G3: Genes|Genomes|Genetics Early Online, published on August 31, 2016 as doi:10.1534/g3.116.033332

Marker-based estimates reveal significant non-additive effects in clonally propagated cassava (*Manihot esculenta*): implications for the prediction of total genetic value and the selection of varieties

Marnin D. Wolfe[*], Peter Kulakow[†], Ismail Y. Rabbi[†], Jean-Luc Jannink[*,‡]

[*] Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA
[†] International Institute for Tropical Agriculture (IITA), Ibadan, Oyo, Nigeria
[‡] USDA-ARS, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA

Running Title: Marker-based estimation of non-additive effects in cassava

Key words: genomic selection, non-additive effects, cassava

Corresponding Author:
Marnin D. Wolfe
Cornell University
Department of Plant Breeding and Genetics
417 Bradfield Hall
306 Tower Road
Ithaca, NY 14853
Email: wolfemd@gmail.com
Phone: 239-595-5081

ABSTRACT

In clonally propagated crops, non-additive genetic effects can be effectively exploited by the identification of superior genetic individuals as varieties. Cassava (*Manihot esculenta* Crantz) is a clonally propagated staple food crop that feeds hundreds of millions. We quantified the amount and nature of non-additive genetic variation for three key traits in a breeding population of cassava from sub-Saharan Africa using additive and non-additive genome-wide marker-based relationship matrices. We then assessed the accuracy of genomic prediction for total (additive plus non-additive) genetic value. We confirmed previous findings based on diallel populations, that non-additive genetic variation is significant for key cassava traits. Specifically, we found that dominance is particularly important for root yield and epistasis contributes strongly to variation in CMD resistance. Further, we showed that total genetic value predicted observed phenotypes more accurately than additive only models for root yield but not for dry matter content, which is mostly additive or for CMD resistance, which has high narrow-sense heritability. We address the implication of these results for cassava breeding and put our work in the context of previous results in cassava, and other plant and animal species.

INTRODUCTION

Understanding genetic architecture requires the decomposition of genetic variance into additive, dominance, and epistatic components (Fisher 1918; Cockerham 1954; Kempthorne 1954). However, partitioning genetic variance components is notoriously difficult, requiring specialized breeding designs (e.g. diallel crosses) and pedigree information, (Lynch and Walsh 1998) often limiting the genetic diversity that can be sampled in any one given study. Genome-wide molecular marker data now enable the accurate measurement of relatedness in the form of genomic realized relationship matrices (GRMs; VanRaden 2008; Heffner *et al.* 2009; Lorenz *et al.* 2011). GRMs, in contrast to pedigrees directly measure Mendelian sampling (variation in relatedness within relatedness classes such as full-siblings; Heffner *et al.* 2009). Further, GRMs can measure relationships even in diverse, nominally unrelated samples expanding the potential for studying inheritance in natural and breeding populations (Lorenz *et al.* 2011).

Estimation of narrow-sense heritability and prediction of breeding values in genomic selection programs is becoming increasingly common using additive formulations of GRMs (Visscher *et al.* 2008). Several recent studies have described dominance and epistatic GRMs for the partitioning of non-additive genetic variance using genome-wide SNP markers (Su *et al.* 2012; Vitezica *et al.* 2013; Muñoz *et al.* 2014; Wang *et al.* 2014). Models using these new formulations have been shown to provide improved partitioning of genetic variances relative to pedigree-based approaches (Su *et al.* 2012; Muñoz *et al.* 2014). These new models can be used not only to estimate genetic variances but also for genomic prediction of total genetic value in genomic selection

69 breeding programs (Su *et al.* 2012; Vitezica *et al.* 2013; Muñoz *et al.* 2014; Wang *et al.*
70 2014).

71 Cassava is a vegetatively propagated, staple food crop that is high in starch and
72 feeds half a billion people worldwide (http://faostat.fao.org). Efforts to improve cassava
73 genetically with cutting edge methodologies including transgenic and genomic selection
74 (GS) approaches are underway thanks to new genomic resources (Prochnik *et al.* 2012;
75 ICGMC, 2015). Prediction with additive models has recently been evaluated (Oliveira *et*
76 *al.* 2012; Ly *et al.* 2013) and genomic selection using standard models is currently being
77 tested **(**http://www.nextgencassava.org**)**. Vegetatively propagated crop (e.g. cassava)
78 breeding can exploit non-additive genetic effects by identifying superior clones as
79 varieties (Ceballos *et al.* 2015).

80 Diallelic studies in cassava indicate that non-additive genetic effects (e.g. specific
81 combining ability) are strong, particularly for root yield traits (Cach *et al.* 2005, 2006;
82 Calle *et al.* 2005; Jaramillo *et al.* 2005; Pérez, Ceballos, Calle, *et al.* 2005; Pérez,
83 Ceballos, Jaramillo, *et al.* 2005; Zacarias and Labuschagne 2010; Kulembeka *et al.* 2012;
84 Tumuhimbise *et al.* 2014; Ceballos *et al.* 2015; Chalwe *et al.* 2015). If the limited
85 number of parents tested thus far represents the broader cassava breeding germplasm,
86 genetic gains, especially for already low-heritability root yield traits will be slow
87 regardless of the breeding scheme employed (e.g. phenotypic vs. pedigree vs. genomic
88 selection). Breeding gains have indeed been slow in cassava (Ceballos *et al.* 2012) and
89 low accuracies have been reported for genomic prediction of yield compared with
90 cassava mosaic disease (CMD) resistance and dry matter (DM) content (Oliveira *et al.*
91 2012; Ly *et al.* 2013). However, cassava varieties are evaluated and disseminated to

92 farmers by clonal propagation, meaning that accurate prediction of total (additive plus
93 non-additive) genetic value could contribute to variety selection.
94 In this study, we test whether certain cassava traits, especially root yield have
95 relatively large non-additive genetic variances that account for low genomic prediction
96 accuracies previously observed. We estimate additive and non-additive variance
97 components using genomic relationship matrices in two datasets of cassava from the
98 International Institute of Tropical Agriculture's (IITA) genomic selection breeding
99 program. Further, we assess the accuracy of predicting total genetic value using the
100 additive and non-additive models. We discuss the origin of non-additive genetic variance
101 in cassava, its potential effect on cassava breeding, and its role in genomic selection
102 strategies for cassava improvement in the future.
103
104 **METHODS**
105 **Germplasm and Phenotyping Trials**
106 We examined additive and non-additive effects in two datasets of cassava that have
107 been genotyped and phenotyped as part of the Next Generation Cassava Breeding
108 Program at IITA, Nigeria (http://www.nextgencassava.org). The IITA's Genetic Gain
109 (GG) collection contains 694 historically important clones, most of which are advanced
110 breeding lines although some are classified as superior landraces. These lines have been
111 selected and maintained clonally since 1970 (Okechukwu and Dixon 2008; Ly *et al.*
112 2013). Most of these materials are derived from the cassava gene pool from West Africa
113 as well as parents derived from the breeding program at Amani Station in Tanzania and

114 hybrids of germplasm introduced from Latin America. Available information on the GG
115 accessions included in our analyses is provided in Table S1.
116 IITA's Genetic Gain trials were conducted in seven locations over 14 years (2000 to
117 2014) in Nigeria for a total of 24,373 observations. Each GG trial comprises a
118 randomized, incomplete block design replicated one or two times per location and year.
119 Since materials have been occasionally lost and new, selected materials are continuously
120 added to the GG, the number of clones trialed in a given year changes gradually across
121 years, generally increasing. The sample sizes, number of replicates and number of clones
122 from the GG in each of the trials (location-year combinations) are provided in Table S2.
123 Theory suggests that founding events and truncation selection can both lead to a
124 conversion of non-additive genetic variation into additive variance. This can happen
125 because of the induction of linkage disequilibrium and reduction in allele frequency (or
126 fixation of alleles) at some loci relative to others (Goodnight 1988; Turelli and Barton
127 2006; Hallander and Waldmann 2007). Consequently, our results might depend on the
128 dataset examined. We therefore analyzed an additional dataset: a collection of 2187
129 clones that are the direct descendants of truncation selection on the GG. Briefly, in 2012
130 the GG and all available historical phenotype data was used as a reference dataset to
131 obtain genomic estimated breeding values (GEBVs) using the genomic BLUP (GBLUP)
132 model (VanRaden 2008; Heffner *et al.* 2009). Selection was based on an index that
133 included mean cassava mosaic disease severity (MCMDS), mean cassava bacterial blight
134 disease severity (MCBBS), dry matter content (DM), harvest index (HI) and fresh root
135 weight (RTWT). This index of GEBVs was used to select 83 members of the GG to cross
136 and generated a collection of 135 full-sib families, which we refer to as the GS Cycle 1

(C1). In the C1, family sizes are 18.3 on average (median 14, range from 3 to 82). Parents have an average of 59.5 progeny (median 38, range from 5 to 406). The pedigree of the C1 is available in Table S3. Further, information about the germplasm analyzed, including data regarding the genetic structure of the population have been published previously (Wolfe *et al.* 2016), however we also provide plots of the first four principal components of the additive genetic relationship matrix (see below) in the supplement (Figure S1).

Cycle 1 progenies were evaluated in a single clonal evaluation trial during the 2013-2014 field season across three-locations (Ibadan, Ikenne, and Mokwa). For the C1 clonal trial, planting material was only available for one plot of five stands per clone, so each clone was only planted in one of the three locations (Table S2). Clones were assigned to each location so as to equally represent each family in every environment.

For both datasets, we analyzed three traits: MCMDS, DM and RTWT. MCMDS was scored on a scale of 1 (no symptoms) to 5 (severe symptoms). We note that the distribution of MCMDS is skewed towards low disease severity (Figure S2). Most GG trials measured dry matter (DM) by the oven drying method although some trials used the specific gravity method. Dry matter content (DM) is expressed as a percentage of the fresh weight of roots. Fresh root weight (RTWT) is measured in kilograms per plot and is natural-log transformed to achieve normally distributed, homoscedastic residuals in all presented analyses. Trait distributions are presented in Figure S2.

Genotype data

We used genotyping-by-sequencing (GBS) to obtain genome-wide SNP marker data (Elshire *et al.* 2011). We used the ApeKI restriction enzyme as recommended by (Hamblin and Rabbi 2014). SNPs were called using the TASSEL V4 GBS pipeline (Glaubitz *et al.* 2014) and aligned to the cassava reference genome, version 5, which is available on Phytozome (http://phytozome.jgi.doe.gov) and described by the International Cassava Genetic Map Consortium (ICGMC, 2015). We removed individuals with >80% missing and markers with >60% missing genotype calls. Also, markers with extreme deviation from Hardy-Weinberg equilibrium (Chi-square > 20) were removed. If there were not at least two reads at a given locus for a given clone, the genotype was set to missing and imputed. SNP marker data was converted to the dosage format (-1 for reference-allele homozygotes, 0 for heterozygotes and +1 for alternative-allele homozygotes) and missing data were imputed with the glmnet algorithm in R (http://cran.r-project.org/web/packages/glmnet/index.html). Similar to the approach of Wong *et al.* (2014), for each marker to be imputed, we pre-selected the 60 markers on the same chromosome in highest LD. We then used these pre-selected markers to predict missing values using the LASSO (default, q=1 in glmnet), with the tuning parameter lambda selected by five-fold cross-validation. We used 114,922 markers that passed these filters with a minor allele frequency greater than 1% to construct genomic relationship matrices as described below.

Genomic Relationship Matrices

In order to capture additive effects variance, we constructed the genomic relationship matrix (**G**) using the formula of VanRaden (2008), method one: $\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_i p_i q_i}$. Here **Z** is a

mean-centered matrix of dimension *n* individuals by *m* SNP markers. To obtain **Z**, we subtract $2(p_i - 0.5)$ from the marker dosages, where the dosages are coded -1 for aa, 0 for Aa, +1 for AA, p_i is the frequency of the second allele (A) at the *i*th locus and $q_i = 1\text{-}p_i$. The a (or 0) allele refers to the reference genome allele. The **G** matrix was calculated using the *A.mat* function in the *rrBLUP* package (Endelman 2011).

We constructed a matrix to capture dominance relationships using the formulation originally proposed by Su et al. (2012). The dominance relationship matrix we will call, **D*** (see below) is $\mathbf{D}^* = \frac{\mathbf{HH}'}{\sum_i 2p_iq_i(1-2p_iq_i)}$. Where **H** is a mean-centered dominance deviation matrix with the same dimensions as **Z**. To obtain **H**, we score heterozygotes as 1 and homozygotes as 0, and subtract the mean ($2p_iq_i$) from the scores. We made a custom modification (available at ftp://ftp.cassavabase.org/manuscripts/) to the *A.mat* function to produce the **D*** matrix.

The **D*** dominance matrix was shown by Vitezica et al. (2013) to produce a partition of genetic variance appropriate for studying genetic architecture because it isolates additive effect variance from dominance effects. However, this partition is not correct for breeding purposes because the additive variance produced is not equivalent to the variation in breeding value. Vitezica et al. (2013) subsequently derived the matrix, **D** defined as $\mathbf{D} = \frac{\mathbf{WW}'}{\sum_i (2p_iq_i)^2}$, where **W** is a marker matrix with markers coded 0 for aa, $2p_i$ for Aa, and $4p_i - 2$ for AA and then centered on the mean, $2p_i^2$. Although our focus in the present study is not on the prediction of breeding value, the matrix **D** has been shown by Zhu et al. (2015) to have the advantage of being uncorrelated (under Hardy-Weinberg equilibrium) with the matrix **G**. For this reason, we tested the **D** and **D*** matrices and

provide comparison of their results. Except where explicitly comparing matrices, we use **D** to indicate the dominance matrix **D*** as in Su et al. (2012).

Finally, we constructed relationship matrices that capture epistasis by taking the Hadamard product (element-by-element multiplication; denoted #) of matrices (Henderson 1985). For simplicity, we only explored additive-by-additive (**A#A**) and additive-by-dominance (**A#D**) relationships in this study.

Variance component and heritability models

Single-step, Multi-environment: We used several approaches to estimate the relative importance of additive and non-additive effects in the Genetic Gain and Cycle 1 datasets. In the first analysis, we analyzed the multi-year, multi-location GG data with a single-step mixed-effects model. Since the entire historical phenotype dataset is large (24,373 observations) and was relatively unbalanced in sample size across years and locations, we only analyzed data from trials with >400 individuals. This filter resulted in a dataset of 7745 observations from three locations (Ibadan, Ubiaja, Mokwa) and eight years (2006-2014, except 2012). All 694 genotyped GG clones were represented in this dataset (Table S2).

The models we fit were similar to those described in Ly et al. (2013). The full model was specified as follows: $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{loc.year}}\boldsymbol{l} + \mathbf{Z}_{\mathbf{rep}}\boldsymbol{r} + \mathbf{Z}_{\mathbf{add}}\boldsymbol{a} + \mathbf{Z}_{\mathbf{dom}}\boldsymbol{d} + \mathbf{Z}_{\mathbf{epi}}\boldsymbol{i} + \boldsymbol{\varepsilon}$. Here, $\boldsymbol{y}$ represents raw phenotypic observations. In our data, the only fixed effect ($\boldsymbol{\beta}$) was an intercept for all traits except RTWT, which contained a covariate accounting for variation in the number of plants harvested per plot. The random effects terms for experimental design terms included a unique intercept for each trial (i.e. location-year

227 combination), $\boldsymbol{l} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_l^2)$, where **I** is the identity matrix and σ_l^2 is the associated
228 variance component as well as a replication effect, nested in location-year combination,
229 $\boldsymbol{r} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_r^2)$.

230 The genetic variance component terms included $\boldsymbol{a} \sim \mathrm{N}(\mathbf{0}, \mathbf{G}\sigma_a^2)$, where **G** is the
231 additive genetic relationship matrix and σ_a^2 is the additive genetic variance component.
232 Similarly, $\boldsymbol{d} \sim \mathrm{N}(\mathbf{0}, \mathbf{D}\sigma_d^2)$, is the dominance effect with covariance **D** equal to the
233 dominance relationship matrix and σ_d^2 equal to the dominance variance. The epistatic
234 term $\boldsymbol{i} \sim \mathrm{N}(\mathbf{0}, \mathbf{E}\sigma_i^2)$ where the covariance matrix **E** took the form either of the **A#A** matrix
235 (additive-by-additive) or the **A#D** matrix (additive-by-dominance) and the epistatic
236 variance σ_i^2 was correspondingly either $\sigma_{A\#A}^2$ or $\sigma_{A\#D}^2$. The final term, $\boldsymbol{\varepsilon}$ is the residual
237 variance, assumed to be random and distributed $\mathrm{N}(\mathbf{0}, \mathbf{I}{\sigma_\varepsilon}^2)$. The terms **X**, $\mathbf{Z}_{\mathbf{loc.year}}$, $\mathbf{Z}_{\mathbf{rep}}$,
238 $\mathbf{Z}_{\mathbf{add}}$, $\mathbf{Z}_{\mathbf{dom}}$ and $\mathbf{Z}_{\mathbf{epi}}$ are incidence matrices relating observations to the levels of each
239 factor. We list the different models fit in Table 1, each of which are variations on the full
240 model described above.

241 The formulation described above was used to fit the subset of the GG historical data
242 described above in a single model. For the C1 progenies only a single season was
243 available and therefore we fit all data together in a single model. Since the C1 trials were
244 conducted across three locations but with no replications we fit the same model for C1 as
245 GG excluding the replication term. The models described above were fit using the
246 *regress* package in R (Clifford and McCullagh 2006). The *regress* function finds REML
247 solutions to mixed models using the Newton-Raphson algorithm.

248 For each trait, in both the C1 and GG we identified a "best fit" model among the
249 models listed in Table 1, based on the lowest Akaike Information Criterion (AIC; $2*k$ –

2*ln(*likelihood*), where *k* = number parameters estimated). In addition, we calculated the Bayesian Information Criterion (BIC; -2*ln(*likelihood*) + *k**ln(*n*), where *n* = number of observations and *k* = number of parameters estimated). We also examined the log-likelihood of each model and the proportion of variance explained by genetic factors (H^2). The precision of variance component estimates and the dependency among estimates was examined using the asymptotic variance-covariance matrix of estimated parameters, provided by *regress* **(V)**. We report standard errors for each variance component, defined as the square root of the diagonal of **V**. We also converted **V** into a correlation matrix (**F,** as in Muñoz *et al.* 2014), where **F** is defined as $\mathbf{L^{-1/2}VL^{-1/2}}$ and **L** is a diagonal matrix containing one over the square root of the diagonal of **V**. We use **F** to assess the dependency of variance components estimates, especially for comparing results among traits and datasets.

Within-trial analyses: We used only a subset of the GG trials to estimate variance components in the single-step multi-environment model described above. In addition, we were able to analyze the entire historical GG data by testing each trial (N=47, unique location-year combinations) separately. This provided us with 47 estimates of additive, dominance and epistatic variance. We examine the distribution of variance components estimates. As in the multi-environment models, within-trial models were fit with *regress* in R.

Genomic prediction and cross-validation

We assessed the influence that modeling non-additive genetic variance components have on genomic prediction using a cross-validation strategy. Because

single-step multi-environment models are computationally intensive, we used a two-step approach here. In the first step, we combined data from all available GG and C1 trials using the following mixed model: $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z_{rep}}\boldsymbol{r} + \mathbf{Z_{clone}}\boldsymbol{g} + \boldsymbol{\varepsilon}$. In this model, β included a fixed effect for the population mean, the location-year combination and for RTWT only, the number of plants harvested per plot. As in the single-step, multi-environment model for GG, we included the random replication effect $\boldsymbol{r} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_r^2)$. In contrast to the previous model, we did not at this stage include a genomic relationship matrix, instead we fit a random effect for clone, $\boldsymbol{g} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_g^2)$, where the covariance structure was the identity matrix, **I**). The BLUP ($\hat{\boldsymbol{g}}$) for the clone effect therefore represents an estimate of the total genetic value for each individual. The mixed model above was solved using the *lmer* function of the *lme4* R package (Bates *et al.* 2014).

In our data, the number of observations per clone ranges from one to 131 with median of two and mean of 5.97 excluding the checks TMEB1 and I30572 which had 941 and 902 observations, respectively. Pooling information from multiple years and locations, especially when there is so much variation in numbers of observations can introduce bias. Much theory, particularly in animal breeding has been developed to address this issue, and we followed the approach recommended by Garrick et al. (2009). Briefly, BLUPs ($\hat{\boldsymbol{g}}$) for clone were deregressed according to $\frac{\hat{\mathbf{g}}}{\boldsymbol{r^2}}$ where $\mathbf{r^2}$ is the reliability $(1 - \frac{\mathbf{PEV}}{\sigma_g^2})$ and PEV is the prediction error variances of the BLUP. In the second step of analysis, where deregressed BLUPs are used as response variables, weights are applied to the diagonal of the error variance-covariance matrix **R**. Weights are calculated as

294 $\frac{1-h^2}{0.1+\frac{1-r^2}{r^2}h^2}$, where h^2 is the proportion of the total variance explained by the clonal

295 variance component, σ_g^2 (Garrick *et al.* 2009).

296 We implemented a 5-fold cross-validation scheme replicated 25 times to test the

297 accuracy of genomic prediction using the genomic relationship matrices and models

298 described above (Table 1). In this scheme, for each replication, we randomly divided the

299 dataset into five equally sized parts (i.e. folds). We used each fold in turn for validation

300 by removing its phenotypes from the training population and then predicting them. We

301 calculated accuracy as the Pearson correlation between the genomic prediction and the

302 BLUP ($\hat{\boldsymbol{g}}$, not-deregressed) from the first step. For each model, we calculated accuracy of

303 the prediction for total genetic value, defined as the sum of the predictions from all

304 available kernels (e.g. additive + dominance + epistasis). Genomic predictions were made

305 using the *EMMREML* R package (Akdemir & Okeke 2015).

306 All raw genotype and phenotype data are available at

307 ftp://ftp.cassavabase.org/manuscripts/ along with custom code used to make de-regressed

308 BLUPs, conduct fold cross-validation, and calculate dominance-relationship matrices.

309

310 **RESULTS**

311 **Partitioning the genetic variance: Single-step, multi-environment models**

312 We used several approaches to estimate genetic variance components in our

313 dataset. The first was to fit single-step models to two datasets: the Genetic Gain (GG) and

314 the Cycle 1 (C1). For each trait, in each dataset, we first identified the best fitting model

315 of the five tested (Table 1) on the basis of lowest AIC. Model comparisons based on AIC

316 and BIC are summarized in Table 2. Key results from the best models for both GG and

C1 are summarized in Table 3 with more detailed results from all models provided in Tables S4 and S5.

The AIC-selected best model for MCMDS included additive-by-dominance epistasis (AxD) in both GG and C1. For RTWT, the model with additive-by-additive epistasis (AxA) fit best in the GG but a simpler dominance only (Dom) model was selected in the C1 dataset. Finally, for DM the additive only (Add) model was best in the GG but additive plus dominance (A+D) was selected in C1 dataset. The BIC criterion places a steeper penalty on increasing the number of parameters. Nevertheless, BIC selected the same model as AIC in all cases except for RTWT in the GG dataset, where the additive plus dominance model was preferred (Table 2, Tables S4 and S5). Based on the guidelines of Raftery (1995), the evidence that the best models for RTWT include dominance and models for MCMDS include non-additive effects, especially epistasis, is very strong (>10 AIC/BIC difference).

We noted that for every trait, when comparing the model achieving the highest broad-sense heritability (H^2), the H^2 was higher in C1 compared to GG. This can be seen most easily in Figure 1, which shows how total explainable genetic variance (H^2) is partitioned among variance components in the C1 and GG (also see Tables 3, S4 and S5). We also noted that the additive only model had the highest H^2 for all traits in the GG dataset, but in C1 models with non-additive components always had at least slightly higher H^2.

On the basis of genetic variance captured, DM had H^2 between 0.25-0.53 and had mostly additive inheritance across all models (Figure 1, Tables 3, S4 and S5). In contrast, non-additive components accounted for the majority of genetic variance for RTWT, with

340 H^2 between 0.21-0.33. Dominance was significant in all models tested for RTWT in both
341 datasets and epistasis was significant in the GG dataset. MCMDS had the highest H^2
342 compared to the other traits (0.66-0.89) and was similar to RTWT in that dominance
343 and/or epistasis were always significant where included. While non-additive genetics
344 were the majority of H^2 in GG, they were much less important in C1 for MCMDS.
345 We examined the asymptotic correlation matrices of parameter estimates (**F)** to
346 ascertain the dependency of variance component estimation. The correlation between
347 genetic variance components was always negative and was, in general, of greater
348 magnitude in the GG compared to the C1 (Tables S6-S11). Correlations between additive
349 and dominance components were greatest in the A+D models (range -0.81 to -0.83 in the
350 GG and -0.5 to -0.61 in the C1). Correlations between additive and dominance
351 components dropped in models with epistasis (range -0.42 to -0.63, GG and -0.26
352 to -0.58, C1). Correlations between additive and AxA epistatic variances (range -0.09
353 to -0.29) and AxD epistasis (range -0.07 to -0.22) were low. Correlations between
354 dominance components and epistasis were higher, ranging from -0.28 to -0.64 with AxA
355 epistasis and -0.36 to -0.69 with AxD epistasis.
356 Comparison of the two alternative dominance matrices **D*** (results described
357 above) and **D** revealed very similar results. In almost every case, AIC and BIC selected
358 the same best-fit model for **D** and **D***. The exception was RTWT in the C1 dataset where
359 the A+D model was preferred over the dominance only model when using the **D** matrix
360 instead of **D*** (Tables S4 and S5). The AIC and BIC are on the whole slightly lower for
361 the models using the **D** matrix, indicating a better fit to the data. As expected, the
362 correlation between additive and dominance parameter estimates is of smaller magnitude

for all analyses with **D** compared with **D*** (Tables S6-S11). However, the correlation between additive and epistatic as well as between dominance and epistatic variances is always of greater magnitude with **D**. Finally and as expected, models using the **D** matrix generally explain the same amount of genetic variance as those with **D*** but partition a smaller portion to dominance (Figure S3). We noted that for RTWT in the C1 dataset, models with **D** actually achieve a slightly higher broad-sense heritability. Because of the similarity of results, we focused the remainder of our analyses and discussion on the results from the **D*** matrix, henceforth referred to only as **D**.

Partitioning the genetic variance: within-trial analyses

We also examined variance partitioning within each of 47 GG trials for the 5 models described in Table 1. This provided a means of testing the entire available dataset for non-additive variances, in contrast to the multi-environment models described above. The mean and variability of model parameters (variance components, heritability, etc.) across these trials are summarized in Table S12 and results for each individual trial-model combination are given in Table S13. Figure 2 provides a visual summary of the proportion of phenotypic variability explained by each genetic variance component on average across the trials. We also compared the mean AIC across trials (Table 2, Table S10) and found them to agree overall with the results of the one-step multi-environment models (Table 3). Specifically, the models that fit best in the one-step models were best on average in the within trial analyses for DM (Add) and MCMDS (AxD). However, for RTWT the within trial AIC-best model was A+D compared to AxD in the one-step multi-environment model. In contrast to the one-step multi-environment model, the BIC agreed

with AIC only for DM. For RTWT and MCMDS the simpler dominance only model was preferred by BIC on average (Table 2).

Genomic Prediction of Additive and Total Genetic Value

We used cross-validation to assess the prediction accuracy for total genetic value from the five models (Table 1) in both datasets. Compared to the single-kernel additive prediction using the additive relationship matrix, multi-kernel total genetic value predictions were an average of 7% better (maximum of 26% improvement; Figure 3, Tables S14-S15). By model, improvements in the correlation between total value and phenotype over the additive only model were 7%, 7% and 8% for A+D, AxA and AxD respectively. The additive only model predictions were on average 12% less accurate in the C1 than in the GG. Total genetic value predictions were less accurate by 12% in the C1 relative to GG. The models we fit for genomic prediction involved the estimation by *EMMREML* of weights, used to create a single kernel that is the weighted average of multiple original kernels and corresponding to the partitioning of genetic variance among the kernels. The average total weight given to non-additive components for both DM and MCMDS was 0.41 but was 0.92 for RTWT.

DISCUSSION

In clonally propagated crops, non-additive genetic effects can be effectively exploited by the identification of superior genetic individuals as varieties. For this reason, we quantified the amount and nature of non-additive genetic variation for key traits in a

409 genomic selection breeding population of cassava from sub-Saharan African. We then
410 assessed the accuracy of genomic prediction of additive compared to total (additive plus
411 non-additive) genetic value. Using several approaches and datasets based on genome-
412 wide marker data, we confirmed previous findings in cassava based on diallel
413 populations: non-additive genetic variation is significant, especially for yield traits (Cach
414 *et al.* 2005, 2006; Calle *et al.* 2005; Jaramillo *et al.* 2005; Pérez, Ceballos, Calle, *et al.*
415 2005; Pérez, Ceballos, Jaramillo, *et al.* 2005; Zacarias and Labuschagne 2010;
416 Kulembeka *et al.* 2012; Tumuhimbise *et al.* 2014; Ceballos *et al.* 2015; Chalwe *et al.*
417 2015). A potential weakness of the marker system we used (GBS) is that it generates a
418 high proportion of missing marker data and it may undercall heterozygotes when read
419 depth is insufficient. The similarity of our findings to previous research, and the
420 important difference in the observations on RTWT versus DM, however, suggest that this
421 weakness did not strongly affect our results. Further, we found that multi-component
422 models incorporating non-additive effects predict observed phenotypes more accurately
423 than additive-only models for root yield but not for dry matter content, which is has
424 primarily additive inheritance or for CMD resistance, which has high narrow-sense
425 heritability. We address the implication of these results for cassava breeding and put our
426 work in the context of previous results in cassava, other plant and animal species below.
427 Our results indicate strong non-additive (mainly dominance) variance for root
428 yields and mostly additive inheritance of root dry matter content. These findings confirm
429 the conclusions of numerous diallelic studies conducted with both Latin American (Cach
430 *et al.* 2005, 2006; Calle *et al.* 2005; Jaramillo *et al.* 2005; Pérez, Ceballos, Calle, *et al.*
431 2005; Pérez, Ceballos, Jaramillo, *et al.* 2005) and African cassava (Zacarias and

Labuschagne 2010; Kulembeka *et al.* 2012; Tumuhimbise *et al.* 2014; Chalwe *et al.* 2015) germplasm (see also Ceballos *et al.* 2015). In agreement with the findings of Ly et al. (2013), we found cassava mosaic disease severity (MCMDS) to be well predicted with an additive only model. However, we found significant dominance and epistatic components in both populations analyzed. This result is in line with previous diallelic studies indicating significant SCA (Tumuhimbise *et al.* 2014; Chalwe *et al.* 2015) and genetic mapping studies that identified a single major effect QTL with a dominant CMD resistance effect (Akano *et al.* 2002; Okogbenin *et al.* 2012; Rabbi *et al.* 2014). In addition, a recent genome-wide association and prediction study of MCMDS, using non-additive genomic relationship matrices (GRMs) found that dominance and especially epistasis explain most of the variance in the region of a large-effect QTL, suggesting multiple interacting loci in the region (Wolfe *et al.* 2016).

The importance of non-additive genetic variance in evolution by natural and artificial selection is controversial (Hill *et al.* 2008; Crow 2010; Hansen 2013). Nevertheless, numerous studies have found and exploited dominance and epistasis in animal breeding, including dairy (Ahlborn-Breier and Hohenboken 1991; Fuerst and Sölkner 1994; Varona *et al.* 1998; Van Tassell *et al.* 2000; Palucci *et al.* 2007) and beef (Rodriguezalmeida *et al.* 1995) cattle. Diallelic studies have indicated significant SCA for maize grain yield (Doerksen *et al.* 2003; Wardyn *et al.* 2007). Aside from cassava, breeding of other non-inbred, clonally propagated species also identify and make use of non-additive effects, including potato (Killick 1977), Eucalyptus (Costa E Silva *et al.* 2004) and loblolly pine (Muñoz *et al.* 2014). More recently, marker-based and GRM-based models have identified significant non-additive effects in pigs (Su *et al.* 2012;

455 Nishio and Satoh 2014), mice (Vitezica *et al.* 2013), beef cattle (Bolormaa *et al.* 2015),
456 dairy cows (Morota *et al.* 2014), maize (Dudley and Johnson 2009), soy (Hu *et al.* 2011),
457 loblolly pine (Muñoz *et al.* 2014) and apple (Kumar *et al.* 2015). Results from the present
458 study suggest that accounting for non-additive in the variety development pipeline should
459 increase the value of hybrids released by cassava breeding programs.

460 One of the more interesting aspects of our study relative to previous ones is the
461 comparison between a parental generation (the Genetic Gain) and their offspring (Cycle
462 1), a collection of full- and half-sib families. From GG to C1, the H^2 generally increased.
463 For RTWT, this is largely attributable to increased non-additive variance in contrast to
464 MCMDS where the increase is concomitant with a drop in non-additive variance. In
465 contrast to our result, theory suggests that reduction (or fixation) of allele frequencies at
466 some loci relative to others in populations undergoing bottlenecks (Goodnight 1988),
467 inbreeding (Turelli and Barton 2006) or truncation selection (Hallander and Waldmann
468 2007) should cause a conversion of non-additive (where present) to additive variance.
469 These results have, however, been based on models with finite numbers of loci in linkage
470 equilibrium. Based on the mean diagonal of the additive genetic relationship matrix, C1
471 (0.66) does not appear notably more inbred than GG (0.64). We also calculated mean
472 pairwise LD (GG = 0.27, C1 = 0.29) and mean LD block size (21.7 kb in GG and 23.1 kb
473 in C1) using standard settings in PLINK (version 1.9, https://www.cog-
474 genomics.org/plink2) and found the two generations to be similar.

475 Probably the strongest explanation for the difference in genetic variance
476 components between GG and C1 is the family structure (135 full-sib families from 83
477 outbred parents). In a population of full-sibs ¾ of the dominance variance is expressed

478 within families and all of it for half-sib populations (Hallauer *et al.* 2010; Ceballos *et al.*
479 2015). Indeed, increasing the number of full-sib relationships is known to increase the
480 non-additive genetic variance detectable in a population (Varona *et al.* 1998; Van Tassel
481 *et al.* 2010).

482 It is also conceivable that maternal plant effects could increase apparent non-
483 additive effects in C1. The C1 clones in contrast to the GG clones are new, and were
484 derived from stem cuttings of seedling plants germinated in the previous field season
485 (2012-2013). The suggestion is therefore that the quality and vigor of the seedling plant,
486 giving rise to the C1 clones may influence their performance in the 2013-2014 trial. We
487 further caution that comparison of GG and C1 may be biased by the disproportionate
488 amount of data from different locations and years available for the GG.

489 In our study, when additive and non-additive kernels were used together, the
490 variance explained by the additive component particularly for RTWT decreased. One
491 interpretation of this result is that the additive component alone absorbs some non-
492 additive variance. Similar results have been obtained by other researchers, leading to
493 similar conclusions (Lu *et al.* 1999; Su *et al.* 2012; Zuk *et al.* 2012; Muñoz *et al.* 2014).
494 We note that this phenomenon occurs whether we use the **D*** matrix, which is correlated
495 with the **G** matrix or the **D** matrix, which is theoretically orthogonal to **G**. We suggest
496 therefore that prediction models that do not explicitly incorporate non-additive
497 components may achieve gains in the short-term that break down over the long-term
498 (Cockerham and Tachida 1988; Walsh 2005; Hansen 2013). Our prediction tests in this
499 study were focused on total genetic values and used the **D*** matrix. However, we
500 hypothesize that including non-additive GRMs, particularly the **D** matrix, when

estimating additive genetic (i.e. breeding) values would provide a less biased, more accurate selection of parents for crossing.

Non-additive variation is prevalent in cassava, especially for low heritability traits. This has many important implications for cassava breeding. It explains, in part, why genetic gains have been slow (Ceballos *et al.* 2012). Inbreeding to convert dominance variance to additive and better control epistatic combinations, as in maize, has been suggested as a solution to non-additive genetics (Ceballos *et al.* 2015). Even for low h^2 traits and without inbred cassava, using the kinds of models presented in this paper, good parents can be selected based on additive predictions and total genetic value can be predicted for the identification of potential commercial varieties, all based on the combination of marker and preliminary field trial data (Heslot *et al.* 2015). This approach has been previously advocated for plant breeding (Oakey *et al.* 2007; Heslot *et al.* 2015) and has proven effective in animal breeding, e.g. (Ahlborn-Breier and Hohenboken 1991; Palucci *et al.* 2007; Su *et al.* 2012; Nishio and Satoh 2014). Non-additive models using genomic relationship matrices can thus improve the efficiency and productivity of variety selection pipelines that are the most labor and time intensive part of selecting good cassava clones after crossing.

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 phenotypic data and to A. Agbona and P. Peteti for data curation.

Literature Cited

(ICGMC), International Cassava Genetic Map Consortium, 2015 High-Resolution Linkage Map and Chromosome-Scale Genome Assembly for Cassava (Manihot esculenta Crantz) from Ten Populations. G3 5: 133–144.

Ahlborn-Breier, G., and W. D. Hohenboken, 1991 Additive and nonadditive genetic effects on milk production in dairy cattle: evidence for major individual heterosis. J. Dairy Sci. 74: 592–602.

Akano, O., O. Dixon, C. Mba, E. Barrera, and M. Fregene, 2002 Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. Theor. Appl. Genet. 105: 521–525.

Bates, D., M. Maechler, B. Bolker, and S. Walker, 2014 Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67: 1–48.

Bolormaa, S., J. E. Pryce, Y. Zhang, A. Reverter, W. Barendse *et al.*, 2015 Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. Genet. Sel. Evol. 47: 1–12.

Cach, N. T., J. I. Lenis, J. C. Perez, N. Morante, F. Calle *et al.*, 2006 Inheritance of useful traits in cassava grown in subhumid conditions. Plant Breed. 125: 177–182.

Cach, N. T., J. C. Perez, J. I. Lenis, F. Calle, N. Morante *et al.*, 2005 Epistasis in the expression of relevant traits in cassava (Manihot esculenta Crantz) for subhumid conditions. J. Hered. 96: 586–592.

Calle, F., J. C. Perez, W. Gaitán, N. Morante, H. Ceballos *et al.*, 2005 Diallel inheritance of relevant traits in cassava (Manihot esculenta Crantz) adapted to acid-soil savannas. Euphytica 144: 177–186.

Ceballos, H., R. S. Kawuki, V. E. Gracen, G. C. Yencho, and C. H. Hershey, 2015 Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. Theor. Appl. Genet.

Ceballos, H., P. Kulakow, and C. Hershey, 2012 Cassava Breeding: Current Status, Bottlenecks and the Potential of Biotechnology Tools. Trop. Plant Biol. 5: 73–87.

Chalwe, A., R. Melis, P. Shanahan, and M. Chiona, 2015 Inheritance of resistance to cassava green mite and other useful agronomic traits in cassava grown in Zambia. Euphytica.

559 Clifford, D., and P. McCullagh, 2006 The regress function. R News 6: 6–10.

560 Cockerham, C. C., 1954 An Extension of the Concept of Partitioning Hereditary Variance
561 for Analysis of Covariances among Relatives When Epistasis Is Present. Genetics
562 39: 859–882.

563 Cockerham, C. C., and H. Tachida, 1988 Permanency of response to selection for
564 quantitative characters in finite populations. Proc. Natl. Acad. Sci. U. S. A. 85:
565 1563–5.

566 Costa E Silva, J., N. M. G. Borralho, and B. M. Potts, 2004 Additive and non-additive
567 genetic parameters from clonally replicated and seedling progenies of Eucalyptus
568 globulus. Theor. Appl. Genet. 108: 1113–1119.

569 Crow, J. F., 2010 On epistasis: why it is unimportant in polygenic directional selection.
570 Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365: 1241–1244.

571 Doerksen, T. K., L. W. Kannenberg, and E. a Lee, 2003 Effect of Recurrent Selection on
572 Combining Ability in Maize Breeding Populations. Crop Sci. 43: 1652–1658.

573 Dudley, J. W., and G. R. Johnson, 2009 Epistatic models improve prediction of
574 performance in corn. Crop Sci. 49: 763–770.

575 Elshire, R. J., J. C. Glaubitz, Q. Sun, J. a Poland, K. Kawamoto *et al.*, 2011 A robust,
576 simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS
577 One 6: e19379.

578 Endelman, J. B., 2011 Ridge Regression and Other Kernels for Genomic Selection with
579 R Package rrBLUP. Plant Genome J. 4: 250.

580 Fisher, R. A., 1918 The Correlation between Relatives on the Supposition of Mendelian
581 Inheritance. Trans. R. Soc. Edinburgh 52: 399–433.

582 Fuerst, C., and J. Sölkner, 1994 Additive and nonadditive genetic variances for milk yield,
583 fertility, and lifetime performance traits of dairy cattle. J. Dairy Sci. 77: 1114–1125.

584 Garrick, D. J., J. F. Taylor, and R. L. Fernando, 2009 Deregressing estimated breeding
585 values and weighting information for genomic regression analyses. Genet. Sel. Evol.
586 41: 55.

587 Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-
588 GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9:
589 e90346.

590 Goodnight, C. J., 1988 Epistasis and the Effect of Founder Events on the Additive
591 Genetic Variance. Evolution (N. Y). 42: 441–454.

592 Hallander, J., and P. Waldmann, 2007 The effect of non-additive genetic interactions on
593 selection in multi-locus genetic models. Heredity (Edinb). 98: 349–359.

594 Hallauer, A. R., M. J. Carena, and J. B. Miranda Filho, 2010 *Quantitative Genetics in*

Maize Breeding. Springer-Verlag, New York.

Hamblin, M. T., and I. Y. Rabbi, 2014 The Effects of Restriction-Enzyme Choice on Properties of Genotyping-by-Sequencing Libraries: A Study in Cassava (). Crop Sci. 54: 2603.

Hansen, T. F., 2013 Why epistasis is important for selection and adaptation. Evolution (N. Y). 67: 3501–3511.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009 Genomic Selection for Crop Improvement. Crop Sci. 49: 1.

Henderson, C. R., 1985 Best Linear Unbiased Prediction of Nonadditive Genetic Merits in Noninbred Populations. J. Anim. Sci. 60: 111–117.

Heslot, N., J. Jannink, and M. E. Sorrells, 2015 Perspectives for genomic selection applications and research in plants. Crop Sci. 55: 1–12.

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4.:

Hu, Z., Y. Li, X. Song, Y. Han, X. Cai *et al.*, 2011 Genomic value prediction for quantitative traits under the epistatic model. BMC Genet. 12: 15.

Jaramillo, G., N. Morante, J. C. Pérez, F. Calle, H. Ceballos *et al.*, 2005 Diallel analysis in cassava adapted to the midaltitude valleys environment. Crop Sci. 45: 1058–1063.

Kempthorne, O., 1954 The Correlation between Relatives in a Random Mating Population. Proc. R. Soc. B Biol. Sci. 143: 103–113.

Killick, R., 1977 Genetic analysis of several traits in potatoes by means of a diallel cross. Ann. Appl. Biol. 86: 279–286.

Kulembeka, H. P., M. Ferguson, L. Herselman, E. Kanju, G. Mkamilo *et al.*, 2012 Diallel analysis of field resistance to brown streak disease in cassava (Manihot esculenta Crantz) landraces from Tanzania. Euphytica 187: 277–288.

Kumar, S., C. Molloy, P. Muñoz, H. Daetwyler, D. Chagné *et al.*, 2015 Genome-enabled estimates of additive and non-additive genetic variances and prediction of apple phenotypes across environments. G3 (Bethesda).

Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 *Genomic Selection in Plant Breeding: Knowledge and Prospects*.

Lu, P. X., D. A. Huber, and T. L. White, 1999 Potential biases of incomplete linear models in heritability estimation and breeding value prediction. Can. J. For. Res. 29: 724–736.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare *et al.*, 2013 Relatedness and Genotype × Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study in Cassava. Crop Sci. 53: 1312.

631 Lynch, M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*.

632 Morota, G., R. Abdollahi-Arpanahi, A. Kranis, and D. Gianola, 2014 Genome-enabled
633 prediction of quantitative traits in chickens using genomic annotation. BMC
634 Genomics 15: 109.

635 Muñoz, P. R., M. F. R. Resende, S. a Gezan, M. Deon, and V. Resende, 2014 Unraveling
636 Additive from Nonadditive Effects Using Genomic Relationship Matrices. Genetics
637 198: 1759–1768.

638 Nishio, M., and M. Satoh, 2014 Impacts of genotyping strategies on long-term genetic
639 response in genomic selection. Anim. Sci. J. 2–7.

640 Oakey, H., A. P. Verbyla, B. R. Cullis, X. Wei, and W. S. Pitchford, 2007 Joint modeling
641 of additive and non-additive (genetic line) effects in multi-environment trials. Theor.
642 Appl. Genet. 114: 1319–1332.

643 Okechukwu, R. U., and a. G. O. Dixon, 2008 Genetic Gains from 30 Years of Cassava
644 Breeding in Nigeria for Storage Root Yield and Disease Resistance in Elite Cassava
645 Genotypes. J. Crop Improv. 22: 181–208.

646 Okogbenin, E., C. Egesi, B. Olasanmi, O. Ogundapo, S. Kahya *et al.*, 2012 Molecular
647 marker analysis and validation of resistance to cassava mosaic disease in elite
648 cassava genotypes in Nigeria. Crop Sci. 52: 2576–2586.

649 Oliveira, E. J., M. D. V. Resende, V. Silva Santos, C. F. Ferreira, G. A. F. Oliveira *et al.*,
650 2012 Genome-wide selection in cassava. Euphytica 187: 263–276.

651 Palucci, V., L. R. Schaeffer, F. Miglior, and V. Osborne, 2007 Non-additive genetic
652 effects for fertility traits in Canadian Holstein cattle (Open Access publication).
653 Genet. Sel. Evol. 39: 181–193.

654 Pérez, J. C., H. Ceballos, F. Calle, N. Morante, W. Gaitán *et al.*, 2005 Within-family
655 genetic variation and epistasis in cassava (Manihot esculenta Crantz) adapted to the
656 acid-soils environment. Euphytica 145: 77–85.

657 Pérez, J. C., H. Ceballos, G. Jaramillo, N. Morante, F. Calle *et al.*, 2005 Epistasis in
658 cassava adapted to midaltitude valley environments. Crop Sci. 45: 1491–1496.

659 Prochnik, S., P. R. Marri, B. Desany, P. D. Rabinowicz, C. Kodira *et al.*, 2012 The
660 Cassava Genome: Current Progress, Future Directions. Trop. Plant Biol. 5: 88–94.

661 Rabbi, I. Y., M. T. Hamblin, P. L. Kumar, M. a Gedil, A. S. Ikpan *et al.*, 2014 High-
662 resolution mapping of resistance to cassava mosaic geminiviruses in cassava using
663 genotyping-by-sequencing and its implications for breeding. Virus Res.

664 Raftery, A. E., 1995 Bayesian Model Selection in Social Research. Sociol. Methodol. 25:
665 111–163.

666 Rodriguezalmeida, F. a, L. D. Vanvleck, R. L. Willham, and S. L. Northcutt, 1995

667 Estimation of Nonadditive Genetic Variances in 3 Synthetic Lines of Beef-Cattle
668 Using an Animal-Model. J. Anim. Sci. 73: 1002–1011.

669 Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 2012 Estimating
670 Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using
671 Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLoS One 7: 1–7.

672 Van Tassel, D. L., L. R. DeHaan, and T. S. Cox, 2010 Missing domesticated plant forms:
673 can artificial selection fill the gap? Evol. Appl. 3: 434–452.

674 Van Tassell, C. P., I. Misztal, and L. Varona, 2000 Method R estimates of additive
675 genetic, dominance genetic, and permanent environmental fraction of variance for
676 yield and health traits of Holsteins. J. Dairy Sci. 83: 1873–1877.

677 Tumuhimbise, R., R. Melis, and P. Shanahan, 2014 Diallel analysis of early storage root
678 yield and disease resistance traits in cassava (Manihot esculenta Crantz). F. Crop.
679 Res. 167: 86–93.

680 Turelli, M., and N. H. Barton, 2006 Will population bottlenecks and multilocus epistasis
681 increase additive genetic variance? Evolution 60: 1763–1776.

682 VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci.
683 91: 4414–23.

684 Varona, L., I. Misztal, J. K. Bertrand, and T. J. Lawlor, 1998 Effect of full sibs on
685 additive breeding values under the dominance model for stature in United States
686 Holsteins. J. Dairy Sci. 81: 1126–1135.

687 Visscher, P. M., W. G. Hill, and N. R. Wray, 2008 Heritability in the genomics era--
688 concepts and misconceptions. Nat. Rev. Genet. 9: 255–266.

689 Vitezica, Z. G., L. Varona, and A. Legarra, 2013 On the additive and dominant variance
690 and covariance of individuals within the genomic selection scope. Genetics 195:
691 1223–1230.

692 Walsh, B., 2005 The struggle to exploit non-additive variation. Aust. J. Agric. Res. 56:
693 873–881.

694 Wang, C., D. Prakapenka, S. Wang, S. Pulugurta, H. B. Runesha *et al.*, 2014 GVCBLUP:
695 a computer package for genomic prediction and variance component estimation of
696 additive and dominance effects. BMC Bioinformatics 15: 270.

697 Wardyn, B. M., J. W. Edwards, and K. R. Lamkey, 2007 The genetic structure of a maize
698 population: The role of dominance. Crop Sci. 47: 467–476.

699 Wolfe, M. D., I. Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki *et al.*, 2016 Genome-wide
700 association and prediction reveals the genetic architecture of cassava mosaic disease
701 resistance and prospects for rapid genetic improvement. Plant Genome 1–248.

702 Wong, W. W. L., J. Griesman, and Z. Z. Feng, 2014 Imputing genotypes using

regularized generalized linear regression models. Stat. Appl. Genet. Mol. Biol. 13: 519–529.

Zacarias, a. M., and M. T. Labuschagne, 2010 Diallel analysis of cassava brown streak disease, yield and yield related characteristics in Mozambique. Euphytica 176: 309–320.

Zhu, Z., A. Bakshi, A. A. E. Vinkhuyzen, G. Hemani, S. H. Lee *et al.*, 2015 Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. Am. J. Hum. Genet. 96: 377–385.

Zuk, O., E. Hechter, S. R. Sunyaev, and E. S. Lander, 2012 The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. 109: 1193–1198.

Figure & Table Legends

Table 1. Additive plus non-additive genetic models tested and their abbreviations.

Table 2. Comparison of models by AIC and BIC. For each trait, five genetic models are compared based on Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Comparisons are done based on single-step multi-environmental models for the Genetic Gain (GG) and Cycle 1 (C1) datasets. In addition, the mean and standard error AIC/BIC from 47 GG trials, each analyzed separately are provided. For each dataset and each trait the lowest AIC and BIC are bolded and highlighted.

Table 3. Best-fitting single-step multi-environment model results. Variance components (± standard errors), narrow-sense heritabilities (h^2), proportion of the total phenotypic variance explained by dominance (d^2), epistasis (i^2_{epi}), and broad-sense heritability (H^2) are provided. Model log-likelihoods are also given. The models shown were selected on the basis of having the lowest Akaike Information Criterion (AIC) relative to other tested models.

Figure 1. Partitioning of broad-sense heritability for single-step multi-environment models in the Genetic Gain and Cycle 1 datasets. Results from each of five models are shown in each panel broken down by trait (rows) and population (columns). Models include additive only (Additive), dominance only (Dominance), Additive plus Dominance (Add + Dom), Additive plus dominance plus either AxA epistasis (AxA Epistasis) or AxD epistasis (AxD Epistasis).

Figure 2. Distribution of genetic variance proportions across Genetic Gain trials. Three models were fitted for each trait in each of 47 Genetic Gain trials. Each panel contains boxplots showing the distribution of proportions of the phenotypic variability explained by a corresponding genetic factor, including the broad-sense heritability (H^2). Red horizontal lines are the median narrow-sense heritability (h^2) from the additive only model. Traits are on columns and three models are on the rows: additive plus dominance (Add + Dom), additive plus dominance plus AxA epistasis (AxA Epistasis) and additive plus dominance plus AxD epistasis (AxD Epistasis).

Figure 3. Accuracy of total genetic value prediction in the Genetic Gain and Cycle 1 datasets. Boxplots showing the distribution over 25 replicates of 5-fold cross-validation of the prediction accuracy of the total genetic value from five different models are shown in each panel. The accuracy within the Genetic Gain (red) and Cycle 1 (blue) are shown. Traits are in the columns. Accuracy is defined as the correlation between the sum of predictions from all genetic variance components in the model and the BLUP from the first stage of analysis where location, year and replicate variability were removed. Models included are: additive only (Additive), dominance only (Dominance), additive plus dominance (Add + Dom), additive plus dominance plus AxA epistasis (AxA Epi.) and additive plus dominance plus AxD epistasis (AxD Epi.).

Supplementary Figure 1. Genetic structure of the IITA Genetic Gain germplasm (red) and the Cycle 1 progenies (blue). Scatterplots represent the first four principle components of the additive genomic relationship matrix.

Supplementary Figure 2. Distribution of raw (left) and BLUP (right) phenotypes.

Supplementary Figure 3. Comparison between the partitioning of broad-sense heritability for models using two alternative dominance matrices, D and D* in the Genetic Gain and Cycle 1 datasets. Results from each of five models are shown in each panel broken down by trait (rows) and population (columns). Models include additive only (Add), dominance only (Dom), Additive plus Dominance (AplusD), Additive plus dominance plus either AxA epistasis (AxA_epi) or AxD epistasis (AxD_epi). Models with dominance terms that used the **D** matrix of Vitezica et al. 2013 are distinguished from models using the **D*** matrix of Su et al. 2012 using either "*" or else "(with **D***)".

Supplementary Table 1. Pedigree and related information for the IITA: Genetic Gain germplasm analyzed in this study.

Supplementary Table 2. Details on design of field trials analyzed. The sample size (Nobs), number of replications (Nreps), number of clones (Nclones) are indicated for two datasets analyzed in this study: the IITA Genetic Gain germplasm and a collection of their progeny, Cycle 1. Whether the trial was included in single-step multi-environment models that we fit in this study is also indicated.

Supplementary Table 3. Pedigree information for the IITA: GS Cycle 1 germplasm analyzed in this study. The GS Cycle 1 are genomic selection germplasm descended from another dataset (IITA: Genetic Gain) also analyzed in this study.

Supplementary Table 4. Results from fitting five different additive and non-additive genetic mixed-models for three key cassava traits in a single-step to data from multiple locations and years for the IITA Genetic Gain dataset. Variance components (± standard errors), narrow-sense heritabilities (h^2), proportion of the total phenotypic variance explained by dominance (d^2), additive-by-additive epistasis ($i^2_{A\#A}$), additive-by-dominance epistasis ($i^2_{A\#D}$) and broad-sense heritability (H^2) are provided. Sample size (N), model log-likelihoods, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are also given. The best model for each trait (lowest AIC) is highlighted in grey. Results for analyses using the **D*** matrix of Su et al. 2012[a] (above) and the **D** matrix of Vitezica et al. 2013[b] (below) are presented. Mixed models fit with the R package *regress*.

Supplementary Table 5. Results from fitting five different additive and non-additive genetic mixed-models for three key cassava traits in a single-step to data from multiple locations for the IITA Cycle 1 dataset. Variance components (± standard errors), narrow-sense heritabilities (h^2), proportion of the total phenotypic variance explained by dominance (d^2), additive-by-additive epistasis ($i^2_{A\#A}$), additive-by-dominance epistasis ($i^2_{A\#D}$) and broad-sense heritability (H^2) are provided. Sample size

(N), model log-likelihoods, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are also given. The best model for each trait (lowest AIC) is highlighted in grey. Results for analyses using the **D*** matrix of Su et al. 2012[a] (above) and the **D** matrix of Vitezica et al. 2013[b] (below) are presented. Mixed models fit with the R package *regress*.

Supplementary Table 6. The asymptotic correlation matrices of parameter estimates for each trait from an additive plus dominance genetic model fit in the IITA's Genetic Gain dataset. Matrices presented here inform about the dependency of variance component estimates and are derived from asymptotic variance-covariance matrix of estimated parameters (**V**), provided by the *regress* R package, which was used to fit each mixed-model. Correlation matrix, $\mathbf{F} = \mathbf{L}^{-1/2}\mathbf{V}\mathbf{L}^{-1/2}$, where **L** is a diagonal matrix containing one over the square root of the diagonal of **V**. Correlations for models using the **D*** matrices of Su et al. 2012[a] (lower off-diagonals) and the **D** matrix of Vitezica et al. 2013[b] (upper off-diagonals) are presented.

Supplementary Table 7. The asymptotic correlation matrices of parameter estimates for each trait from an additive plus dominance plus additive-by-additive epistasis genetic model fit in the IITA's Genetic Gain dataset. Matrices presented here inform about the dependency of variance component estimates and are derived from asymptotic variance-covariance matrix of estimated parameters (**V**), provided by the *regress* R package, which was used to fit each mixed-model. Correlation matrix, $\mathbf{F} = \mathbf{L}^{-1/2}\mathbf{V}\mathbf{L}^{-1/2}$, where **L** is a diagonal matrix containing one over the square root of the diagonal of **V**. Correlations for models using the **D*** matrices of Su et al. 2012[a] (lower off-diagonals) and the **D** matrix of Vitezica et al. 2013[b] (upper off-diagonals) are presented.

Supplementary Table 8. The asymptotic correlation matrices of parameter estimates for each trait from an additive plus dominance plus additive-by-dominance epistasis genetic model fit in the IITA's Genetic Gain dataset. Matrices presented here inform about the dependency of variance component estimates and are derived from asymptotic variance-covariance matrix of estimated parameters (**V**), provided by the *regress* R package, which was used to fit each mixed-model. Correlation matrix, $\mathbf{F} = \mathbf{L}^{-1/2}\mathbf{V}\mathbf{L}^{-1/2}$, where **L** is a diagonal matrix containing one over the square root of the diagonal of **V**. Correlations for models using the **D*** matrices of Su et al. 2012[a] (lower off-diagonals) and the **D** matrix of Vitezica et al. 2013[b] (upper off-diagonals) are presented.

Supplementary Table 9. The asymptotic correlation matrices of parameter estimates for each trait from an additive plus dominance genetic model fit in the IITA's Cycle 1 dataset. Matrices presented here inform about the dependency of variance component estimates and are derived from asymptotic variance-covariance matrix of estimated parameters (**V**), provided by the *regress* R package, which was used to fit each mixed-model. Correlation matrix, $\mathbf{F} = \mathbf{L}^{-1/2}\mathbf{V}\mathbf{L}^{-1/2}$, where **L** is a diagonal matrix containing one

over the square root of the diagonal of **V**. Correlations for models using the **D*** matrices of Su et al. 2012[a] (lower off-diagonals) and the **D** matrix of Vitezica et al. 2013[b] (upper off-diagonals) are presented.

Supplementary Table 10. The asymptotic correlation matrices of parameter estimates for each trait from an additive plus dominance plus additive-by-additive epistasis genetic model fit in the IITA's Cycle 1 dataset. Matrices presented here inform about the dependency of variance component estimates and are derived from asymptotic variance-covariance matrix of estimated parameters (**V**), provided by the *regress* R package, which was used to fit each mixed-model. Correlation matrix, $\mathbf{F} = \mathbf{L}^{-1/2}\mathbf{V}\mathbf{L}^{-1/2}$, where **L** is a diagonal matrix containing one over the square root of the diagonal of **V**. Correlations for models using the **D*** matrices of Su et al. 2012[a] (lower off-diagonals) and the **D** matrix of Vitezica et al. 2013[b] (upper off-diagonals) are presented.

Supplementary Table 11. The asymptotic correlation matrices of parameter estimates for each trait from an additive plus dominance plus additive-by-dominance epistasis genetic model fit in the IITA's Cycle 1 dataset. Matrices presented here inform about the dependency of variance component estimates and are derived from asymptotic variance-covariance matrix of estimated parameters (**V**), provided by the *regress* R package, which was used to fit each mixed-model. Correlation matrix, $\mathbf{F} = \mathbf{L}^{-1/2}\mathbf{V}\mathbf{L}^{-1/2}$, where **L** is a diagonal matrix containing one over the square root of the diagonal of **V**. Correlations for models using the **D*** matrices of Su et al. 2012[a] (lower off-diagonals) and the **D** matrix of Vitezica et al. 2013[b] (upper off-diagonals) are presented.

Supplementary Table 12. Summary of results from five additive and non-additive genetic mixed-models for three traits across 47 trials conducted on the IITA Genetic Gain germplasm. The mean (± standard errors) across 47 trials for each trait and model fitted is given for the following model parameters: variance components, narrow-sense heritabilities (h^2), proportion of the total phenotypic variance explained by dominance (d^2), additive-by-additive epistasis ($i^2_{A\#A}$), additive-by-dominance epistasis ($i^2_{A\#D}$) and broad-sense heritability (H^2), trial sample size (N), model log-likelihoods and Akaike Information Criterion (AIC) are also given. The best model for each trait (lowest AIC) is highlighted in grey. Models were fit with the R package *regress*.

Supplementary Table 13. Results from five additive and non-additive genetic mixed-models for three traits across 47 trials conducted on the IITA Genetic Gain germplasm. The following model parameters are given for each trial-trait-model combination tested: variance components, narrow-sense heritabilities (h^2), proportion of the total phenotypic variance explained by dominance (d^2), additive-by-additive epistasis ($i^2_{A\#A}$), additive-by-dominance epistasis ($i^2_{A\#D}$) and broad-sense heritability (H^2), trial sample size (N), model log-likelihoods and Akaike Information Criterion (AIC) are also given. Models were fit with the R package *regress*.

Supplementary Table 14. Results from 25 replicates of 5-fold cross-validation in the Genetic Gain population for three traits and five additive and non-additive genetic mixed-models. Mean (± standard errors) across the 25 replicates are given for prediction accuracy of each kernel plus total genetic value (sum across all kernels), variance components (V_g and V_e) and kernel weights. Models were fit with the R package *EMMREML*.

Supplementary Table 15. Results from 25 replicates of 5-fold cross-validation in the Cycle 1 population for three traits and five additive and non-additive genetic mixed-models. Mean (± standard errors) across the 25 replicates are given for prediction accuracy of each kernel plus total genetic value (sum across all kernels), variance components (V_g and V_e) and kernel weights. Models were fit with the R package *EMMREML*.

Table 1. Additive plus non-additive genetic models tested and their abbreviations.

Model	Relationship Matrices / Variance Components
Add	Additive
Dom	Dominance
A+D	Additive + Dominance
AxA	Additive + Dominance + A#A Epistasis
AxD	Additive + Dominance + A#D Epistasis

Table 2. Comparison of models by AIC and BIC. For each trait, five genetic models are compared based on Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Comparisons are done based on single-step multi-environmental models for the Genetic Gain (GG) and Cycle 1 (C1) datasets. In addition, the mean and standard error AIC/BIC from 47 GG trials, each analyzed separately are provided. For each dataset and each trait the lowest AIC and BIC are bolded and highlighted.

		Genetic Gain (GG)		Cycle 1 (C1)		Genetic Gain (Within Trials)	
Trait	**Model**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
DM	Add	**18921.9**	**18947.9**	7094.3	7110.8	**1335.9 ± 140.3**	**1355.2 ± 141.0**
	Dom	18947.7	18973.7	7176.6	7193.2	1338.6 ± 140.7	1357.8 ± 141.4
	A+D	18922.4	18954.9	**7083.9**	**7106.0**	1337.4 ± 140.3	1360.2 ± 141.1
	AxA	18923.0	18962.0	7085.0	7112.6	1339.0 ± 140.3	1365.2 ± 141.2
	AxD	18922.6	18961.6	7085.9	7113.5	1339.0 ± 140.2	1365.3 ± 141.1
RTWT	Add	-4716.1	-4688.5	-315.2	-298.0	-310.7 ± 42.9	-287.1 ± 42.6
	Dom	-4731.0	-4703.4	**-361.0**	**-343.8**	-311.9 ± 42.5	**-288.3 ± 42.2**
	A+D	-4740.8	**-4706.2**	-360.4	-337.4	**-311.9 ± 42.9**	-284.3 ± 42.6
	AxA	**-4744.2**	-4702.7	-358.4	-329.7	-311.0 ± 43.0	-279.4 ± 42.6
	AxD	-4743.6	-4702.1	-358.4	-329.7	-311.1 ± 43.0	-279.5 ± 42.6
MCMDS	Add	1255.7	1283.4	2417.8	2435.3	38.2 ± 47.1	62.1 ± 47.2
	Dom	1207.6	1235.4	2746.4	2763.8	20.1 ± 47.3	**44.1 ± 47.4**
	A+D	1202.1	1236.9	2396.3	2419.5	20.8 ± 47.3	48.8 ± 47.4
	AxA	1180.7	1222.4	2391.7	2420.8	19.6 ± 47.0	51.7 ± 47.1
	AxD	**1173.0**	**1214.7**	**2378.9**	**2408.0**	**18.1 ± 47.0**	50.2 ± 47.2

Table 3. Best-fitting single-step multi-environment model results. Variance components (± standard errors), narrow-sense heritabilities (h^2), proportion of the total phenotypic variance explained by dominance (d^2), epistasis (i^2_{epi}), and broad-sense heritability (H^2) are provided. Model log-likelihoods are also given. The models shown were selected on the basis of having the lowest Akaike Information Criterion (AIC) relative to other tested models.

Dataset	Genetic Gain (GG)			Cycle 1 (C1)		
Trait	DM	RTWT	MCMDS	DM	RTWT	MCMDS
Best Model	Add	AxA	AxD	A+D	Dom	AxD
$\sigma^2_{loc.year}$	0.025 (4.8)	0.056 (0.03)	0.051 (0.02)	8.38 (8.4)	0.006 (0.007)	0.054 (0.055)
σ^2_{rep}	6.16 (5.4)	0.014 (0.01)	0.000 (0)	- -	- -	- -
σ^2_{add}	10.44 (1)	0.029 (0.012)	0.32 (0.1)	17.3 (2.5)	- -	1.780 (0.178)
σ^2_{dom}	- -	0.020 (0.011)	0.000 (0.08)	3.4 (1.5)	0.116 (0.018)	0.172 (0.082)
σ^2_{epi}	- -	0.033 (0.014)	0.556 (0.09)	- -	- -	0.514 (0.101)
σ^2_{error}	15.36 (0.33)	0.17 (0.003)	0.34 (0.006)	10.7 (0.7)	0.25 (0.011)	0.26 (0.023)
h^2	0.33	0.09	0.25	0.43	-	0.64
d^2	-	0.06	0.00	0.08	0.31	0.06
i^2	-	0.10	0.44	-	-	0.18
H^2	0.33	0.25	0.69	0.52	0.31	0.89
loglik	-9457	2378.1	-581	-3538	184	-1184

Figure 1. Partitioning of broad-sense heritability for single-step multi-environment models in the Genetic Gain and Cycle 1 datasets. Results from each of five models are shown in each panel broken down by trait (rows) and population (columns). Models include additive only (Additive), dominance only (Dominance), Additive plus Dominance (Add + Dom), Additive plus dominance plus either AxA epistasis (AxA Epistasis) or AxD epistasis (AxD Epistasis).

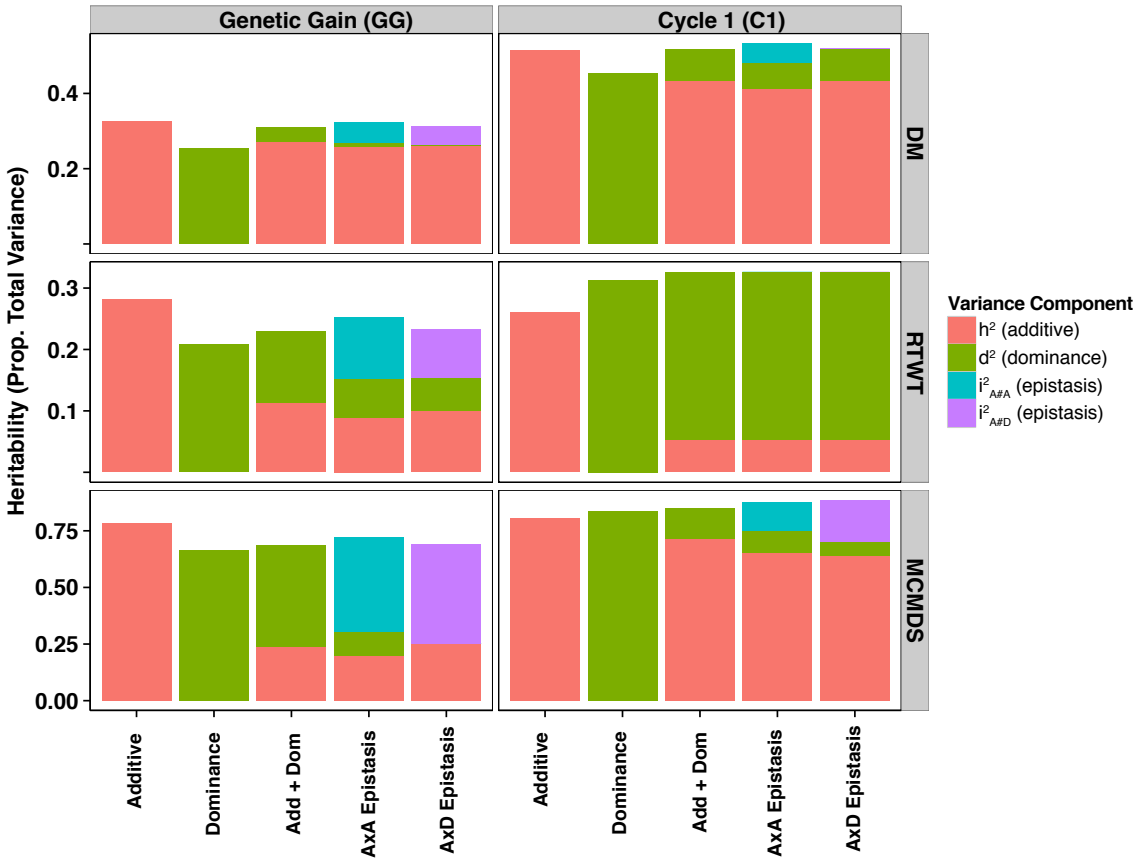

Figure 2. Distribution of genetic variance proportions across Genetic Gain trials. Three models were fitted for each trait in each of 47 Genetic Gain trials. Each panel contains boxplots showing the distribution of proportions of the phenotypic variability explained by a corresponding genetic factor, including the broad-sense heritability (H^2). Red horizontal lines are the median narrow-sense heritability (h^2) from the additive only model. Traits are on columns and three models are on the rows: additive plus dominance (Add + Dom), additive plus dominance plus AxA epistasis (AxA Epistasis) and additive plus dominance plus AxD epistasis (AxD Epistasis).

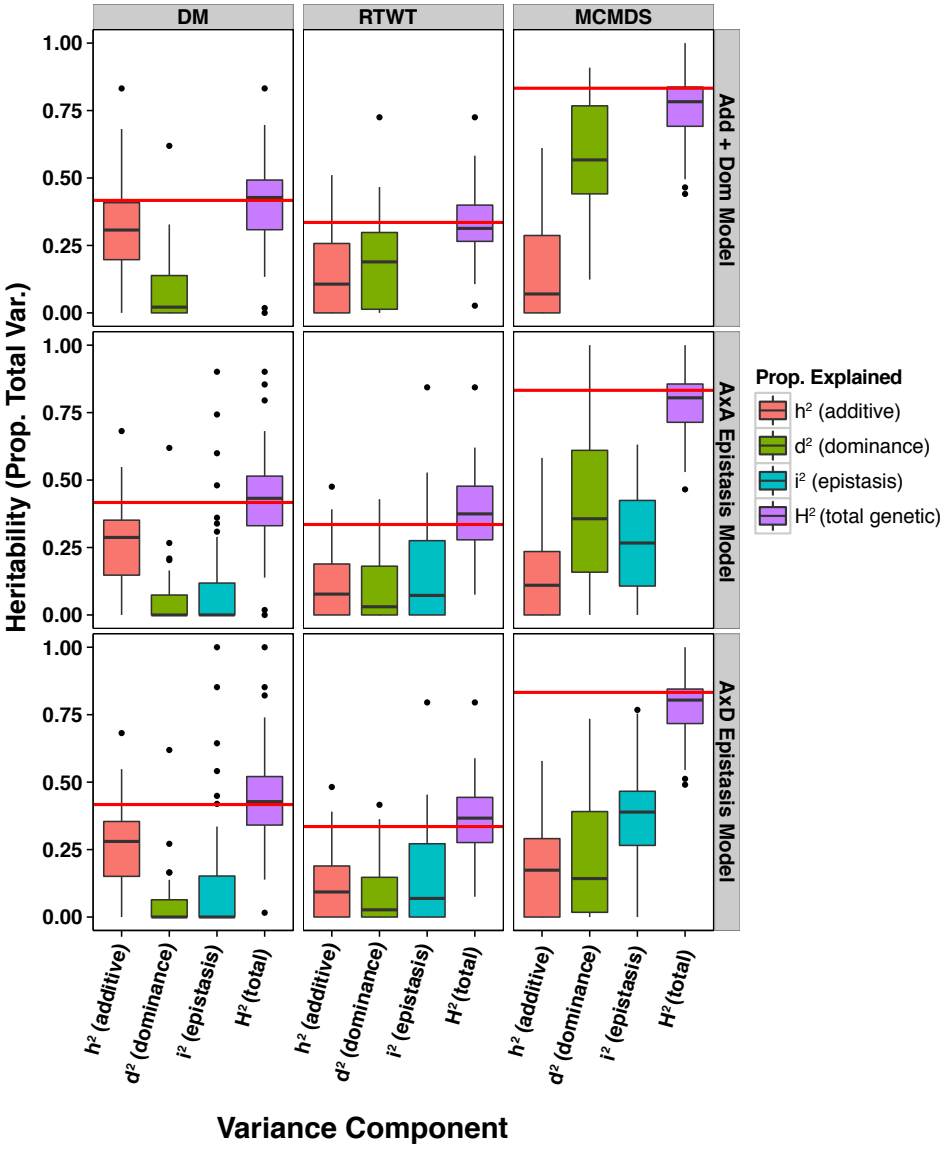

Figure 3. Accuracy of total genetic value prediction in the Genetic Gain and Cycle 1 datasets. Boxplots showing the distribution over 25 replicates of 5-fold cross-validation of the prediction accuracy of the total genetic value from five different models are shown in each panel. The accuracy within the Genetic Gain (red) and Cycle 1 (blue) are shown. Traits are in the columns. Accuracy is defined as the correlation between the sum of predictions from all genetic variance components in the model and the BLUP from the first stage of analysis where location, year and replicate variability were removed. Models included are: additive only (Additive), dominance only (Dominance), additive plus dominance (Add + Dom), additive plus dominance plus AxA epistasis (AxA Epi.) and additive plus dominance plus AxD epistasis (AxD Epi.).