Genome-wide association and prediction reveals the genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement

Authors:
Marnin D. Wolfe*, Ismail Y. Rabbi, Chiedozie Egesi, Martha Hamblin, Robert Kawuki, Peter Kulakow, Roberto Lozano, Dunia Pino Del Carpio, Punna Ramu, Jean-Luc Jannink

Marnin D. Wolfe, Martha Hamblin, Roberto Lozano, Dunia Pino Del Carpio, Punna Ramu and Jean-Luc Jannink, Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA; Ismail Y. Rabbi and Peter Kulakow, International Institute for Tropical Agriculture (IITA), Ibadan, Oyo, Nigeria; Chiedozie Egesi, National Root Crops Research Institute (NRCRI), Umudike, Umuahia, Nigeria; Robert Kawuki, National Crops Resources Research Institute (NaCRRI), Namulonge, Uganda; Jean-Luc Jannink, USDA-ARS, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA. Marnin D. Wolfe and Ismail Y. Rabbi contributed equally to this work.
*Correspondence author Marnin Wolfe (wolfemd@gmail.com)

Received _____

Abbreviations:
Genome-wide association analysis (GWAS); genomic selection (GS); Cassava mosaic disease (CMD); genotype-by-sequencing (GBS); International Institute of Tropical Agriculture (IITA); National Root Crops Research Institute (NRCRI); National Crops Resources Research Institute (NaCRRI); African Cassava Mosaic Virus (ACMV); East African Cassava Mosaic Virus (EACMV); genomic estimated breeding values (GEBVs); Cassava mosaic disease severity (CMDS)

ABSTRACT

Cassava (*Manihot esculenta*) is a crucial, under-researched crop feeding millions worldwide, especially in Africa. Cassava mosaic disease (CMD) has plagued production in Africa for over a century. Bi-parental mapping studies suggest primarily a single major gene mediates resistance. To be certain and to potentially identify new loci we conducted the first genome-wide association mapping study in cassava with 6128 African breeding lines. We also assessed the accuracy of genomic selection to improve CMD resistance. We found a single region on chromosome 8 accounts for most resistance but also identified 13 small effect regions. We found evidence that two epistatic loci and/or alternatively multiple resistance alleles exist at major QTL. We identified two peroxidases and one thioredoxin as candidate genes. Genomic prediction of additive and total genetic merit was accurate for CMD and will be effective both for selecting parents and identifying highly resistant clones as varieties.

Cassava (*Manihot esculenta* Crantz) is a crucial staple food crop, usually grown by smallholder farmers and feeding over half a billion people worldwide, especially in sub-Saharan Africa (http://faostat.fao.org). Breeding cycles are long in this outcrossing, clonally-propagated crop and genetic gains from breeding have been small over the last century compared with other crops (Ceballos et al., 2004, 2012). With a recently sequenced genome (Prochnik et al., 2012) and a high-density SNP-based genetic linkage maps ((ICGMC), 2014), it is for the first time possible to study the genetic architecture of key traits using modern genome-wide association analysis (GWAS) and to improve those traits with genomic selection (GS) (Oliveira et al., 2012; Ly et al., 2013).

Cassava mosaic disease (CMD) is the longest running and thus-far most impactful of the challenges cassava farmers face in sub-Saharan Africa(Fauquet et al., 1990). The disease is caused by several related species of geminiviruses and transmitted both through infected cuttings and by a vector, the common whitefly (*Bemisia tabaci* G.). Development and deployment of resistant cultivars is the most effective control method for this devastating disease. Following an unsuccessful world-wide search for resistance in cultivated germplasm (*M. esculenta*) in the 1930s, cassava breeders at the Amani research station in Tanzania resorted to interspecific hybridization with Ceara rubber tree (*M. glaziovii* Müll. Arg) and other related wild species in the 1930s (Hahn et al., 1979, 1980a; Fauquet et al., 1990). Moderate polygenic resistance combined with reasonable root yields was achieved through several cycles of backcross of Ceara rubber to the cultivated cassava (Hahn et al., 1980b). One of these interspecific hybrids, clone 58308, was subsequently

68 used to initiate cassava breeding breeding at the International Institute of Tropical
69 Agriculture (IITA) in the 1970s and resulted in the Tropical Manihot Selections
70 (TMS) varieties (Hahn et al., 1980b).

71 More recently, a strong qualitative and dominant monogenic resistance
72 known as *CMD2* was discovered in a Nigerian landrace (TMEB3) in the 1980's
73 (Akano et al., 2002). Multiple bi-parental QTL analyses have been conducted,
74 initially using SSR markers (Akano et al., 2002; Lokko et al., 2005; Okogbenin et al.,
75 2007, 2012a; Mohan et al., 2013) but more recently genome-wide SNPs (Rabbi et al.,
76 2014a; b) to understand the genetic basis of this type of qualitative resistance to
77 CMD. Although some studies hint at additional resistance loci (Okogbenin et al.,
78 2012a; Mohan et al., 2013) most evidence points solely to the *CMD2* locus (Rabbi et
79 al., 2014a; b). However, these bi-parental mapping efforts relied on a handful of
80 unique parental genotypes from West Africa and therefore only examined a narrow
81 slice of African cassava germplasm diversity (Rabbi et al., 2014b).

82 A limited genetic base for the dominant resistance implies potential
83 vulnerability if the cassava mosaic geminivirus can evolve to overcome it. This
84 possibility necessitates diversification of resistance sources to ensure durability. In
85 order to determine with greater certainty whether there are additional sources of
86 CMD resistance in the continent's breeding germplasm, we undertook a large
87 Genome-Wide Association Study (GWAS) using over six thousand cassava
88 accessions from West and East Africa genotyped at more than 40,000 SNP loci using
89 genotype-by-sequencing (GBS) approach (Elshire et al., 2011). The entire collection
90 represents five sub-populations (Table 1) that are part of an ongoing international

genomic selection-based breeding project in cassava (**http://www.nextgencassava.org**). In addition, we combined GWAS and genomic prediction in order to not only dissect the genetic architecture of resistance to CMD but also to assess the potential for population improvement by genomic selection (GS). We used a variety of approaches to localize and identify candidate genes for future investigation. The potential for GS to improve CMD resistance and for non-additive models to predict total genetic merit of clones for the selection of superior CMD resistant varieties were assessed. Finally, multi-kernel genomic prediction models were used to study the relative importance of qualitative and quantitative resistance sources.

MATERIALS & METHODS

Germplasm collection

The germplasm included in this study represent the reference populations used to develop genomic prediction models as part of a collaborative project between Cornell University and three breeding institutions: The International Institute of Tropical Agriculture (IITA) in Ibadan, Nigeria, the National Root Crops Research Institute (NRCRI) in Umudike, Nigeria and the National Crops Resources Research Institute (NaCRRI) in Namulonge, Uganda. The IITA's Genetic Gain population is comprised of 694 historically important, mostly advanced breeding lines that have been selected and maintained clonally since 1970 (Okechukwu and Dixon, 2008; Ly et al., 2013). Most of these materials are derived from the cassava gene-pool from West Africa and early introductions of CMD tolerant parents derived from the inter-specific hybridization program at the Amani Station in Tanzania (Hahn et al., 1980b). It also includes hybrids of germplasm introduced from Latin America (see Table S1 for a list of accessions and details on pedigree where available). The NRCRI population contains 626 clones from their breeding program, 189 of which are also part of IITA's Genetic Gain. The remainder of the NRCRI collection includes a large number of materials either directly from or derived with parentage from the International Center for Tropical Agriculture (CIAT) in Cali, Columbia (Table S2).

There are two major clades of cassava mosaic virus species, African Cassava Mosaic Virus (ACMV) and East African Cassava Mosaic Virus (EACMV) (Legg and Fauquet, 2004). EACMV is generally more severe in its symptoms and is present in

125 west Africa but only in low proportion to ACMV, usually occurring as a dual infection
126 (Legg and Fauquet, 2004; Rabbi et al., 2014b). This fact makes it all the more
127 important to include east African cassava breeding germplasm in a more
128 comprehensive screen of the genetic architecture of CMD resistance. The NaCRRI in
129 Uganda has a population of 414 clones that represent the genetic diversity of East
130 African cassava gene pool. The pedigree of this population arises from 49 parents
131 coming from IITA, CIAT in Columbia and Amani Research Station in Tanzania (Table
132 S3). The population was generated in part by making crosses of parents with
133 qualitative resistance to parents with quantitative resistance as well as quantitative
134 x quantitative and qualitative x qualitative resistances.
135 We also analyzed a large genotyped and phenotyped multi-parental
136 population of individuals from two cycles of genomic selection (GS) conducted at
137 IITA. The GS program at IITA will be described briefly here and in detail as part of
138 other publications. In 2012 the IITA Genetic Gain population was used as the
139 reference population from which genomic estimated breeding values (GEBVs) were
140 obtained using the genomic BLUP method (GBLUP) (Heffner et al., 2009). Selection
141 of clones from the Genetic Gain was based on a selection index including CMD and
142 cassava bacterial blight disease severity and yield components (dry matter content,
143 harvest index and fresh root yield). In the end, 83 parents gave rise to 2187
144 progenies, which we will call IITA Cycle 1. In 2013, the GEBVs for Cycle 1 were
145 obtained, again using the Genetic Gain as a reference population and 84 Cycle 1 plus
146 13 (97 total) Genetic Gain clones were selected as parents, giving rise in 2014 to

147 2466 progenies (Cycle 2). The pedigrees of IITA Cycle 1 and Cycle 2 are available in
148 Tables S4-S5.
149
150 **Phenotyping Trials**
151 Phenotypic data were combined from trials conducted at multiple locations
152 in Nigeria and Uganda. The data are contributed from all three breeding programs
153 (IITA, NRCRI and NaCRRI). IITA's Genetic Gain trials were conducted in seven
154 locations over 14 years (2000 to 2014) in Nigeria. Each Genetic Gain trial comprises
155 a randomized, unblocked design replicated one or two times per location and year.
156 NRCRI's population was phenotyped in two years, 2013 and 2014. During the 2012-
157 2013 season the trial was conducted in one location, Umudike, Nigeria. In 2013-
158 2014 the population was planted in three locations (Umudike, Kano, and Otobi).
159 NRCRI's trial design was a randomized incomplete block with three replications per
160 location/year and 10 plants per plot. Trials at NaCRRI were conducted in two years:
161 2012-2013 and 2013-2014. In both years plots were 10 plants in two rows of 5 with
162 randomized incomplete blocks. During the first year, a single location (Namulonge,
163 Uganda) was used with only one replicate. During the second season, two
164 replications were used at each of three locations: Namulonge, Kasese and Ngetta.
165 Genomic selection (GS) Cycle 1 (C1) progenies were observed as seedlings in the
166 2012-2013 field season with phenotyping conducted only for early disease
167 expression and seedling vigor. Cycle 1 progenies were subsequently cloned and
168 phenotyped in a three-location (Ibadan, Ikenne, and Mokwa) trial in 2013-2014
169 with all phenotypes scored. 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. Clones were assigned to each location so as to equally represent each family in every environment. The GS Cycle 2 (C2) individuals were observed in a seedling trial during the 2013-2014 field season. We note that expression of disease symptoms in cassava seedlings may not be representative of expression in clonal evaluations. This is in part because seedling symptoms can arise solely from whitefly transmission, making it probable that some asymptomatic plants are in fact escapes rather than resistant genotypes. Table 1 summarizes the phenotypes and phenotyping trials available for each sub-population. We also provide details about the sample sizes and replication numbers for each location/year of data analyzed (Table S6) and per accession (Table S7)

Cassava mosaic disease severity (CMDS) was scored on a scale of 1 to 5, with 1 representing no symptoms and 5 indicating the most severe symptoms. CMDS was scored at up to five time points (1, 3, 6, 9 and 12 months after planting) depending on the trial. Additionally, we analyze the season-wide mean CMDS score (MCMDS), which is used for making selections and the area under the disease progress curve (see below; AUDPC). The distribution of raw phenotypic data used in each population and for each trait can be seen in Figures S1-S6.

Statistical Models and Analyses of Phenotypes

Our interest in this study was to identify key aspects of the genetic architecture of cassava in Africa rather than location- or year-specific QTLs. We condensed up to 38854 observations on 6198 genotyped and phenotyped clones to

single BLUPs for each. To do this, we fit the following mixed linear model with the *lme4* package in R:

$$y_{l,i,j} = \mu + c_l + \beta_i + r_{j(i)} + \varepsilon_{l,i,j} \quad [1]$$

Here, $y_{l,i,j}$ represents raw phenotypic observations, μ is the grand mean, c_l is a random effects term for clone with $c_l \sim N(0, \sigma_l^2)$, β_i is a fixed effect for the combination of location and year harvested, $r_{j(i)}$ is a random effect for replication nested within location-year combination assumed to be distributed $N(0, \sigma_r^2)$ and finally, $\varepsilon_{l,i,j}$ is the residual variance, assumed to be random and distributed $N(0, \sigma^2)$. Because the number of observations per clone varies greatly in our dataset (from 1 to 941, median of 2; Table S7), we expect BLUPs are differentially shrunken to the mean. To counter this, we de-regressed BLUPs according to the following formula:

$$deregressed\ BLUP = \frac{BLUP}{1 - \frac{PEV}{\sigma_l^2}} \quad [2]$$

Where PEV is the prediction error variance for each clone and σ_l^2 is the clonal variance component. The distribution of deregressed BLUPs used as response variables in GWAS can be seen in Figures S7-S13.

We also calculated areas under disease progress curves (AUDPC) for each clone using data from 1, 3 and 6 months after planting. To do this, we treated severity scores from any time point as the same trait with a second variable indicating the time point of the score. We then ran the model indicated in [1] but with c_l indicating the clone-time point combination. This gave us a deregressed

215 BLUP for each clone at each time point. We calculated areas under these curves
216 using the trapezoid rule as implemented by the *auc* function in the *flux* R package
217 (http://cran.r-project.org/web/packages/flux/index.html). We excluded 9 and 12
218 months data because they were only scored at NRCRI, thus including them would
219 have limited the ability to compare results for this trait between populations.
220
221 **Genotype Data**
222 Genotyping of SNP markers was done by the genotyping-by-sequencing
223 procedure (Elshire et al., 2011) using the ApeKI restriction enzyme recommended
224 by (Hamblin and Rabbi, 2014) and read lengths of 100bp. Marker genotypes were
225 called with the TASSEL GBS pipeline V4(Glaubitz et al., 2014) and aligned to the
226 cassava version 5 reference genome, available on Phytozome
227 (http://phytozome.jgi.doe.gov) and described by the International Cassava Genetic
228 Map Consortium (2014). Individuals with >80% missing SNP calls and markers with
229 more than 60% missing were removed. Also, markers with extreme deviation from
230 Hardy-Weinberg equilibrium (Chi-square > 20) were removed. Allele calls were
231 maintained if depth was ≥2 otherwise the call was set to missing. Marker data was
232 converted to dosage format (0, 1, 2) and missing data were imputed with the glmnet
233 algorithm in R (http://cran.r-project.org/web/packages/glmnet/index.html) as
234 described in(Wong et al., 2014). In order to judge the resolution of association
235 analyses we calculated pair-wise linkage disequilibrium (LD) between all markers
236 with a MAF of 5% on each chromosome using PLINK (version 1.9, https://www.cog-

237 genomics.org/plink2). We examine the rate of decay with increasing physical
238 distance between markers.
239
240 **Population Structure and Genome-Wide Association Analyses**
241 In order to examine the patterns of relatedness within and among our
242 populations and to control population structure, we constructed a genomic-
243 relationship matrix according to the formulation of Van Raden (VanRaden, 2008);
244 see also 25), as implemented in PLINK, using all markers with greater than 1%
245 minor allele frequency (MAF). We also use this relationship matrix for genomic
246 prediction (see below).
247 We conducted principal components analysis (PCA) on SNP markers with
248 MAF > 5% using the *prcomp* function in R. PCA on SNP markers is often used to
249 identify major patterns of relatedness (population structure) in a sample and the
250 first few PCs can be used as covariates to control false-positive rates in GWAS(Price
251 et al., 2006). Because the genomic selection progenies (C1 and C2) are by far the
252 largest part of our dataset and because these individuals are descended from the
253 IITA Genetic Gain population, we excluded these from the initial PCA. We then
254 projected these individuals into the genetic space defined by the three training
255 populations (NRCRI, IITA, NaCRRI) using the *predict* function in R. This allows us to
256 visualize and quantify the relatedness in our populations based on the founders only
257 and unbiased by the large size of the C1 and C2 collections.
258 Because GWAS has not previously been done in this or any other cassava
259 collection, we tested several different models for controlling population structure.

In particular, we compared the genome-wide inflation of p-values between a general linear model (GLM) with no population structure controls, a GLM with 5 principal components (GLM + 5 PCs)(Price et al., 2006), and a mixed linear model (MLM), which fits a random effect for clone with $\sim N(0,\sigma_g^2 K)$, where σ_g^2 is the clonal variance component and K is the relationship matrix described above (Kang et al., 2010). MLM were conducted using the P3D and compression method (Zhang et al., 2010). All GWAS were conducted in TASSEL (version 5, [27]). We compare the observed $-\log_{10}$(p-values) against the expectation using QQ-plots. We used visual inspection of QQ-plots to judge which model most effectively reduced the genome-wide inflation of $-\log_{10}$(p-values) typically attributed with population structure. We consider association tests significant when more extreme than the Bonferroni threshold (with experiment-wise type I error rate of 0.05).

Because marker effects, LD patterns, and allele frequencies may differ within as well as across sub-populations, we conducted GWAS population-wide as well as within each sub-population. In each analysis, we used markers that segregated with MAF > 5% in that specific sub-population. Bonferroni thresholds were calculated according to the number of markers analyzed in each sub-population.

We also examined the proportion of variance in the deregressed BLUPs explained by the kinship matrix, K using the variance components estimated when TASSEL fits the MLM model.

Candidate Genes

Because the underlying mechanisms of plant disease resistance are of general interest and identification of causal polymorphism may aid in transgenic approaches and/or marker assisted selection, we identified candidate genes in CMD associated regions. Significant SNPs from the GWAS results corresponding to four time points (1, 3, 6 and 9 months after planting) were selected for the analysis. We considered SNPs that were both above the Bonferroni threshold and were located within exons or introns of cassava genes. The SNP position on the genome was compared with the gene interval position using the annotation list from Phytozome 10. Gene ontology annotation for each time point and combining all the datasets was done with Panther (http://go.pantherdb.org/). We have generated whole genome sequences (WGS) from one CMD resistant clone (I011412) and two CMD tolerant clones (I30572 and TMEB1). TMEB1 is a landrace from Ogun State, Nigeria also called Antiota, that is not likely to contain the qualitative resistance allele and is usually classified as tolerant or only partially susceptible to CMD (Raji et al., 2008; Rabbi et al., 2014b). Similarly, I30572 is an improved variety whose parents were the *M. glaziovii*-derived clone 58308 and a south American Cassava (Branca de Santa Catarina) and is therefore known to have only the quantitative resistance source (Fauquet et al., 1990). *PCR*-free libraries were generated from these clones and sequenced at 20X coverage using Illumina HiSeq. Additionally two resistant clones (TMEB3 and TMEB7) were obtained from Phytozome (http://phytozome.jgi.doe.gov). TMEB3 is itself the original landrace parent from which the qualitative resistance source has been derived and TMEB7 has been shown to be nearly genetically identical to TMEB3. We therefore define TMEB3,

TMEB7 and I011412 as "resistant" lines while TMEB1 and I30572 will be referred to, for simplicity as "susceptible" primarily on the basis of whether they do or do not have the qualitative resistance source *CMD2*. These sequences were aligned against the cassava V5 reference genome assembly to call the variants to identify the genomic difference between resistance and susceptible clones in candidate gene loci. Since the genotypes compared were few in number, we called SNPs manually using an exon annotated sequence and the Integrative Genomics Viewer software (IGV; http://www.broadinstitute.org/igv/).

Genomic Prediction of Additive and Total Genetic Merit

We used a multi-random effects (a.k.a. multi-kernel or multi-relationship matrix) genomic prediction model to compare the variance explained and prediction accuracy achieved from the major CMD QTL (CMD2) compared to the rest of the genome. Specifically, we created relationship matrices either from all markers, markers significantly associated with CMD2 from GWAS results, or all markers *not* in the region of the QTL.

For additive relationships, we used the formulation described above for controlling population-structure (VanRaden, 2008). Dominance relationships can be captured as $\mathbf{D} = \frac{HH'}{\sum_i 2p_i q_i (1-2p_i q_i)}$ (Su et al., 2012; Muñoz et al., 2014). Where H is the SNP marker matrix (individuals on rows, markers along columns), heterozygotes are given as $(1 - 2p_i q_i)$ and homozygotes are $(0 - 2p_i q_i)$. We made a custom modification (available upon request) to the *A.mat* function in the *rrBLUP* package

(Endelman, 2011) to produce the **D** matrix. Relationship matrices that capture epistasis can also be calculated by taking the hadamard product (element-by-element multiplication; denoted #) of two or more matrices (Henderson, 1985). For simplicity, we tested an additive-by-dominance (**A#D**) matrix in this study.

We tested four models. Model 1 used all markers and only a single, additive kernel ($Additive_{All\_Markers}$). Model 2 used all markers but three kernels, $Additive_{All\_Markers} + Dominance_{All\_Markers} + Epistasis_{All\_Markers}$. Model 3 used two additive kernels, one constructed from the 163 CMD2 significant markers ($Additive_{CMD2}$) and the other from all markers outside of the chromosomal region bounded by CMD2 markers ($Additive_{Non\text{-}CMD2}$). Model 4 had four kernels: $Additive_{CMD2} + Dominance_{CMD2} + Epistasis_{CMD2} + Additive_{Non\text{-}CMD2}$.

We assessed the influence that modeling non-additive genetic variance components have on genomic prediction using a cross-validation strategy (see below). We used the deregressed BLUPs for MCMDS as described above. In our data, the number of observations per clone ranges from one to 941 (checks, TMEB1 and I30572) with median of two and mean of 10.6 (Table S7). Pooling information from multiple years and locations, especially when there is so much variation in numbers of observations can introduce bias. Much theoretical development, particularly in animal breeding has been done to address this issue, and we followed the approach recommended by Garrick et al. (2009)

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 $\frac{1-h^2}{0.1+\frac{1-r^2}{r^2}h^2}$, where h^2 is the proportion of the total variance explained by the clonal variance component, σ_g^2 derived in the first step **(Garrick et al. 2009)**.

We implemented a 5-fold cross-validation scheme replicated 25 times to test the accuracy of genomic prediction using the genomic relationship matrices and models described above. In each replication, we randomly assign each individual to one of five groups (folds). We then select one fold, remove the corresponding individuals from the training set and use the remaining four folds to predict the fold that was left out. We iterate this process over each of the five folds to produce a prediction for each individual that was made while its phenotypes were unobserved. For each replicate of each model, we calculated accuracy as the Pearson correlation between the genomic prediction made when phenotypes were excluded from the training sample and the BLUP (ĝ, not-deregressed) from the first step. For each model, we calculated accuracy both of the prediction from the additive kernel (where present) and the total genetic merit prediction, defined as the sum of the predictions from all available kernels (e.g. additive + dominance + epistasis). Genomic predictions were made using the *EMMREML* R package. For simplicity, we tested only the trait MCMDS in the IITA Genetic Gain population.

In addition, we assessed the predictability of CMD based on random forest regression (RF), a non-linear, machine-learning approach that excels at capturing non-additive especially interaction-type genetic effects(Jannink et al., 2010). We used RF only with the significant CMD2 associated markers as predictors to assess

372 additional evidence for interaction at this locus on the basis of prediction accuracy

373 achieved. We used the same cross-validation scheme described above.

374

375

376

RESULTS

Genotyping Data

SNP marker data was generated using genotyping-by-sequencing (GBS) (Elshire et al., 2011). Overall coverage was 0.07x (range 0.05-0.2). There were 114,922 markers that passed initial filters with a MAF > 1%. Of these, 95,047 are mapped to the genome. Of mapped markers used for GWAS (MAF > 5%), there was an average of 2293 SNPs per chromosome or one marker every 9.5 kb. The mean MAF (0.21-0.22), mean heterozygosity (0.32-0.35) and number of markers analyzed (40,539-42,113) were similar between sub-populations (Table 1). Most chromosomes in most populations had mean $r^2 > 0.2$ extending 10 to 50 Kb. The r^2 between markers 4.5-5.5kb apart was 0.3 on average (median 0.13) suggesting at least some LD between most causals and at least one marker but also that increased density in future studies will provide additional mapping resolution (Figs S14-S19).

Population stratification and structure

Principal components analysis of our SNP dataset revealed subtle differentiation among African cassava clones analyzed. This can been seen from a plot of the first four PCs (cumulative variance explained = 15%). The Nigerian sub-populations (NRCRI, IITA Genetic Gain, Cycle 1 and Cycle 2) occupy similar genetic space, but the Ugandan sub-population (NaCRRI) is somewhat distinct on PC1 and PC2. This may be consistent with a history of germplasm sharing and recurrent use of elite parents among African breeding institutes.

399 We tested several standard GWAS models for controlling inflation of p-values
400 caused by population structure including a general linear model (GLM, no
401 correction); a GLM with the first 5 PCs of the SNP matrix as covariates; and a mixed-
402 linear model using the marker-estimated kinship matrix. Visual inspection of QQ
403 plots (Fig. 2 inset, Figs. S20-25) indicated that the MLM was most consistent for
404 reducing $-\log_{10}$(p-values) towards the expected level (i.e. controlling false-positives,
405 removing population structure effects). All subsequent results are therefore based
406 on mixed-model associations. From the variance components estimated when fitting
407 MLMs we found that kinship matrices explained on average 57% (range 31-94%) of
408 the phenotypic variance (Table S1).
409
410
411 **Overall Genome-wide Associations**
412 Association tests were performed for CMD symptom severity at one, three,
413 six, nine and twelve months after planting (where measured) in the five sub-
414 populations (Table 1) and in analysis that combined all accessions. We identified
415 311 markers in total that pass a Bonferroni significance threshold (Fig. 2, Table S8).
416 However, many significant SNPs were detected because of rare marker genotypes
417 that were phenotypically extreme (Figure S26). The F-test implemented by TASSEL
418 is sensitive to imbalanced sample size between groups and we wish to be
419 conservative and only consider significant results that we can be confident in.
420 Therefore, we only consider SNPs where each genotype class (e.g. aa, Aa, AA) is
421 represented by at least 10 individuals. This reduced the number of significant

422 markers to 198, on 14 chromosomes, mostly concentrated at a single region of
423 chromosome 8. Significant results were found within each sub-population, with
424 more signals associated with greater sample size (e.g. Cycle 1). Variance explained
425 by significant markers ranged from 0.5% to 22% (median 3.5%) (Table S9).
426
427 **Chromosome 8 contains the major resistance locus, *CMD2***
428 There were 163 significant markers on chromosome 8 (between 3.56-11.38
429 megabases; Fig. 3a) with the top marker (S8_7762525) explaining 5-22% of the
430 variance depending on the sub-population. The frequency of the resistance-
431 associated allele at S8_7762525 is 56% overall (range: 44% in IITA Genetic Gain to
432 63% in IITA Cycle 2 progenies).
433 The resistance allele at S8_7762525 is incompletely dominant (Fig. 3 inset);
434 homozygotes with the alternate allele were closer to CMD free than heterozygotes.
435 To formally test this, we conducted a post-hoc test for an additive effect at this
436 marker that explained 15% of the variance compared to a test of additive plus
437 dominance effect that explained 20%, and a test of dominance alone that accounted
438 for only 2%.
439 We confirmed that our major QTL is the *CMD2* locus by aligning previously
440 published SSR marker primers (SSRY28, NS158 and SSRY106) (Akano et al., 2002;
441 Lokko et al., 2005; Okogbenin et al., 2007, 2012a; Mohan et al., 2013) to the
442 reference genome using E-PCR (http://www.ncbi.nlm.nih.gov/tools/epcr/). Our
443 significant markers on chromosome 8 co-locate with these markers (Fig. S28a).
444 Additionally, scaffolds bearing the significant QTL reported in Rabbi et al. (2014a;

b) are located in this region. However, while Rabbi et al.'s (2014a; b) strongest association was on scaffold 5214, corresponding to Chr. 8 position 6511133, the strongest association for the present study is on scaffold 6906 (7454373-7836749), more than a megabase away. This discrepancy is due to the fact that the SNP markers in scaffold 6906 did not segregate in the resistant parents of the bi-parental mapping populations.

Dissecting resistance originating from alleles or loci on Chromosome 8

The significance region on chromosome 8 is large (~8 Mb; Fig. 3a). In fact, the region appears as two, sometimes equally significant peaks in some sub-populations (Fig. S27). We scanned the region for haplotype blocks with PLINK (version 1.9, https://www.cog-genomics.org/plink2) and found it was not characterized by a single, or even a few large, but many small LD blocks (Fig. S29). A second locus (*CMD3)* has been reported on the same chromosome as *CMD2*(Okogbenin et al., 2012b). The authors reported the marker NS198 to be 36 cM from *CMD2* and associated with very strong resistance in the progeny of TMS972205. E-PCR collocated NS198 on chromosome 8, five megabases (position 997099) outside our significance region (Fig. S28). Thus our results suggest a second QTL (i.e. *CMD3*), if present, is much closer to *CMD2* than previously believed.

We used several approaches to evaluate the evidence for multiple QTL in the region. We conducted a post-hoc test for interactions between the top-marker on chromosome 8 and every other marker on the chromosome. There were significant

interactions, explaining up to 42% of the variance, 1-3 megabases from the top GWAS hits, but none in the region surrounding S8_7762525 (Fig. 3b, Fig. S30).

We implemented a multi-locus mixed-model (MLMM (Segura et al., 2012)), which uses a forward-backward stepwise model selection approach to determine which and how many marker cofactors are required to explain the associated variance in the region. The MLMM for MCMDS in the population-wide sample selected five markers (S8_7762525, S8_6380064, S8_6632472, S8_7325389, S8_4919667) spanning the significance region (Fig. S31). Of the five, the first was our top marker S8_7762525, the fourth is only about 400 Kb away, and the remaining three cover the secondary peak and the region of statistical interaction. These markers are mostly in linkage equilibrium (Table S10) and collectively explain up to 40% of the variance. The selection of markers distributed across the region by MLMM including both putative peaks to explain the phenotypic-association in the region is additional evidence in support of a multi-locus hypothesis.

LD decays in the region to low levels ($r^2<0.25$) and is virtually zero between significant markers on the left, e.g. S8_5064191 and those on the right, e.g. S8_762525 of the significance region (Fig. 3c). Combined with our genome-wide analysis of LD decay rates (Figs. S14-S19) this LD decay makes it unlikely that a single locus or allele is responsible for the associated region.

We examined the two-locus genotype effects (e.g. between S8_7762525 and the SNP with the most significant interaction test, S8_4919667) and found a usually dominant effect of the secondary resistance allele (e.g. S8_4919667) in the

heterozygous and homozygous resistant background at the primary peak (Fig. 4; Fig. S32). We found little evidence of secondary peak effects in the homozygous susceptible primary peak background. Clones that are homozygous resistant at both loci are superior to all other cassava clones, expressing virtually no symptoms (Fig. 4c).

Other Loci Associated with CMD Resistance

We identified thirty-five markers on 13 chromosomes that explained 0.5–10% (median 4%) of the variance (Table S9). Many of these had recessive and usually rare susceptibility alleles (Fig. S26). Marker S4_637212 explained 4% of the variance (CMD6S, Genetic Gain) and had an additive effect. Marker S11_20888811's recessive resistance allele appears to lower CMD symptoms 4% more than *CMD2* (S8_7762625) but only 14 clones are homozygous resistant at this locus (Fig. S26). Further work on this locus is urgently needed to determine it's possible impact as its frequency increases. There were four significant markers on chromosome 14 with mostly dominant effects and explaining up to 5% of the variance. Two previously published SSR markers (SSRY44, NS146) (Mohan et al., 2013) are located within 1.4 megabases of these SNPs (Fig. S28b). Four markers, spread across seven megabases of chromosome 9, with recessive susceptibility loci, explained up to 10% (S9_14551208) of the variance. These markers co-located with SSRY40, originally reported as *CMD1* and associated with quantitative resistance (Fregene et al., 2000; Mohan et al., 2013).

513 **Candidate Genes**

514 We intersected our association-results with available gene annotations and
515 related data and identified 105 unique genes within the association peaks, with 79,
516 61, 56 and 9 genes identified at one, three, six and nine months after planting,
517 respectively (Table S11, Fig. S33). There were no significant differences between
518 gene ontology categories between time points. Most of the annotated genes are
519 involved in metabolic processes (Fig. S34). Thirty-five out of the 105 genes are
520 known to respond to cassava mosaic virus infection (Allie et al., 2014) (Table S11).
521 Among these genes we found ones known to be susceptibility or resistance
522 factors, a number of which are also involved in plant-geminivirus interaction
523 processes (Hanley-Bowdoin et al., 2013). We found two peroxidases
524 Cassava4.1_029175 and Cassava4.1_011768 within the primary QTL region
525 (scaffold 6906, ~7.7Mb); peroxidases are pathogenesis-related proteins (PRs),
526 involved in host response to infection(van Loon et al., 2006). In the secondary GWAS
527 peak (scaffold 5214, 5-6Mb) six SNPs were in a protein disulfide-isomerase like 2-2
528 ortholog, a thioredoxin (PDIL2-2, cassava4.1_007986). In barley, an ortholog of
529 PDIL2-2 (*HvPDIL5-1*) is a known virus susceptibility factor as are *PDI* gene family
530 members across the animal and plant kingdoms (Yang et al., 2014). We also
531 identified the Ubiquitin-conjugating enzyme E2 ortholog (UBC5) gene
532 (cassava4.1_017202) under the secondary GWAS peak (scaffold 5214, 5-6 Mb
533 region). Genes like UBC5 in the ubiquitinylation pathway have been known to
534 influence plant virus infection response (Becker et al., 1993).

We analyzed the coding sequence of the three genes mentioned above in three CMD resistant cassava genotypes known to possess the qualitative resistance allele(s) (TMEB3, TMEB7 and I011412) and in two susceptible/tolerant ones known to possess only quantitative resistance sources (I30572 and TMEB1). We identified SNPs within the coding regions and identified amino acid changes (Table S12). Two non-synonymous mutations were found on exons 7 and 9 of Cassava4.1_007986, homozygous in the susceptible group but heterozygous in the resistant clones (Table S12). The peroxidase, Cassava4.1_011768 did not show any non-synonymous mutations specific to the resistant/susceptible group. However, Cassava4.1_02917, showed three non-synonymous mutations that were specific to the susceptible group.

Genomic Prediction of Additive and Total Genetic Merit

We tested four prediction models using cross-validation: (1) $\text{Additive}_{\text{All\_Markers}}$, (2) $\text{Additive}_{\text{All\_Markers}} + \text{Dominance}_{\text{All\_Markers}} + \text{Epistasis}_{\text{All\_Markers}}$, (3) $\text{Additive}_{\text{CMD2}} + \text{Additive}_{\text{Non-CMD2}}$, (4) $\text{Additive}_{\text{CMD2}} + \text{Dominance}_{\text{CMD2}} + \text{Epistasis}_{\text{CMD2}} + \text{Additive}_{\text{Non-CMD2}}$. Mean cross-validation accuracy averaged 0.53 for additive and 0.55 for total value across models (Table 2, Figure 5). Including non-additive effects, using all markers (model 2) shifted 60% of the variance to dominance and epistasis and decreased the accuracy of the additive prediction from 0.53 (model 1) to 0.51, but gave increased total prediction accuracy of 0.55. An additive only model giving separate weight to CMD2 and non-CMD2 regions (model 3) had the highest total prediction accuracy (0.58), with most accuracy coming from CMD2 (0.54) vs. non-

CMD2 (0.29) but most variance absorbed by non-CMD regions. Modifying model 3 to allow the CMD2 region additive, dominance and epistatic effects (model 4) slightly decreased total prediction accuracy (0.57) relative to model 3, with most accuracy coming from the additive CMD2 kernel (0.52), but with 51.7% non-additive variance, 33.6% non-CMD2 variance and only 14.7% additive CMD2.

DISCUSSION

The present study solidifies our understanding of the genetic resistance to CMD that is available in African cassava germplasm and demonstrates the efficacy of genomic selection at improving CMD resistance. After conducting the first genome-wide association study for this species with markers anchored to chromosomes, we are able to confirm that the basis of genetic resistance to CMD is indeed narrow, arising chiefly from a single region of chromosome 8 that collocates with the loci *CMD2* (Akano et al., 2002) and *CMD3* (Okogbenin et al., 2012b). The lack of new major effect loci is a key outcome of our study. Even after analyzing a broad sample of the breeding germplasm from West and East Africa. However, we also identified 13 regions of small effect including one on chromosome 9 that collocates with *CMD1* (Fregene et al., 2000).

Another key result of our analysis is that the most highly resistant cassava clones, those that never show disease symptoms, are only identified using models of epistasis in the significance region on chromosome 8. We propose two alternative hypotheses to explain this result. As suggested both in our analyses and previous studies (Okogbenin et al., 2012b) there may be multiple interacting loci in the region (i.e. *CMD2* and *CMD3*). Alternatively, our results may arise from a complex haplotype structure, where observed levels of resistance come from a single locus with one moderate and another strong resistance allele segregating in the population. An example of the later scenario is resistance to tomato yellow leaf curl which initially mapped to two genes Ty-1 and Ty-3 on the same chromosome, but was later revealed by fine-mapping to be one gene with multiple alleles (Verlaan et al., 2013).

In order to facilitate functional studies of the qualitative resistance source(s) on chromosome 8, we used our GWAS results to identify three candidate genes. Interestingly, there are no major resistance genes (e.g. NBS-LRR) in our region of interest (Lozano et al., 2015). We found two peroxidases, which have recently been shown to down-regulate in response to cassava mosaic geminivirus infection in susceptible genotypes (Allie et al., 2014) and a thioredoxin, which can be important for plant defense activation (Bashandy et al., 2010; Ballaré, 2014). We note that our genome assembly contains gaps ((ICGMC), 2014) and is based on a South American accession (Prochnik et al., 2012) that may not possess the causal gene(s). Significant work remains to identify the causal mechanism of qualitative resistance to CMD.

Finally, we demonstrate the potential of genomic selection for CMD resistance breeding. In agreement with our association analyses, we found most of the variance and the prediction accuracy was attributable to the chromosome 8 QTL(s). While additive models will allow us to accurately select parents for cassava breeding, we found non-additive prediction of total genetic merit to be even more accurate. Prediction of total genetic merit will therefore enable cassava breeders to more easily identify clones with superior disease resistance to be elite varieties, effectively exploiting dominance and epistasis for crop improvement. Further, it is significant that, while accuracy is low for the quantitative (non-major gene) components, it is not zero. Thus it should be possible to do genomic selection to simultaneously improve both qualitative (i.e. *CMD2/CMD3*) and quantitative (i.e. polygenic background) resistance.

610 The results we present in this study will represent progress towards discovering
611 the mechanistic basis for major gene resistance to CMD and will also aid breeders
612 seeking to pyramid useful alleles and achieve symptom-free cassava varieties either
613 by marker assisted or genomic selection. In only two years we have conducted two
614 rounds of selection and recombination, twice as fast as conventional phenotypic
615 selection, and have increased the resistance-allele frequency at our top marker from
616 44% to 63%. The present study is an example of the possibilities for rapidly
617 improving and dynamically breeding a crop that is crucial for hundreds of millions,
618 particularly in underdeveloped regions of the world.
619
620
621
622

ACKNOWELDGEMENTS

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, NRCRI and NaCRRI for collection of phenotypic data and to A. Agbona, A. Ogbonna, E. Uba and R. Mukisa for data curation. This work is dedicated to the memory of Martha Hamblin.

REFERENCES

(ICGMC), I.C.G.M.C. 2014. High-Resolution Linkage Map and Chromosome-Scale Genome Assembly for Cassava (Manihot esculenta Crantz) from Ten Populations. G3 Genes| Genomes| Genet.Available at http://g3journal.org/cgi/doi/10.1534/g3.114.015008 (verified 16 December 2014).

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(4): 521–525Available at http://www.ncbi.nlm.nih.gov/pubmed/12582500 (verified 29 October 2013).

Allie, F., E. Pierce, M. Okoniewski, and C. Rey. 2014. Transcriptional analysis of South African cassava mosaic virus-infected susceptible and tolerant landraces of cassava highlights differences in resistance, basal defense and cell wall associated genes during infection. BMC Genomics 15: 1006Available at http://www.biomedcentral.com/1471-2164/15/1006?utm_source=dlvr.it&utm_medium=tumblr (verified 17 December 2014).

Ballaré, C.L. 2014. Light regulation of plant defense. Annu. Rev. Plant Biol. 65: 335–63Available at http://www.annualreviews.org/doi/abs/10.1146/annurev-arplant-050213-040145.

Bashandy, T., J. Guilleminot, T. Vernoux, D. Caparros-Ruiz, K. Ljung, Y. Meyer, and J.-P. Reichheld. 2010. Interplay between the NADP-linked thioredoxin and glutathione systems in Arabidopsis auxin signaling. Plant Cell 22(February): 376–391.

Becker, F., E. Buschfeld, J. Schell, and A. Bachmair. 1993. Altered response to viral infection by tobacco plants perturbed in ubiquitin system. Plant J. 3(6): 875–881Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.1993.00875.x/abstract.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19): 2633–5Available at http://www.ncbi.nlm.nih.gov/pubmed/17586829.

Ceballos, H., C. a Iglesias, J.C. Pérez, and A.G.O. Dixon. 2004. Cassava breeding: opportunities and challenges. Plant Mol. Biol. 56(4): 503–16Available at http://www.ncbi.nlm.nih.gov/pubmed/15630615.

Ceballos, H., P. Kulakow, and C. Hershey. 2012. Cassava Breeding: Current Status, Bottlenecks and the Potential of Biotechnology Tools. Trop. Plant Biol. 5(1): 73–87Available at http://link.springer.com/10.1007/s12042-012-9094-9 (verified 4 January 2014).

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(5): e19379Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract (verified 21 May 2013).

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J. 4(3): 250Available at https://www.crops.org/publications/tpg/abstracts/4/3/250 (verified 22 July 2014).

Fauquet, C., D. Fargette, and C. Munihor. 1990. African Cassava Mosaic Virus : Etiology , Epidemiology , and Control. Plant Dis. 74(6): 404–411.

Fregene, M., a. Bernal, M. Duque, a. Dixon, and J. Tohme. 2000. AFLP analysis of African cassava (

674 Manihot esculenta Crantz) germplasm resistant to the cassava mosaic disease (CMD). TAG Theor.
675 Appl. Genet. 100(5): 678–685Available at http://link.springer.com/10.1007/s001220051339.

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

678 Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014.
679 TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2):
680 e90346Available at
681 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3938676&tool=pmcentrez&rendertype=a
682 bstract (verified 10 July 2014).

683 Hahn, S., A. Howland, and E. Terry. 1980a. CORRELATED RESISTANCE OF CASSAVA TO MOSAIC
684 AND BACTERIAL BLIGHT DISEASES. Euphytica 29: 305–311Available at
685 http://link.springer.com/article/10.1007/BF00025127 (verified 30 July 2014).

686 Hahn, S., E. Terry, and K. Leuschner. 1980b. BREEDING CASSAVA FOR RESISTANCE TO
687 CASSAVA MOSAIC DISEASE. Euphytica 29: 673–683Available at
688 http://link.springer.com/article/10.1007/BF00023215 (verified 26 July 2014).

689 Hahn, S., E. Terry, K. Leuschner, I. Akobundu, C. Okali, and R. Lal. 1979. Cassava improvement in
690 Africa. F. Crop. Res. 2: 193–226Available at
691 http://www.sciencedirect.com/science/article/pii/0378429079900248 (verified 14 August 2014).

692 Hamblin, M.T., and I.Y. Rabbi. 2014. The Effects of Restriction-Enzyme Choice on Properties of
693 Genotyping-by-Sequencing Libraries: A Study in Cassava (). Crop Sci. 54(6): 2603Available at
694 https://www.crops.org/publications/cs/abstracts/54/6/2603 (verified 1 December 2014).

695 Hanley-Bowdoin, L., E.R. Bejarano, D. Robertson, and S. Mansoor. 2013. Geminiviruses: masters at
696 redirecting and reprogramming plant processes. Nat. Rev. Microbiol. 11(11): 777–88Available at
697 http://www.ncbi.nlm.nih.gov/pubmed/24100361.

698 Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic Selection for Crop Improvement. Crop Sci.
699 49(1): 1Available at https://www.crops.org/publications/cs/abstracts/49/1/1 (verified 19 September
700 2013).

701 Henderson, C.R. 1985. Best Linear Unbiased Prediction of Nonadditive Genetic Merits in Noninbred
702 Populations. J. Anim. Sci. 60(1): 111–117.

703 Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to
704 practice. Brief. Funct. Genomics 9(2): 166–77Available at
705 http://www.ncbi.nlm.nih.gov/pubmed/20156985 (verified 22 May 2013).

706 Kang, H.M., J.H. Sul, S.K. Service, N. a Zaitlen, S.-Y. Kong, N.B. Freimer, C. Sabatti, and E. Eskin. 2010.
707 Variance component model to account for sample structure in genome-wide association studies. Nat.
708 Genet. 42(4): 348–54Available at http://www.ncbi.nlm.nih.gov/pubmed/20208533 (verified 3 August
709 2010).

710 Legg, J.Q., and C.M. Fauquet. 2004. Cassava mosaic geminiviruses. Plant Mol. Biol. 56: 585–599.

711 Lokko, Y., E. Danquah, and S. Offei. 2005. Molecular markers associated with a new source of resistance
712 to the cassava mosaic disease. African J. … 4(September): 873–881Available at
713 http://www.ajol.info/index.php/ajb/article/view/71131 (verified 26 July 2014).

van Loon, L.C., M. Rep, and C.M.J. Pieterse. 2006. Significance of inducible defense-related proteins in infected plants. Annu. Rev. Phytopathol. 44: 135–162.

Lozano, R., M.T. Hamblin, S. Prochnik, and J.-L. Jannink. 2015. Identification and distribution of the NBS-LRR gene family in the Cassava genome. BMC Genomics 16(1): 1–14Available at http://www.biomedcentral.com/1471-2164/16/360.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu, A.G.O. Dixon, P. Kulakow, and J.-L. Jannink. 2013. Relatedness and Genotype × Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study in Cassava. Crop Sci. 53(4): 1312Available at https://www.crops.org/publications/cs/abstracts/53/4/1312 (verified 20 September 2013).

Mohan, C., P. Shanmugasundaram, M. Maheswaran, N. Senthil, D. Raghu, and M. Unnikrishnan. 2013. Mapping New Genetic Markers Associated with CMD Resistance in Cassava (Manihot esculenta Crantz) Using Simple Sequence Repeat Markers. J. Agric. Sci. 5(5): 57–65Available at http://www.ccsenet.org/journal/index.php/jas/article/view/21210 (verified 4 December 2014).

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

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(2): 181–208Available at http://www.tandfonline.com/doi/abs/10.1080/15427520802212506 (verified 18 July 2014).

Okogbenin, E., C.N. Egesi, B. Olasanmi, O. Ogundapo, S. Kahya, P. Hurtado, J. Marin, O. Akinbo, C. Mba, H. Gomez, C. de Vicente, S. Baiyeri, M. Uguru, F. Ewa, and M. Fregene. 2012a. Molecular Marker Analysis and Validation of Resistance to Cassava Mosaic Disease in Elite Cassava Genotypes in Nigeria. Crop Sci. 52(6): 2576Available at https://www.crops.org/publications/cs/abstracts/52/6/2576 (verified 9 June 2014).

Okogbenin, E., C. Egesi, B. Olasanmi, O. Ogundapo, S. Kahya, P. Hurtado, J. Marin, O. Akinbo, C. Mba, H. Gomez, C. de Vicente, S. Baiyeri, M. Uguru, F. Ewa, and M. Fregene. 2012b. Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. Crop Sci. 52(December): 2576–2586Available at https://dl.sciencesocieties.org/publications/cs/abstracts/52/6/2576 (verified 6 December 2014).

Okogbenin, E., M. Porto, and C. Egesi. 2007. Marker-assisted introgression of resistance to cassava mosaic disease into Latin American germplasm for the genetic improvement of cassava in Africa. Crop Sci. 47: 1895–1904Available at https://dl.sciencesocieties.org/publications/cs/abstracts/47/5/1895 (verified 4 November 2013).

Oliveira, E.J., M.D.V. Resende, V. Silva Santos, C.F. Ferreira, G.A.F. Oliveira, M.S. Silva, L.A. Oliveira, and C.I. Aguilar-Vildoso. 2012. Genome-wide selection in cassava. Euphytica 187(2): 263–276Available at http://link.springer.com/10.1007/s10681-012-0722-0 (verified 19 September 2013).

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N. a Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38(8): 904–9Available at http://www.ncbi.nlm.nih.gov/pubmed/16862161 (verified 23 May 2014).

Prochnik, S., P.R. Marri, B. Desany, P.D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D.S. Rokhsar, and S. Rounsley. 2012. The Cassava Genome: Current

Progress, Future Directions. Trop. Plant Biol. 5(1): 88–94Available at http://link.springer.com/10.1007/s12042-011-9088-z (verified 29 May 2014).

Rabbi, I., M. Hamblin, M. Gedil, P. Kulakow, M. Ferguson, A.S. Ikpan, D. Ly, and J.-L. Jannink. 2014a. Genetic Mapping Using Genotyping-by-Sequencing in the Clonally Propagated Cassava. Crop Sci. 54: 1–13Available at https://www.crops.org/publications/cs/abstracts/0/0/cropsci2013.07.0482 (verified 20 June 2014).

Rabbi, I.Y., M.T. Hamblin, P.L. Kumar, M. a Gedil, A.S. Ikpan, J.-L. Jannink, and P. a Kulakow. 2014b. High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. Virus Res.Available at http://www.ncbi.nlm.nih.gov/pubmed/24389096 (verified 9 June 2014).

Raji, A., O. Ladeinde, and A. Dixon. 2008. Screening landraces for additional sources of field resistance to cassava mosaic disease and green mite for integration into the cassava improvement program. J. Integr. Plant Biol. 50(3): 311–318.

Segura, V., B.J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, and M. Nordborg. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44(7): 825–30Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3386481&tool=pmcentrez&rendertype=abstract (verified 23 May 2014).

Su, G., O.F. Christensen, T. Ostersen, M. Henryon, and M.S. Lund. 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLoS One 7(9): 1–7.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11): 4414–23Available at http://www.ncbi.nlm.nih.gov/pubmed/18946147 (verified 18 October 2013).

Verlaan, M.G., S.F. Hutton, R.M. Ibrahem, R. Kormelink, R.G.F. Visser, J.W. Scott, J.D. Edwards, and Y. Bai. 2013. The Tomato Yellow Leaf Curl Virus Resistance Genes Ty-1 and Ty-3 Are Allelic and Code for DFDGD-Class RNA-Dependent RNA Polymerases. PLoS Genet. 9(3).

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(5): 519–529Available at http://www.degruyter.com/view/j/sagmb.2014.13.issue-5/sagmb-2012-0044/sagmb-2012-0044.xml.

Yang, J., S.H. Lee, M.E. Goddard, and P.M. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88(1): 76–82Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3014363&tool=pmcentrez&rendertype=abstract (verified 29 May 2014).

Yang, P., T. Lüpken, A. Habekuss, G. Hensel, B. Steuernagel, B. Kilian, R. Ariyadasa, A. Himmelbach, J. Kumlehn, U. Scholz, F. Ordon, and N. Stein. 2014. PROTEIN DISULFIDE ISOMERASE LIKE 5-1 is a susceptibility factor to plant viruses. Proc. Natl. Acad. Sci. U. S. A. 111(6): 2104–9Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3926060&tool=pmcentrez&rendertype=abstract.

Zhang, Z., E. Ersoz, C. Lai, and R. Todhunter. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42(4): 355–360Available at http://www.nature.com/ng/journal/v42/n4/abs/ng.546.html (verified 16 December 2014).

796
797

FIGURE LEGENDS

Figure 1. Plot of the first four principal components of the SNP marker matrix. The three main training populations were used in the PCA and are shown here. A random sample of IITA GS Cycles 1 and 2 were projected into the genetic space and are displayed here.

Figure 2. Manhattan plot from mixed-linear models summarizing genome-wide association results for all traits in all sub-populations. Bonferroni significance threshold is shown in red. An example QQ-plot (MCMDS in the population-wide analysis) is shown inset to demonstrate the differences between various population structure controls.

Figure 3. Plots dissecting the major effect QTL on chromosome 8. (A) Manhattan plot summarizing genome-wide association results for all CMD-related traits in the population-wide mixed-linear model analysis, zoomed to chromosome 8 only. (B) Manhattan plot showing linear model tests for interactions (blue dots) between the top marker (S8_7762525, blue vertical line) and every other maker on chromosome 8. Red dots are for the main effect of the second marker (main effects of S8_7762525 are not shown). (C) LD between S8_7762525 (blue vertical line) and every other maker (blue dots), plus LD between the marker with the strongest interaction effect (S8_4919667; red vertical line) and every other marker. Bar plot showing mean and standard error for MCMDS between each genotype class at the top marker, S8_7762525 (Inset).

Figure 4. Plots demonstrating the combined effect of the genotype at the top marker (S8_7762525) and the most epistatic marker (S8_4919667). (A) Boxplot showing the distribution of mean CMD severity scores (MCMDS) for each two-locus genotype. (B) Disease progress curves showing mean and standard error CMD severity across 1, 3 and 6 months after planting for each two-locus genotype. (C) Zoomed Manhattan plot showing the location of the two markers being compared; S8_7762525 (red line) and S8_4919667 (blue line).

Figure 5. Cross-validated genomic prediction results for MCMDS. (A-C) Box plots of accuracies from 25 reps of 5-fold cross-validation. (A) Accuracies of the additive models (#1 and #3) using either a single kernel ($Additive_{All\_Markers}$) or two-kernels ($Additive_{CMD2}$ and $Additive_{Non\text{-}CMD2}$). (B) Accuracies of the additive predictions from the models that included dominance and additive-by-dominance epistasis (models #2 and #4). A single additive accuracy is calculated from the model #2 ($Additive_{All\_Markers}$) and two accuracies for model #4 ($Additive_{CMD2}$ and $Additive_{Non\text{-}CMD2}$). The accuracy of total genetic merit prediction from models (#2, all markers; #4, CMD2 + Non-CMD2) with dominance and epistasis are shown in (C). Kernel weights corresponding to the partitioning of the genetic variance for the epistatic models are shown in (D, all markers model #2) and (E, CMD2 + Non-CMD2 model #4).

Table 1. Summary of phenotype and genotype datasets analyzed.

	Population					
Trait	NRCRI	NaCRRI	IITA: Genetic Gain	GS Cycle 1	GS Cycle 2	Trait Description
CMD1S	X		X	X	X	Cassava mosaic disease (CMD) severity rated on a scale from 1 (no symptoms) to 5 (extremely severe). One month after planting (MAP).
CMD3S	X	X	X	X		CMD Severity at 3 MAP
CMD6S	X		X	X		CMD Severity at 6 MAP
CMD9S	X					CMD Severity at 9 MAP
CMD12S	X					CMD Severity at 12 MAP
MCMDS	X	X	X	X	X	Mean across all growing season observations of Cassava Mosaic Disease Severity
AUDPC	X		X	X		Area under disease severity progress curves (1, 3 and 6 MAP).
	2	2	14	2	1	Years
	3	3	10	3	1	Locations
	Clonal	Clonal	Clonal	Seed/Clonal	Seed	Propagation
	626	414	2187	694	2466	N Clones
	41820	41060	42113	41369	40539	N Markers (MAF > 0.05)
	0.21	0.21	0.22	0.21	0.22	Mean MAF (mapped markers, with MAF > 0.05)
	0.32	0.33	0.34	0.33	0.35	Mean observed Heterozygote frequency

NRCRI	National Root Crops Research Institute (NRCRI) in Umudike, Nigeria
NaCRRI	National Crops Resources Research Institute (NaCRRI) in Namulonge, Uganda
IITA: Genetic Gain	International Institute of Tropical Agriculture in Ibadan, Nigeria,
IITA: Cycle 1	Genomic selection progenies of 76 IITA: Genetic Gain clones.
IITA: Cycle 2	Genomic selection progenies of 158 IITA: Cycle 1 clones.

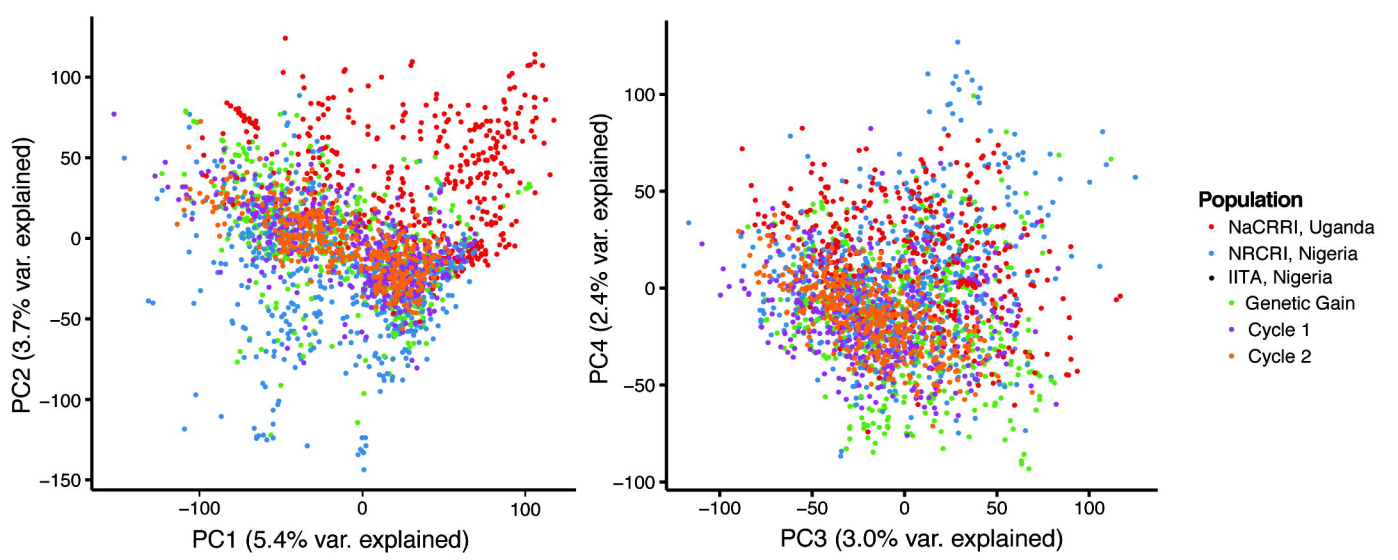
844

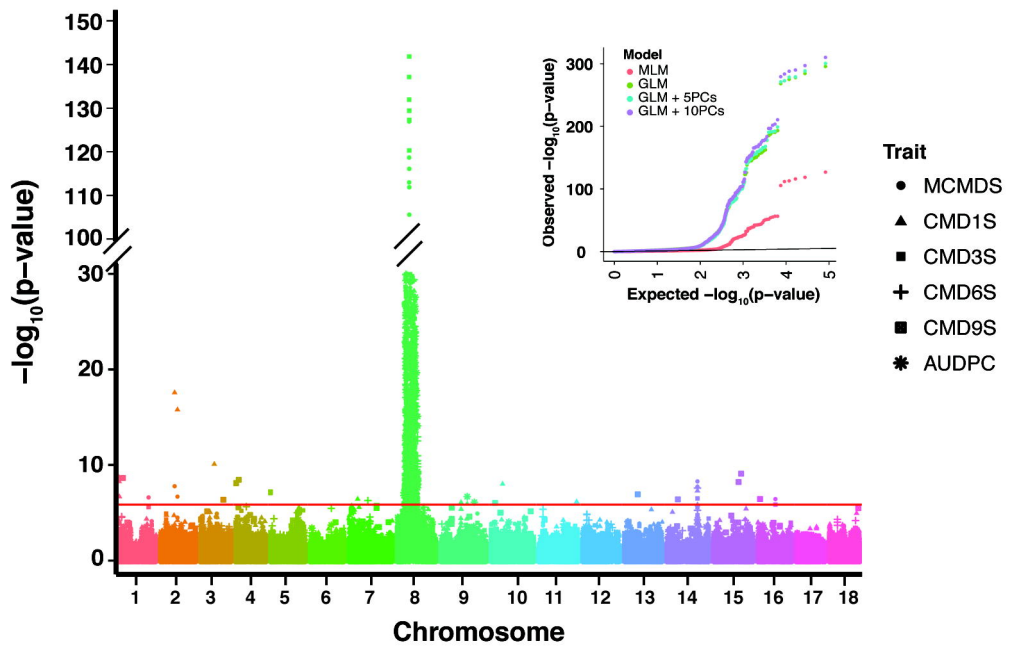
845

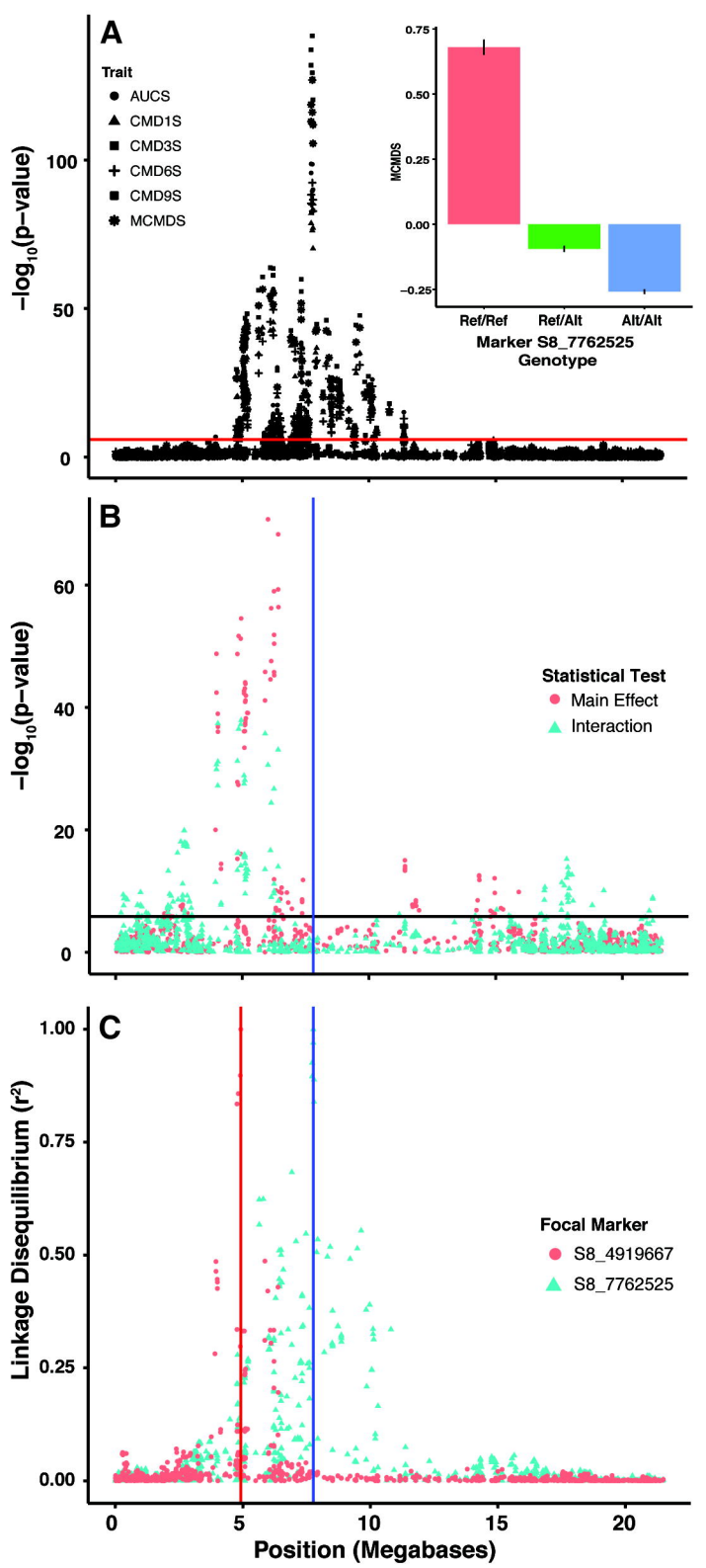
Table 2. Summary of cross-validation results for MCMDS. Mean kernel weights as well as additive and total prediction accuracies are reported for each of four models tested.

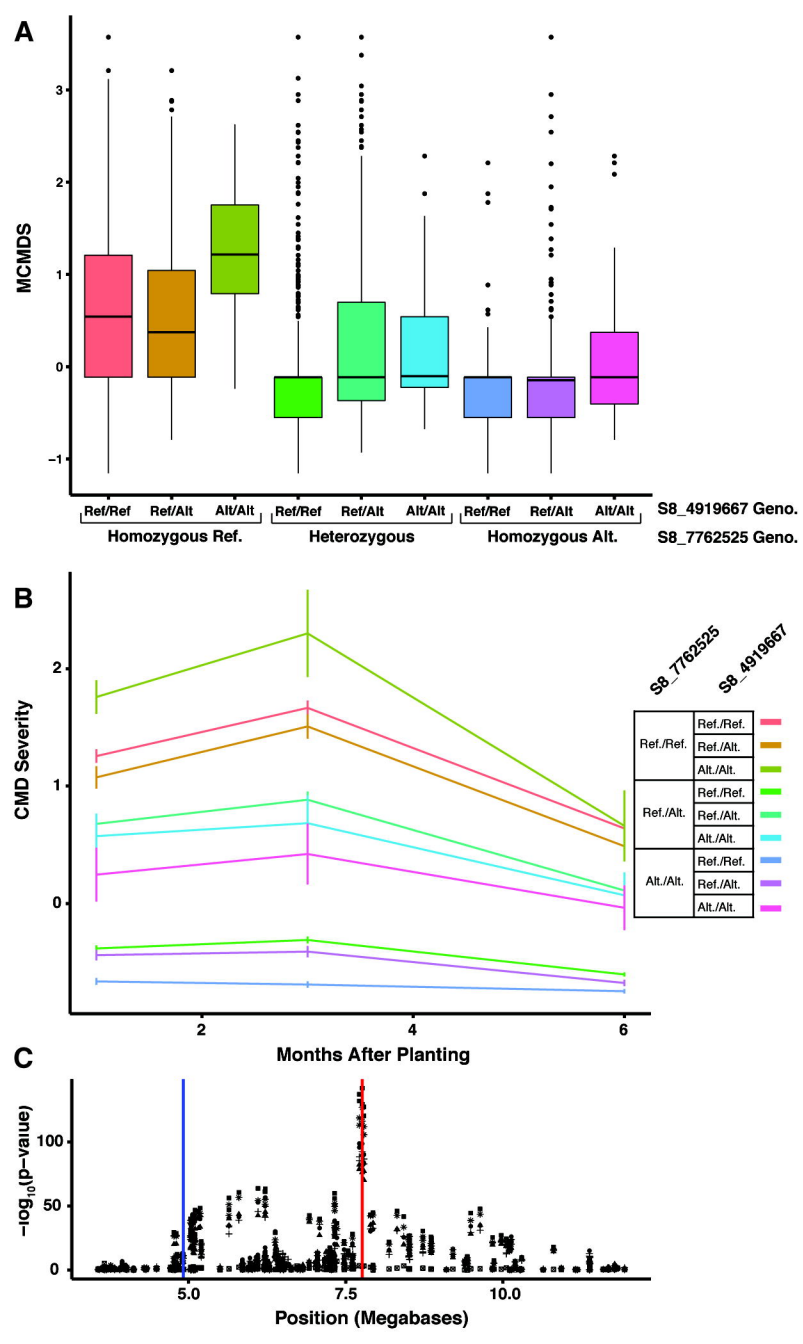
	Fraction of the variance explained (Kernel Weights)				Additive Component Accuracy	Total Sum Accuracy
1)	**$Additive_{All\_Markers}$** 1				0.53	0.53
2)	**$Add_{All\_Markers}$** 0.396	**$Dom_{All\_Markers}$** 0.015	**$Epi_{All\_Markers}$** 0.589		0.51	0.55
3)	**$Additive_{CMD2}$** 0.3		**$Additive_{Non\text{-}CMD2}$** 0.7		0.54_{CMD2} / $0.29_{Non\text{-}CMD2}$	0.58
4)	**Add_{CMD2}** 0.147	**Dom_{CMD2}** 0.019	**Epi_{CMD2}** 0.498	**$Add_{Non\text{-}CMD2}$** 0.336	0.52_{CMD2} / $0.25_{Non\text{-}CMD2}$	0.57

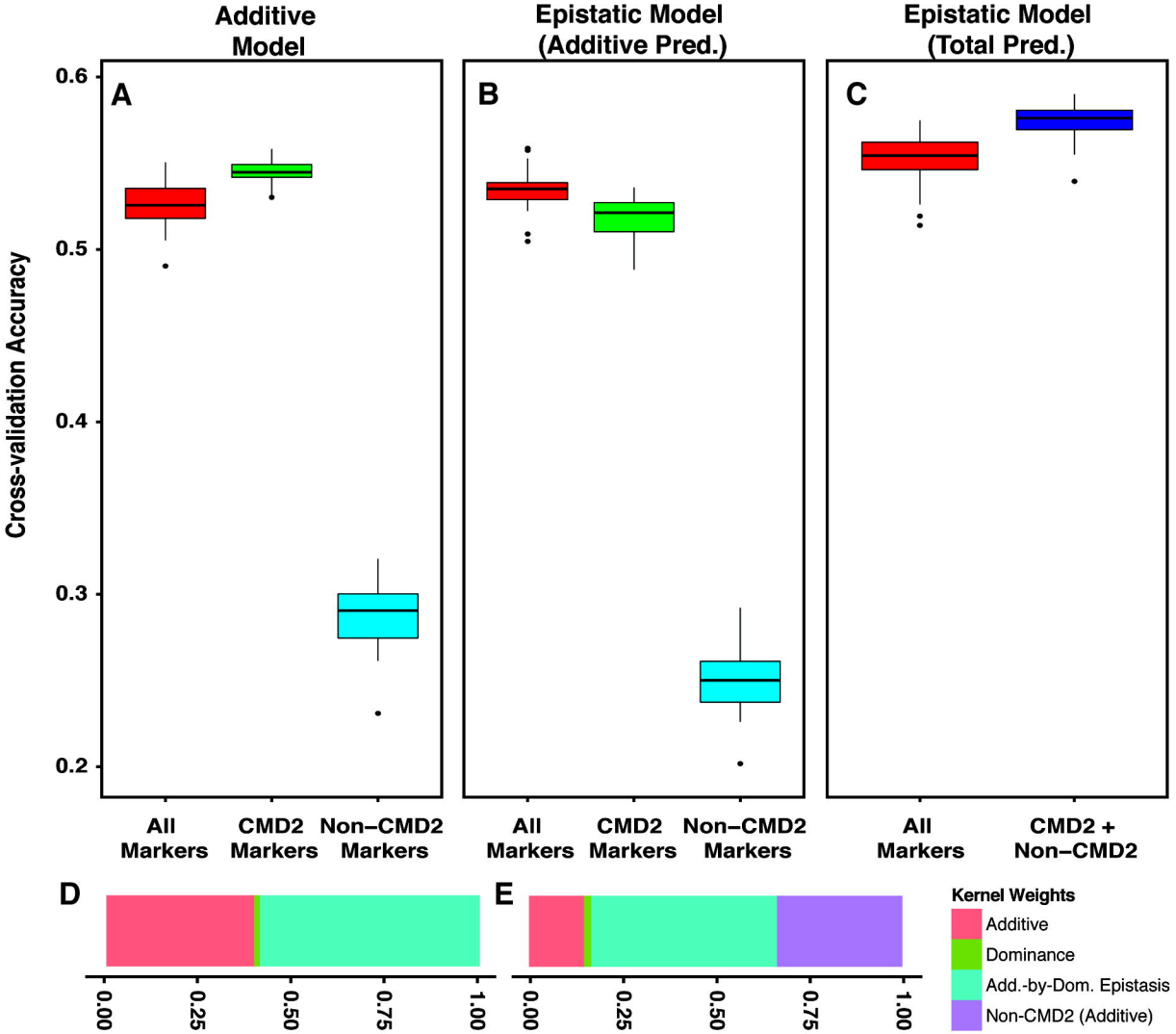
846


PC2 (3.7% var. explained)
PC1 (5.4% var. explained)
PC4 (2.4% var. explained)
PC3 (3.0% var. explained)
Population
NaCRRI, Uganda
NRCRI, Nigeria
IITA, Nigeria
Genetic Gain
Cycle 1
Cycle 2


−log10(p−value)
150
140
130
120
110
100
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Chromosome
Model
MLM
GLM
GLM + 5PCs
GLM + 10PCs
Observed −log10(p−value)
300
200
100
0
0
1
2
3
4
5
Expected −log10(p−value)
Trait
MCMDS
CMD1S
CMD3S
CMD6S
CMD9S
AUDPC


A
Trait
AUCS
CMD1S
CMD3S
CMD6S
CMD9S
MCMDS
−log10(p−value)
100
50
0
MCMDS
0.75
0.50
0.25
0.00
−0.25
Ref/Ref
Ref/Alt
Alt/Alt
Marker S8_7762525
Genotype
B
−log10(p−value)
60
40
20
0
Statistical Test
Main Effect
Interaction
C
Linkage Disequilibrium (r2)
1.00
0.75
0.50
0.25
0.00
Focal Marker
S8_4919667
S8_7762525
0
5
10
15
20
Position (Megabases)


A
MCMDS
3
2
1
0
−1
Ref/Ref
Ref/Alt
Alt/Alt
Ref/Ref
Ref/Alt
Alt/Alt
Ref/Ref
Ref/Alt
Alt/Alt
S8_4919667 Geno.
S8_7762525 Geno.
Homozygous Ref.
Heterozygous
Homozygous Alt.
B
CMD Severity
2
1
0
2
4
6
Months After Planting
S8_7762525
S8_4919667
Ref./Ref.
Ref./Ref.
Ref./Alt.
Alt./Alt.
Ref./Alt.
Ref./Ref.
Ref./Alt.
Alt./Alt.
Alt./Alt.
Ref./Ref.
Ref./Alt.
Alt./Alt.
C
−log10(p−value)
100
50
0
5.0
7.5
10.0
Position (Megabases)


Additive Model
Epistatic Model (Additive Pred.)
Epistatic Model (Total Pred.)
A
B
C
Cross-validation Accuracy
0.6
0.5
0.4
0.3
0.2
All Markers
CMD2 Markers
Non−CMD2 Markers
All Markers
CMD2 Markers
Non−CMD2 Markers
All Markers
CMD2 + Non−CMD2
D
E
0.00
0.25
0.50
0.75
1.00
Kernel Weights
Additive
Dominance
Add.-by-Dom. Epistasis
Non-CMD2 (Additive)