bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Accuracies of univariate and multivariate genomic prediction models in African Cassava.

Authors:

Uche Godfrey Okeke, Deniz Akdemir, Ismail Rabbi, Peter Kulakow, Jean-Luc Jannink*.

Uche Godfrey Okeke, Jean-Luc Jannink, Section of Plant breeding and Genetics, School of Integrative Plant Sciences, College of Agriculture and Life sciences, Cornell University, 14853, Ithaca, New York, USA; Deniz Akdemir, QuantGen Group, Michigan State University; Ismail Rabbi, Peter Kulakow, International Institute of Tropical Agriculture (IITA), PMB 5320, Oyo Road, Ibadan, Nigeria.

*Corresponding author: Jean-Luc Jannink (jeanluc.work@gmail.com)

List of abbreviations:

Genomic Selection (GS); Best Linear Unbiased Prediction (BLUP); Estimated Breeding Values (EBVs); Estimated genetic Values (EGVs); Genomic Estimated Breeding Values (GEBVs); Single Nucleotide polymorphisms (SNPs); Genotype-by-environment interactions (GxE); Gene-by-gene interactions (GxG); Gene-by-gene-by-environment interactions (GxGxE); Univariate single environment one-step model (uT); Univariate multi environment one-step model (uE); Multi-trait single environment one-step model (MT);Multivariate single trait multi environment model (ME), seedling vigor (VIGOR), Number of storage roots per hectare at harvest (RTNO), Fresh weight of harvested roots in tons per hectares (T/ha) per plant (FYLD), dry matter (DM), plot mean cassava mosaic disease severity (MCMDS), plot mean cassava green mite (MCGM) severity.

Keywords:

Genomic selection, plant genetic evaluation, cassava breeding, single trait models, univariate multi-environment models, multi-trait (MT) models, multivariate multi-environment (ME) models, GxE interactions, prediction accuracies, Genomic estimated breeding values (GEBVs).

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Background:

Genomic selection (GS) promises to accelerate genetic gain in plant breeding programs especially for long cycle crops like cassava. To practically implement GS in cassava breeding, it is useful to evaluate different GS models and to develop suitable models for an optimized breeding pipeline.

Methods:

We compared prediction accuracies from a single-trait (uT) and a multi-trait (MT) mixed model for single environment genetic evaluation (Scenario 1) while for multi-environment evaluation accounting for genotype-by-environment interaction (Scenario 2) we compared accuracies from a univariate (uE) and a multivariate (ME) multi-environment mixed model. We used sixteen years of data for six target cassava traits for these analyses. All models for Scenario 1 and Scenario 2 were based on the one-step approach. A 5-fold cross validation scheme with 10-repeat cycles were used to assess model prediction accuracies.

Results:

In Scenario 1, the MT models had higher prediction accuracies than the uT models for most traits and locations analyzed amounting to 32 percent better prediction accuracy on average. However for Scenario 2, we observed that the ME model had on average (across all locations and traits) 12 percent better predictive power than the uE model.

Conclusion:

We recommend the use of multivariate mixed models (MT and ME) for cassava genetic evaluation. These models may be useful for other plant species.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background:

Cassava (*Manihot esculenta* Crantz)[1] is a staple food for over 700 million people in Africa, South America and Asia [2]. Cassava also has immense industrial potential. Pure white cassava starch is easy to extract and contains low levels of fat (about 1.5%), protein (about 0.6%) and phosphorus (about 4%), which are desirable attributes for the food industry [3,4]. Given the issues of climate change and rapid population growth in countries that rely heavily on cassava, rapid genetic improvement of cassava is critically needed. To enable rapid genetic improvement of cassava, genetic evaluation protocols hinged on Best Linear Unbiased Prediction (BLUP) analysis [5,6] and selection on a merit index [7,8] has been recommended [9] to maximize gain from selection.

Genomic selection (GS) [10] offers crops like cassava tremendous opportunity for accelerated genetic gains [11] by making use of whole genome SNP markers accessible with methods like the genotyping-by-sequencing (GBS) [12]. These whole genome SNP markers could be dense enough to be in linkage disequilibrium with most quantitative trait loci (QTL) affecting traits of interest. Using GS, selection is imposed at these QTL without actually identifying the QTL or the functional polymorphisms [10]. Also these markers will help to better track relatedness due to mendelian sampling [24] in a breeding population especially where pedigree records are not fully available thus yielding an improvement in selection accuracies [25].

GS models for plant genetic evaluation:

Genetic evaluation [9] starts with accurately estimating the genetic value of an individual for a wide range of traits using its own performance records, progeny performance

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

records or records from relatives or a combination of the three [13]. This has usually been carried out using single trait (uT) BLUP methodology [14] for obtaining estimated breeding values (EBVs) for one trait at a time. In plant and animal breeding, breeders usually select on the basis of multiple traits that are often genetically correlated although these correlations may range from very weak to very strong. The uT model for traits measured in a single environment assumes zero genetic and residual covariances between these traits such that information from other traits are not utilized when obtaining EBVs of the evaluated individuals for the traits in the analysis. However, the optimal estimation procedure to combine information from multiple trait records and obtain EBVs is the multi-trait BLUP methodology (MT) [15,16]. The MT model does not assume zero genetic and residual covariances but rather provides an estimate for these and also uses this information when obtaining individual EBVs for the traits in the analysis. The MT model has several advantages over the uT model including:

- Improved prediction accuracies for individual traits in the model because of more information (direct or indirect), better connectedness of the data [17], and exploitation of genetic and residual correlations in the model especially when traits with varying heritabilities are analyzed jointly.
- Simplified index selection because optimal weight factors for the total merit index are the economic weights [17].
- Efficient procedures for obtaining genetic and residual covariances and EBVs across location, country or region evaluations on information from same or related individuals [18, 19].
- Better selection accuracies by accounting for selection bias when all target traits under

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

selection are included in the model [20] in addition to the use of all individuals (selected or not) in the relationship matrix.

In most plant breeding programs, selected individuals are evaluated in multi-environment trials (METs). The goal is to sample the influence on selection candidates of the range of environments for which varieties will be targeted. Addressing the problem of the analysis of METs brings into focus another potential use for MT models [30]. Here, phenotypes of the same trait, but measured at different locations are parameterized as different traits in the MT model [31], producing what we call a multi-environment BLUP (ME) model. Like the MT model, the ME model estimates genetic covariances between a single trait measured at multiple environments which may lead to more accurate estimates of individual EBVs for the trait at all the environments where data has been recorded. For ME models used for modelling MET data, residual covariances are set to zero reflecting the assumption that no mechanism generates error covariances between a trait measured in different environments [18]. In contrast, the typical univariate BLUP model for modelling METs data, termed the univariate multi-environment model (uE), fits a multikernel mixed model with the genotypic effect as one kernel and the genotype-by-environment (GxE) effect as the second kernel and maybe environment as third kernel [26]. This model yields a GxE variance for a MET and individuals can be ranked on their performance at different locations. It is expected that for traits with high correlation between environments, the uE model should offer more accurate EBVs, while EBV accuracies for traits with lower correlation between environments will benefit more from the ME model. Different variants of the ME model have been used for modeling environment covariance structures in plant [32-35] and in animal

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

breeding [36,37]. Genetic covariances from the ME model offer a convenient tool for assessing the impact of GxE on a trait. The genetic covariances relate directly to the extent of GxE at all locations in the analysis. A low genetic correlation of the EBVs between a trait at different locations from the ME model indicate GxE impact on that trait [9,38-41].

Selecting the GS model to be employed in a practical cassava breeding program is not trivial hence the need to compare different GS models that will be useful in the different stages of cassava breeding with METs data. Finally, fitting multivariate BLUP models is not trivial. Even with software that can in principle fit these models, model convergence is not guaranteed and may require several attempts [21-23] and hence univariate models may be more practical if benefits of the multivariate models are not substantial.

The objectives of this paper are to:

- Compare multi-trait (MT) and single trait (uT) mixed models for METs data using cross-validated prediction accuracies.
- Compare the multivariate multi-environment (ME) model to a single-trait multi-environment (uE) model using cross-validated prediction accuracies while assessing GxE impact on analyzed traits via genetic covariances from the ME model fit.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

METHODS AND MATERIALS:

Cassava phenotype data:

We used historical phenotype data from different trials conducted by the cassava breeding program at the Institute of Tropical Agriculture (IITA), Ibadan, Nigeria in our analysis. The Genetic Gain population represents a collection of clones selected from the 1970s to 2007 by the cassava breeding program at the IITA [48,49]. Some of these clones are West African landraces and some are of East African origin. Clones in the Genetic Gain population have gone through advanced stages of the cassava breeding process up to on farm variety testing trials. The data used in our analysis comprises data collected on clonal evaluation trials (CETs) which are augmented design trials with about 2 known checks and unreplicated plots with 5 plant stands per clone. These data were collected from three target locations in Nigeria: Ibadan (7.40° N, 3.90° E), Mokwa (9.3° N, 5.0° E), and Ubiaja (6.66° N, 6.38° E). These locations represent regions which encompass about 35 percent of the cassava production base in Nigeria. Data sets were collected from 2000 to 2015 and included trials with most of the 764 clones of the Genetic Gain population. Six target agronomic traits were used in the analysis including seedling vigor (VIGOR), Number of storage roots per hectare at harvest (RTNO), Fresh weight of harvested roots expressed in tons per hectares (T/ha) per plant (FYLD), Percentage dry matter (DM) of storage roots, which measures root dry weight as the percentage of the root fresh weight, plot mean cassava mosaic disease severity (MCMDS), rated on a scale from 1 (no symptoms) to 5 (extremely severe), and plot mean cassava green mite (MCGM) severity, rated on a scale from 1 (no symptoms) to 5 (extremely severe). Cassava mosaic disease is caused by a *Begomovirus* that belongs to the Geminiviridae family, and is carried and transmitted by the whitefly *Bemisia tabaci*. The

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cassava green mite is *Mononychellus tanajoa* [50]. These traits are target traits used in the selection index for selection decisions in the IITA cassava breeding program. Phenotype data metrics are shown in Table 1.

Cassava genotype data:

DNA was extracted using DNeasy Plant Mini Kits (Qiagen) from 739 clones from the 2013 Genetic Gain trial at IITA and was quantified using PicoGreen. Genotyping-by-sequencing (GBS) was used for genotyping [12] these clones. Six 95-plex and one 75-plex *Ape*KI libraries were constructed and sequenced on Illumina HiSeq, one lane per library. Single nucleotide polymorphisms (SNPs) were called from the sequence data using the TASSEL pipeline version 4.0 [51], using an alignment to the *Manihot esculenta* version 6 reference genome [52]. Individuals with greater than 80% missing SNP calls and markers with more than 60% missing were removed. The marker data was converted to dosage format (0, 1, 2) and missing genotypic data were imputed using a LASSO regression method (Ariel Chan, personal communication, 2014) implemented using the R glmnet package [53]. The final data set consisted of 183,201 SNPs scored in 764 clones.

Statistical analysis:

We restructured the cassava phenotype data described above into two types of data common in most plant breeding programs. The first set was achieved by pooling data from multiple years in all locations (multi-environment trials data - METs). We termed this scenario the single environment genetic evaluation (Scenario 1). This resulting predictive power from this data were assessed for our three locations Ubiaja, Mokwa and Ibadan. The second scenario

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was achieved by using same METs data but in this case extracting location specific information by attempting to model GxE interaction. We termed this scenario the multi-environment genetic evaluation (Scenario 2). The goal of the latter data is to access the value of evaluate the impact of GxE and check if this yields better predictive value of the breeding value of a clone.

Pseudo-true genetic values for model accuracy computations:

For validating the models in this study, we define first a univariate single trait mixed model for each trait at each location separately (to preserve the variation embedded in each location) using an identity covariance matrix among clone effects, which assumes no relationship among all clones. The univariate mixed model was as follows:

$$y = Xb+Zu+e; \qquad u \sim N(0, \sigma^2_u I); \qquad e \sim N(0, \sigma^2 I) \qquad \text{(A)}$$

$$\hat{u} = (\sigma^2_u I_u) Z^T V^{-1}(y - X\hat{b})$$

where y is a vector of observations, b is a vector of fixed effects with design matrix X (relating observations to fixed effects in this case including grand mean and a nested effect of Trial-within-Year-within-Location and the ratio of plants harvested to number planted); u is a vector of clonal genetic effects with design matrix Z (relating observations to clones). This model was fit using the *lmer* function in the R lme4 package [55] and resulting BLUP values ($\hat{u}$), which we refer to as Estimated Genotypic Values (EGVs), were used as pseudo-true genetic effects for prediction accuracy computations.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

GS models for Scenario 1:

We define two mixed models fitted here as follows:

The single trait mixed model (uT):

$$y = X\beta + Zu + e \qquad (1)$$

where $y = (y'_a, y'_b, y'_c, \ldots, y'_z)'$; $e = (e'_a, e'_b, e'_c, \ldots, e'_z)'$

$u \sim N(0, \sigma^2_u K)$; $e \sim N(0, \sigma^2 I)$

$var(y) = V = Z(\sigma^2_u K)Z^T + \sigma^2 I$; $\hat{u} = (\sigma^2_u K)Z^T V^{-1}(y - X\hat{b})$

where y is the response vector of a trait at locations *a, b.. z*, β is the vector of fixed effects with design matrix X (relating observations to fixed effects namely the grand mean, nesting of Trial-within-Year-within-Location and ratio of plants harvested to number planted); u is the vector of random additive genomic effects with design matrix Z (relating trait values to clones) and K is the additive genomic relationship matrix generated from SNP markers as in VanRaden, 2008 [63] method two. Estimation of the parameters in model (1) were performed using Restricted Maximum Likelihood (REML) procedure implemented in the *airemlf90* FORTRAN program of the *blupf90* package [62] which estimates (co)variance parameters and then obtains best linear unbiased estimates (BLUEs) of fixed effects and BLUPs of random effects by solving the mixed model equations (MME) [5,6].

The multitrait Mixed Model:

$$y = X\beta + Zu + e \qquad (2)$$

where $y = (y'_{1a}, y'_{2a}, y'_{3a}, y'_{1b}, y'_{2b} \ldots, y'_{dz})'$; $u = (u'_1, u'_2, \ldots, u'_d)'$;

$$e = (e'_{1a}, e'_{2a}, e'_{3a}, e'_{1b}, e'_{2b} \ldots, e'_{dz})'$$

where y is a long vector of d traits in locations *a, b … z*, recorded for n clones, **X** and **Z** design

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

matrices are block diagonal matrices represented as *diag*(X_a, X_b,..., X_z) and *diag*(Z_a, Z_b,..., Z_z) respectively allowing for missing clones and observations. **X** is a design matrix for fixed effects *β* (with components as in model 1 above for every location and trait) and **Z** is a design matrix for random genetic effects *u*. Following a multivariate normal distribution (N_m), the marginal density of *y* is given as:

$$(y \mid \beta,\ R,\ G) \sim N_m(X\beta,\ V) \qquad (3)$$

$$V = Z(G \otimes K)Z^T + I_d \otimes R); \qquad \hat{u} = (G \otimes K)Z^T V^{-1}(y - X\hat{\beta})$$

where *d* is number of traits being analyzed, G and R are *d* × *d* symmetric genomic and error covariance matrices respectively, *K* remains the additive genomic relationship matrix for *n* clones generated from SNP markers as in VanRaden, 2008 [63] method two, *I* is a *d*-traits identity matrix and u are the genomic estimated breeding values (GEBVs) the clones and for the traits in the analysis. Estimation of the parameters in model (3) were also performed using REML procedure implemented in the *airemlf90* program [62] from which BLUEs of fixed effects and BLUPs of random effects were also obtained by solving the mixed model equations (MME) [5,6].

GS models for Scenario 2:

We also defined two mixed models here with the aim of modeling genotype-by-environment interaction effects and these are as follows:

The univariate multi-environment model (uE):

$$y = X\beta + Z_1u + Z_2w + e \qquad (4)$$

where $y = (y'_a, y'_b, y'_c, \ldots, y'_z)'$; $e = (e'_a, e'_b, e'_c, \ldots, e'_z)'$

$$u \sim N(0, \sigma^2_u K); \qquad w \sim N(0, \sigma^2_w K_w) \qquad e \sim N(0, \sigma^2 I)$$

$$var(y) = V = Z_1(\sigma^2_u K)Z_1^T + Z_2(\sigma^2_w K_w)Z_2^T + \sigma^2 I; \qquad \hat{u} = (\sigma^2_u K)Z_1^T V^{-1}(y - X\hat{\beta})$$

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\hat{w} = (\sigma^2{}_w K_w) Z_2{}^T V^{-1}(y - X\hat{b})$$

where *y* is a long vector of trait values for all locations, *β* is the vector of fixed effects with design matrix *X* (relating observations to fixed effects namely the grand mean, nested Trial-within-Year-within-Location and ratio of plants harvested to number planted); *u* is the vector of random additive genomic effects with design matrix Z_1 (relating trait values to clones), *w* is the vector of random marker-based clone-by-location interaction effects with design matrix Z_2 (relating trait values to clones-location combinations) and *K* is the additive genomic relationship matrix generated from SNP markers as in VanRaden, 2008 [63] method two, K_w is a block diagonal covariance matrix represented as $diag(K_a, K_b, K_c, \ldots, K_z)$ with the *K* matrices in this block being genomic relationship matrices calculated as in VanRaden, 2008 [63] method two for clones in locations *a,b...z*. In this model, the genomic value of a clone were estimated as $\hat{u} + \hat{w}$. We also fit model(4) using the *airemlf90* tool of the *blupf90* package [62] with a custom R-script for a 5-fold cross validation procedure with 10 repeat runs using the *airemlf90* function.

The multivariate multi-environment (ME) model:

We fit the ME model in a single step procedure using the following model:

$$y = X\beta + Zu + e \qquad (5)$$

where $y = (y'_a, y'_b, y'_c \ldots, y'_z)'$; $u = (u'_a, u'_b, \ldots, u'_z)'$; $e = (e'_a, e'_b, e'_c, \ldots, e'_z)'$

where *y* is a long vector of same trait in locations *a, b … z*, recorded for *n* clones, **X** and **Z** design matrices are block diagonal matrices represented as *diag*(X_a, X_b,..., X_z) and *diag*(Z_a, Z_b,..., Z_z) respectively allowing for missing clones and observations. **X** is a design matrix for fixed effects *β* (with components as the grand mean and year effect) and **Z** is a design matrix for random genomic effects *u*. Following a multivariate normal distribution (N_m), the marginal

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

density of *y* is given as:

$$(y \mid b,\ R,\ G)\ \sim N_m(X\beta,\ V) \qquad (6)$$

$$V\ =\ Z(G\ \otimes K)Z^T\ +\ I_d \otimes R)\,; \qquad \hat{u}\ = (K \otimes G)Z^T V^{-1}(y - X\hat{\beta})$$

where *d* is number of traits being analyzed, G and R are *d* × *d* symmetric genomic and error covariance matrices respectively, *K* remains the additive genomic relationship matrix for *n* clones generated from SNP markers as in VanRaden, 2008 [63] method two, *I* is a *d*-traits identity matrix and *u* are the genomic estimated breeding values (GEBVs) the clones and for the traits in the analysis. Estimation of the parameters in model (6) were also performed using REML procedure implemented in the *airemlf90* program [62] from which BLUEs of fixed effects and BLUPs of random effects were also obtained by solving the mixed model equations (MME) [5,6]. In this model, the error covariance matrix R is a diagonal matrix allowing heterogeneous variances of a trait for different locations but the covariances are restricted to zero following the assumption that no mechanism generates errors between a trait at multiple locations.

Comparison of prediction accuracies:

We used a 5-fold cross validation scheme with 10 repeats for comparisons between the univariate and and multivariate models. The same folds were used for the models in each scenario. We hereafter refer to predicted BLUPs or genomic effects from these models as Genomic EBVs (GEBVs). Prediction accuracies were calculated as a correlation of the validation fold GEBVs to their corresponding EGVs.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results:

Scenario 1: MT vs uT models:

In Scenario 1, we observed that the prediction accuracies of the MT model were higher than those from the uT models for most traits and locations in our analysis (Table 2). On average (across traits and locations), the MT model had 35 percent better predictive power for VIGOR, 34 percent for RTNO, 19 percent for DM, 33 percent for MCMDS, 54 percent for FYLD and 15 percent for MCGM compared to the uT model. The MT models had 32 percent higher predictive power than the uT models for all traits and locations in the model.

Scenario 2: ME vs uE models:

In Scenario 2 , we observed different patterns of prediction accuracies of the uE and ME models. The ME model had higher prediction accuracies for DM and MCMDS at all locations. On average (across locations), the uE model had 2 percent better predictive power for VIGOR and 1 percent for RTNO while the ME model had 32 percent better predictive power for DM, 24 percent for MCMDS, 5 percent for FYLD and lastly 4 percent for MCGM. The ME models had 12 percent higher predictive power than the uE model for all traits and locations in the model.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion:

Scenario 1: MT vs uT model:

Few comparisons between MT and uT models have previously been reported in simulation studies and also real data sets [57-59]. Guo et al., 2014 [59] and Calus et al., 2011 [58] in their studies with simulated data sets reported similar accuracies with small differences between their MT and uT models where accuracies for the MT models for low heritability traits were slightly higher when genomic correlations between the traits increased. VanRaden et al., 2014 [57] in Table 10 of their paper (using Holstein and Jersey breed datasets from the US Dairy National evaluation program) also reported similar accuracies with small differences between their MT and uT models for all the traits in their analysis. In several traits, their uT model accuracies were slightly higher than those of their MT model. Accuracies from the MT model will not be clearly better than those from the uT model for traits with high heritability and for traits whose complete phenotypic data are available [59]. Improvement in prediction accuracies for the MT model is accrued mostly for low heritability traits analysed jointly with high heritability traits that have medium to high genomic correlations and low residual correlations [58, 59]. Our results were consistent with other studies [58,59] where our MT model had higher accuracies for most traits and locations in our analysis as a result of joint analysis of low heritability traits with other traits of higher heritabilities. We also observed from pooling data that more efforts on a trial design that replicates lines across multiple locations may improve the predictive power of these models. For parental selections in specific locations, we recommend the use of MT models. However, for varietal selections where non-additive genomic effects are useful, further analysis will have to be performed using data

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

from fully replicated trials at multiple locations (termed uniform yield trials in cassava breeding) to inform this decision. Further studies on the selection gains based on these models are also recommended to facilitate a more accurate decision.

Scenario 2: ME vs uE model:

Again few comparisons of the ME and uE models have been done in plant breeding literature. However, Burgueno et al., 2012 [35] conducted extensive modeling for multi-environment trials using pedigree and genomic markers and incorporated many covariance structures including diagonal, factor analytic (FA), identity and unstructured covariances for both the genomic and error components in their models. They observed higher prediction accuracies for their genomic ME model with a diagonal genomic covariance structure and a diagonal error covariance structure ($ME_{D\text{-}D}$) compared to their genomic ME model with a FA genomic covariance structure and diagonal error covariance ($ME_{FA\text{-}D}$) for most of the locations in their analysis based on their cross-validation scheme (CV1) [35]. This $ME_{D\text{-}D}$ can be likened to our one-step uE model which combined data for a trait measured at multiple locations together and assumes same genomic and error variances for this trait at all locations analyzed and treating multiple location data as just replicates. Our results were in line with this study for the traits VIGOR and RTNO at all our locations where the uE model had higher prediction accuracies than the ME models and differed from this study for the traits DM and MCGM at all locations, FYLD at Ubiaja and Ibadan where the ME models had higher prediction accuracies than those of the uE model. Interestingly on average across locations, the ME models had better predictive power for the most valuable traits in cassava (a combination of DM and FYLD; so called dry yield). Like in other crops, there is need to genetically improve these

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

production (yield) traits in order to drive the economies of the regions that rely on cassava production. Genetic improvement can be made on these cassava target traits using a genetic evaluation system based on multivariate mixed models. The ME model exploits the positive genomic and residual correlations captured in the their *G* and *R* matrices for prediction. The differences between the prediction accuracies of these models (ME and uE) were mainly due to the the estimation of these covariances since their genomic variances were very similar. In addition, the genomic covariances for the ME models are a reflection of GxE interactions of the trait of interest.

Given these genomic covariances from the ME model's G matrix, we observed differing degrees of GxE (observing the genomic correlations which were less than 0.9) at the locations Ubiaja, Mokwa and Ibadan (Table 3B). Interestingly, the performance of clones for the traits VIGOR, DM and CMD were most similar between Ubiaja and Ibadan while for the production traits RTNO and FYLD clonal performances were most similar between the locations Ubiaja and Mokwa. Lastly for the performance of clones for MCGM, Ibadan and Mokwa were most similar. A very naive implication of this is that there are no defined clusters for regional breeding by trait and regions as genomic correlations from the ME model partitioned these traits into different categories with different location pairs (Table 3B). From these estimates, improvement for RTNO and FYLD at Ubiaja may translate to about 50 percent stability for these traits at Mokwa but at Ibadan these may translate to about 35 percent. This makes a case for decentralized breeding especially for these yield component traits. These also revealed that cassava DM and MCMDS were highly repeatable across the locations in our study suggesting that more progress can easily be made from selection on

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

these traits and that genotypes selected for these traits will perform appreciably similar across the locations in our analysis and other very closely related locations not in our analysis. However for RTNO, FYLD, VIGOR and MCGM, their GxE profiles reveal that the environment exacts more influence on these traits making their improvement more challenging. Breeding for good varieties which combine these core traits may be targeted towards regional or environment clusters with specific genotypes selected for these clusters. We could also obtain from the GEBVs of the ME model fits a typical reflection of how the clones in our model interacted with the different locations analyzed trait by trait. However, ME GEBVs lack information from trait-trait covariances which has been a difficult hurdle to cross when selecting for multiple traits simultaneously using MET data.

Parameter estimates and implications for cassava breeding:

The estimates of genomic correlations and heritabilities shown in Table 3A have interesting implications for cassava genetic improvement. Most of the target cassava traits (VIGOR, RTNO, FYLD and MCGM) have low heritabilities and hence should benefit from joint analysis with traits of higher heritabilities (DM and MCMDS) in an MT model. The genomic correlations for the core production traits including DM, RTNO and FYLD indicates a promising horizon for improvement. RTNO and FYLD have high positive genomic correlation (0.69) as well as a medium correlation for RTNO and DM (0.17) while that for FYLD and DM is rather weak or almost neutral (-0.02). This indicates that these traits can be improved concurrently with efforts towards fully replicated trials and more accurate parental selections given the low heritabilities for FYLD and RTNO. VIGOR can also be improved concurrently with these production traits as it is also positively correlated with these (Table 3A). The disease trait

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(MCMDS) shows strong negative genomic correlations with VIGOR and the production traits which raises legitimate concerns towards cassava breeding in Africa where the cassava mosaic disease (CMD) pressure is high. Good progress have been made to fix genes for CMD resistance [64] but a good question to answer is whether this is bringing about negative results and limitations for other traits especially the production traits. This leads to the effort for searching for more quantitative resistance for the diseases CMD and the pest CGM which may better allow for improvement of the traits VIGOR, RTNO, DM and FYLD. Here, the merit index from MT breeding values will offer a good alternative towards a balanced selection goal because it takes into account the genomic correlations of these traits.

Conclusion:

The effectiveness and success of a breeding program is generally evaluated by its impact in its locale mainly on its ability to provide highly adapted and productive varieties to the farming community it serves. This is a huge task to a breeder who will try to fulfil this mandate by making subtle and accurate evaluations of the performance of all the genotypes in his breeding program and subsequently making critical selection decisions based on these evaluations. In our case at the Cassava breeding program at IITA, we are scouting for highly predictive models for use in accurate evaluations of the performance of all our genotypes and hence efficient implementation of GS in our program. In this paper, we have compared methods utilized for this process (in Scenario 1) and compared the efficiencies of these models using prediction accuracies, an indispensable parameter for prediction of selection gains. We also compared models accounting for GxE for genotype evaluations at multiple environments in Scenario 2. We have done these using sixteen years data measured on core

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

or target cassava traits. Genetic evaluation which consists of accurate estimation of breeding values for all individuals in a population for the desired traits of interest and subsequently combination of these breeding values in a total merit index is the crux of genetic improvement. However, the models implemented for estimation of breeding values affect the total merit index and subsequently response to selection. We have shown here that there is value in accounting for covariances using multivariate mixed models for METs data. This value will translate into superior total merit index and hence more selection gain when compared to univariate mixed models. We have also shown that the genomic covariance matrix for ME models can show the effects of GxE for all the traits in the analysis thus helping the breeder to make better decisions. More importantly, we have shown that the MT model has on average 32 percent better predictive power than the uT model while offering more accurate breeding or genetic values for use in total merit indices. We therefore recommend the use of multivariate models (MT and ME) for genetic evaluation in the breeding program where ME is used to assess GxE interactions and MT breeding or genetic values used to set up total merit indices for selection purposes.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Declarations:

Author's contributions:

UGO designed, carried out study and drafted manuscript, DA provided statistical assistance and advice, JLJ supervised the study and revised manuscript, IR and PK supervised data generation for this study and revised manuscript. All authors read and approved manuscript.

Acknowledgements:

We acknowledge the Bill & Melinda Gates Foundation and UKaid (Grant 1048542; http://www.gatesfoundation.org) and support from the CGIAR Research Program on Roots, Tubers and Bananas (http://www.rtb.cgiar.org). We give special thanks to A. G. O. Dixon for his development of many of the breeding lines and historical data we analyzed. Thanks also to A. I. Smith and technical teams at IITA for collection of other phenotypic data and to A. Agbona for data curation.

Competing interests

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References:

[1] The International Plant Names Index (IPNI). http://www.ipni.org/ipni/idPlantNameSearch.do?id=351790-1. Accessed October 31, 2015.

[2] Taylor, N., Chavarriaga, P., Raemakers, K., Siritunga, D., & Zhang, P. (2004). Development and application of transgenic technologies in cassava. *Plant Molecular Biology*, *56*(4), 671-688.

[3] Moorthy, S. N. (2002). Physicochemical and functional properties of tropical tuber starches: a review. *Starch-Stärke, 54*(12), 559-592.

[4] Balagopalan, C. (2002). Cassava utilization in food, feed and industry.*Cassava: Biology, production and utilization*, 301-318.

[5] Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pp.423-447.

[6] Henderson, C. R. "Estimation of variances and covariances under multiple trait models." Journal of Dairy Science 67, no. 7 (1984): 1581-1589.

[7] Smith, H.F. 1936. A discriminant function of plant selection. Ann. Eugenics, 7: 240-250.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[8] Hazel, Lanoy Nelson. "The genetic basis for constructing selection indexes." Genetics 28, no. 6 (1943): 476-490.

[9] Ducrocq, V.; Wiggans, G.; Garrick, D. J.; Ruvinsky, A.; CABI, Wallingford, UK, The genetics of cattle, 2015, Ed. 2, pp 380-384.

[10] Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Prediction of total genetic value using genome wide dense marker maps. Genetics, 157, 1819–1829.

[11] Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. "Invited review: Genomic selection in dairy cattle: Progress and challenges." Journal of dairy science 92, no. 2 (2009): 433-443.

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

[13] VanRaden, P.M. and Wiggans, G.R. (1991). Derivation, calculation, and use of national animal model information. Journal of Dairy Science 74, 2737-2746.

[14] Goddard, Mike. "Genomic selection: prediction of accuracy and maximisation of long term response." Genetica 136.2 (2009): 245-257.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[15] Van der Werf, J., Van Arendonk, J.A.M., de Vries, A.G., 1992. Improving selection of pigs using correlated characters. 43rd Ann. EAAP meeting, selection of pigs using correlated characters. 43rd Ann. EAAP meeting, Madrid, Spain.

[16] Ducrocq, V., 1994. Multiple trait prediction: principles and problems. in Proc. 5th World Congr. Genet. App. Livest. Prod., Guelph, Ontario, Canada. Vol. 18, 455-462.

[17] Colleau, Jean-Jacques, Vincent Ducrocq, Didier Boichard, and Hélène Larroque. "Approximate multi-trait BLUP evaluation to combine functional traits information." Interbull Bulletin 23 (1999): 151.

[18] Schaeffer, L. R. "Multiple-country comparison of dairy sires." Journal of Dairy Science 77, no. 9 (1994): 2671-2678.

[19] Schaeffer, L. R. "Multiple trait international bull comparisons." Livestock Production Science 69, no. 2 (2001): 145-153.

[20] Thompson, R., and K. Meyer. "A review of theoretical aspects in the estimation of breeding values for multi-trait selection." Livestock Production Science 15, no. 4 (1986): 299-313.

[21] Ducrocq V, Boichard D, Barbat A, Larroque H. Implementation of an approximate multi-trait BLUP evaluation to combine production traits and functional traits into a total merit

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

index. In Proceedings of the 52nd Annual Meeting of the European Association for Animal Production: 26-29 August 2001; Budapest; 2001.

[22] Lassen J, Sørensen MK, Madsen P, Ducrocq V. A stochastic simulation study on validation of an approximate multitrait model using preadjusted data for prediction of breeding values. J Dairy Sci. 2007;90:3002–11.

[23] Lassen J, Sørensen MK, Madsen P, Ducrocq V. An approximate multi-trait model for genetic evaluation in dairy cattle with a robust estimation of genetic trends. Genet Sel Evol. 2007;39:353–67.

[24] Segelke, D., Reinhardt, F., Liu, Z. and Thaller, G., 2014. Prediction of expected genetic variation within groups of offspring for innovative mating schemes.*Genetics Selection Evolution*, *46*(1), p.1.

[25] Habier, D., Fernando, R. L., & Dekkers, J. C. (2008, 02). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. doi:10.1534/genetics.107.081190

[26] Malosetti, M., Ribaut, J.M. and van Eeuwijk, F.A., 2014. I. 6 The statistical analysis of multienvironment data: modelling genotype-by-environment interaction and its genetic basis. *Drought phenotyping in crops: From theory to practice*, *4*(44), p.53.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[27] Halsey, M.E., Olsen, K.M., Taylor, N.J. and Chavarriaga-Aguirre, P., 2008. Reproductive Biology of Cassava (Crantz) and Isolation of Experimental Field Trials. *Crop science*, *48*(1), pp.49-58.

[28] Kawano, K., Amaya, A., Daza, P. and Rios, M., 1978. Factors affecting efficiency of hybridization and selection in cassava. *Crop Science*, *18*(3), pp.373-376.

[29] Kawano, K. 1980. Cassava. p. 225–233. *In* W.R. Fehr and H.H. Hadley (ed.) Hybridization of crop plants. ASA and CSSA, Madison, WI.

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

[31] Falconer, D.S. 1952. The problem of environment and selection. Am. Nat. 86:293-298

[32] Crossa, J., J. Burgueño, P.L. Cornelius, G. McLaren, R. Trethowan, and A. Krishnamachari. 2006. Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop Sci.46:1722–1733. doi:10.2135/cropsci 2005.11-0427

[33] Burgueño, J., J. Crossa, P.L. Cornelius, and R.-C. Yang. 2008. Using factor analytic models for joining environments and genotypes without crossover genotype × environment

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

interaction. Crop Sci. 48:1291–1305. doi:10.2135/cropsci 2007.11.0632

[34] Burgueño, J., J. Crossa, J. Miguel Cotes, F. San Vicente, and B. Das. 2011. Prediction assessment of linear mixed models for multienvironment trials. Crop Sci. 51:944–954. doi:10.2135/cropsci 2010.07.0403

[35] Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science, 52(2), 707-719.

[36] de los Campos, G., and D. Gianola. 2007. Factor analysis models for structuring covariance matrices of additive genetic effects: A Bayesian implementation. Genet. Sel. Evol. 39:481–494. doi:10.1186/1297-9686-39-5-481

[37] Meyer, Karin. "Factor-analytic models for genotype x environment type problems and structured covariance matrices." *Genet Sel Evol* 41, no. 11 (2009).

[38] Tsuruta, S., Lourenco, D. A. L., Misztal, I., & Lawlor, T. J. (2015). Genotype by environment interactions on culling rates and 305-day milk yield of Holstein cows in 3 US regions.*Journal of Dairy Science*.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[39] Kolmodin, R., F. Strandberg, P. Madsen, J. Jensen, and H. Jorji-ani. 2002. Genotype by environment interaction in Nordic dairy Acta Agric. Scand. Anim. cattle studied using reaction norms. Sci. 52:11–24.

[40] Windig, J. J., M. P. L. Calus, and R. F. Veerkamp. 2005. Influence of herd environment on health and fertility and their relationship with milk production. J. Dairy Sci. 88:335–347.

[41] Cooper, M., and I.H. DeLacy. 1994. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theor. Appl. Genet. 88(5): 561–572.

[42] Malosetti, Marcos, Jean Marcel Ribaut, Mateo Vargas, José Crossa, and Fred A. Van Eeuwijk. "A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.)." *Euphytica* 161, no. 1-2 (2008): 241-257.

[43] Alimi, N. A., M. C. A. M. Bink, J. A. Dieleman, J. J. Magán, A. M. Wubs, A. Palloix, and F. A. van Eeuwijk. "Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper." *Theoretical and applied genetics* 126, no. 10 (2013): 2597-2625.

[44] Garrick, D. J., Taylor, J. F., & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses.Genet Sel Evol, 41(55),

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

10-1186.

[45] Ostersen, Tage, Ole F. Christensen, Mark Henryon, Bjarne Nielsen, Guosheng Su, and Per Madsen. "Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs." *Genetics Selection Evolution* 43, no. 1 (2011): 1-6.

[46] Zhou, Xiang, and Matthew Stephens. "Efficient multivariate linear mixed model algorithms for genome-wide association studies." Nature methods 11, no. 4 (2014): 407-409.

[47] Deniz Akdemir and Okeke Uche Godfrey (2014). EMMREML: Fitting mixed models with known covariance structures. R package version 2.0. http://CRAN.R-project.org/package=EMMREML

[48] Maziya-Dixon, B., A.G.O. Dixon, and A.-R.A. Adebowale. 2007. Targeting different end uses of cassava: Genotypic variations for cyanogenic potentials and pasting properties. Int. J. Food Sci. Technol. 42:969–976. doi:10.1111/j.1365-2621.2006.01319.x

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

[50] Ly, Delphine, et al. "Relatedness and genotype× environment interaction affect prediction

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

accuracies in genomic selection: A study in cassava." Crop Science53.4 (2013): 1312-1325.

[51] Glaubitz, J., T. Casstevens, R. Elshire, J. Harriman, and E.S. Buckler. 2012. TASSEL 3.0 genotyping by sequencing (GBS) pipeline documentation. Edward S. Buckler, USDA-ARS, Ithaca, NY. http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf (accessed 3. Jan. 2014)

[52] Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., ... & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, *40*(D1), D1178-D1186. (http://www.phytozome.net; accessed 1 July, 2014).

[53] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33, no. 1 (2010): 1.

[54] R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (January, 2014)

[55] Douglas Bates, Martin Maechler, Ben Bolker and Steven Walker (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5. http://CRAN.R-project.org/package=lme4.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[56] Pfeiffer, Christina, Birgit Fuerst-Waltl, Hermann Schwarzenbacher, Franz Steininger, and Christian Fuerst. "A comparison of methods to calculate a total merit index using stochastic simulation." Genetics Selection Evolution 47, no. 1 (2015): 36.

[57] VanRaden, P. M., Tooker, M. E., Wright, J. R., Sun, C., & Hutchison, J. L. (2014). Comparison of single-trait to multi-trait national evaluations for yield, health, and fertility. *Journal of dairy science*, *97*(12), 7952-7962.

[58] Calus, M. P., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet Sel Evol*, *43*(1), 1-14.

[59] Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., & Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC genetics*, *15*(1), 30.

[60] Ceballos, H., Kawuki, R.S., Gracen, V.E., Yencho, G.C. and Hershey, C.H., 2015. Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. *Theoretical and Applied Genetics*, *128*(9), pp.1647-1667.

[61] Wimmer, V., Albrecht, T., Auinger, H.J. and Schön, C.C., 2012. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*,*28*(15), pp.2086-2087.

[62] Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. Lee. 2002. BLUPF90 and related programs (BGF90). Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier,

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

France. CD-ROM Communication 28:07

[63] VanRaden, P.M., 2008. Efficient methods to compute genomic predictions.*Journal of dairy science*, *91*(11), pp.4414-4423.

[64] Jennings, D.L., 1972. Breeding for resistance to African cassava mosaic disease: Progress and prospects.

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: Cassava phenotype data metrics showing number of records, number of observed clones and mean of traits for 3 locations Ubiaja, Mokwa and Ibadan.

	Ubiaja	**Mokwa**	**Ibadan**
No. of records	7806	5345	6010
No. of Clones	739	573	691
VIGOR	6.51	6.52	6.11
RTNO	31.71	64.05	52.74
FYLD	12.61	16.51	15.84
DM	31.95	29.01	35.8
MCMDS	1.59	1.21	1.93
MCGM	3.56	2.99	3

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2: Cross validation prediction accuracies for GS models in scenarios 1 and 2.

Table 2 shows prediction accuracies for MT and uT models (GS scenario 1) and for ME and uE models (GS scenario 2).

GS Scenario 1						
	Single trait single env. model (uT)			Multi-trait models (MT)		
	Ubiaja	Mokwa	Ibadan	Ubiaja	Mokwa	Ibadan
VIGOR	0.25 ± 0.01	0.11 ± 0.01	0.38 ± 0.02	0.39 ± 0.02	0.09 ± 0.03	0.52 ± 0.03
RTNO	0.27 ± 0.01	0.19 ± 0.01	0.31 ± 0.01	0.39 ± 0.02	0.21 ± 0.03	0.43 ± 0.02
DM	0.59 ± 0.01	0.33 ± 0.01	0.49 ± 0.01	0.69 ± 0.01	0.39 ± 0.02	0.60 ± 0.02
MCMDS	0.49 ± 0.01	0.43 ± 0.01	0.56 ± 0.02	0.68 ± 0.02	0.58 ± 0.04	0.71 ± 0.01
FYLD	0.32 ± 0.01	0.13 ± 0.02	0.29 ± 0.01	0.46 ± 0.01	0.26 ± 0.03	0.42 ± 0.03
MCGM	0.36 ± 0.01	0.50 ± 0.01	0.56 ± 0.01	0.44 ± 0.02	0.54 ± 0.01	0.65 ± 0.01
GS Scenario 2						
	Single trait multi-env. model (uE)			Multi-env model (ME)		
	Ubiaja	Mokwa	Ibadan	Ubiaja	Mokwa	Ibadan
VIGOR	0.22 ± 0.01	0.10 ± 0.01	0.37 ± 0.01	0.24 ± 0.01	0.12 ± 0.02	0.32 ± 0.01
RTNO	0.29 ± 0.01	0.11 ± 0.01	0.34 ± 0.01	0.27 ± 0.02	0.13 ± 0.01	0.34 ± 0.02
DM	0.49 ± 0.01	0.20 ± 0.02	0.40 ± 0.01	0.60 ± 0.01	0.35 ± 0.01	0.50 ± 0.01
MCMDS	0.40 ± 0.01	0.23 ± 0.01	0.53 ± 0.01	0.48 ± 0.01	0.39 ± 0.02	0.57 ± 0.01
FYLD	0.38 ± 0.01	0.10 ± 0.02	0.35 ± 0.01	0.37 ± 0.01	0.12 ± 0.03	0.36 ± 0.02
MCGM	0.31 ± 0.01	0.48 ± 0.01	0.56 ± 0.01	0.38 ± 0.02	0.47 ± 0.01	0.55 ± 0.01

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3A: Genetic correlations and heritabilities for analyzed traits.

Table 3A shows plot-based heritabilities for six cassava traits on diagonals respectively, genetic correlations from the MT model as lower triangular values.

	VIGOR	RTNO	DM	MCMDS	FYLD	MCGM
VIGOR	**0.16**					
RTNO	0.53	**0.15**				
DM	0.22	0.17	**0.40**			
MCMDS	-0.64	-0.56	-0.17	**0.64**		
FYLD	0.52	0.69	-0.02	-0.47	**0.16**	
MCGM	-0.07	-0.037	-0.08	0.13	-0.05	**0.12**

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3B: Genetic correlations from the multi-environment analysis. Table 3B show genetic correlations from the ME model for six core cassava traits.

VIGOR		**Ubiaja**	**Mokwa**
	Ubiaja		
	Mokwa	0.39	
	Ibadan	0.66	0.21
RTNO			
	Ubiaja		
	Mokwa	0.54	
	Ibadan	0.36	0.38
DM			
	Ubiaja		
	Mokwa	0.57	
	Ibadan	0.81	0.77
MCMDS			
	Ubiaja		
	Mokwa	0.8	
	Ibadan	0.87	0.68
FYLD			
	Ubiaja		
	Mokwa	0.52	
	Ibadan	0.31	0.33
MCGM			
	Ubiaja		
	Mokwa	0.34	
	Ibadan	0.24	0.53

bioRxiv preprint doi: https://doi.org/10.1101/116301; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.