1 **Genome-wide association and prediction reveals the genetic architecture of**
2 **cassava mosaic disease resistance and prospects for rapid genetic**
3 **improvement**
4
5 Authors:
6 Marnin D. Wolfe*, Ismail Y. Rabbi, Chiedozie Egesi, Martha Hamblin, Robert Kawuki,
7 Peter Kulakow, Roberto Lozano, Dunia Pino Del Carpio, Punna Ramu, Jean-Luc
8 Jannink
9

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

18
19
20 Received _____
21
22 Abbreviations:
23 Genome-wide association analysis (GWAS); genomic selection (GS); Cassava mosaic
24 disease (CMD); genotype-by-sequencing (GBS); International Institute of Tropical
25 Agriculture (IITA); National Root Crops Research Institute (NRCRI); National Crops
26 Resources Research Institute (NaCRRI); African Cassava Mosaic Virus (ACMV); East
27 African Cassava Mosaic Virus (EACMV); genomic estimated breeding values
28 (GEBVs); Cassava mosaic disease severity (CMDS)
29

30 **ABSTRACT**

31   Cassava (*Manihot esculenta*) is a crucial, under-researched crop feeding

32 millions worldwide, especially in Africa. Cassava mosaic disease (CMD) has plagued

33 production in Africa for over a century. Bi-parental mapping studies suggest

34 primarily a single major gene mediates resistance. To be certain and to potentially

35 identify new loci we conducted the first genome-wide association mapping study in

36 cassava with 6128 African breeding lines. We also assessed the accuracy of genomic

37 selection to improve CMD resistance. We found a single region on chromosome 8

38 accounts for most resistance but also identified 13 small effect regions. We found

39 evidence that two epistatic loci and/or alternatively multiple resistance alleles exist

40 at major QTL. We identified two peroxidases and one thioredoxin as candidate

41 genes. Genomic prediction of additive and total genetic merit was accurate for CMD

42 and will be effective both for selecting parents and identifying highly resistant

43 clones as varieties.

44

45   Cassava (*Manihot esculenta* Crantz) is a crucial staple food crop, usually

46   grown by smallholder farmers and feeding over half a billion people worldwide,

47   especially in sub-Saharan Africa (http://faostat.fao.org). Breeding cycles are long in

48   this outcrossing, clonally-propagated crop and genetic gains from breeding have

49   been small over the last century compared with other crops (Ceballos et al., 2004,

50   2012). With a recently sequenced genome (Prochnik et al., 2012) and a high-density

51   SNP-based genetic linkage maps ((ICGMC), 2014), it is for the first time possible to

52   study the genetic architecture of key traits using modern genome-wide association

53   analysis (GWAS) and to improve those traits with genomic selection (GS) (Oliveira

54   et al., 2012; Ly et al., 2013).

55   Cassava mosaic disease (CMD) is the longest running and thus-far most

56   impactful of the challenges cassava farmers face in sub-Saharan Africa(Fauquet et

57   al., 1990). The disease is caused by several related species of geminiviruses and

58   transmitted both through infected cuttings and by a vector, the common whitefly

59   (*Bemisia tabaci* G.). Development and deployment of resistant cultivars is the most

60   effective control method for this devastating disease. Following an unsuccessful

61   world-wide search for resistance in cultivated germplasm (*M. esculenta*) in the

62   1930s, cassava breeders at the Amani research station in Tanzania resorted to

63   interspecific hybridization with Ceara rubber tree (*M. glaziovii* Müll. Arg) and other

64   related wild species in the 1930s (Hahn et al., 1979, 1980a; Fauquet et al., 1990).

65   Moderate polygenic resistance combined with reasonable root yields was achieved

66   through several cycles of backcross of Ceara rubber to the cultivated cassava (Hahn

67   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

91   genomic selection-based breeding project in cassava

92   (**http://www.nextgencassava.org**). In addition, we combined GWAS and genomic

93   prediction in order to not only dissect the genetic architecture of resistance to CMD

94   but also to assess the potential for population improvement by genomic selection

95   (GS). We used a variety of approaches to localize and identify candidate genes for

96   future investigation. The potential for GS to improve CMD resistance and for non-

97   additive models to predict total genetic merit of clones for the selection of superior

98   CMD resistant varieties were assessed. Finally, multi-kernel genomic prediction

99   models were used to study the relative importance of qualitative and quantitative

100   resistance sources.

101

102   **MATERIALS & METHODS**

103   **Germplasm collection**

104   The germplasm included in this study represent the reference populations

105   used to develop genomic prediction models as part of a collaborative project

106   between Cornell University and three breeding institutions: The International

107   Institute of Tropical Agriculture (IITA) in Ibadan, Nigeria, the National Root Crops

108   Research Institute (NRCRI) in Umudike, Nigeria and the National Crops Resources

109   Research Institute (NaCRRI) in Namulonge, Uganda. The IITA's Genetic Gain

110   population is comprised of 694 historically important, mostly advanced breeding

111   lines that have been selected and maintained clonally since 1970 (Okechukwu and

112   Dixon, 2008; Ly et al., 2013). Most of these materials are derived from the cassava

113   gene-pool from West Africa and early introductions of CMD tolerant parents derived

114   from the inter-specific hybridization program at the Amani Station in Tanzania

115   (Hahn et al., 1980b). It also includes hybrids of germplasm introduced from Latin

116   America (see Table S1 for a list of accessions and details on pedigree where

117   available). The NRCRI population contains 626 clones from their breeding program,

118   189 of which are also part of IITA's Genetic Gain. The remainder of the NRCRI

119   collection includes a large number of materials either directly from or derived with

120   parentage from the International Center for Tropical Agriculture (CIAT) in Cali,

121   Columbia (Table S2).

122   There are two major clades of cassava mosaic virus species, African Cassava

123   Mosaic Virus (ACMV) and East African Cassava Mosaic Virus (EACMV) (Legg and

124   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

170   available for one plot of five stands per clone, so each clone was only planted in one

171   of the three locations. Clones were assigned to each location so as to equally

172   represent each family in every environment. The GS Cycle 2 (C2) individuals were

173   observed in a seedling trial during the 2013-2014 field season. We note that

174   expression of disease symptoms in cassava seedlings may not be representative of

175   expression in clonal evaluations. This is in part because seedling symptoms can

176   arise solely from whitefly transmission, making it probable that some asymptomatic

177   plants are in fact escapes rather than resistant genotypes. Table 1 summarizes the

178   phenotypes and phenotyping trials available for each sub-population. We also

179   provide details about the sample sizes and replication numbers for each

180   location/year of data analyzed (Table S6) and per accession (Table S7)

181   Cassava mosaic disease severity (CMDS) was scored on a scale of 1 to 5, with 1

182   representing no symptoms and 5 indicating the most severe symptoms. CMDS was

183   scored at up to five time points (1, 3, 6, 9 and 12 months after planting) depending

184   on the trial. Additionally, we analyze the season-wide mean CMDS score (MCMDS),

185   which is used for making selections and the area under the disease progress curve

186   (see below; AUDPC). The distribution of raw phenotypic data used in each

187   population and for each trait can be seen in Figures S1-S6.

188

189   **Statistical Models and Analyses of Phenotypes**

190   Our interest in this study was to identify key aspects of the genetic

191   architecture of cassava in Africa rather than location- or year-specific QTLs. We

192   condensed up to 38854 observations on 6198 genotyped and phenotyped clones to

193   single BLUPs for each. To do this, we fit the following mixed linear model with the

194   *lme4* package in R:

195

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

197

198   Here, $y_{l,i,j}$ represents raw phenotypic observations, μ is the grand mean, c_l is a

199   random effects term for clone with $c_l \sim N(0, \sigma_l^2)$, β_i is a fixed effect for the

200   combination of location and year harvested, $r_{j(i)}$ is a random effect for replication

201   nested within location-year combination assumed to be distributed $N(0, \sigma_r^2)$ and

202   finally, $\varepsilon_{l,i,j}$ is the residual variance, assumed to be random and distributed $N(0, \sigma^2)$.

203   Because the number of observations per clone varies greatly in our dataset (from 1

204   to 941, median of 2; Table S7), we expect BLUPs are differentially shrunken to the

205   mean. To counter this, we de-regressed BLUPs according to the following formula:

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

207   Where PEV is the prediction error variance for each clone and σ_l^2 is the clonal

208   variance component. The distribution of deregressed BLUPs used as response

209   variables in GWAS can be seen in Figures S7-S13.

210   We also calculated areas under disease progress curves (AUDPC) for each

211   clone using data from 1, 3 and 6 months after planting. To do this, we treated

212   severity scores from any time point as the same trait with a second variable

213   indicating the time point of the score. We then ran the model indicated in [1] but

214   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

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.

260   In particular, we compared the genome-wide inflation of p-values between a general

261   linear model (GLM) with no population structure controls, a GLM with 5 principal

262   components (GLM + 5 PCs)(Price et al., 2006), and a mixed linear model (MLM),

263   which fits a random effect for clone with $\sim N(0,\sigma_g^2 K)$, where σ_g^2 is the clonal variance

264   component and K is the relationship matrix described above (Kang et al., 2010).

265   MLM were conducted using the P3D and compression method (Zhang et al., 2010).

266   All GWAS were conducted in TASSEL (version 5, [27]). We compare the observed –

267    $\log_{10}$(p-values) against the expectation using QQ-plots. We used visual inspection of

268   QQ-plots to judge which model most effectively reduced the genome-wide inflation

269   of $-\log_{10}$(p-values) typically attributed with population structure. We consider

270   association tests significant when more extreme than the Bonferroni threshold

271   (with experiment-wise type I error rate of 0.05).

272   Because marker effects, LD patterns, and allele frequencies may differ within as

273   well as across sub-populations, we conducted GWAS population-wide as well as

274   within each sub-population. In each analysis, we used markers that segregated with

275   MAF > 5% in that specific sub-population. Bonferroni thresholds were calculated

276   according to the number of markers analyzed in each sub-population.

277   We also examined the proportion of variance in the deregressed BLUPs

278   explained by the kinship matrix, K using the variance components estimated when

279   TASSEL fits the MLM model.

280

281   **Candidate Genes**

282   Because the underlying mechanisms of plant disease resistance are of general

283   interest and identification of causal polymorphism may aid in transgenic

284   approaches and/or marker assisted selection, we identified candidate genes in CMD

285   associated regions. Significant SNPs from the GWAS results corresponding to four

286   time points (1, 3, 6 and 9 months after planting) were selected for the analysis. We

287   considered SNPs that were both above the Bonferroni threshold and were located

288   within exons or introns of cassava genes. The SNP position on the genome was

289   compared with the gene interval position using the annotation list from Phytozome

290   10. Gene ontology annotation for each time point and combining all the datasets was

291   done with Panther (http://go.pantherdb.org/). We have generated whole genome

292   sequences (WGS) from one CMD resistant clone (I011412) and two CMD tolerant

293   clones (I30572 and TMEB1). TMEB1 is a landrace from Ogun State, Nigeria also

294   called Antiota, that is not likely to contain the qualitative resistance allele and is

295   usually classified as tolerant or only partially susceptible to CMD (Raji et al., 2008;

296   Rabbi et al., 2014b). Similarly, I30572 is an improved variety whose parents were

297   the *M. glaziovii*-derived clone 58308 and a south American Cassava (Branca de

298   Santa Catarina) and is therefore known to have only the quantitative resistance

299   source (Fauquet et al., 1990). *PCR*-free libraries were generated from these clones

300   and sequenced at 20X coverage using Illumina HiSeq. Additionally two resistant

301   clones (TMEB3 and TMEB7) were obtained from Phytozome

302   (http://phytozome.jgi.doe.gov). TMEB3 is itself the original landrace parent from

303   which the qualitative resistance source has been derived and TMEB7 has been

304   shown to be nearly genetically identical to TMEB3. We therefore define TMEB3,

305   TMEB7 and I011412 as "resistant" lines while TMEB1 and I30572 will be referred

306   to, for simplicity as "susceptible" primarily on the basis of whether they do or do not

307   have the qualitative resistance source *CMD2*. These sequences were aligned against

308   the cassava V5 reference genome assembly to call the variants to identify the

309   genomic difference between resistance and susceptible clones in candidate gene

310   loci. Since the genotypes compared were few in number, we called SNPs manually

311   using an exon annotated sequence and the Integrative Genomics Viewer software

312   (IGV; http://www.broadinstitute.org/igv/).

313

314

315   **Genomic Prediction of Additive and Total Genetic Merit**

316   We used a multi-random effects (a.k.a. multi-kernel or multi-relationship

317   matrix) genomic prediction model to compare the variance explained and

318   prediction accuracy achieved from the major CMD QTL (CMD2) compared to the

319   rest of the genome. Specifically, we created relationship matrices either from all

320   markers, markers significantly associated with CMD2 from GWAS results, or all

321   markers *not* in the region of the QTL.

322   For additive relationships, we used the formulation described above for

323   controlling population-structure (VanRaden, 2008). Dominance relationships can be

324   captured as $\mathbf{D} = \frac{HH'}{\sum_i 2p_iq_i(1-2p_iq_i)}$ (Su et al., 2012; Muñoz et al., 2014). Where H is the

325   SNP marker matrix (individuals on rows, markers along columns), heterozygotes

326   are given as $(1 - 2p_iq_i)$ and homozygotes are $(0 - 2p_iq_i)$. We made a custom

327   modification (available upon request) to the *A.mat* function in the *rrBLUP* package

328   (Endelman, 2011) to produce the **D** matrix. Relationship matrices that capture

329   epistasis can also be calculated by taking the hadamard product (element-by-

330   element multiplication; denoted #) of two or more matrices (Henderson, 1985). For

331   simplicity, we tested an additive-by-dominance (**A#D**) matrix in this study.

332   We tested four models. Model 1 used all markers and only a single, additive

333   kernel (Additive$_{All\_Markers}$). Model 2 used all markers but three kernels,

334   Additive$_{All\_Markers}$ + Dominance$_{All\_Markers}$ + Epistasis$_{All\_Markers}$. Model 3 used two

335   additive kernels, one constructed from the 163 CMD2 significant markers

336   (Additive$_{CMD2}$) and the other from all markers outside of the chromosomal region

337   bounded by CMD2 markers (Additive$_{Non-CMD2}$). Model 4 had four kernels:

338   Additive$_{CMD2}$ + Dominance$_{CMD2}$ + Epistasis$_{CMD2}$ + Additive$_{Non-CMD2}$.

339   We assessed the influence that modeling non-additive genetic variance

340   components have on genomic prediction using a cross-validation strategy (see

341   below). We used the deregressed BLUPs for MCMDS as described above. In our data,

342   the number of observations per clone ranges from one to 941 (checks, TMEB1 and

343   I30572) with median of two and mean of 10.6 (Table S7). Pooling information from

344   multiple years and locations, especially when there is so much variation in numbers

345   of observations can introduce bias. Much theoretical development, particularly in

346   animal breeding has been done to address this issue, and we followed the approach

347   recommended by Garrick et al. (2009)

348   In the second step of analysis, where deregressed BLUPs are used as

349   response variables, weights are applied to the diagonal of the error variance-

350   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

351   of the total variance explained by the clonal variance component, σ_g^2 derived in the

352   first step **(Garrick et al. 2009)**.

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

354   the accuracy of genomic prediction using the genomic relationship matrices and

355   models described above. In each replication, we randomly assign each individual to

356   one of five groups (folds). We then select one fold, remove the corresponding

357   individuals from the training set and use the remaining four folds to predict the fold

358   that was left out. We iterate this process over each of the five folds to produce a

359   prediction for each individual that was made while its phenotypes were

360   unobserved. For each replicate of each model, we calculated accuracy as the Pearson

361   correlation between the genomic prediction made when phenotypes were excluded

362   from the training sample and the BLUP (ĝ, not-deregressed) from the first step. For

363   each model, we calculated accuracy both of the prediction from the additive kernel

364   (where present) and the total genetic merit prediction, defined as the sum of the

365   predictions from all available kernels (e.g. additive + dominance + epistasis).

366   Genomic predictions were made using the *EMMREML* R package. For simplicity, we

367   tested only the trait MCMDS in the IITA Genetic Gain population.

368   In addition, we assessed the predictability of CMD based on random forest

369   regression (RF), a non-linear, machine-learning approach that excels at capturing

370   non-additive especially interaction-type genetic effects(Jannink et al., 2010). We

371   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

377   **RESULTS**

378   **Genotyping Data**

379   SNP marker data was generated using genotyping-by-sequencing (GBS)

380   (Elshire et al., 2011). Overall coverage was 0.07x (range 0.05-0.2). There were

381   114,922 markers that passed initial filters with a MAF > 1%. Of these, 95,047 are

382   mapped to the genome. Of mapped markers used for GWAS (MAF > 5%), there was

383   an average of 2293 SNPs per chromosome or one marker every 9.5 kb. The mean

384   MAF (0.21-0.22), mean heterozygosity (0.32-0.35) and number of markers analyzed

385   (40,539-42,113) were similar between sub-populations (Table 1). Most

386   chromosomes in most populations had mean r^2 > 0.2 extending 10 to 50 Kb. The r^2

387   between markers 4.5-5.5kb apart was 0.3 on average (median 0.13) suggesting at

388   least some LD between most causals and at least one marker but also that increased

389   density in future studies will provide additional mapping resolution (Figs S14-S19).

390

391   **Population stratification and structure**

392   Principal components analysis of our SNP dataset revealed subtle

393   differentiation among African cassava clones analyzed. This can been seen from a

394   plot of the first four PCs (cumulative variance explained = 15%). The Nigerian sub-

395   populations (NRCRI, IITA Genetic Gain, Cycle 1 and Cycle 2) occupy similar genetic

396   space, but the Ugandan sub-population (NaCRRI) is somewhat distinct on PC1 and

397   PC2. This may be consistent with a history of germplasm sharing and recurrent use

398   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

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;

445   b) are located in this region. However, while Rabbi et al.'s (2014a; b) strongest

446   association was on scaffold 5214, corresponding to Chr. 8 position 6511133, the

447   strongest association for the present study is on scaffold 6906 (7454373-7836749),

448   more than a megabase away. This discrepancy is due to the fact that the SNP

449   markers in scaffold 6906 did not segregate in the resistant parents of the bi-parental

450   mapping populations.

451

452   **Dissecting resistance originating from alleles or loci on Chromosome 8**

453   The significance region on chromosome 8 is large (~8 Mb; Fig. 3a). In fact,

454   the region appears as two, sometimes equally significant peaks in some sub-

455   populations (Fig. S27). We scanned the region for haplotype blocks with PLINK

456   (version 1.9, https://www.cog-genomics.org/plink2) and found it was not

457   characterized by a single, or even a few large, but many small LD blocks (Fig. S29). A

458   second locus (*CMD3)* has been reported on the same chromosome as

459   *CMD2*(Okogbenin et al., 2012b). The authors reported the marker NS198 to be 36

460   cM from *CMD2* and associated with very strong resistance in the progeny of

461   TMS972205. E-PCR collocated NS198 on chromosome 8, five megabases (position

462   997099) outside our significance region (Fig. S28). Thus our results suggest a

463   second QTL (i.e. *CMD3*), if present, is much closer to *CMD2* than previously believed.

464   We used several approaches to evaluate the evidence for multiple QTL in the

465   region. We conducted a post-hoc test for interactions between the top-marker on

466   chromosome 8 and every other marker on the chromosome. There were significant

467   interactions, explaining up to 42% of the variance, 1-3 megabases from the top

468   GWAS hits, but none in the region surrounding S8_7762525 (Fig. 3b, Fig. S30).

469   We implemented a multi-locus mixed-model (MLMM (Segura et al., 2012)),

470   which uses a forward-backward stepwise model selection approach to determine

471   which and how many marker cofactors are required to explain the associated

472   variance in the region. The MLMM for MCMDS in the population-wide sample

473   selected five markers (S8_7762525, S8_6380064, S8_6632472, S8_7325389,

474   S8_4919667) spanning the significance region (Fig. S31). Of the five, the first was

475   our top marker S8_7762525, the fourth is only about 400 Kb away, and the

476   remaining three cover the secondary peak and the region of statistical interaction.

477   These markers are mostly in linkage equilibrium (Table S10) and collectively

478   explain up to 40% of the variance. The selection of markers distributed across the

479   region by MLMM including both putative peaks to explain the phenotypic-

480   association in the region is additional evidence in support of a multi-locus

481   hypothesis.

482   LD decays in the region to low levels ($r^2<0.25$) and is virtually zero between

483   significant markers on the left, e.g. S8_5064191 and those on the right, e.g.

484   S8_762525 of the significance region (Fig. 3c). Combined with our genome-wide

485   analysis of LD decay rates (Figs. S14-S19) this LD decay makes it unlikely that a

486   single locus or allele is responsible for the associated region.

487   We examined the two-locus genotype effects (e.g. between S8_7762525 and

488   the SNP with the most significant interaction test, S8_4919667) and found a usually

489   dominant effect of the secondary resistance allele (e.g. S8_4919667) in the

490   heterozygous and homozygous resistant background at the primary peak (Fig. 4;

491   Fig. S32). We found little evidence of secondary peak effects in the homozygous

492   susceptible primary peak background. Clones that are homozygous resistant at both

493   loci are superior to all other cassava clones, expressing virtually no symptoms (Fig.

494   4c).

495

496   **Other Loci Associated with CMD Resistance**

497   We identified thirty-five markers on 13 chromosomes that explained 0.5–

498   10% (median 4%) of the variance (Table S9). Many of these had recessive and

499   usually rare susceptibility alleles (Fig. S26). Marker S4_637212 explained 4% of the

500   variance (CMD6S, Genetic Gain) and had an additive effect. Marker S11_20888811's

501   recessive resistance allele appears to lower CMD symptoms 4% more than *CMD2*

502   (S8_7762625) but only 14 clones are homozygous resistant at this locus (Fig. S26).

503   Further work on this locus is urgently needed to determine it's possible impact as

504   its frequency increases. There were four significant markers on chromosome 14

505   with mostly dominant effects and explaining up to 5% of the variance. Two

506   previously published SSR markers (SSRY44, NS146) (Mohan et al., 2013) are located

507   within 1.4 megabases of these SNPs (Fig. S28b). Four markers, spread across seven

508   megabases of chromosome 9, with recessive susceptibility loci, explained up to 10%

509   (S9_14551208) of the variance. These markers co-located with SSRY40, originally

510   reported as *CMD1* and associated with quantitative resistance (Fregene et al., 2000;

511   Mohan et al., 2013).

512

Candidate Genes

513

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).

535   We analyzed the coding sequence of the three genes mentioned above in

536 three CMD resistant cassava genotypes known to possess the qualitative resistance

537 allele(s) (TMEB3, TMEB7 and I011412) and in two susceptible/tolerant ones known

538 to possess only quantitative resistance sources (I30572 and TMEB1). We identified

539 SNPs within the coding regions and identified amino acid changes (Table S12). Two

540 non-synonymous mutations were found on exons 7 and 9 of Cassava4.1_007986,

541 homozygous in the susceptible group but heterozygous in the resistant clones

542 (Table S12). The peroxidase, Cassava4.1_011768 did not show any non-synonymous

543 mutations specific to the resistant/susceptible group. However, Cassava4.1_02917,

544 showed three non-synonymous mutations that were specific to the susceptible

545 group.

546

547 **Genomic Prediction of Additive and Total Genetic Merit**

548   We tested four prediction models using cross-validation: (1)

549 $\text{Additive}_{\text{All\_Markers}}$, (2) $\text{Additive}_{\text{All\_Markers}}$ + $\text{Dominance}_{\text{All\_Markers}}$ + $\text{Epistasis}_{\text{All\_Markers}}$, (3)

550 $\text{Additive}_{\text{CMD2}}$ + $\text{Additive}_{\text{Non-CMD2}}$, (4) $\text{Additive}_{\text{CMD2}}$ + $\text{Dominance}_{\text{CMD2}}$ + $\text{Epistasis}_{\text{CMD2}}$ +

551 $\text{Additive}_{\text{Non-CMD2}}$. Mean cross-validation accuracy averaged 0.53 for additive and 0.55

552 for total value across models (Table 2, Figure 5). Including non-additive effects,

553 using all markers (model 2) shifted 60% of the variance to dominance and epistasis

554 and decreased the accuracy of the additive prediction from 0.53 (model 1) to 0.51,

555 but gave increased total prediction accuracy of 0.55. An additive only model giving

556 separate weight to CMD2 and non-CMD2 regions (model 3) had the highest total

557 prediction accuracy (0.58), with most accuracy coming from CMD2 (0.54) vs. non-

558   CMD2 (0.29) but most variance absorbed by non-CMD regions. Modifying model 3 to

559   allow the CMD2 region additive, dominance and epistatic effects (model 4) slightly

560   decreased total prediction accuracy (0.57) relative to model 3, with most accuracy

561   coming from the additive CMD2 kernel (0.52), but with 51.7% non-additive

562   variance, 33.6% non-CMD2 variance and only 14.7% additive CMD2.

563

564

565 **DISCUSSION**

566   The present study solidifies our understanding of the genetic resistance to CMD

567 that is available in African cassava germplasm and demonstrates the efficacy of

568 genomic selection at improving CMD resistance. After conducting the first genome-

569 wide association study for this species with markers anchored to chromosomes, we

570 are able to confirm that the basis of genetic resistance to CMD is indeed narrow,

571 arising chiefly from a single region of chromosome 8 that collocates with the loci

572 *CMD2* (Akano et al., 2002) and *CMD3* (Okogbenin et al., 2012b). The lack of new

573 major effect loci is a key outcome of our study. Even after analyzing a broad sample

574 of the breeding germplasm from West and East Africa. However, we also identified

575 13 regions of small effect including one on chromosome 9 that collocates with *CMD1*

576 (Fregene et al., 2000).

577   Another key result of our analysis is that the most highly resistant cassava

578 clones, those that never show disease symptoms, are only identified using models of

579 epistasis in the significance region on chromosome 8. We propose two alternative

580 hypotheses to explain this result. As suggested both in our analyses and previous

581 studies (Okogbenin et al., 2012b) there may be multiple interacting loci in the region

582 (i.e. *CMD2* and *CMD3*). Alternatively, our results may arise from a complex haplotype

583 structure, where observed levels of resistance come from a single locus with one

584 moderate and another strong resistance allele segregating in the population. An

585 example of the later scenario is resistance to tomato yellow leaf curl which initially

586 mapped to two genes Ty-1 and Ty-3 on the same chromosome, but was later

587 revealed by fine-mapping to be one gene with multiple alleles (Verlaan et al., 2013).

588   In order to facilitate functional studies of the qualitative resistance source(s) on

589   chromosome 8, we used our GWAS results to identify three candidate genes.

590   Interestingly, there are no major resistance genes (e.g. NBS-LRR) in our region of

591   interest (Lozano et al., 2015). We found two peroxidases, which have recently been

592   shown to down-regulate in response to cassava mosaic geminivirus infection in

593   susceptible genotypes (Allie et al., 2014) and a thioredoxin, which can be important

594   for plant defense activation (Bashandy et al., 2010; Ballaré, 2014). We note that our

595   genome assembly contains gaps ((ICGMC), 2014) and is based on a South American

596   accession (Prochnik et al., 2012) that may not possess the causal gene(s). Significant

597   work remains to identify the causal mechanism of qualitative resistance to CMD.

598   Finally, we demonstrate the potential of genomic selection for CMD resistance

599   breeding. In agreement with our association analyses, we found most of the variance

600   and the prediction accuracy was attributable to the chromosome 8 QTL(s). While

601   additive models will allow us to accurately select parents for cassava breeding, we

602   found non-additive prediction of total genetic merit to be even more accurate.

603   Prediction of total genetic merit will therefore enable cassava breeders to more

604   easily identify clones with superior disease resistance to be elite varieties,

605   effectively exploiting dominance and epistasis for crop improvement. Further, it is

606   significant that, while accuracy is low for the quantitative (non-major gene)

607   components, it is not zero. Thus it should be possible to do genomic selection to

608   simultaneously improve both qualitative (i.e. *CMD2/CMD3*) and quantitative (i.e.

609   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

623 **ACKNOWELDGEMENTS**

632

633

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).

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

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

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

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

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

729 Okechukwu, R.U., and a. G.O. Dixon. 2008. Genetic Gains from 30 Years of Cassava Breeding in Nigeria
730   for Storage Root Yield and Disease Resistance in Elite Cassava Genotypes. J. Crop Improv. 22(2):
731   181–208Available at http://www.tandfonline.com/doi/abs/10.1080/15427520802212506 (verified 18
732   July 2014).

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

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

743 Okogbenin, E., M. Porto, and C. Egesi. 2007. Marker-assisted introgression of resistance to cassava mosaic
744   disease into Latin American germplasm for the genetic improvement of cassava in Africa. Crop Sci.
745   47: 1895–1904Available at https://dl.sciencesocieties.org/publications/cs/abstracts/47/5/1895
746   (verified 4 November 2013).

747 Oliveira, E.J., M.D.V. Resende, V. Silva Santos, C.F. Ferreira, G.A.F. Oliveira, M.S. Silva, L.A. Oliveira,
748   and C.I. Aguilar-Vildoso. 2012. Genome-wide selection in cassava. Euphytica 187(2): 263–
749   276Available at http://link.springer.com/10.1007/s10681-012-0722-0 (verified 19 September 2013).

750 Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N. a Shadick, and D. Reich. 2006. Principal
751   components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38(8):
752   904–9Available at http://www.ncbi.nlm.nih.gov/pubmed/16862161 (verified 23 May 2014).

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

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

757 Rabbi, I., M. Hamblin, M. Gedil, P. Kulakow, M. Ferguson, A.S. Ikpan, D. Ly, and J.-L. Jannink. 2014a.
758   Genetic Mapping Using Genotyping-by-Sequencing in the Clonally Propagated Cassava. Crop Sci.
759   54: 1–13Available at https://www.crops.org/publications/cs/abstracts/0/0/cropsci2013.07.0482
760   (verified 20 June 2014).

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

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

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

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

776 VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11): 4414–
777   23Available at http://www.ncbi.nlm.nih.gov/pubmed/18946147 (verified 18 October 2013).

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

781 Wong, W.W.L., J. Griesman, and Z.Z. Feng. 2014. Imputing genotypes using regularized generalized linear
782   regression models. Stat. Appl. Genet. Mol. Biol. 13(5): 519–529Available at
783   http://www.degruyter.com/view/j/sagmb.2014.13.issue-5/sagmb-2012-0044/sagmb-2012-0044.xml.

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

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

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

796
797

798   **FIGURE LEGENDS**

799

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

804

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

810

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

822

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

830

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

843

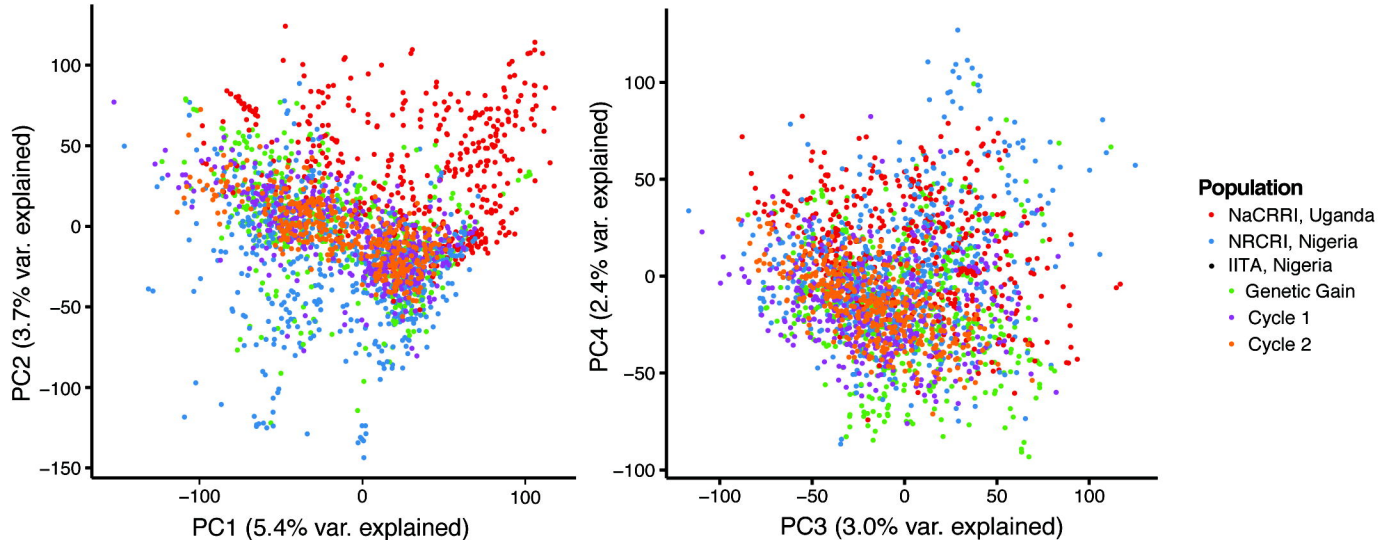**Table 1**. Summary of phenotype and genotype datasets analyzed.

Trait	NRCRI	NaCRRI	IITA: Genetic Gain	GS Cycle 1	GS Cycle 2	Trait Description
			Population			
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

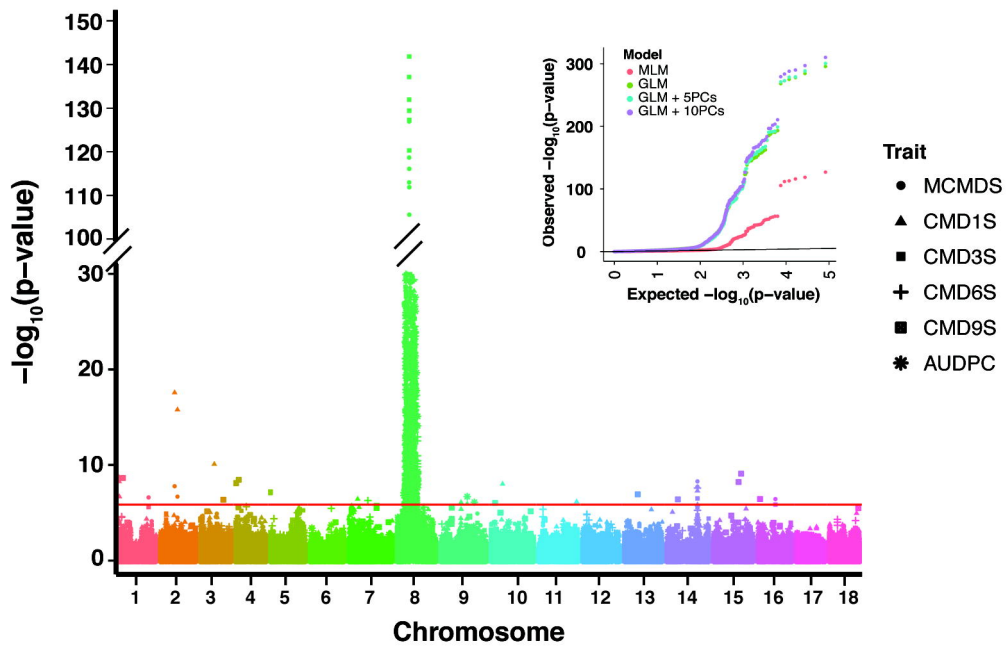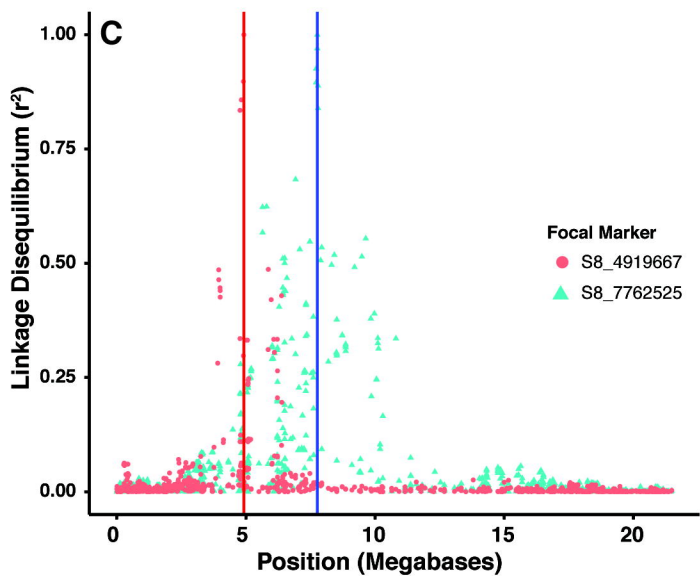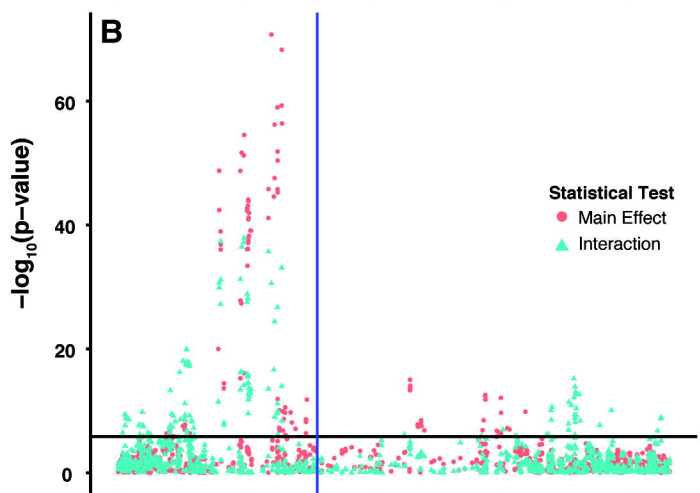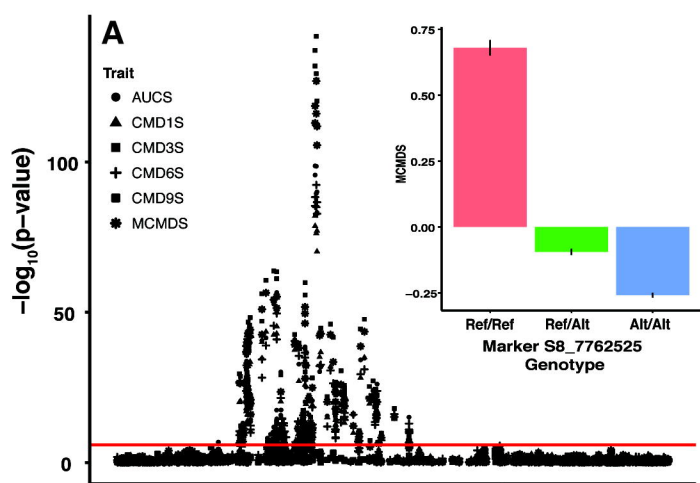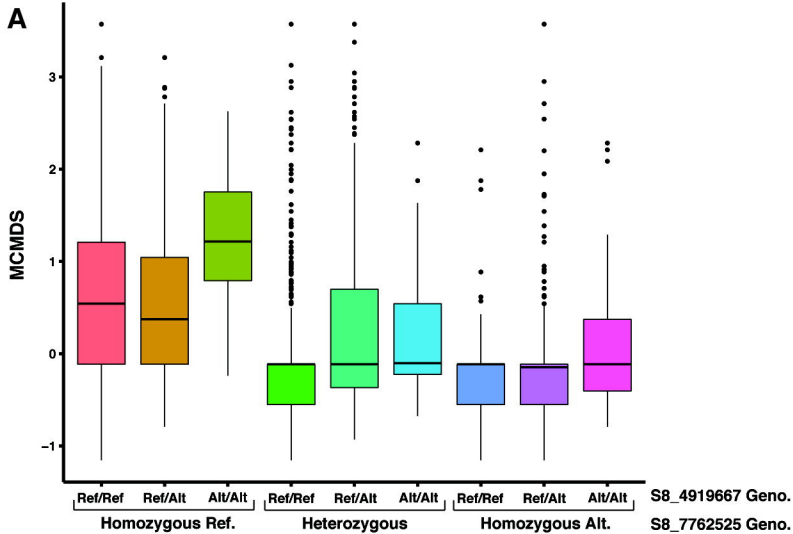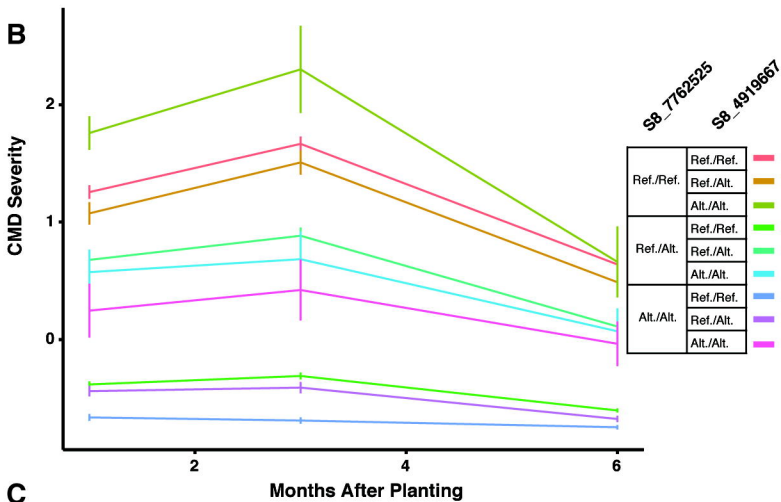|	
NRCRI	National Root Crops Research Institute (NRCRI) in Umudike, Nigeria
NaCRRI	National Crops Resources Research Institute (NaCRRI) in Namulonge, Uganda
IITA: Genetic Gain	International Institute of Tropical Agriculture in Ibadan, Nigeria,
IITA: Cycle 1	Genomic selection progenies of 76 IITA: Genetic 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) $\text{Additive}_{\text{All\_Markers}}$ 1				0.53	0.53
2) $\text{Add}_{\text{All\_Markers}}$ 0.396	$\text{Dom}_{\text{All\_Markers}}$ 0.015	$\text{Epi}_{\text{All\_Markers}}$ 0.589		0.51	0.55
3) $\text{Additive}_{\text{CMD2}}$ 0.3		$\text{Additive}_{\text{Non-CMD2}}$ 0.7		0.54_{CMD2} / $0.29_{\text{Non-CMD2}}$	0.58
4) Add_{CMD2} 0.147	Dom_{CMD2} 0.019	Epi_{CMD2} 0.498	$\text{Add}_{\text{Non-CMD2}}$ 0.336	0.52_{CMD2} / $0.25_{\text{Non-CMD2}}$	0.57

846

Population
- NaCRRI, Uganda
- NRCRI, Nigeria
- IITA, Nigeria
- Genetic Gain
- Cycle 1
- Cycle 2

Additive Model | Epistatic Model (Additive Pred.) | Epistatic Model (Total Pred.)

A All Markers, CMD2 Markers, Non−CMD2 Markers

B All Markers, CMD2 Markers, Non−CMD2 Markers

C All Markers, CMD2 + Non−CMD2

Cross-validation Accuracy

D **E**

Kernel Weights
- Additive
- Dominance
- Add.-by-Dom. Epistasis
- Non-CMD2 (Additive)