Comparison between breeding stages in the two locations showed high FYLD prediction accuracy in CETUKG and a high DM in AYTKIB. One would have expected opposite results because of trait selection preferences between the two locations as well as the larger plot size associated with AYTs. We are not certain why clones in early generation evaluated in smaller plots had higher yield than clones in advanced generation evaluated in larger plots. The difference could be attributed to high genetic variation and high heritability estimates for CET clones than AYT 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 et al. (2012) using data from two consecutive breeding cycles of sugar beets and Michel et al. (2016) study on genomic selection using multiple breeding cycles in bread wheat. Ceballos et al. (2016) recommends the used of phenotypic information from clones in advanced breeding stage during genomic selection because of their “stable” genotypic performance, in this
study we observed that phenotypic records of clones in early breeding stage predicted test set better for most traits than advanced ones. From this study we suggest that consideration should be given to the within the breeding stage clones particularly from early generation when forming training populations. Additionally, CET clones are generally tested in smaller plots because of limited number of planting stakes. The use of smaller plots for trait evaluation can further help reduce phenotyping cost. Additional investigation using different cassava population is needed to validate our results.

**Within and Cross Location Prediction**

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

Reduced common ancestors overtime results in reduced kinship, which reduces accuracy (de los Campos et al., 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 to within cycles in wheat. Our results are also in agreement with Ly et al. (2013) findings. They observed a decrease in prediction accuracies across locations in cassava. Therefore, based on this population, cross-location prediction may not be useful for programs implementing GS. Although generally, it would be useful to implement GS across locations, using populations with structure and weak SNP-QTL LD associations across populations could limit GS-assisted breeding in cassava and other species. This limitation has also been observed in livestock, Hayes et al. (2009), reported that pooling animals from different populations did not improve trait predictions due to non-persistent association between SNPs and QTL across breeds (populations).

In our study we did not find consistent relationship between heritability and prediction accuracy estimates across breeding cycles. In fact, we observed same accuracies for FYLD in both CET and PYT at Ukiriguru. Other studies have reported positive relationship between heritability and prediction accuracy across breeding cycles. For example, Sallam et al. (2015) reported high prediction accuracy for fusarium head blight resistance than yield in cross breeding cycles. This is because highly heritable traits have a less complex genetic architecture and 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.

**Across-Program Prediction**

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

Further, we used UKG+KIB+UG to predict UKG and KIB to determine if adding UG clones to TARI clones would improve trait prediction accuracies. There was no improvement in trait prediction. In fact, there were decreases for CBSDS, CBSDRS, and SHTWT (r=0.16 to 0.07, r=0.14 to 0.07, r=0.16 to 0.08, respectively) when UKG+KIB+UG was used to predict KIB. We observed similar decreases for CBSDRS, SHTWT, and FYLD (r=0.23 to 0.15, r=0.20 to 0.14, 0.30 to 0.26, respectively) when UKG+KIB+UG was used to predict UKG. Decrease in accuracies 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 population from North Dakota state University barley program and a second population from University of Minnesota barley breeding program to form training population. Other researchers have also reported low accuracies in cross-population predictions (Wolfe et al., 2017; Crossa et al., 2010; Endelman, 2011). Our results and evidences from other studies suggests that breeders can achieve better and a more reliable prediction accuracy estimates using smaller population with closely related genotypes than using large population with unrelated individuals. Population structure, environmental condition differences, differences in experimental design, direction of trait selection in each of the TARI breeding program, and variation in heritability could have impacted the accuracy estimates. Based on published results, the Ugandan training population had slightly higher accuracies for most traits compared to TARI (Wolfe et al., 2017; Ozimati et al., 2018). More locations and replication of clones across environments could have improved their accuracies. The poor prediction results across-programs reported in our study will
make it harder for breeders to use training data from different locations, breeding programs, and countries.

**Genetic Architecture of Disease Resistance**

Genetic associations with MCMDS were distributed across all chromosomes except 4, 5, 6, 8, 11, 13, and 18 (Fig. 3, Table 6, and Supplementary File1). Previous GWAS study for cassava mosaic disease severity resistance revealed CMD2 locus on chromosome 12 in other cassava populations (Wolfe et al., 2016). SNPs S12_5529819 and S12_11351823 delimit this QTL region (6.3-8.7 Mbps). S12_7270219 was the most significant marker tagging the QTL in all the TARI sub populations and accounted for 14 to 37% of the phenotypic variance explained (PVE). This result validates the presence of previously reported CMD2 locus. In addition to CMD2 locus, we also detected another major CMD QTL (*QTL-cmds16-1*) on chromosome 16 tagged by marker S16_421670 when we used PYTUKG population. This region accounted for 87% of PVE (Fig. 3, Table 6, and Supplementary File1). There are no known CMD resistance genes in this region. However, *Manes.16G036200.1*, *Manes.16G036300.1*, and *Manes.16G036400.1* genes have been annotated in this region and 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 association with both MCBSDS and MCMDS resistance (Fig. 3, Table 6, and Supplementary File 1). Marker S12_7929439 tagging *cbsd12|cmd-2* QTL accounted for 8% PVE for MCMDS and 33% MCBSDS in several subsets of TARI population. In the same region, a single marker (S12_929320) was found associated with MCMDS and MCBSDS resistance QTL (*QTL-cbsd12|cmd-1*) in PYTKIB dataset and accounted for 37% and 8% PVE, respectively. *QTL-cbsd4|cmd-1* tagged by marker S4_24670203 was detected in K4_cluster1 dataset and explained 14% of PVE. *Manes.04G113900.1* gene occur in the marker region detected on chromosome 4. This gene is known to encode for phospholipased α and activates the plant responses to pathogen attacks.
There are no known annotated genes in the region of QTL-cbsd12|cmd-1 and QTL-cbsd12|cmd-2 QTL. Further studies need to be conducted to confirm the presence of QTL conferring resistance to both CMD and CBSD on chromosome 4 and 12.

On chromosome 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 UKG and PYTKIB sub populations, respectively. QTL-cbsd9-1 accounted for 9% PVE while QTL-cbsd11-2 accounted for 5% PVE. The annotated gene within the QTL-cbsd11-2 is Manes.11G120800.1. This gene is known to encode for a protein kinase. An overexpression of similar gene in tobacco stimulated plant defense response (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. Wen-Yuan 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 defense mechanism to confer CBSD resistance. However, further investigation to validate the presence of CBSD resistant is needed.

Nine markers representing four loci on chromosomes 2, 3, 8, and 10 were significantly associated with CBSDRS responses (Table 6, Supplementary File 1). Two loci on chromosomes 3 (QTL-cbsdrs3-1) and 8 (QTL-cbsdrs8-1), occurring in AYTKIB accessions, accounted for 30% and 32% of the phenotypic variation (chromosomes 3 and 8, respectively). While the other two minor loci on chromosomes 2 (QTL-cbsdrs2-1) and 10 (QTL-cbsdrs10-1), detected in UKG accessions, explained 8% and 11% of the phenotypic variation (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 they drive pathways for biotrophic pathogen resistance in wheat. No annotated genes within the region significant for resistance to CBSDRS on chromosome 3 are in cassava reference genome.

This article is protected by copyright. All rights reserved.
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 CBSDS 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). From this study we observed that using fewer but closely related individuals particularly from same clusters improves the discovery of QTL compared to the use of large number of individuals. These results are similar to the GWAS findings of Bradbury et al. (2011) who reported that accounting for individual relatedness in barley improved detection of true QTL.

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. TARI population differentiation could have resulted from existing restrictions on clonal movement between programs. This restriction was imposed to contain the spread of cassava foliar diseases between Lake and Coastal Zone (Legg and Thresh, 2000). Although, there is long tradition of germplasm sharing between Tanzania and Uganda, the introgression occurring in Uganda TP is large and restricted to chromosome 1 while those in Tanzania are spread across the entire 18 chromosomes. This could suggest that the Uganda introgression could be more recent than in Tanzania. This may explain the population differentiation between sub populations with the introgression versus ones without.

There was no relationship between increase in plot size and decrease in error variance across all traits. For some traits, larger plots were preferable while 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 accuracies were observed for some traits, contrary to other plant studies (Combs and Bernardo, 2013; Lian et al., 2014). We are not certain whether this inverse relationship is due to noise in the data or is true and therefore further assessment is needed to determine this relationship.

Prediction accuracies within and between locations (KIB and UKG) were generally low compared to other cassava populations. Although larger training population sizes have been associated with improved accuracies, in this study, adding clones from KIB to UKG and vice versa did not improve prediction accuracy of either population. Similarly, adding the UG clones to either KIB or UKG did not improve accuracies of either. The lack of relatedness between germplasm, population structure, and the impact of genotype-by-environment interaction negatively impacted accuracy estimates.

Generally, across breeding cycle GS is more difficult in plants than in animals. This is because every year, plant breeders largely use new parents, some with unknow background from other breeding programs or competitors while animal breeders work in closed populations. This could make it challenging for the breeders to keep materials adequately related across breeding cycle GS. The successful use of GS is dependent on the close relationship between individuals in training and test set. Clones in similar breeding stages were more valuable than mixing clones from different breeding stages when constructing training population. This because clones in the same generation are likely to share ancestor few generations back and therefore marker-QTL linkage are preserved due to limited number of recombination events (Habier et al., 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 and Cohen, 2012; Mohammadi et al., 2015).

Consistent with the findings of other researchers, we conclude that clones from within-breeding cycles are presently better option as candidates for GS training population (Song et al., 2017; Michel et al., 2016; Cericola et al., 2017; Riedelsheimer et al., 2013; Albrecht et al., 2014; Lehermeier et al., 2014). Also, it may not be useful to constitute training populations from programs with divergent populations or populations separated by barriers like the existing restriction on clone movement.
between the two Tanzania programs. Across-breeding cycle GS especially those in more advanced breeding stages needs to be investigated further because prediction accuracies are very low. The impact of GxE modeling on unreplicated clones across environments need to be investigated further. Moreover, clones in the early breeding stage provided more reliable trait prediction accuracies because of their inherent genetic variation. Therefore, these clones are better candidates for training population construction.

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

**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 NaCRRRI cassava breeding programs, who helped in the 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 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/)

References

Akdemir, D., and O.U. Godfrey. 2015. EMMREML: fitting mixed models with known

cassava root yield in Adamawa State, Nigeria. African Journal of Agricultural Research,
6(13):2931-2934.

Bates, D., M. Maechler, B. Bolker, S. Walker, R.H.B. Christensen, H. Singmann, B. Dai, and


Bradbury, P., T. Parker, M.T. Hamblin, and J. L. Jannink. 2011. Assessment of power and false
discovery rate in genome-wide association studies using the BarleyCAP germplasm. Crop

Bredeson, J., V. Lyons, J. B. Prochnik, S. E. Wu, G. A., Ha, C. M. Edsinger-Gonzales, E. and
D. S. Rokhsar. 2016. Sequencing wild and cultivated cassava and related species reveals
extensive interspecific hybridization and genetic diversity. Nature Biotechnology

Browning, B.L., and S.R. Browning. 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.

Ceballos, H., P. Kulakow, and C. Hershey. 2012. Cassava breeding: current status,

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.

mode of gene action in cassava for resistance to cassava green mite and cassava mealy
Combs, E. and R. Bernardo. 2013. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. The Plant Genome, 6(1).


This article is protected by copyright. All rights reserved.


Figure Legends

**Figure 1.** PCA plot of the TARI clones alone (A) and the combined TARI and UG clones (B) based on SNP marker matrix and the loadings values (eigenvector coefficients) for each marker on PC1 against marker positions on all cassava chromosomes for TARI alone (C) and combined TARI and UG (D). For TARI PCA, 121K SNPs were used and for combined PCA, we used 37K SNPs common to both TARI and UG. For both cases, marker loadings were based on common markers (37K).
Figure 2. Phenotypic (left) and genetic (right) correlations between nine traits in Cassava using combined TARI data. Blue and red colors 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 $P$-value < 0.05 were left blank.
**Figure 3.** The Manhattan (A) and quantile-quantile (B) plots from mixed linear models summarizing genome-wide association results for three significant traits in all subsets of combined TARI population. The QQ plot demonstrate the differences between various population structure controls. CET=clonal evaluation trial, PYT=preliminary yield trial, AYT=advanced yield trial, K3 & K4 =cluster 3 and 4 generated using kinship relationship matrix. The horizontal solid line indicates the genome-wide Bonferroni significance threshold (−log10 (P-value) = 7.27), the dashed line indicates FDRpt1 and Dotted line indicate=FDRpt2. Additional detail for significant markers, Pvalue, and pFDR values are detailed in supplementary File 1 (Sheet 1). Additional plot for all the other traits below threshold are in Supplementary Figure 3.
Table 1. The trial structure for the materials used for training population. Reps are the number of replicates in each trial. The number of individual clones in each of the three replicates after sprouting are represented by Rep1, Rep2, and Rep3.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Stage</th>
<th>Location</th>
<th>No. Reps</th>
<th>Clones/Rep</th>
<th>Plot Area (M²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CET1</td>
<td>CET</td>
<td>Ukiriguru</td>
<td>2</td>
<td>108 91</td>
<td>5 5</td>
</tr>
<tr>
<td>CET2</td>
<td>CET</td>
<td>Ukiriguru</td>
<td>2</td>
<td>213 214</td>
<td>5 5</td>
</tr>
<tr>
<td>PYT1</td>
<td>PYT</td>
<td>Ukiriguru</td>
<td>2</td>
<td>96 81</td>
<td>10 10</td>
</tr>
<tr>
<td>PYT1</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>64 65</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT2</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>74 77</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT3</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>45 48</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT4</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>21 23</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT5</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>24 23</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT6</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>15 15</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT7</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>20 19</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT8</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>55 51</td>
<td>14 14</td>
</tr>
<tr>
<td>PYT9</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>58 63</td>
<td>14 14</td>
</tr>
<tr>
<td>AYT1</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>26 25</td>
<td>42 20</td>
</tr>
<tr>
<td>AYT2</td>
<td>PYT</td>
<td>Kibaha</td>
<td>3</td>
<td>12 12 12</td>
<td>42 20</td>
</tr>
<tr>
<td>AYT3</td>
<td>PYT</td>
<td>Kibaha</td>
<td>3</td>
<td>16 13 15</td>
<td>42 20</td>
</tr>
<tr>
<td>AYT4</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>23 23</td>
<td>42 20</td>
</tr>
<tr>
<td>AYT5</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>25 20</td>
<td>42 20</td>
</tr>
<tr>
<td>AYT6</td>
<td>PYT</td>
<td>Kibaha</td>
<td>2</td>
<td>21 21</td>
<td>42 20</td>
</tr>
</tbody>
</table>

Table 2. Heritability, genetic and residual variance components for four plot sizes
<table>
<thead>
<tr>
<th>Plot Size</th>
<th>Variance Component</th>
<th>MCMD S</th>
<th>MCBSD S</th>
<th>CBSDR S</th>
<th>MCGM S</th>
<th>RTNO</th>
<th>SHTW T</th>
<th>HI</th>
<th>FYLD</th>
<th>DM</th>
</tr>
</thead>
<tbody>
<tr>
<td>5m² (CETUKG)</td>
<td>$\sigma^2_G$</td>
<td>0.042</td>
<td>0.097</td>
<td>0.059</td>
<td>0.009</td>
<td>2195</td>
<td>12.71</td>
<td>0.01</td>
<td>11.01</td>
<td>7.11</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_{TR}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.184</td>
<td>0.000</td>
<td>9516</td>
<td>9.10</td>
<td>0.00</td>
<td>0.46</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_e$</td>
<td>0.049</td>
<td>0.215</td>
<td>0.142</td>
<td>0.000</td>
<td>7192</td>
<td>28.66</td>
<td>0.02</td>
<td>25.12</td>
<td>15.20</td>
</tr>
<tr>
<td></td>
<td>$H^2$</td>
<td>0.46</td>
<td>0.31</td>
<td>0.06</td>
<td>0.23</td>
<td>0.31</td>
<td>0.20</td>
<td>0.30</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>10m² (PYTUKG)</td>
<td>$\sigma^2_G$</td>
<td>0.050</td>
<td>0.008</td>
<td>0.037</td>
<td>0.003</td>
<td>2205</td>
<td>0.215</td>
<td>0.00</td>
<td>1.80</td>
<td>3.72</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_{TR}$</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1571</td>
<td>4.155</td>
<td>0.00</td>
<td>0.06</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_e$</td>
<td>0.005</td>
<td>0.129</td>
<td>0.131</td>
<td>0.000</td>
<td>7714</td>
<td>10.605</td>
<td>0.01</td>
<td>5.80</td>
<td>16.84</td>
</tr>
<tr>
<td></td>
<td>$H^2$</td>
<td>0.91</td>
<td>0.06</td>
<td>0.22</td>
<td>0.22</td>
<td>0.02</td>
<td>0.15</td>
<td>0.24</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>14m² (PYTKIB)</td>
<td>$\sigma^2_G$</td>
<td>0.349</td>
<td>0.042</td>
<td>0.029</td>
<td>0.003</td>
<td>1376</td>
<td>14.89</td>
<td>0.00</td>
<td>50.09</td>
<td>5.23</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_{TR}$</td>
<td>0.026</td>
<td>0.013</td>
<td>0.000</td>
<td>0.000</td>
<td>9203</td>
<td>28.08</td>
<td>0.00</td>
<td>24.69</td>
<td>1.74</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_e$</td>
<td>0.432</td>
<td>0.211</td>
<td>0.118</td>
<td>0.011</td>
<td>4391</td>
<td>105.48</td>
<td>0.01</td>
<td>129.2</td>
<td>12.28</td>
</tr>
<tr>
<td></td>
<td>$H^2$</td>
<td>0.43</td>
<td>0.18</td>
<td>0.24</td>
<td>0.10</td>
<td>0.17</td>
<td>0.25</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20m² (AYTKIB)</td>
<td>$\sigma^2_G$</td>
<td>0.28</td>
<td>0.031</td>
<td>0.051</td>
<td>0.036</td>
<td>4760</td>
<td>36.00</td>
<td>0.00</td>
<td>59.84</td>
<td>2.0885</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_{TR}$</td>
<td>0.04</td>
<td>0.001</td>
<td>0.036</td>
<td>0.000</td>
<td>6688</td>
<td>70.08</td>
<td>0.00</td>
<td>66.79</td>
<td>0.9732</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2_e$</td>
<td>0.30</td>
<td>0.122</td>
<td>0.121</td>
<td>0.117</td>
<td>5530</td>
<td>197.90</td>
<td>0.00</td>
<td>211.4</td>
<td>10.252</td>
</tr>
<tr>
<td></td>
<td>$H^2$</td>
<td>0.44</td>
<td>0.20</td>
<td>0.25</td>
<td>0.12</td>
<td>0.80</td>
<td>0.18</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\sigma^2_G$ = Genetic variance among clones, $\sigma^2_{TR}$ = Variance among reps in trials, $\sigma^2_e$ = 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; and AYTKIB=Advanced Yield Trial at Kibaha.

Table 3. Summary of cross-validated prediction accuracies by trait and breeding stages

<table>
<thead>
<tr>
<th>Trait</th>
<th>CETUKG</th>
<th>PYTUKG</th>
<th>PYTKIB</th>
<th>AYTKIB</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCMDS</td>
<td>0.11*ns</td>
<td>0.06*ns</td>
<td>0.17*ns</td>
<td>0.32*ns</td>
<td>0.14</td>
</tr>
<tr>
<td>MCBSDS</td>
<td>0.26*</td>
<td>0.33*</td>
<td>0.08</td>
<td>0.24ns</td>
<td>0.23</td>
</tr>
<tr>
<td>CBSDRS</td>
<td>0.26ns</td>
<td>0.10*ns</td>
<td>0.09ns</td>
<td>0.10ns</td>
<td>0.14</td>
</tr>
<tr>
<td>MCGMS</td>
<td>0.22*</td>
<td>0.09ns</td>
<td>0.21ns</td>
<td>0.43*</td>
<td>0.24</td>
</tr>
<tr>
<td>RTNO</td>
<td>0.28*</td>
<td>0.27*ns</td>
<td>0.16ns</td>
<td>0.03ns</td>
<td>0.19</td>
</tr>
<tr>
<td>SHTWT</td>
<td>0.22ns</td>
<td>0.03ns</td>
<td>0.12ns</td>
<td>0.08ns</td>
<td>0.11</td>
</tr>
<tr>
<td>HI</td>
<td>0.22*</td>
<td>0.38*</td>
<td>0.13ns</td>
<td>0.19ns</td>
<td>0.23</td>
</tr>
<tr>
<td>FYLD</td>
<td>0.36*</td>
<td>0.35*</td>
<td>0.10ns</td>
<td>0.09ns</td>
<td>0.23</td>
</tr>
<tr>
<td>DM</td>
<td>0.14ns</td>
<td>0.18ns</td>
<td>0.28*</td>
<td>0.43*</td>
<td>0.26</td>
</tr>
<tr>
<td>Mean</td>
<td>0.23</td>
<td>0.20</td>
<td>0.15</td>
<td>0.20</td>
<td>0.19</td>
</tr>
</tbody>
</table>

ns = non-significant prediction accuracies (r), * accuracy significantly different from zero \( (P \leq 0.05) \). MCMDS=mean cassava mosaic disease severity, MCBSDS=mean cassava brown streak disease severity, CBSDRS=mean Cassava brown streak disease root severity, MCGMS=Mean Cassava green mite severity, RTNO=root number, SHTWT=shoot weight, HI=harvest index, FYLD=Fresh root yield, DM=Dry matter.
Table 4. Average prediction trait accuracies for cross-location and combined TARI

<table>
<thead>
<tr>
<th>Trait</th>
<th>TARI</th>
<th>UKG to UKG</th>
<th>KIB to UKG</th>
<th>KIB to KIB</th>
<th>UKG to KIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCMDS</td>
<td>0.15*</td>
<td>0.09</td>
<td>0.02(^{ns})</td>
<td>0.18*</td>
<td>0.06(^{ns})</td>
</tr>
<tr>
<td>MCBSDS</td>
<td>0.23*</td>
<td>0.25</td>
<td>0.11*</td>
<td>0.08(^{ns})</td>
<td>0.09(^{ns})</td>
</tr>
<tr>
<td>CBSDRS</td>
<td>0.09*</td>
<td>0.21</td>
<td>0.16*</td>
<td>0.05(^{ns})</td>
<td>0.10*</td>
</tr>
<tr>
<td>MCGMS</td>
<td>0.25*</td>
<td>0.22</td>
<td>0.23*</td>
<td>0.23*</td>
<td>0.25*</td>
</tr>
<tr>
<td>RTNO</td>
<td>0.16*</td>
<td>0.24</td>
<td>0.08(^{ns})</td>
<td>0.09(^{ns})</td>
<td>0.11*</td>
</tr>
<tr>
<td>SHTWT</td>
<td>0.11(^{ns})</td>
<td>0.22</td>
<td>0.19(^{ns})</td>
<td>0.09(^{ns})</td>
<td>0.01(^{ns})</td>
</tr>
<tr>
<td>HI</td>
<td>0.16(^{ns})</td>
<td>0.18</td>
<td>0.15*</td>
<td>0.11*</td>
<td>0.12*</td>
</tr>
<tr>
<td>FYLD</td>
<td>0.18*</td>
<td>0.29</td>
<td>0.13*</td>
<td>0.08(^{ns})</td>
<td>0.05(^{ns})</td>
</tr>
<tr>
<td>DM</td>
<td>0.23*</td>
<td>0.13</td>
<td>0.16*</td>
<td>0.28*</td>
<td>0.24*</td>
</tr>
</tbody>
</table>

Mean | 0.17  | 0.20   | 0.14   | 0.13    | 0.10    |

\(^{ns}\) = non-significant prediction accuracies, * accuracy significantly different from zero (P \leq 0.05). The values in the bracket are the accuracies for within the location prediction.

Table 5. Summary of cross-validated prediction accuracies by trait and cross programs

<table>
<thead>
<tr>
<th>Scenario</th>
<th>MCMDS</th>
<th>MCBSD</th>
<th>CBSDR</th>
<th>RTNO</th>
<th>SHTWT</th>
<th>HI</th>
<th>FYLD</th>
<th>DM</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIB_to_KIB</td>
<td>0.19(^{ns})</td>
<td>0.16(^{ns})</td>
<td>0.14(^{ns})</td>
<td>0.16(_s^{ns})</td>
<td>0.15(_s^{ns})</td>
<td>0.12(_s^{ns})</td>
<td>0.09(_s^{ns})</td>
<td>0.29*</td>
<td>0.16</td>
</tr>
<tr>
<td>UKG_to_UKG</td>
<td>0.01(^{ns})</td>
<td>0.28*</td>
<td>0.23*</td>
<td>0.26*</td>
<td>0.20*</td>
<td>0.17(_s^{ns})</td>
<td>0.30*</td>
<td>0.13(_s^{ns})</td>
<td>0.20</td>
</tr>
<tr>
<td>KIB+UKG_to_KIB</td>
<td>0.17(^{ns})</td>
<td>0.14(^{ns})</td>
<td>0.14(_s^{ns})</td>
<td>0.16(_s^{ns})</td>
<td>0.14(_s^{ns})</td>
<td>0.14(_s^{ns})</td>
<td>0.09(_s^{ns})</td>
<td>0.30*</td>
<td>0.16</td>
</tr>
</tbody>
</table>

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/csc2.20003.

This article is protected by copyright. All rights reserved.
<table>
<thead>
<tr>
<th>Trait Target</th>
<th>Ukgi+Kib to Ukgi</th>
<th>Kib+Ug to Kib</th>
<th>Ukgi+Ukgi to Ukgi</th>
<th>Kib+Ukgi+Ug to Kib</th>
<th>Ukgi+</th>
<th>Kib+Ug to Ukgi</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01* ns</td>
<td>0.20* ns</td>
<td>0.02* ns</td>
<td>0.17* ns</td>
<td>-0.01*</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>0.18* ns</td>
<td>0.06* ns</td>
<td>0.15* ns</td>
<td>0.11* ns</td>
<td>0.31*</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.27* ns</td>
<td>0.14* s</td>
<td>0.24* ns</td>
<td>0.06* ns</td>
<td>0.17*</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.19* ns</td>
<td>0.08* s</td>
<td>0.19* ns</td>
<td>0.15* ns</td>
<td>0.16*</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.08* s</td>
<td>0.08* s</td>
<td>0.14* ns</td>
<td>0.06* ns</td>
<td>0.20*</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.15* s</td>
<td>0.15* s</td>
<td>0.28* ns</td>
<td>0.15* s</td>
<td>0.17*</td>
<td>0.13</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Significant markers associated with MCMDS, MCBSDS, and CBSDRS resistance detected in TARI TP QTL

<table>
<thead>
<tr>
<th>QTL-Subregion</th>
<th>Trait</th>
<th>C hr. (Mb)</th>
<th>Tag-SNP</th>
<th>Position</th>
<th>Allele</th>
<th>Fr Eq</th>
<th>Log10P value</th>
<th>PVE (%)</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>QTL-cbsd2s2-1</td>
<td>CBSDRS</td>
<td>9.15 - 9.16</td>
<td>S2_9258 334</td>
<td>92583 34</td>
<td>A/C</td>
<td>0.1</td>
<td>0.5</td>
<td>11</td>
<td>6.843</td>
</tr>
<tr>
<td>QTL-cbsd3s3-1</td>
<td>CBSDRS</td>
<td>10.1 - 10.6</td>
<td>S3_1016 5675</td>
<td>10165 70</td>
<td>A/G</td>
<td>0.0</td>
<td>0.5</td>
<td>20</td>
<td>6.137</td>
</tr>
<tr>
<td>QTL-cbsd8s8-1</td>
<td>CBSDRS</td>
<td>1.74</td>
<td>S8_1736 3721</td>
<td>17363 721</td>
<td>A/G</td>
<td>0.0</td>
<td>0.5</td>
<td>21</td>
<td>6.35</td>
</tr>
<tr>
<td>QTL-cbsd10s10-1</td>
<td>CBSDRS</td>
<td>0.16</td>
<td>S10_160 775</td>
<td>16077 5</td>
<td>G/A</td>
<td>0.0</td>
<td>0.5</td>
<td>16</td>
<td>5.24</td>
</tr>
<tr>
<td>QTL-cbsd9s1-1</td>
<td>MCBSDS</td>
<td>10.7 - 10.8</td>
<td>S9_1070 7044</td>
<td>10707 044</td>
<td>G/A</td>
<td>0.5</td>
<td>0.5</td>
<td>57</td>
<td>6.457</td>
</tr>
<tr>
<td>QTL-cbsd11s11-1</td>
<td>MCBSDS</td>
<td>8 - 22.9</td>
<td>S11_229 42418</td>
<td>22942 418</td>
<td>A/G</td>
<td>0.4</td>
<td>0.5</td>
<td>08</td>
<td>6.011</td>
</tr>
<tr>
<td>QTL-cmds16s16-1</td>
<td>MCMDS</td>
<td>6.3 - 8.7</td>
<td>S16_421 670</td>
<td>42167</td>
<td>T/C</td>
<td>0.4</td>
<td>0.9</td>
<td>86</td>
<td>9.844</td>
</tr>
<tr>
<td>QTL-cmds12s12-1</td>
<td>MCMDS</td>
<td>24.6 - 24.7</td>
<td>S12_727 0219</td>
<td>72702 19</td>
<td>T/C</td>
<td>0.4</td>
<td>0.9</td>
<td>24</td>
<td>11.246</td>
</tr>
<tr>
<td>QTL-cbsd4/cmd-</td>
<td>MCBSDS</td>
<td>4</td>
<td>S4_2467 24670</td>
<td>T/C</td>
<td>0.2</td>
<td>0.5</td>
<td>44</td>
<td>14</td>
<td>K4_Clu</td>
</tr>
<tr>
<td>I</td>
<td>MCMDS</td>
<td>0203</td>
<td>203</td>
<td>G</td>
<td>4</td>
<td>14</td>
<td>ster1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>------</td>
<td>-----</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QTL-cbsd12/cmd-1</td>
<td>MBSDS</td>
<td>12</td>
<td>0.93</td>
<td>S12_929 92932</td>
<td>T/ 0.0</td>
<td>0</td>
<td>5.941</td>
<td>37</td>
<td>8 PYTKI B</td>
</tr>
<tr>
<td></td>
<td>CMDS</td>
<td></td>
<td>320</td>
<td>G/ 6</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QTL-cbsd12/cmd-2</td>
<td>MBSDS</td>
<td>12</td>
<td>7.93</td>
<td>S12_792 79294</td>
<td>G/ 0.3</td>
<td>0</td>
<td>8.749</td>
<td>8</td>
<td>33 Several</td>
</tr>
<tr>
<td></td>
<td>CMDS</td>
<td></td>
<td>7.95</td>
<td>C/ 7</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>