Published May 13, 2016

ORIGINAL RESEARCH

Genome-Wide Association and Prediction Reveals Genetic Architecture of Cassava Mosaic Disease Resistance and Prospects for Rapid Genetic Improvement

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

Abstract

Cassava (*Manihot esculenta* Crantz) 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. Biparental mapping studies suggest primarily a single major gene mediates resistance. To investigate this genetic architecture, we conducted the first genome-wide association mapping study in cassava with up to 6128 genotyping-by-sequenced African breeding lines and 42,113 reference genome-mapped single-nucleotide polymorphism (SNP) markers. We found a single region on chromosome 8 that accounts for 30 to 66% of genetic resistance in the African cassava germplasm. Thirteen additional regions with small effects were also identified. Further dissection of the major quantitative trait locus (QTL) on chromosome 8 revealed the presence of two possibly epistatic loci and/or multiple resistance alleles, which may account for the difference between moderate and strong disease resistances in the germplasm. Search of potential candidate genes in the major QTL region identified two peroxidases and one thioredoxin. Finally, we found genomic prediction accuracy of 0.53 to 0.58 suggesting that genomic selection (GS) will be effective both for improving resistance in breeding populations and identifying highly resistant clones as varieties.

Core Ideas

- Cassava mosaic disease resistance has a narrow genetic basis in breeding germplasm.
- Evidence suggests two possibly epistatic loci and/or multiple resistance alleles exist at the major QTL.
- Genomic prediction is accurate both for selecting parents and identifying highly CMD-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

M.D. Wolfe, M. Hamblin, R. Lozano, D.P.D. Carpio, P. Ramu, and J.L. Jannink, Dep. of Plant Breeding and Genetics, Cornell Univ., Ithaca, NY, USA; I.Y. Rabbi and P. Kulakow, International Institute for Tropical Agriculture, Ibadan, Oyo, Nigeria; C. Egesi, National Root Crops Research Institute, Umudike, Umuahia, Nigeria; R. Kawuki, National Crops Resources Research Institute, Namulonge, Uganda; J.L. Jannink, USDA–ARS, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA. M.D. Wolfe and I.Y. Rabbi contributed equally to this work. Received 13 Nov. 2015. Accepted 11 Mar. 2016. *Correspondence author (wolfemd@gmail.com).

Abbreviations: AUDPC, area under the disease progress curve; BLUP, best linear unbiased prediction; CIAT, International Center for Tropical Agriculture; CMD, cassava mosaic disease; CMDS, cassava mosaic disease severity; GBS, genotype-by-sequencing; GEBV, genomic estimated breeding value; GLM, general linear model; GS, genomic selection; GWAS, genome-wide association study; IITA, International Institute of Tropical Agriculture; LD, linkage disequilibrium; MAF, minor allele frequency; MCMDS, mean cassava mosaic disease severity; MLM, mixed linear model; MLMM, multilocus mixed model; NaCRRI, National Crops Resources Research Institute; NRCRI, National Root Crops Research Institute; PCA, principal components analysis; PCR, polymerase chain reaction; QTL, quantitative trait loci; SNP, single-nucleotide polymorphism; SSR, simple-sequence repeat; TMS, Tropical Manihot Selections.

Published in Plant Genome
Volume 9. doi: 10.3835/plantgenome2015.11.0118

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 SNP-based genetic linkage maps (International Cassava Genetic Map Consortium, 2014), it is for the first time possible to study the genetic architecture of key traits using modern genome-wide association study (GWAS) and to improve those traits with GS (Oliveira et al., 2012; Ly et al., 2013).

Cassava mosaic disease 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 worldwide search for resistance in cultivated germplasm in the 1930s, cassava breeders at the Amani research station in Tanzania resorted to interspecific hybridization with Ceara rubber tree (*M. glaziovii* Mull. 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 used to initiate cassava breeding at the International Institute of Tropical Agriculture (IITA) in the 1970s and resulted in the Tropical Manihot Selections (TMS) varieties (Hahn et al., 1980b).

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

A limited genetic base for the dominant resistance implies potential vulnerability if the cassava mosaic geminivirus can evolve to overcome it. This possibility necessitates diversification of resistance sources to ensure durability. To determine with greater certainty whether there are additional sources of CMD resistance in the continent's breeding germplasm, we undertook a large GWAS using >6000 cassava accessions from West and East Africa genotyped at more than 40,000 SNP loci using genotype-by-sequencing (GBS) (Elshire et al., 2011). The entire collection represents five subpopulations (Table 1) that are part of an ongoing international GS-based breeding project in cassava (http://www.nextgencassava.org). In addition, we combined GWAS and genomic prediction to not only dissect the genetic architecture of resistance to CMD but also to assess the potential for population improvement by 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 nonadditive models to predict total genetic merit of clones for the selection of superior CMD resistant varieties were assessed. Finally, multikernel 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 for three African plant 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 population (also referred to as the Genetic Gain) 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 interspecific hybridization program at the Amani Station in Tanzania (Hahn et al., 1980b). It also includes hybrids of germplasm introduced from Latin America (see Supplemental 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 (Supplemental Table S2).

The NaCRRI in Uganda has a population of 414 clones that represents the genetic diversity of the East African cassava gene pool. The pedigree of this population arises from 49 parents coming from IITA, CIAT in Columbia, and Amani Research Station in Tanzania (Supplemental Table S3). The population was generated, in part, by making crosses of parents with qualitative resistance to parents with quantitative resistance as well as quantitative × quantitative and qualitative × qualitative resistances. We note that there are two major clades of cassava mosaic virus species, African cassava mosaic virus and East African cassava mosaic virus (Legg and

Table 1. Summary of phenotype and genotype datasets analyzed.

Trait†	Population‡ NRCRI	NaCRRI	IITA: Genetic Gain	GS Cycle 1¶	GS Cycle 2#	Trait Description§
CMD1S	X		X	X	X	CMD severity rated on a scale from 1 (no symptoms) to 5 (extremely severe) at 1 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).
Years	2	2	14	2	1	
Locations	3	3	10	3	1	
Propagation	Clonal	Clonal	Clonal	Seed and clonal	Seed	
No. clones	626	414	694	2187	2466	
No. markers (MAF > 0.05)	41820	41060	42113	41369	40539	
Mean MAF (mapped markers, with MAF > 0.05)	0.21	0.21	0.22	0.21	0.22	
Mean observed heterozygote frequency	0.32	0.33	0.34	0.33	0.35	

† CMD, cassava mosaic disease; MCMDS, mean cassava mosaic disease severity; AUDPC, area under disease progress curves; IITA, International Institute of Tropical Agriculture.

‡ NRCRI, National Root Crops Research Institute in Umudike, Nigeria; NaCRRI, National Crops Resources Research Institute in Namulonge, Uganda; GS, genomic selection.

§ MAP, months after planting; MAF, minor allele frequency.

¶ Genomic selection progenies of 76 IITA: Genetic Gain clones.

Genomic selection progenies of 158 IITA: Cycle 1 clones.

Fauquet, 2004). East African cassava mosaic virus is generally more severe in its symptoms and is present in West Africa but only at low levels, usually occurring as a dual infection with African cassava mosaic virus (Legg and Fauquet, 2004; Rabbi et al., 2014b). This fact makes it all the more important to include East African cassava breeding germplasm in a more comprehensive screen of the genetic architecture of CMD resistance.

We also analyzed a large genotyped and phenotyped multiparental population of individuals from two cycles of GS conducted at IITA. The GS program at IITA will be described briefly here and in detail as part of other publications. In 2012 the IITA Genetic Gain population was used as the reference population from which genomic estimated breeding values (GEBVs) were obtained using the genomic best linear unbiased prediction (BLUP) (Heffner et al., 2009). Selection of clones from the Genetic Gain was based on a selection index including CMD and cassava bacterial blight disease severity and yield components (dry matter content, harvest index, and fresh root yield). In the end, 83 parents gave rise to 2187 progenies, which we will call IITA Cycle 1. In 2013, the GEBVs for Cycle 1 were obtained, again using the Genetic Gain as a reference population, and 84 Cycle 1 plus 13 (97 total) Genetic Gain clones were selected as parents, giving rise to 2466 progenies (Cycle 2). The pedigrees of IITA Cycle 1 and Cycle 2 are available in Supplemental Table S4 and S5.

Phenotyping Trials

Phenotypic data were combined from trials conducted at multiple locations in Nigeria and Uganda. The data are contributed from all three breeding programs (IITA, NRCRI, and NaCRRI). The IITA's Genetic Gain trials were conducted in seven locations over 14 yr (2000–2014) in Nigeria. Each Genetic Gain trial comprises a randomized, unblocked design replicated one or two times per location and year. The NRCRI's population was phenotyped in 2 yr, 2013 and 2014. During the 2012–2013 season, the trial was conducted in one location, Umudike, Nigeria. In the 2013–2014 season, the population was planted in three locations (Umudike, Kano, and Otobi). The NRCRI's trial design was a randomized incomplete block with three replications per location per year and five plants per plot. Trials at NaCRRI were conducted in two growing seasons: 2012–2013 and 2013–2014. In both years, plots were 10 plants in two rows of five with randomized incomplete blocks. During the first year, a single location (Namulonge, Uganda) was used with only one replicate. During the second season, two replications were used at each of three locations: Namulonge, Kasese, and Ngetta.

Genomic selection Cycle 1 (C1) progenies were observed as seedlings in the 2012–2013 field season with phenotyping conducted only for early disease expression and seedling vigor. Cycle 1 progenies were subsequently cloned and phenotyped in a three-location (Ibadan, Ikenne, and Mokwa) trial in the 2013–2014 season 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 subpopulation. We also provide details about the sample sizes and replication numbers for each location and year of data analyzed (Supplemental Table S6) and per accession (Supplemental 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. Cassava mosaic disease severity was scored at up to five time points (1, 3, 6, 9, and 12 mo 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 Supplemental Fig. S1 through 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 38,854 observations on 6,198 genotyped and phenotyped clones to single BLUPs for each. To do this, we fit the following mixed linear model (MLM) with the lme4 (Bates et al., 2015) package in R (R Development Core Team, 2010):

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

Here, y_{lij} represents raw phenotypic observations, μ is the grand mean, c_l is a random effects term for clone with $c_l \sim N\left(0,\sigma_l^2\right)$, 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 $N\left(0,\sigma_r^2\right)$, and finally, ε_{lij} is the residual variance, assumed to be random and distributed $N\left(0,\sigma_e^2\right)$. Because the number of observations per clone varies greatly in our dataset (from 1 to 941, median of 2; Supplemental Table S7), we expect BLUPs are differentially shrunken to the mean. To counter this, we deregressed BLUPs according to the following formula:

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

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 Supplemental Fig. S7 through S13.

We also calculated AUDPCs for each clone using data from 1, 3, and 6 mo 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 Eq. [1] but with c_l indicating the clone–time point combination. This gave us a deregressed BLUP for each clone at each time point. We calculated areas under these curves using the trapezoid rule as implemented by the auc function in the flux R package (http://cran.r-project.org/web/packages/flux/index.html). We excluded 9 and 12 mo data because they were only scored at NRCRI, thus, including them would have limited the ability to compare results for this trait between populations.

Genotype Data

Genotyping of SNP markers was done by the GBS procedure (Elshire et al., 2011) using the ApeKI restriction enzyme recommended by Hamblin and Rabbi (2014) and read lengths of 100 bp. Marker genotypes were called with the TASSEL GBS pipeline V4 (Glaubitz et al., 2014) and aligned to the cassava version 5 reference genome available on Phytozome (http://phytozome.jgi.doe.gov) and described by the International Cassava Genetic Map Consortium (2014). Individuals with >80% missing SNP calls and markers with >60% missing were removed. Also, markers with extreme deviation from Hardy–Weinberg equilibrium ($\chi^2 > 20$) were removed. Allele calls were maintained if depth was $\geq$2, otherwise the call was set to missing. Marker data was converted to dosage format (0, 1, 2) and missing data were imputed with the glmnet algorithm in R (http://cran.r-project.org/web/packages/glmnet/index.html) as described in Wong et al. (2014). To judge the resolution of association analyses, we calculated pair-wise linkage disequilibrium (LD) between all markers with a minor allele frequency (MAF) $\geq$5% on each chromosome using PLINK (version 1.9, Chang et al., 2015). We examined the rate of decay with increasing physical distance between markers.

Population Structure and Genome-Wide Association Analyses

To examine the patterns of relatedness within and among our populations and to control population structure, we constructed a genomic-relationship matrix according to the formulation of Yang et al. (2011), as implemented in PLINK, using all markers with >1% MAF.

We conducted principal components analysis (PCA) on SNP markers with MAF >5% using the prcomp function in R. Principal component analysis on SNP markers is often used to identify major patterns of relatedness (population structure) in a sample, and the first few PCs can be used as covariates to control false-positive rates

in GWAS (Price et al., 2006). Because the GS progenies (C1 and C2) are by far the largest part of our dataset and because these individuals are descended from the IITA Genetic Gain population, we excluded these from the initial PCA. We then projected these individuals into the genetic space defined by the three training populations (NRCRI, IITA, and NaCRRI) using the predict function in R. This allowed us to visualize and quantify the relatedness in our populations based on the founders only and unbiased by the large size of the C1 and C2 collections.

Because GWAS has not previously been done in this or any other cassava 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 five and one with 10 PCs (GLM + 5 PCs, GLM + 10 PCs; Price et al., 2006), and a MLM, which fits a random effect for clone with $\sim N\left(0,\sigma_g^2\mathbf{K}\right)$, where σ_g^2 is the clonal variance component and **K** is the relationship matrix described above (Kang et al., 2010). The MLMs were conducted using the 'population parameters previously determined' and compression method (Zhang et al., 2010). All GWAS were conducted in TASSEL (version 5; Bradbury et al., 2007). We compared the observed $-\log_{10}(p\text{-values})$ against the expectation using quantile–quantile plots. We used visual inspection of quantile–quantile plots to judge which model most effectively reduced the genome-wide inflation of $-\log_{10}(p\text{-values})$ 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 subpopulations, we conducted GWAS population wide as well as within each subpopulation. In each analysis, we used markers that segregated with MAF >5% in that specific subpopulation. Bonferroni thresholds were calculated according to the number of markers analyzed in each subpopulation.

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 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 mo 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 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). Polymerase chain reaction (PCR)–free libraries were generated from these clones and sequenced at 20× 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 for simplicity, TMEB1 and I30572 will be referred to 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 multiple random effects (also know as multiple-kernel or multiple-relationship matrix) genomic prediction model to compare the variance explained and prediction accuracy achieved from the major CMD QTL (CMD2) compared with 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 of VanRaden (2008) as implemented by the A.mat function in the rrBUP package (Endelman, 2011). Dominance relationships can be captured as $\mathbf{D}=\frac{\mathbf{HH}'}{\sum_i 2p_iq_i\left(1-2p_iq_i\right)}$ (Su et al., 2012; Muñoz et al., 2014), where **H** is the SNP marker matrix (individuals in rows, markers along columns), heterozygotes are given as $(1-2p_iq_i)$ and homozygotes are $(0-2p_iq_i)$. We made a custom modification (available on 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 nonadditive 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 (Supplemental 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 fivefold 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 selected one fold, removed the corresponding individuals from the training set, and used the remaining four folds to predict the fold that was left out. We iterated 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 ($\hat{g}$, 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, a nonlinear, machine-learning approach that excels at capturing nonadditive especially interaction-type genetic effects (Jannink et al., 2010). We used random forest regression only with the significant CMD2-associated markers as predictors to assess additional evidence for interaction at this locus on the basis of prediction accuracy achieved. We used the same cross-validation scheme described above.

RESULTS

Genotyping Data

The SNP marker data was generated using GBS (Elshire et al., 2011). Overall coverage was 0.07× (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 subpopulations (Table 1). Most chromosomes in most populations had mean $r^2 > 0.2$ extending 10 to 50 kb. Given one marker every ≈9.5 kb, we estimated r^2 on average 0.3 (median 0.13) between markers 4.5 to 5.5 kb apart, that is, the approximate maximum distance between an untyped causal locus and the nearest marker. This suggests 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 (Supplemental Fig. 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%; Fig. 1). The Nigerian subpopulations (NRCRI, IITA Genetic Gain, Cycle 1 and Cycle 2) occupy similar genetic space, but the Ugandan subpopulation (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.

We tested several standard GWAS models for controlling inflation of *p*-values caused by population structure including a GLM (no correction), a GLM with the first five PCs of the SNP matrix as covariates, and a MLM using the marker-estimated kinship matrix. Visual inspection of quantile–quantile plots (Fig. 2 inset; Supplemental Fig. S20–25) indicated that the MLM was most consistent for reducing $-\log_{10}$(*p*-values) toward the expected level (i.e., controlling false-positives, removing population structure effects). All subsequent results are therefore based on mixed-model associations. From the variance components estimated when fitting MLMs, we found that kinship matrices explained on average 57% (range 29–95%) of the phenotypic variance (Supplemental Table S9).

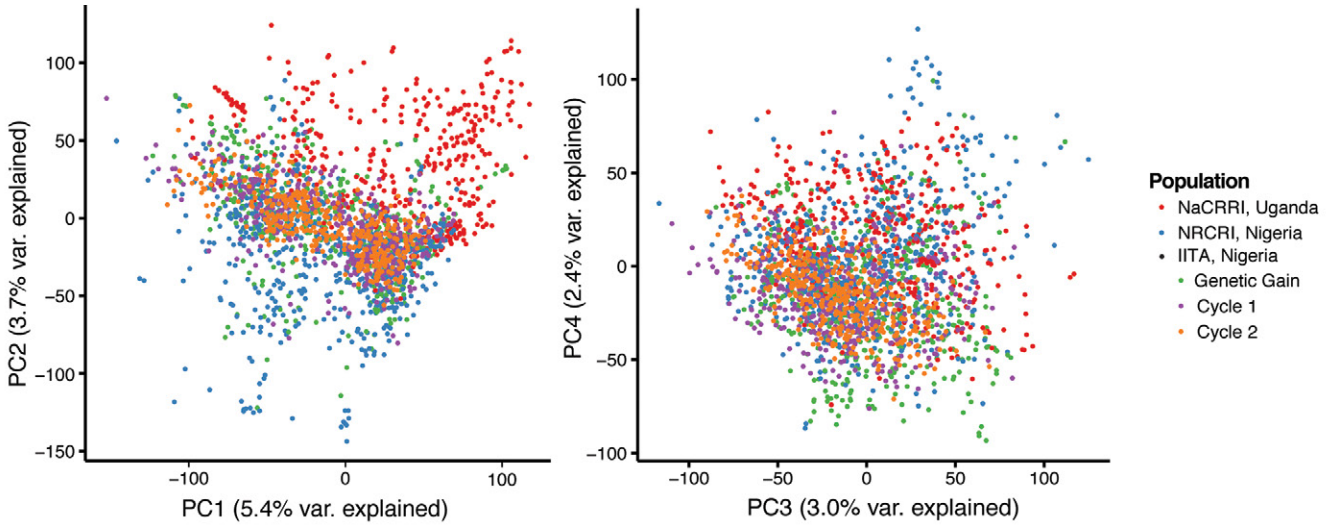

Fig. 1. Plot of the first four principal components (PCs) of the single-nucleotide polymorphism marker matrix. The three main training populations were used in the PC analysis and are shown here. A random sample of International Institute for Tropical Agriculture genomic selection Cycles 1 and 2 were projected into the genetic space and are displayed here.

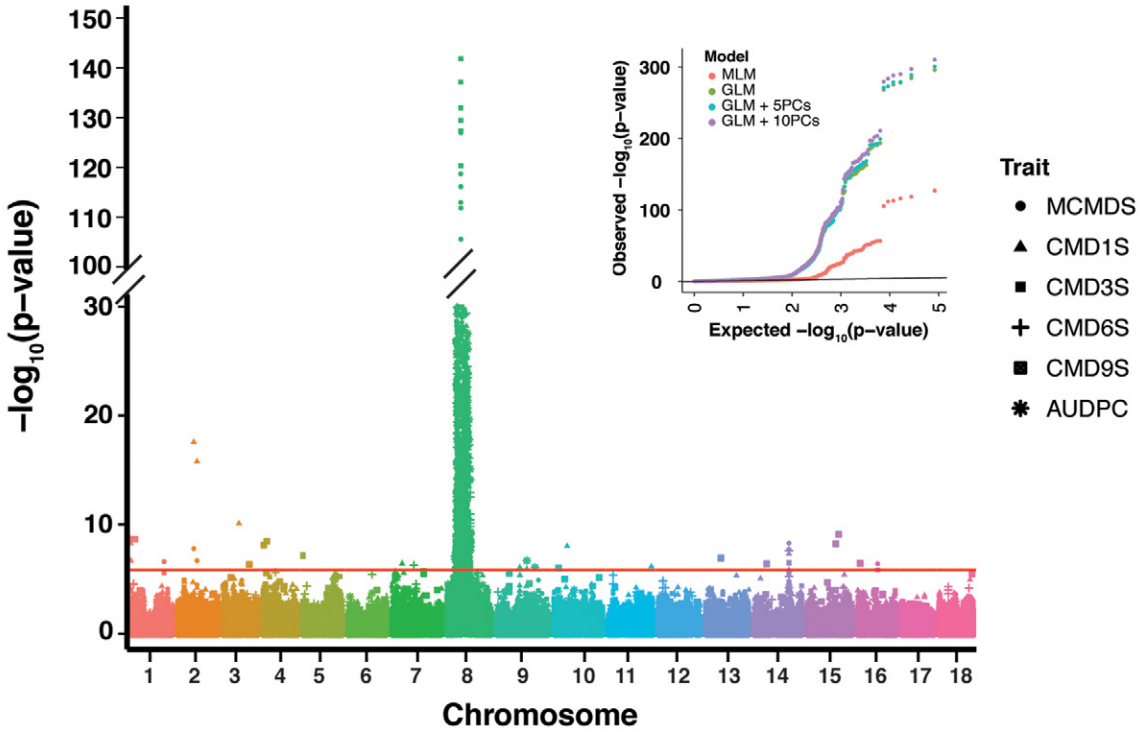

Fig. 2. Manhattan plot from mixed linear models summarizing genome-wide association results for all traits in all subpopulations. Bonferroni significance threshold is shown in red. An example quantile–quantile plot (mean cassava mosaic disease severity [MCMDS] in the population-wide analysis) is shown inset to demonstrate the differences between various population structure controls.

Overall Genome-Wide Associations

Association tests were performed for CMD symptom severity at 1, 3, 6, 9, and 12 mo after planting (where measured) in the five subpopulations (Table 1) and in an analysis that combined all accessions. We identified 311 markers in total that pass a Bonferroni significance threshold (Fig. 2; Supplemental Table S8). However, many significant SNPs were detected because of rare marker genotypes that were phenotypically extreme (Supplemental Fig. S26). The *F*-test implemented by TASSEL is sensitive to imbalanced sample size between groups, and we wish to be conservative and only consider significant results that we can be confident in. Therefore, we only considered SNPs where each genotype class (e.g., aa, Aa, AA) is represented by at least 10 individuals. This reduced the number of significant markers to 198 on 14 chromosomes, mostly concentrated at a single region of chromosome 8. Significant results were found within

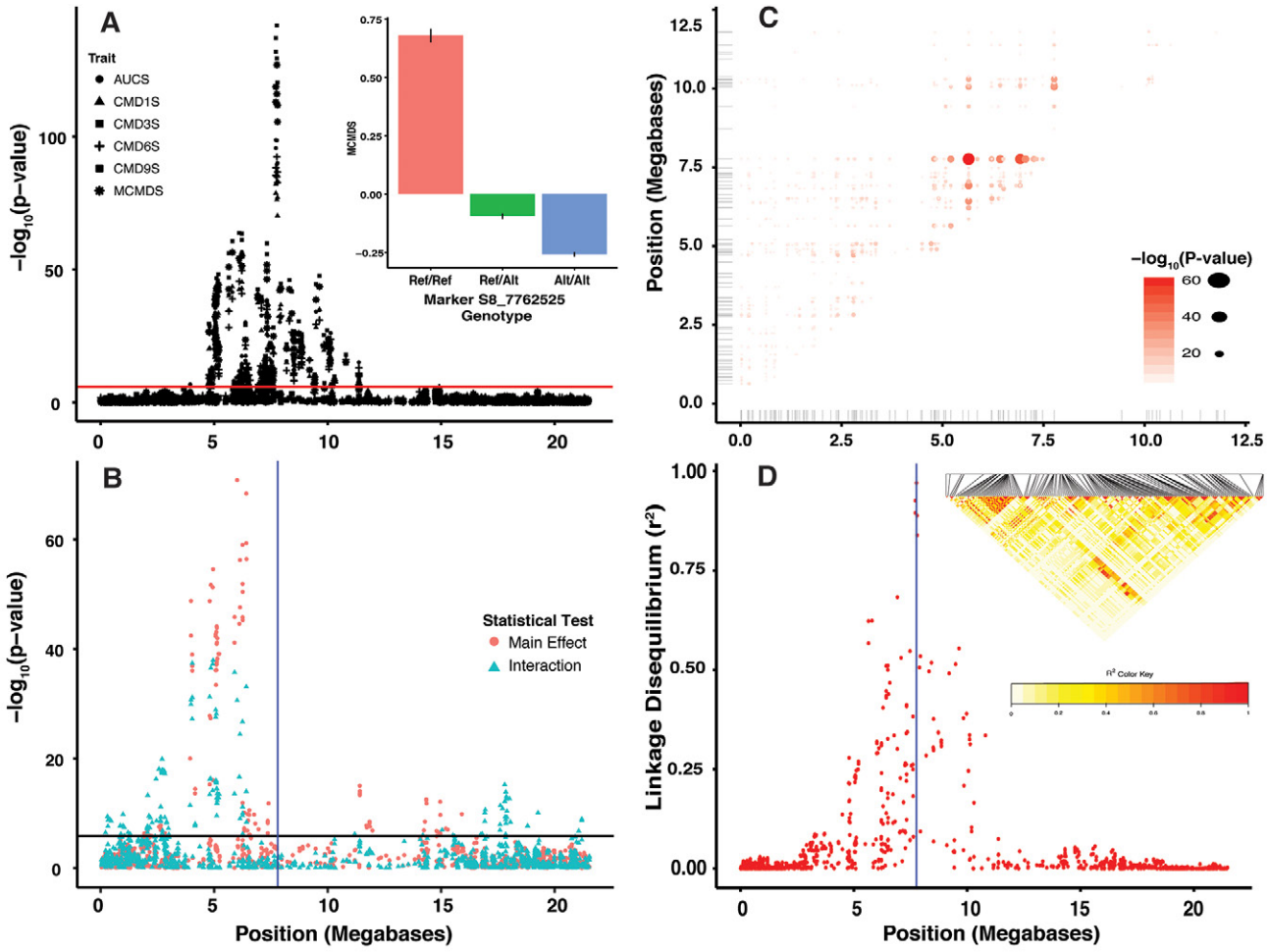

Fig. 3. Plots dissecting the major-effect quantitative trait loci on chromosome 8. (A) Manhattan plot summarizing genome-wide association study (GWAS) results for all cassava mosaic disease (CMD)-related traits in the population-wide mixed linear model analysis zoomed to chromosome 8 only. Bar plot showing mean and standard error for mean cassava mosaic disease severity (MCMDS) between each genotype class at the top marker, S8_7762525 (inset). (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) Results of pairwise statistical epistasis tests with position in megabases on chromosome 8 of each pair of markers tested represented on the axes. Significance is indicated by the size and color of each data point. (D) Linkage disequilibrium (LD) between S8_7762525 (blue vertical line) and every other maker (red dots). Heatmap showing LD between each pair of markers that passed the Bonferonni threshold in GWAS (inset).

each subpopulation with more signals associated with greater sample size (e.g., Cycle 1). Variance explained by significant markers ranged from 0.5 to 22% (median 3.5%) (Supplemental Table S8).

Chromosome 8 Contains the Major Resistance Locus *CMD2*

There were 163 significant markers on chromosome 8 (between 3.56–11.38 Mb; Fig. 3a) with the top marker (S8_7762525) explaining 5 to 22% of the variance depending on the subpopulation. The frequency of the resistance-associated allele at S8_7762525 is 56% overall (range: 44% in IITA Genetic Gain to 63% in IITA Cycle 2 progenies).

The resistance allele at S8_7762525 is incompletely dominant (Fig. 3 inset); homozygotes with the alternate allele were closer to CMD free than heterozygotes. To formally test this, we conducted post hoc tests using the lm function in R, in which we included either an additive or dominance deviation or both effects. The additive-only test explained 15% of the variance compared with a test of additive plus dominance effect that explained 20% and the test of dominance alone that accounted for only 2%.

We confirmed that our major QTL is the *CMD2* locus by aligning previously published SSR marker primers (SSRY28, NS158, and SSRY106) (Akano et al., 2002; Lokko et al., 2005; Okogbenin et al., 2007, 2012; Mohan et al., 2013) to the reference genome using electronic PCR (http://www.ncbi.nlm.nih.gov/tools/epcr/). Our significant markers on chromosome 8 colocate with these markers (Supplemental Fig. S28a). 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 chromosome 8 position 6,511,133,

the strongest association for the present study is on scaffold 6906 (7,454,373–7,836,749), 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 biparental mapping populations.

Multiallelic or Multilocus Resistance 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 subpopulations (Supplemental Fig. S27). We scanned the region for haplotype blocks with PLINK and found it was not characterized by a single, or even a few large, but many small LD blocks (Supplemental Fig. S29). A second locus (*CMD3*) has been reported on the same chromosome as *CMD2* (Okogbenin et al., 2012). The authors reported the marker NS198 to be 36 cM from *CMD2* and associated with very strong resistance in the progeny of TMS972205. Electronic PCR colocated NS198 on chromosome 8, 5 Mb (position 997,099) outside our significance region (Supplemental 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 further evaluate the QTL 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 to 3 Mb from the top GWAS hit but none in the region surrounding S8_7762525 (Fig. 3b). We also used PLINK to conduct pairwise epistasis tests for MCMDS after filtering to 286 markers on chromosome 8 with MAF >5% and LD <0.4. In this case, the most significant interaction was identified between markers S8_7762556 (31 bp from our top hit) and S8_5645072 in the secondary QTL region (Fig. 3c; Supplemental Fig. S30). This analysis explains a maximum of 19% of the variance and agrees with our other epistasis scan in the general location of a statistical interaction effect.

We implemented a multilocus 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, and S8_4919667) spanning the significance region (Supplemental Fig. S31). Of the five, the first was our top marker S8_7762525, the fourth is only ~400 Kb away, and the remaining three cover the secondary peak and the region of statistical interaction. These markers are mostly in linkage equilibrium (Supplemental 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 multilocus hypothesis.

Linkage disequilibrium decays in the region to low levels ($r^2 < 0.25$) and is virtually zero between significant markers on the left, for example, S8_5064191, and those on the right, for example, S8_7762525 of the significance region (Fig. 3d) but still extends relatively extensively around the top hit. Combined with our genome-wide analysis of LD decay rates (Supplemental Fig. S14–S19), this pattern explains, in part, the complexity of the association signal we see in the region and supports the possibility of multiple loci and alleles in the region.

In examining the two-locus genotype effects (e.g., between S8_7762525 and the SNP with the most significant interaction test, S8_4919667), we found the reference allele of the secondary resistance locus (e.g., S8_4919667) has an effect even in homozygous susceptible background of S8_7762525, although this depended on the marker considered (Fig. 4a; Supplemental Fig. S32). Clones that are homozygous resistant at both loci are superior to all other cassava clones, expressing virtually no symptoms (Fig. 4b).

Other Loci Associated with Cassava Mosaic Disease Resistance

We identified 35 markers on 13 chromosomes that explained 0.5 to 10% (median 4%) of the variance (Supplemental Table S9). Many of these had recessive and usually rare susceptibility alleles (Supplemental 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 (Supplemental Fig. S26). 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 and NS146) (Mohan et al., 2013) are located within 1.4 Mb of these SNPs (Supplemental Fig. S28b). Four markers, spread across 7 Mb of chromosome 9 with recessive susceptibility loci, explained up to 10% (S9_14551208) of the variance. These markers colocated with SSRY40 originally reported as *CMD1* and associated with quantitative resistance (Fregene et al., 2000; Mohan et al., 2013).

Candidate Genes

We intersected our association results with available gene annotations and related data and identified 105 unique genes within the association peaks with 79, 61, 56, and nine genes identified at 1, 3, 6, and 9 mo after planting, respectively (Supplemental Table S11, Supplemental Fig. S33). There were no significant differences between gene ontology categories between time points. Most of the annotated genes are involved in metabolic processes (Supplemental Fig. S34). Thirty-five out of the 105 genes are known to respond to cassava mosaic virus infection (Allie et al., 2014) (Supplemental Table S11).

Among these genes, we found ones known to be susceptibility or resistance factors, a number of which

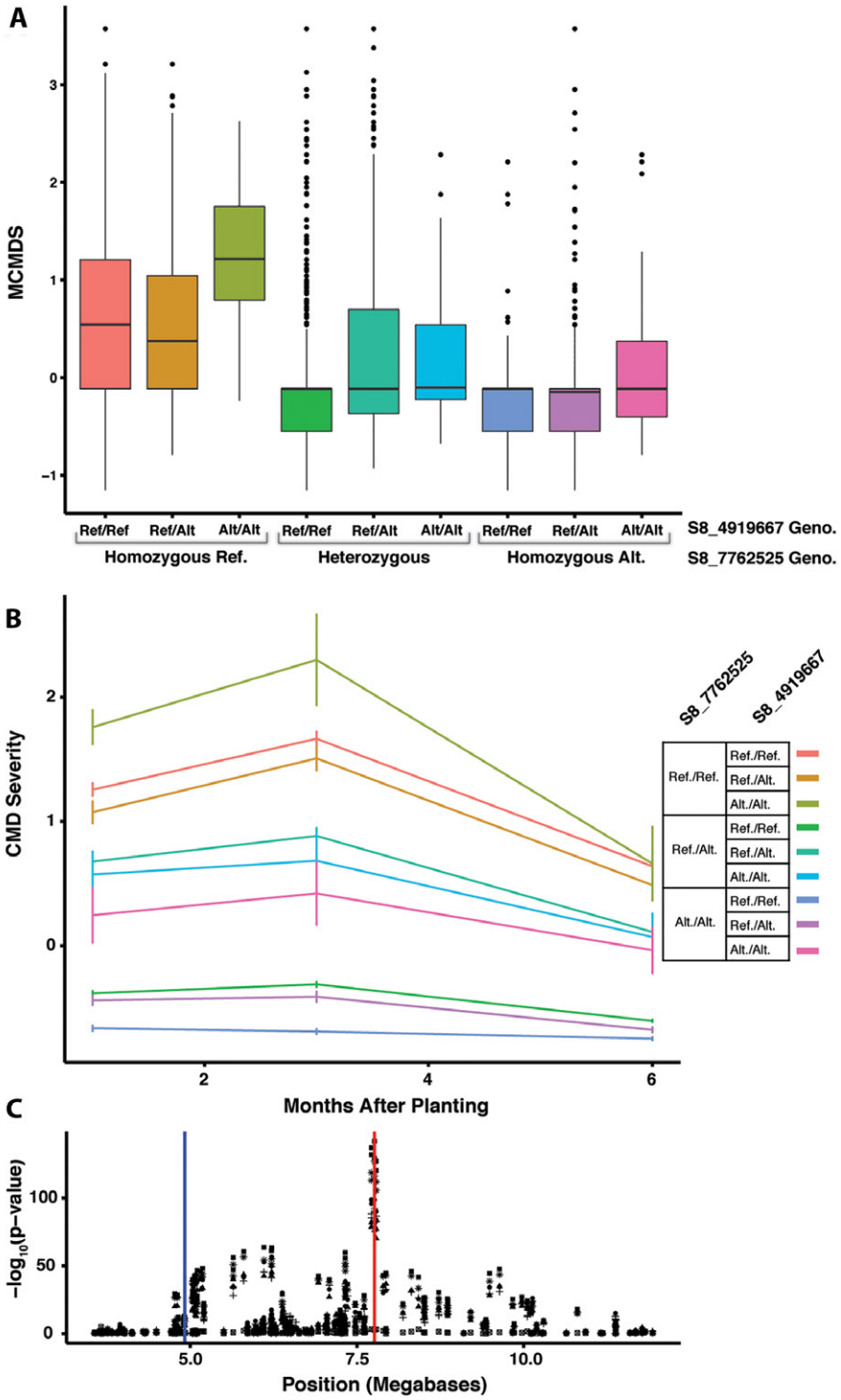

Fig. 4. Plots demonstrating the combined effect of the genotype at the top marker (S8_7762525) and the most epistatic marker (S8_4919667). (A) Boxplot showing the distribution of mean cassava mosaic disease (CMD) severity scores (MCMDS) for each two-locus genotype. (B) Disease progress curves showing mean and standard error CMD severity across 1, 3, and 6 mo 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)

are also involved in plant–geminivirus interaction processes (Hanley-Bowdoin et al., 2013). We found two peroxidases, Cassava4.1_029175 and Cassava4.1_011768, within the primary QTL region (scaffold 6906, ~7.7 Mb); peroxidases are pathogenesis-related proteins involved in host response to infection (van Loon et al., 2006).

In the secondary GWAS peak (scaffold 5214, 5–6 Mb), six SNPs were in a protein disulfide-isomerase-like 2-2 ortholog, a thioredoxin (PDIL2-2, cassava4.1_007986). In barley (*Hordeum vulgare* L.), an ortholog of PDIL2-2 (*HvPDIL5-1*) is a known virus susceptibility factor as are *PDI* gene family members across the animal and

plant kingdoms (Yang et al., 2014). We also identified the Ubiquitin-conjugating enzyme E2 ortholog (UBC5) gene (cassava4.1_017202) under the secondary GWAS peak (scaffold 5214, 5–6 Mb region). Genes like UBC5 in the ubiquitinylation pathway have been known to 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 alleles (TMEB3, TMEB7, and I011412) and in two susceptible and tolerant ones known to possess only quantitative resistance sources (I30572 and TMEB1). We identified SNPs within the coding regions and identified amino acid changes (Supplemental Table S12). Two nonsynonymous mutations were found on exons 7 and 9 of Cassava4.1_007986, homozygous in the susceptible group but heterozygous in the resistant clones (Supplemental Table S12). The peroxidase Cassava4.1_011768 did not show any nonsynonymous mutations specific to the resistant–susceptible group. However, Cassava4.1_02917 showed three nonsynonymous mutations that were specific to the susceptible group.

Genomic Prediction of Additive and Total Genetic Merit

We tested four prediction models using cross-validation: (i) $Additive_{All\_Markers}$, (ii) $Additive_{All\_Markers}$ + $Dominance_{All\_Markers}$ + $Epistasis_{All\_Markers}$, (iii) $Additive_{CMD2}$ + $Additive_{Non\text{-}CMD2}$, and (iv) $Additive_{CMD2}$ + $Dominance_{CMD2}$ + $Epistasis_{CMD2}$ + $Additive_{Non\text{-}CMD2}$. Mean cross-validation accuracy averaged 0.53 for additive and 0.55 for total value across models (Table 2). Including nonadditive 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-CMD2 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% nonadditive variance, 33.6% non-CMD2 variance, and only 14.7% additive CMD2.

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

Model	Fraction of the variance explained (kernel weights)†				Additive component accuracy	Total sum accuracy
1	$Add_{All\_Markers}$ 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

† Add, additive; Dom, dominance; Epi, epistasis.

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 GS at improving CMD resistance. After conducting the first GWAS 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 colocates with the loci *CMD2* (Akano et al., 2002) and *CMD3* (Okogbenin et al., 2012). 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 colocates 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 best 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., 2012), there may be multiple possibly epistatic 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).

To facilitate functional studies of the qualitative resistance sources on chromosome 8, we used our GWAS results to identify three candidate genes. Interestingly, there are no major resistance genes (e.g., nucleotide binding site–leucine rich repeat) in our region of interest (Lozano et al., 2015). We found two peroxidases, which have recently been shown to downregulate 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 (International Cassava Genetic Map Consortium, 2014) and is based on a South American accession (Prochnik et al., 2012) that may not possess the causal genes. Significant work remains to identify the causal mechanism of qualitative resistance to CMD.

Finally, we demonstrate the potential of GS 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 QTLs. While additive-only models achieve the highest accuracy of predicting phenotypes (e.g., Model 3), nonadditive models suggest that a large proportion of the apparent additive variance is actually dominance or epistasis. Partitioning genetic variance in prediction models may therefore provide more accurate estimates of the breeding (i.e., additive) value of cassava clones while simultaneously providing for the selection of clones with superior disease resistance to be elite varieties. The accuracies for disease resistance we achieved are consistent with those for other diseases in other plant species (e.g., sugarcane yellow leaf virus; Gouy et al., 2013).

It is significant that while accuracy is low for the quantitative (nonmajor gene) components, it is not zero. This is consistent with results for quantitative resistance in other plant species, including fusarium head blight in barley (Lorenz et al., 2012) and wheat (*Triticum aestivum* L.) (Rutkoski et al., 2012) as well as rust in sugarcane (*Saccharum officinarum* L.) (Gouy et al., 2013). It should therefore be possible to do GS to simultaneously improve both qualitative (i.e., *CMD2/CMD3*) and quantitative (i.e., polygenic background) resistance. This is important because quantitative disease resistance is important in many plant species (St. Clair, 2010) and has been shown to improve the durability of resistance when combined with major gene resistance, for example in pepper (*Capsicum annuum* L.) (Quenouille et al., 2014). Similar analyses to integrate results from GWAS into genomic prediction have been performed in wheat (Bentley et al., 2014; Zhao et al., 2014) and in rice (*Oryza sativa* L.) (Spindel et al., 2016). As in our case, this integration usually results in small increases in prediction accuracy.

The results we present in this study will represent progress toward discovering the mechanistic basis for major gene resistance to CMD and will also aid breeders seeking to pyramid useful alleles and achieve symptom-free cassava varieties either by marker-assisted selection or GS. In only 2 yr, we have conducted two rounds of selection and recombination, more than twice as fast as conventional phenotypic selection, and have increased the resistance-allele frequency at our top marker from 44 to 63%. The present study is an example of the possibilities for rapidly improving and dynamically breeding a crop that is crucial for hundreds of millions particularly in underdeveloped regions of the world.

Acknowledgments

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

Akano, O., O. Dixon, C. Mba, E. Barrera, and M. Fregene. 2002. Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. Theor. Appl. Genet. 105:521–525 doi:10.1007/s00122-002-0891-7

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 15:1006. doi:10.1186/1471-2164-15-1006

Ballaré, C.L. 2014. Light regulation of plant defense. Annu. Rev. Plant Biol. 65:335–63. doi: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 22:376–391. doi:10.1105/tpc.109.071225

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67:1–48. doi:10.18637/jss.v067.i01

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

Bentley, A.R., M. Scutari, N. Gosman, S. Faure, F. Bedford, P. Howell, J. Cockram, G.A. Rose, T. Barber, J. Irigoyen, R. Horsnell, C. Pumfrey, E. Winnie, J. Schacht, K. Beauchêne, S. Praud, A. Greenland, D. Balding, and I.J. Mackay. 2014. Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. Theor. Appl. Genet. 127:2619–2633. doi:10.1007/s00122-014-2403-y

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 23:2633–2635. doi:10.1093/bioinformatics/btm308

Ceballos, H., C.A. Iglesias, J.C. Pérez, and A.G.O. Dixon. 2004. Cassava breeding: Opportunities and challenges. Plant Mol. Biol. 56:503–516. doi:10.1007/s11103-004-5010-5.

Ceballos, H., P. Kulakow, and C. Hershey. 2012. Cassava Breeding: Current status, bottlenecks and the potential of biotechnology tools. Trop. Plant Biol. 5:73–87. doi:10.1007/s12042-012-9094-9

Chang C.C., C.C. Chow, L.C.A.M. Tellier, S. Vattikuti, S.M. Purcell, and J.J. Lee. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience 4:7. doi:10.1186/s13742-015-0047-8

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

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R Package rrBLUP. Plant Genome J. 4:250. doi:10.3835/plantgenome2011.08.0024

Fauquet, C., D. Fargette, and C. Munihor. 1990. African cassava mosaic virus: Etiology, epidemiology, and control. Plant Disease 74:404–411. doi:10.1094/PD-74-0404

Fregene, M., A. Bernal, M. Duque, A. Dixon, and J. Tohme. 2000. AFLP analysis of African cassava (*Manihot esculenta* Crantz) germplasm resistant to the cassava mosaic disease (CMD). Theor. Appl. Genet. 100:678–685. doi:10.1007/s001220051339

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55. doi:10.1186/1297-9686-41-55

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9:e90346 doi:10.1371/journal.pone.0090346

Gouy, M., Y. Rousselle, D. Bastianelli, P. Lecomte, L. Bonnal, D. Roques, J.C. Efile, S. Rocher, J. Daugrois, L. Toubi, S. Nabeneza, C. Hervouet, H. Telismart, M. Denis, A. Thong-Chane, J.C. Glaszmann, J.Y. Hoarau, S. Nibouche, and L. Costet. 2013. Experimental assessment of the accuracy of genomic selection in sugarcane. Theor. Appl. Genet. 126:2575–2586. doi:10.1007/s00122-013-2156-z

Hahn, S., A. Howland, and E. Terry. 1980a. Correlated resistance of cassava to mosaic and bacterial blight diseases. Euphytica 29:305–311. doi:10.1007/BF00025127

Hahn, S., E. Terry, and K. Leuschner. 1980b. Breeding cassava for resistance to cassava mosaic disease. Euphytica 29:673–683. doi:10.1007/BF00023215
Hahn, S., E. Terry, K. Leuschner, I. Akobundu, C. Okali, and R. Lal. 1979. Cassava improvement in Africa. Field Crops Res. 2:193–226. doi:10.1016/0378-4290(79)90024-8
Hamblin, M.T., and I.Y. Rabbi. 2014. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in cassava (*Manihot esculenta*). Crop Sci. 54:2603–2608. doi:10.2135/cropsci2014.02.0160
Hanley-Bowdoin, L., E.R. Bejarano, D. Robertson, and S. Mansoor. 2013. Geminiviruses: Masters at redirecting and reprogramming plant processes. Nat. Rev. Microbiol. 11:777–788. doi:10.1038/nrmicro3117
Heffner, E.L., M.E. Sorrells, and J.L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1–12. doi:10.2135/cropsci2008.08.0512
Henderson, C.R. 1985. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. J. Anim. Sci. 60:111–117.
International Cassava Genetic Map Consortium (ICGMC). 2014. High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from ten populations. G3: Genes, Genomes, Genet. doi:10.1534/g3.114.015008
Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomics 9:166–177. doi:10.1093/bfgp/elq001
Kang, H.M., J.H. Sul, S.K. Service. N.A. Zaitlen, S.Y. Kong, N.B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348–354 doi:10.1038/ng.548
Legg, J.Q., and C.M. Fauquet. 2004. Cassava mosaic geminiviruses. Plant Mol. Biol. 56:585–599. doi:10.1007/s11103-004-1651-7
Lokko, Y., E. Danquah, and S. Offei. 2005. Molecular markers associated with a new source of resistance to the cassava mosaic disease. African J. Biotech. 4:873–881.
Lorenz, J., K.P. Smith, and J.L. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52:1609–1621. doi:10.2135/cropsci2011.09.0503
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 16:360. doi:10.1186/s12864-015-1554-9
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. 53:1312–1325. doi:10.2135/cropsci2012.11.0653
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. 5:57–65 doi:10.5539/jas.v5n5p57
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 198:1759–1768. doi:10.1534/genetics.114.171322
Okechukwu, R.U., and G.O. Dixon. 2008. Genetic gains from 30 years of cassava breeding in Nigeria for storage root yield and disease resistance in elite cassava genotypes. J. Crop Improv. 22:181–208. doi:10.1080/15427520802212506
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. 2012. Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. Crop Sci. 52:2576–2586. doi:10.2135/cropsci2011.11.0586
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. 47:1895–1904. doi:10.2135/cropsci2006.10.0688
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 187:263–276 doi:10.1007/s10681-012-0722-0
Price, A.L., N.J. Patterson, R.M. Plenge, and M.E. Weinblatt. N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909 doi:10.1038/ng1847.
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. 5:88–94. doi:10.1007/s12042-011-9088-z
Quenouille, J., E. Paulhiac, B. Moury, and A. Palloix. 2014. Quantitative trait loci from the host genetic background modulate the durability of a resistance gene: A rational basis for sustainable resistance breeding in plants. Heredity (Edinb). 112:579–587. doi:10.1038/hdy.2013.138
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. 54:1384–1396. doi:10.2135/cropsci2013.07.0482
Rabbi, I.Y., M.T. Hamblin, and 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. 186:87–96. doi:10.1016/j.virusres.2013.12.028
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. 50:311–318. doi:10.1111/j.1744-7909.2007.00606.x
R Development Core Team 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.L. Jannink, and M. Sorrells. 2012. Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. Plant Genome J. 5:51–61 doi:10.3835/plantgenome2012.02.0001
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. 44:825–830 doi:10.1038/ng.2314
Spindel, J., H. Begum, D. Akdemir, B. Collard, E. Redoña, J. Jannink, and S. McCouch. 2016. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116:395–408. doi:10.1038/hdy.2015.113
St. Clair, D.A. 2010. Quantitative disease resistance and quantitative resistance loci in breeding. Annu. Rev. Phytopathol. 48:247–268. doi:10.1146/annurev-phyto-080508-081904
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:e45293. doi:10.1371/journal.pone.0045293
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. doi:10.1146/annurev.phyto.44.070505.143425
VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980
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:e1003399. doi:10.1371/journal.pgen.1003399
Wong, W.W.L., J. Griesman, and Z.Z. Feng. 2014. Imputing genotypes using regularized generalized linear regression models. Stat. Appl. Genet. Mol. Biol. 13:519–529. doi:10.1515/sagmb-2012-0044
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. 88:76–82. doi:10.1016/j.ajhg.2010.11.01
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. USA 111:2104–2109. doi:10.1073/pnas.1320362111
Zhang, Z., E. Ersoz, C. Lai, and R. Todhunter. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42:355–360. doi:10.1038/ng.546
Zhao, Y., M.F. Mette, M. Gowda, C.F.H. Longin, and J.C. Reif. 2014. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity 112:638–645. doi:10.1038/hdy.2014.1