High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding

Ismail Y. Rabbi [a,*], Martha T. Hamblin [b], P. Lava Kumar [a], Melaku A. Gedil [a], Andrew S. Ikpan [a], Jean-Luc Jannink [b,c], Peter A. Kulakow [a]

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

ARTICLE INFO

ABSTRACT

Cassava mosaic disease (CMD), caused by different species of cassava mosaic geminiviruses (CMGs), is the most important disease of cassava in Africa and the Indian sub-continent. The cultivated cassava species is protected from CMD by polygenic resistance introgressed from the wild species *Manihot glaziovii* and a dominant monogenic type of resistance, named *CMD2*, discovered in African landraces. The ability of the monogenic resistance to confer high levels of resistance in different genetic backgrounds has led recently to its extensive usage in breeding across Africa as well as pre-emptive breeding in Latin America. However, most of the landraces carrying the monogenic resistance are morphologically very similar and come from a geographically restricted area of West Africa, raising the possibility that the diversity of the single-gene resistance could be very limited, or even located at a single locus. Several mapping studies, employing bulk segregant analysis, in different genetic backgrounds have reported additional molecular markers linked to supposedly new resistance genes. However, it is not possible to tell if these are indeed new genes in the absence adequate genetic map framework or allelism tests. To address this important question, a high-density single nucleotide polymorphism (SNP) map of cassava was developed through genotyping-by-sequencing a bi-parental mapping population ($N = 180$) that segregates for the dominant monogenic resistance to CMD. Virus screening using PCR showed that CMD symptoms and presence of virus were strongly correlated ($r = 0.98$). Genome-wide scan and high-resolution composite interval mapping using 6756 SNPs uncovered a single locus with large effect ($R^2 = 0.74$). Projection of the previously published resistance-linked microsatellite markers showed that they co-occurred in the same chromosomal location surrounding the presently mapped resistance locus. Moreover, their relative distance to the mapped resistance locus correlated with the reported degree of linkage with the resistance phenotype. Cluster analysis of the landraces first shown to have this type of resistance revealed that they are very closely related, if not identical. These findings suggest that there is a single source of monogenic resistance in the crop's genepool tracing back to a common ancestral clone. In the absence of further resistance diversification, the long-term effectiveness of the single gene resistance is known to be precarious, given the potential to be overcome by CMGs due to their fast-paced evolutionary rate. However, combining the quantitative with the qualitative type of resistance may ensure that this resistance gene continues to offer protection to cassava, a crop that is depended upon by millions of people in Africa against the devastating onslaught of CMGs.

1. Introduction

1.1. Cassava mosaic disease

Cassava (*Manihot esculenta* Crantz, family Euphorbiaceae) is a starchy root crop that supplies carbohydrate energy to millions of people in the tropics (Ceballos et al., 2004) and it is being used increasingly as an industrial crop (Jansson et al., 2009). Though its remarkable ability to tolerate unfavourable conditions such as

* Corresponding author at: IITA, Headquarters & West Africa Hub, PMB 5320, Oyo Road, Ibadan 200001, Oyo State, Nigeria. Tel.: +234 2 7517472x2719; fax: +44 208 7113785.

drought and poor soils makes it a food security crop in many parts of sub-Saharan Africa (SSA), on-farm productivity of cassava has remained stagnant for many years due to several production constraints. Cassava mosaic disease (CMD), caused by several species of cassava mosaic geminiviruses (CMGs), is the most economically important constraint to cassava in SSA and the Indian sub-continent (Herrera-Campo et al., 2011). Though significant efforts have been expended on combating this disease, it still causes huge losses to production. The most striking example of the devastating potential of CMD to undermine food security in Africa is the severe pandemic that started as an epidemic in Uganda in the 1990s and led farmers to abandon the crop in many parts of the country (Otim-Nape and Thresh, 1998), and later spread to most countries in East and Central Africa (Legg and Fauquet, 2004). The pandemic is characterized by rapid spread through super-abundant *Bemicia tabaci* vectors (Legg and Ogwal, 1998) and is associated with a recombinant strain of the *East African cassava mosaic virus – Uganda* (EACMV-UG) along with *African cassava mosaic virus* (ACMV) belonging to the genus *Begomovirus*, within the *Geminiviridiae* family (Harrison et al., 1997).

Important control measures against CMD include rogueing of symptomatic plants, use of virus-free planting materials and deployment of resistant varieties. The first two options are not only labour intensive and difficult to implement but also require continuous and long-term intervention. Use of resistant varieties is the most effective solution in mitigating the negative effect of CMD in farmers' fields because this approach not only reduces yield losses due to the disease, but also reduces levels of the virus inoculum in the farming system particularly in varieties that suppress virus accumulation.

1.2. Sources of resistance to the disease

Currently deployed resistance against CMD in Africa is of two types: (i) quantitative resistance derived from *Manihot glaziovii*; and (ii) qualitative resistance conferred by a single resistance gene(s). The quantitative resistance was introgressed into cultivated cassava following an unsuccessful worldwide search for resistant clones in the 1930s (Nichols, 1947; Jennings, 1976). Genetic studies reveal that the polygenic resistance from *M. glaziovii* is recessive with a heritability of about 60% (Jennings, 1976). The second type of resistance, which is conditioned by a single-gene with a dominant effect, was discovered in the 1980s in landraces from Nigeria and other West African countries (Akano et al., 2002; Fregene et al., 2001). These landraces, which display near-immunity against nearly all species of CMGs, are currently maintained in the IITA germplasm collection referred to as the Tropical *Manihot esculenta* (TMe) series. Diversity studies using molecular markers have previously shown that most of the original landraces bearing this qualitative resistance to CMD are genetically very similar if not identical (Fregene et al., 2000; Lokko et al., 2006). This suggests that the genetic base of this type of resistance in the African cassava genepool may be narrow, or even just a single locus. In contrast, the relative ease with which the highly heritable monogenic resistance can be transferred between germplasm through simple crosses, has resulted in its extensive usage in breeding across Africa as well as pre-emptive breeding in Latin America (Okogbenin et al., 2007). The long-term stability of this single-gene type of resistance in diverse geographical regions with heterogeneous species and recombinants of CMGs is uncertain given the high evolutionary rate of geminiviruses (Duffy and Holmes, 2009).

Several genetic mapping studies have been conducted to find molecular markers linked to the qualitative resistance in the African germplasm. The first study identified two markers, a microsatellite (SSRY28) and an RFLP (GY1) that flank a single locus named *CMD2* at distances of 9 and 8 cM, respectively (Akano et al., 2002). Subsequent to the discovery of the *CMD2* locus, later studies have

reported several additional markers that are linked to new resistance genes in other genetic backgrounds, including landraces and improved varieties derived from them (Lokko et al., 2005; Okogbenin et al., 2012). However, nearly all of these studies relied on the bulk segregant analysis (BSA) approach and/or very sparse maps for interval mapping analysis. The BSA approach provides little or no information regarding the chromosomal location of the identified markers, making it difficult to ascertain the number of unique loci/genes associated with a trait. When sparse maps are used, the confidence interval surrounding a QTL is usually large, making it difficult to determine the precise QTL location. For example, the *CMD2*-containing linkage group of Akano et al. (2002) had a total of five markers. Lokko et al. (2005) used a linkage map with just 45 markers, of which only three were in the linkage group containing the resistance locus.

The objective of this study was to provide a comprehensive framework for describing the breadth of the genetic base of the single-gene resistance to CMD in the African cassava germplasm. Firstly, a full-sib mapping population segregating for qualitative resistance to CMD was developed and phenotyped for three growing seasons. The population was genotyped using genotype-by-sequencing (GBS), and a dense genetic linkage map with more than 8000 single nucleotide polymorphism (SNPs) was constructed. Using this resource, a high-resolution genetic mapping of the CMD resistance locus was carried out. The markers previously reported to be linked to CMD resistance were then projected onto the newly generated genetic map. This revealed their genomic locations, and the spatial relationship between them and the mapped resistance locus from the present study. To confirm the relationship among the CMD resistant landraces, cluster analysis was carried out using genome-wide SNP markers.

2. Materials and methods

2.1. Mapping population development, phenotyping and genotyping

A full-sib mapping population segregating for dominant monogenic resistance to CMD was generated by crossing two non-inbred clones. Both parents are elite lines developed by IITA in Nigeria. The female parent, IITA-TMS-011412, is highly resistant to CMD and also rich in pro-vitamin A. Cloned in 1974, the male parent, IITA-TMS-4(2)1425, is an improved variety from a cross between a landrace from Nigeria (TME109, locally known as Oyarugbafunfun) and variety 58308, a hybrid derived directly from recombination of the *M. glaziovii* × *M. esculenta* triple-backcrosses (Jennings, 1976). Variety 58,308 was the main source of the quantitative resistance to CMD and produced some of the first generation Tropical Manihot Selection lines (see the discussion section for more background). IITA-TMS-4(2)1425 shows considerable susceptibility to CMD (Fig. 1).

The 180 F1 seeds produced were germinated in sterilized garden soil and transplanted one month after sowing. At maturity, the seedlings were cloned, regardless whether they were infected or not, and planted at Ibadan, Nigeria (7.40° North latitude, 3.90° East longitude) using a randomized complete block design. Each clone was planted in two replicated plots of five stands per plot with plant spacing of 1 m × 0.5 m for three 12-month long cropping seasons established in 2011, 2012 and 2013. Generation-to-generation propagation through cloning was based on use of 12-cm long stem cuttings in both the infected and non-infected F1s. A local landrace that is highly susceptible to CMD, TME117, was planted as spreader row every fifth plot and as border row surrounding the experimental field to facilitate whitefly-mediated inoculation of the F1 population. The Ibadan site is known for high CMD pressure and

Fig. 1. CMD symptom on the mapping population parents, (a) IITA-TMS-011412 and (b) IITA-TMS-4(2)1425; (c) The frequency distribution of CMD scores using the BLUPs calculated from the three-year data.

the planting period coincide with high whitefly activity providing high probability of natural exposure of plant population to CMD inoculum. Individual plants were evaluated for CMD symptoms at one and three months after planting using a scale ranging from 1 to 5, with one for symptomless plants while five is given for most severe symptoms (severe mosaic and distortion of leaves).

The entire population was screened for presence of CMGs, particularly for the presence of ACMV and EACMCV, the two predominant species prevailing in West Africa, to confirm virus infection in the infected plants. The third fully expanded leaf from the top was sampled from each of the 5 plants per plot; they were pooled and DNA was extracted and analyzed for ACMV and EACMCV using a multiplex PCR protocol (Alabi et al., 2008).

2.2. Genotyping-by-sequencing

DNA was extracted from 180 F1 individuals and the two parents using a modified Dellaporta method (Dellaporta et al., 1983). Genotyping-by-sequencing as described by Elshire et al. (2011) was carried out at the Institute of Genomic Diversity, Cornell University. Briefly, DNAs from the F1 individuals and parents were digested individually with ApeKI restriction enzyme, which recognizes a five base-pair sequence (GCWGC, where W is either A or T). This enzyme was chosen because of its partial sensitivity to DNA methylation, thus avoiding repetitive elements regions, and frequency of DNA-cutting (Elshire et al., 2011). Two 95-plex GBS sequencing libraries were prepared by ligating the digested DNA to unique nucleotide adapters (barcodes) followed by standard PCR. Sequencing was performed using Illumina HiSeq2000. The sequencing reads from different genotypes were de-convoluted using the barcodes and aligned to the version 4.1 of the cassava reference genome (www.phytozome.org/cassava) by using Burrow Wheelers Alignment tool (Li and Durbin, 2009). SNPs were extracted using the GBS pipeline implemented in TASSEL software (Bradbury et al., 2007), and genotypes were called using a custom R script.

2.3. Data analysis

The pseudo-testcross linkage mapping strategy that is employed in the analysis of full-sib mapping populations requires unambiguous scoring of the parental genotypes at each marker. To ensure this, the parental DNAs were sequenced redundantly four times and their Illumina reads were pooled to increase the number and accuracy of the called SNPs. Following alignment of the reads against the reference genome, the SNPs that segregated in the parents as $ab \times ab$ (both parents heterozygous), $aa \times ab$ (male parent heterozygous), and $ab \times aa$ (female parent heterozygous) were extracted. Prior to linkage analysis, standard quality control was used to filter

out SNPs from paralogous sequences (i.e. loci which appear as heterozygous in both parents and all progenies). Also filtered were loci showing significant deviation from expected genotypic frequencies based on chi-square test (threshold for removal: $P \leq 0.05$) as well as those with missing information in more than 20% of the genotyped individuals in the mapping population.

2.3.1. Mapping of GBS-derived ApeKI SNPs

Genetic linkage maps were constructed using JoinMap version 4.1 (Van Ooijen, 2006). Following calculation of pair-wise recombination frequencies, linkage groups were identified using the logarithm of odds (LOD) score of independence between pairs of loci at a threshold of 10. Due to the large number of markers per linkage group, the maximum-likelihood mapping algorithm implemented in Joinmap 4.1 was used to find the order of the markers in the linkage groups. This method is suitable for dealing with large datasets compared to the regression mapping method (Cheema and Dicks, 2009) and incorporates several numerical methods: simulated annealing for estimating the best map order by minimizing the sum of recombination frequencies in adjacent segments; Gibbs sampling for estimation of multipoint recombination frequency, given the current map order; and spatial sampling of loci to reduce the influence of unknown or dominant genotypes as well as potential errors. Simulated annealing was carried out using a chain length of 30,000 with an acceptance probability threshold of 0.25. Gibbs sampling for maximum likelihood estimation of multipoint recombination frequencies (Jansen et al., 2001) was done using chain length of 50,000 after a burn-in length of 20,000.

2.3.2. Phenotypic data analysis

Because the categorical disease severity scores fitted a bi-modal distribution, for statistical analyses (ANOVA and QTL mapping) the trait was converted to a binary variable (either resistant or susceptible). Individuals with categorical CMD severity score larger than one were classified as Affected; all others were classified as Unaffected. A logistic regression model using generalized linear model was used to estimate the effect of the genotype, replication, environment and genotype-by-environment interaction as follows:

$$y_{ijkl} = \mu + \beta_i + R_{ij} + G_k + \beta_i^* G_k + e_{ijkl}$$

where y_{ijkl} was the phenotype; μ the mean, β_i the year effect; R_{ij} the replication effect; G_k the clone effect; $\beta_i^* G_k$ is the interaction between clone and year and e_{ijkl} is the residual. Mixed model was used to obtain best linear unbiased predictors (BLUPs) for each genotype for the combined three-year data. The mixed model was computed using the R package *lme4* (Vazquez et al., 2010), considering the effects of the genotypes as random, while replications within environments were regarded as fixed because trials were

carried out for three years. Broad-sense heritability for CMD resistance at one and three months after planting was calculated using the formula

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

where σ_g^2 and σ_e^2 are the variance components for the genotype effect and the residual error, respectively, based on individual plants. Correlations were calculated among disease resistance score BLUPs for three growing seasons (2011, 2012 and 2013).

2.3.3. High-resolution mapping of the CMD resistance locus

Mapping of the CMD resistance locus in the present population was carried out with the BLUPs obtained for each year, and across the combined analysis of the 2011, 2012 and 2013 data. QTL analysis was performed using three complementary approaches. Because of high marker density, single marker-trait association for all 6756 SNPs was carried out. The markers were considered as fixed effects in a linear model implemented in the GLM function TASSEL (Bradbury et al., 2007). The genome-wide significance threshold for the F-statistic was determined by the Bonferroni method (Bland and Altman, 1995). Secondly, standard interval mapping (intervals step of 2.5 cM) was carried out using the regression mapping function "scanone" implemented in R/qtl package (Broman et al., 2003). The genome-wide significance ($\alpha = 1\%$) for declaring a significant QTL locus was determined using 1000 permutations. The 95% Bayesian credible interval for the CMD resistance locus was determined using the function "bayesint" implemented in R/QTL. The proportion of phenotypic variance explained by the resistance locus was obtained by fitting a linear model. Thirdly, QTL analysis using the Composite Interval Mapping (CIM) method was carried out with the number of marker covariates set to three. The number of markers for use as co-factors was determined using the automatic co-factor selection function (*stepwiseqtl*) implemented in R/qtl. The CIM method enabled a reduction in residual variation and thereby increases the resolution of the QTL location and with it the possibility of detecting any additional genomic regions that underly resistance to CMD.Anchoring CMD-resistance-linked markers in the present high-density genetic map

The high-density SNP map developed in this study was used to anchor published loci associated with qualitative resistance to CMD (Table 1). Primer sequences that flank five of these markers (*viz.* SSRY28, NS198, SSRY106, NS158 and NS169) were used in BLAST searches of the cassava reference genome sequence (www.phytozome.org/cassava). Marker positions were interpolated onto the genetic map on the basis of the scaffolds harbouring them (Table 1).

To obtain a linkage/recombination profile of SNPs along the linkage group that bears the CMD resistance locus, pairwise estimates of linkage disequilibrium (r^2; Flint-Garcia et al., 2003) were calculated for the SNPs from the entire mapping population using the software package Haploview v. 3.31 (Barrett et al., 2005) and plotted in a matrix form.

2.3.4. Genetic relatedness of the CMD-resistant landraces

In addition to the mapping analysis in the bi-parental population, the genetic relatedness was examined among the TME clones that were originally identified to be sources of the resistance to CMGs. A total of 2069 GBS markers from 34 clones, including 29 landraces and five TMS clones with the quantitative resistance (as an out-group), were obtained from a previous study (Ly et al., 2013), and used to perform hierarchical clustering using the Euclidean distance between the genotypes. These distances were used to construct a relationship dendrogram of the clones.

3. Results

3.1. Segregation for resistance to CMD in the mapping population

The frequency distribution of the CMD severity scores in the mapping population revealed a bi-modal pattern with two peaks (Fig. 1): nearly half of the progenies and the female parent (IITA-TMS-011412) were resistant to CMD and showed no symptoms while the remainder of the F1 individuals showed disease symptoms ranging from mild (score 2) to severe (score 5). Resistance to CMD found in the female parent is therefore likely to be a single gene with dominant effect. There was very little variation within plots with respect to CMD symptom expression: all stands either showed similar symptoms in infected susceptible plots or no symptoms at all in the resistant plots. The consistency in symptom expression is largely due to the clonal origin from infected cuttings which ensures transmission of viruses across cropping cycles. Moreover, there was very little year-to-year variation in terms of CMD incidences: The disease ratings in the 2011, 2012 and 2013 growing seasons were highly correlated (Pearson's correlation coefficient, $r > 0.91$). This was reflected in the large broad-sense heritability of the resistance trait as measured at one and three months after planting (H^2 of 0.89 and 0.93, respectively). Analysis of variance using the logistic regression model showed a highly significant effect for clone ($p = <2E-16$), while the other factors such as environment (year), clone $\times$ environment interaction and replication were not significant.

Table 1
Summary of the known markers tagging CMD resistance in cassava and their linkage group.

Marker	Primer sequence	Linkage group	Scaffold (v4.1)	Study	Resistance source
S5214_780931	GBS-SNP	16	scaffold05214	Present study	IITA-TMS-011412
S5214_30911	GBS-SNP	16	Scaffold05214	Rabbi et al. (in press)	IITA-TMS-961089A
SSRY28 (*CMD2*)	Fw:TTGACATGAGTGATATTTTCTTGAG Rev:GCTGCGTGCAAAACTAAAAT	16	scaffold05214	Akano et al. (2002), Lokko et al. (2005), Okogbenin et al. (2012)	TME3; TME7; IITA-TMS-972205
SSR NS158	Fw:GTGCGAAATGGAAATCAATG Rev:TGAAATAGTGATACATGCAAAAGGA	16	scaffold06906	Okogbenin et al. (2007)	TME3
SSR NS169	Fw:GTGCGAAATGGAAATCAATG Rev:GCCTTCTCAGCATATGGAGC	16	scaffold06906	Okogbenin et al. (2007)	TME3
RFLP RME-1	Fw:ATGTTAATGTAATGAAAGAGC Rev:AGAAGAGGGTAGGAGTTATGT	16	No match	Okogbenin et al. (2007)	TME3
SSR NS198	Fw:TGCAGCATATCAGGCATTTC Rev:TGGAAGCATGCATCAAATGT	16	Scaffold04175	Okogbenin et al. (2012)	IITA-TMS-972205
SSRY106	Fw:GGAAACTGCTTGCACAAAGA Rev:CAGCAAGACCATCACCAGTTT	16	scaffold07933	Lokko et al. (2005)	TME7

Table 2
Screening results of the mapping population for the presence of ACMV and EACMV.

| Disease status	Virus not detected	ACMV only	ACMV + EACMV
Symptomatic	1	65	15
Asymptomatic	92	6	1

3.2. Screening of the population for cassava mosaic geminiviruses using PCR

The PCR-based screening of the mapping population detected one or more of the CMGs in 87 of 180 plots assayed, while 93 plots were negative (Table 2). Only two species of CMGs, ACMV and EACMCV, were detected in the trial, which is consistent with known CMGs prevalence in West Africa. ACMV was detected in all the 87 virus-positive plots, whereas EACMCV was detected as a co-infection with ACMV in 16 plots (17% incidence). No case of single infection by EACMCV was detected. CMGs were detected in 79 of 81 symptomatic plots indicating a strong positive correlation between the visual scoring of disease and the PCR results. In addition, PCR also detected occurrence of ACMV in 7 asymptomatic plots and ACMV and EACMCV in one plot. Mean severity of CMD symptoms in ACMV infected plants was 3.1 and plants infected with ACMV and EACMCV was 3.2, which suggest apparent lack of synergistic effect in heightening symptom severity in dually infected plants.

3.3. Genotyping of SNP markers and construction of a dense genetic map

In all, 17,682 SNPs were obtained from the SNP calling pipeline. The SNP data were subsequently filtered for markers with more than 20% missing values across the genotyped individuals. Also removed were loci that deviated from the expected genotypic frequencies at Chi-square significance threshold of $P < 0.05$. Linkage analysis was done using 8704 SNPs that passed these two QC filters.

A high-density genetic linkage map was constructed using the Maximum-Likelihood approach implemented in Joinmap 4.1 (Table 3). A total of 6756 SNP markers were mapped across 19 linkage groups with between 115 and 559 SNPs (average = 256). With an average inter-SNP distance of 0.52 cM, this is the densest map developed for cassava so far (Fig. 2). Despite the high density of the GBS-derived SNPs mapped, several regions without markers were observed. Most notable were a single region on linkage
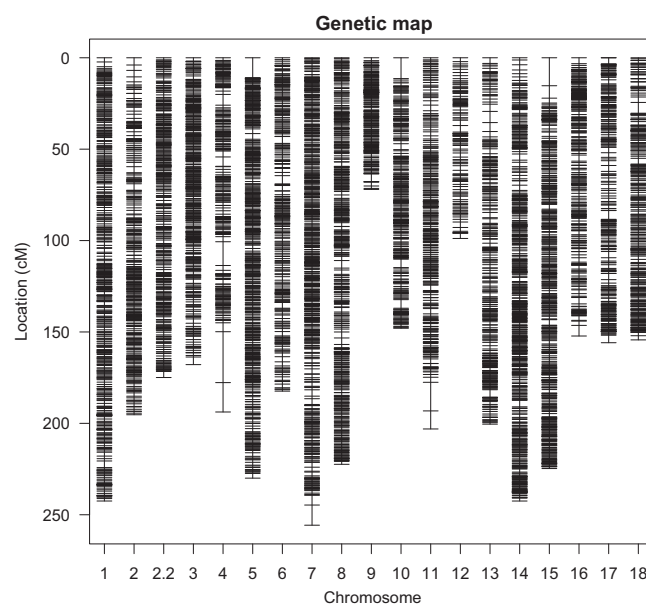
Fig. 2. Overview of genetic map developed from the 6756 *ApeKI*-derived SNPs.

group 4 and two regions in linkage group 18. A possible reason for such gaps could be lack of polymorphic markers as a result of identity-by-descent of this region in the two parents.

Most of the 12,977 scaffolds that constitute the current 533 Mb of the version 4.1 cassava genome assembly are fairly small; nearly 95% of them are 200 Kb or smaller. A total of 1093 unique scaffolds were anchored in the present map, and ranged from 19 to 89 in the different linkage groups. Despite their relatively small number, the anchored scaffolds covered a total size of 313.3 Mb, and accounted for 58.7% of the current cassava genome assembly. The complete genetic map developed from this work is available in Supplementary Table 1.

3.4. High-resolution mapping of CMD resistance locus

Based on the qualitative nature of the resistance to CMD in the present mapping population, a single locus was expected to underlie the resistance phenotype. The high-density genetic map developed with 6756 GBS SNPs permitted a genome-wide search

Table 3
Summary statistics of the genetic linkage map developed from ApekI SNP markers.

| Linkage group	Number of SNPs	Length (cM)	Average distances (cM)	Number of scaffolds anchored	Cumulative scaffold size (base-pairs)
1	473	242	0.51	66	30,226,735
2	344	195	0.57	52	20,822,503
2.2	366	175	0.48	59	20,298,754
3	419	168	0.40	55	12,608,655
4	255	194	0.76	57	13,843,915
5	543	230	0.42	61	15,175,445
6	275	182	0.67	73	15,911,788
7	559	256	0.46	66	21,949,504
8	454	222	0.49	78	18,054,518
9	207	72	0.35	19	6,148,875
10	299	148	0.50	55	15,132,961
11	304	203	0.67	41	13,797,633
12	115	99	0.87	21	6,111,492
13	302	200	0.67	60	17,048,413
14	543	242	0.45	89	19,282,711
15	451	225	0.50	49	18,831,996
16	281	152	0.54	64	16,048,872
17	275	156	0.57	70	17,372,231
18	291	154	0.53	58	14,665,500
Total	6756	256[a]	0.52[a]	1093	313,332,501

[a] Average values per linkage group.
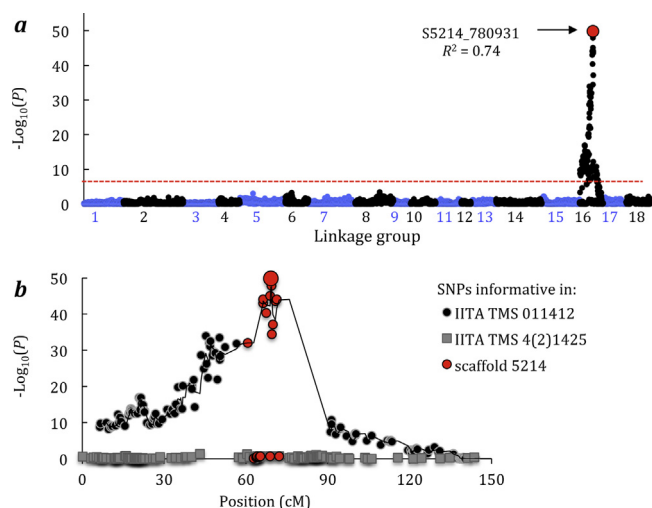
Fig. 3. Single-marker association with qualitative resistance to CMD. (a) Genome-wide P-values for 6756 SNPs across 19 linkage groups showing the strongest association signal was located in linkage group 16. The x-axis shows the SNPs along each chromosome; y-axis is the $-\log 10$ (P-value) for the association. (b) An association plot for linkage group 16 showing SNPs that are informative in resistant (IITA-TMS-011412) and susceptible (IITA-TMS-4(2)1425) parents. The SNPs from scaffold 5214 with strongest association to resistance phenotype are highlighted.

for the locus underlying the qualitative resistance to CMD. A strong association was detected in linkage group 16 that peaked at around 69.12 cM (Fig. 3). This association decreases on both sides of the peak as a result of increasing recombination between the markers and the underlying resistance gene. Most SNPs showing highly significant association ($P < 1E-40$) came from the 1.46 Mb-long scaffold 5214, with marker S5214_780931 at the peak; this marker explained 74% of the disease resistance variance. Additionally, only those SNPs that are informative in the CMD-resistant female parent show the significant associations while those segregating only in the susceptible male parent do not (Fig. 3). The presently mapped resistance locus occurs in the vicinity of the previously mapped *CMD2* locus (Akano et al., 2002); marker S5214_780931 is just 623.24 kb away from a microsatellite marker, SSRY28 (between 157,470 bp to 157,616 bp), that was first reported to be closely linked to the *CMD2* locus (Akano et al., 2002), indicating that the same gene may account for the observed resistance to CMD in the present mapping population.

The results from the interval-mapping based QTL analysis recapitulated those from single-marker trait associations, and uncovered a single peak with a maximum LOD value of 43 on linkage group 16 (Supplementary Fig. 1). Despite employing the Composite Interval Mapping method, no additional peaks exceeding the significance threshold were detected, confirming that only a single locus conferred the qualitative resistance in the present population. The SNP marker S5214_780931 is flanked by S5214_472282 and S5214_1084049 was the closest to the QTL peak

(Table 4). The approximate 95% Bayesian credible interval for the mapped locus spans from 68.8–70.74 cM region along LG16, irrespective of the scoring time (one or three months after planting) or growing season (2011, 2012 or 2013). The profiles of LOD scores along the linkage group 16 were very similar for the disease scores recorded at one and three months after planting as well as the three seasons of data, supporting the high heritability observed for this trait. Comparison of the phenotypes of the F1s against the resistance-linked marker S5214_780931 genotypes showed a small proportion of recombinants that carry the SNP allele linked to resistance but show susceptibility to the disease and vice versa.

3.5. Genomic localization of markers flanking previously mapped qualitative resistance against CMD

A major objective of the present study was to use the high-density SNP genetic map to anchor seven molecular markers previously reported to be linked to single dominant gene resistance to CMD in four other genetic backgrounds (Table 1). The SSR and RFLP makers used in those studies were therefore anchored in the present map via their harbouring scaffolds. These markers came from scaffold 05214 (SSRY28), scaffold 06906 (NS158 and NS169), scaffold 04175 (NS198) and scaffold 07933 (SSRY106) all of which fall within the same genomic region of linkage group 16 of the present map (Fig. 4). The primer sequences for the RFLP marker RME-1 did not identify a suitable match in the reference genome. In a parallel study using another bi-parental mapping population derived from a cross between another improved variety that is nearly immune to CMD (IITA-TMS-961089A) and a susceptible landrace (TME117), another SNP marker was identified from the same scaffold (S5214_30911). It was strongly associated with the disease resistance and explained 60% of phenotypic variation (Rabbi et al., in press). The reported percentage of variation explained by these markers shows a gradient that peaks around scaffold 05214, the region that is likely to contain the concerned resistance gene (Fig. 4). Markers away from this region have been reported to be less linked to the resistance gene by the low percentage of variation that they explain, a trend that agrees with the GWA results, particularly considering the segregating markers from the female parent (Fig. 3).

3.6. Recombination pattern in linkage group 16

To visualize the degree of linkage and recombination pattern between the presently mapped CMD resistance locus and other previously published SSR markers along the linkage group, the chromosome-wide pattern of LD on linkage group 16 was examined (Fig. 4). The resulting haplotype map is useful in providing a framework for interpreting the results of the previous mapping studies, particularly the proportion of phenotypic variation explained by the various microsatellite markers and their relative locations in the present map. Though different parents are used in the present and previous mapping studies, these populations have all
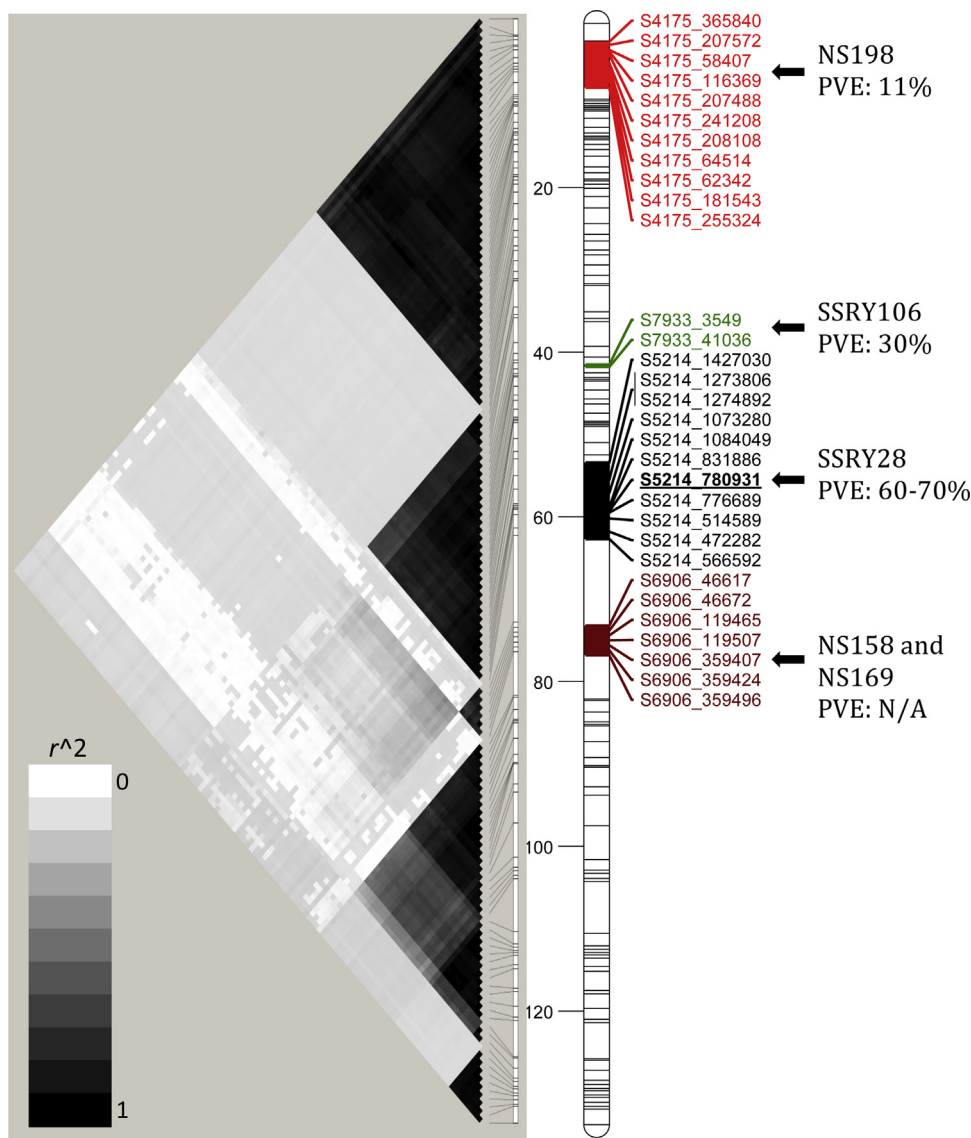
Table 4
Markers flanking the presently mapped CMD resistance locus calculated from disease severity scored at one- and three-months after planting. The table also presents the linkage group, position and interval mapping-based percentage of phenotype variation explained by the closest marker.

Trait	SNP	Linkage group	Position (cM)	Peak LOD[a]	R^2	H^2
CMD1S[b]	S5214_1084049	16	68.88	45.37		
	S5214_780931	16	70.00	45.59	0.708	0.892
	S5214_472282	16	70.74	44.99		
CMD3S[b]	S5214_1084049	16	68.88	42.98		
	S5214_780931	16	69.31	43.20	0.696	0.928
	S5214_472282	16	70.74	42.33		

[a] Logarithm of odds score for presence of QTL.
[b] Values calculated using the three-year BLUPs; R^2 = percentage variation explained by QTL; H^2 = Heritability using the three year-data.

Fig. 4. Graphical display of the variation in the linkage disequilibrium (r^2) along linkage group 16 calculated for every pair of SNPs in the bi-parental mapping population (left); and the genetic map (right). The location of the mapped CMD resistance locus (underlined SNP, *viz.* S5214_780931) in scaffold 05214 relative to SSRs other scaffolds (S7933, S4175 and S6906) containing microsatellite markers reported to be linked to resistance to CMD is shown on the right. The percentage of phenotypic variation explained by the (PVE, measured by r^2) are also presented. The dark shading corresponds to stronger LD (higher r^2). Names of other SNPs in the linkage group were omitted due to space constraints; but are available in the Supplementary (Supplementary Table 1).

undergone a single round of meiosis and are thus expected to have a similar extent of linkage disequilibrium across their chromosomes.

The region bearing the CMD resistance locus was characterized by two large haplotype blocks (dark-grey shading between 0 to 32 cM and 36 to 62 cM, respectively). Moderate LD was detected between these blocks (light-grey shading). The first block encompasses scaffold 4175, which harbours microsatellite marker NS198, reported by Okogbenin et al. (2012) to explain 11% of variation in CMD resistance in a bi-parental population. SNPs from this scaffold and those near the *CMD2* locus in scaffold 05214 also show moderate LD ($r^2 \sim 0.10$).

The CMD resistance locus occurs in the second LD block. Scaffold 5214 harbours the SNPs that were strongly associated to the resistance as well as the microsatellite marker SSRY28 reported previously (Akano et al., 2002). Though discovered from different genetic backgrounds, the resistance-linked SNPs and SSRY28 explain between 60% and 70% of the disease resistance variance. These markers are just 623.24 kb apart. Another scaffold in this block (07933) which harbours microsatellite marker SSRY106, was

reported to explain nearly 40% of variation in CMD resistance (Lokko et al., 2005).

3.7. Genetic relatedness of CMD-resistant landraces using genome-wide SNPs

In addition to genetic mapping of the bi-parental population, cluster analysis was performed with key landraces known to possess strong resistance to CMD (Fig. 5). To estimate the "residual genetic distance" between identical clones resulting from genotyping error, several clones – some of which have different names as a result of adoption in different regions – were redundantly genotyped. These pairs include the male parent (IITA-TMS-4(2)1425) and "Kibandameno-white"; IITA-TMS-30572 known as "Migyera" in Uganda; and TME12 (A and B). Most of the CMD resistant landraces, including those that were first discovered in Nigeria, are clearly very closely related, or even perhaps identical, based on comparison of the distance between them and the residual distance between the redundantly genotyped clones, confirming
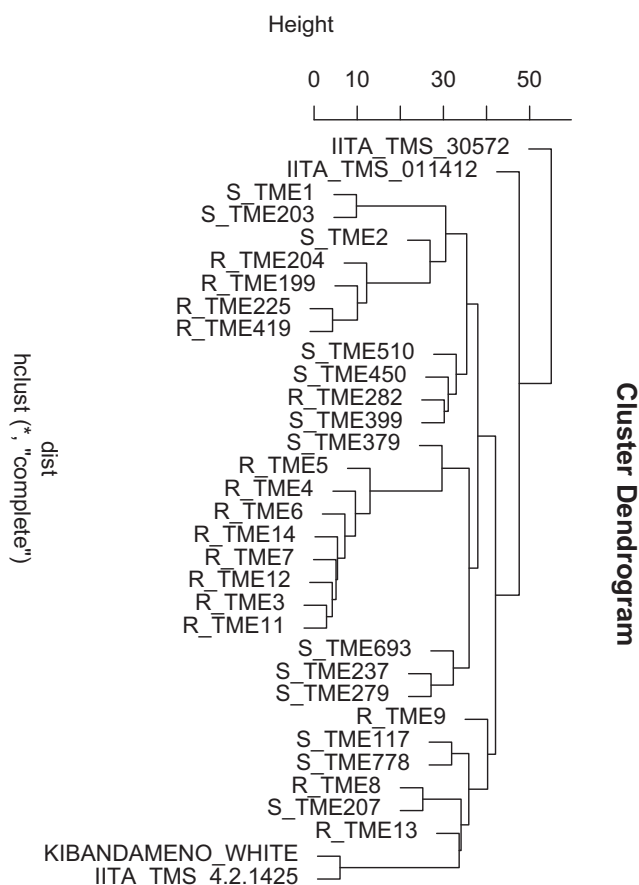
Height

Fig. 5. A hierarchical clustering tree based on dissimilarities between a selection of landraces using of 2069 SNP markers. Most of the lower TME series landraces that are resistant to CMD form a single, tight cluster (bottom). The prefix "R" denotes CMD-resistant and "S" denotes CMD-susceptible varieties.

previous studies using AFLPs (Fregene et al., 2000). These landraces (TME3, TME4, TME6, TME7, TME11, TME12 and TME14) have a very similar morphological appearance, most prominent of which is red petioles.

4. Discussion

4.1. A historical perspective of development and diffusion of early CMD resistant varieties across of Africa

Breeding for resistance to CMD has been a major goal of cassava improvement programmes in Africa for more than 70 years. Prior to the discovery of the single-gene resistance (Akano et al., 2002), the primary defence against the disease was the polygenic resistance introgressed into cultivated cassava (*M. esculenta*) from *M. glaziovii* after three cycles of backcrossing (Nichols, 1947). Examining the breeding history of improved varieties with the quantitative resistance to CMD reveals a very narrow genetic base tracing back to a single derivative of the *M. glaziovii* × *M. esculenta* crosses (Jennings, 2003). None of these descendants is immune to infection by CMGs (Nichols, 1947) although some express mild and sometimes transient symptoms as a result of incomplete systemic infection that leads to reversion of symptoms (Jennings, 1976; Fargette et al., 1994), while others are quite susceptible to the disease. These TMS varieties that trace back to *M. glaziovii* have been widely adopted in Africa and helped revive cassava farming following the devastation of many local landraces in East Africa (Legg and Fauquet 2004). Some of the adopted TMS varieties are IITA-TMS-60142,

IITA-TMS-30337 and IITA-TMS-30572. Since the 1990s, IITA has been making crosses to combine the single dominant gene (*CMD2*) with the multigenic *M. glaziovii* resistance and have produced clones with near immunity to CMD (Legg and Fauquet, 2004).

4.2. Mapping resolution of the CMD2 locus

Genetic mapping of qualitative resistance to CMD in this study uncovered a single locus on linkage group 16 with a large peak LOD (>40). This locus, S5214_780931, explained 74% of the phenotypic variance, and is co-located with the SSRY28 on scaffold5214, indicating the dominant monogenic resistance gene from the female parent is likely to be the *CMD2* locus of Akano et al. (2002). The dense genetic map obtained from GBS has enabled a higher level of mapping resolution of the CMD resistance locus. Linkage group 16 has a total of 281 SNPs, and the approximate 95% Bayesian credible interval around the mapped locus is just 1.1 cM. This resolution was not obtainable using the traditional marker from previous mapping efforts.

4.3. All markers linked to qualitative resistance occur in the same chromosome region

The high-density SNP genetic map was used to anchor markers previously reported to be linked to the dominant monogenic resistance to CMD. These markers (Table 1) were interpolated in the present map via the scaffolds that harbour them. All of them (except RME1 and RME4 for which a suitable matches in the current reference genome were not found) came from scaffolds occurring in the same region of linkage group 16. Overall, there is a general congruence between the proportion of genetic variation reported for these microsatellite markers, the marker-trait association profile in linkage group 16 (Fig. 3) and the pattern of linkage disequilibrium (measure as r^2) between microsatellite locations and the putative position of the resistance locus in the present population (Fig. 4). This pattern supports the idea that most of the microsatellite markers reported to be linked to varying degrees with CMD resistance are associated with a single resistance locus that is most likely the *CMD2* gene. Moreover, the strength of the linkage decreased with increasing distance from the gene location. In a parallel mapping study using another improved variety that is nearly immune to CMD (IITA-TMS-961089A) and a susceptible landrace (TME117), a strong QTL signal was found in scaffold05214 that also explained 60% of resistance variation, implicating the same *CMD2* locus (Rabbi et al., in press). Considering the results of the present study and those of Akano et al. (2002), Okogbenin et al. (2012), and Lokko et al. (2005), it is highly likely that a there is a single gene, or a cluster of resistance genes (Michelmore and Meyers, 1998).

Indeed, the conservation of QTL positions among the different sources of CMD resistance is not surprising given that the majority of the highly resistant cultivars developed recently in Africa trace to a few landrace accessions from Nigeria that were used over the last 12 or so years of resistance breeding in which the merits of the TME materials were appreciated. According to the phylogeny (Fig. 5) it was confirmed that several landraces that were first discovered to be nearly immune to CMD were genetically very similar. These findings agree with a previous study (Fregene et al., 2000). Other duplicates, which are also morphologically very similar, were identified. These landraces were collected from South West and Central Nigeria and have different local names. The resistance in the landrace TME9, which also occurs in a separate clade on the phylogenetic tree, is likely to be *CMD2*. This landraces is found in the pedigree of IITA-TMS-961089A that was used in a parallel mapping study; resistance-linked markers also came from scaffold 05214 (Rabbi et al., in press). It is postulated that all of these landraces, which come from West Africa, trace back to a single CMD resistant

mutant that was selected by farmers and rapidly diffused in the region through varietal dissemination efforts.

4.4. Screening of cassava mosaic geminiviruses in the mapping population

The CMG screening results showed that ACMV was the predominant species in the field, whereas EACMV occurred at lower frequency and only with ACMV as a dual infection. The preponderance of ACMV reflects the known proportions of the two viruses in West Africa and contrasts with the predominance of EACMV-UG in the pandemic regions of Eastern Africa (Legg and Fauquet, 2004). The severe form of EACMV-UG has displaced other CMG species and strains during the recent pandemic that swept through the region (Legg et al., 2006). An important question is whether the CMD2-type of resistance available in the West-African cassava germplasm can protect against these more virulent strains of CMG that does not contribute to the disease pressure in West Africa. Germplasm screening in Uganda have demonstrated that the qualitative resistance locus in the TME landraces is highly effective against this virus (Kawuki et al., 2011). Similar results have been obtained through biolistic inoculation methods using pseudo-recombinants of ACMV and EACMV (Sserubombwe et al., 2008). Some of the screened clones include TME5 and TME14, which cluster together with TME3, the original source of the *CMD2* locus (Fig. 5). The evidence suggests that *CMD2* locus is a monogenic resistance with broad specificity against cassava-infecting geminiviruses.

4.5. Prospects for long-term efficacy of CMD2 resistance gene

Despite the apparent broad specificity of the CMD2 gene, whether the locus can confer durable resistance according to the classical definition (Johnson, 1984) depends on whether it will remain effective over a prolonged period of widespread use under conditions conducive to the disease. It is well known that monogenic resistance can be ephemeral and subjected to the well known boom-and-bust cycles (McDonald and Linde, 2002) as a result of the rapid genetic evolution in the pathogen though there is evidence of durable resistance from single gene actions (Johnson, 1984; Stuthman et al., 2007). Since its discovery in the 1980s and subsequent extensive use in breeding programmes in sub-Saharan Africa, there has been no report of breakdown of the *CMD2*-type of resistance in the field, indicating that the locus could be durable against CMGs. Still, the long-term effectiveness of the resistance locus needs to be augmented by combining it with the quantitative resistance derived from *M. glaziovii*. Indeed, crosses combining these two types of resistance have given rise to progeny which are near immune to all known CMG species, including the virulent EACMV-Uganda (Legg and Fauquet, 2004; Monde et al., 2012). Additionally, more efforts are needed for a comprehensive screening of the resistant landraces to identify any additional sources that may exist. Though increased resolution was achieved in the mapping analysis, a larger mapping population is required for fine-mapping and cloning of the concerned locus. This will provide insights on the mechanism of the resistance that so far seems to be highly effective against various CMG species.

Acknowledgments

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2013.12.028.

References

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

Alabi, O.J., Kumar, P.L., Naidu, R.A., 2008. Multiplex PCR method for the detection of African cassava mosaic virus and East African cassava mosaic Cameroon virus in cassava. J. Virol. Met. 154, 111–120.

Barrett, J.C., Fry, B., Maller, J.D.M.J., Daly, M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–265.

Bland, J.M., Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. Br. Med. J. 310, 170.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633–2635.

Broman, K.W., Wu, H., Sen, Ś., Churchill, G.A., 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19, 889–890.

Ceballos, H., Iglesias, C.A., Pérez, J.C., Dixon, A.G., 2004. Cassava breeding: opportunities and challenges. Plant Mol. Biol. 56, 503–516.

Cheema, J., Dicks, J., 2009. Computational approaches and software tools for genetic linkage map estimation in plants. Brief. Bioinform. 10, 595–608.

Dellaporta, S.L., Wood, J., Hicks, J.B., 1983. A plant DNA minipreparation: version II. Plant Mol. Biol. Rep. 1, 19–21.

Duffy, S., Holmes, E.C., 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J. Gen. Virol. 90, 1539–1547.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379.

Fargette, D., Jeger, M., Fauquet, C., Fishpool, L.D.C., 1994. Analysis of temporal disease progress of African cassava mosaic virus. Phytopathology 84, 91–98.

Flint-Garcia, S.A., Thornsberry, J.M., Buckler, I.V.E.S., 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54, 357–374.

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

Fregene, M., Okogbenin, E., Mba, C., Angel, F., Suarez, M.C., Janneth, G., et al., 2001. Genome mapping in cassava improvement: challenges, achievements and opportunities. Euphytica 120, 159–165.

Harrison, B.D., Zhou, X., Otim-Nape, G.W., Liu, Y., Robinson, D.J., 1997. Role of a novel type of double infection in the geminivirus-induced epidemic of severe cassava mosaic in Uganda. Ann. Appl. Biol. 131, 437–448.

Herrera-Campo, B.V., Hyman, G., Belloti, A., 2011. Threats to cassava production: known and potential geographic distribution of four key biotic constraints. Food Sec. 3, 329–345.

Jansson, C., Westerbergh, A., Zhang, J., Hu, X., Sun, C., 2009. Cassava, a potential biofuel crop in (the) People's Republic of China. Appl. Energy 86, S95–S99.

Jennings, D.L.,1976. Breeding for Resistance to African Cassava Mosaic Disease: Progress and Prospects. In: Interdisiplinary Workshop. IDRC, Muguga (Kenya).

Jennings, D.L., 2003. Historical perspective on breeding for resistance to cassava Brown Streak Virus Disease. In: Hillocks, R.J. (Ed.), Cassava Brown Streak Virus Disease: Past, Present, and Future. Proceedings of an International Workshop. Mombasa, Kenya, 27–30 October 2002. Natural Resources International Limited, Aylesford, UK, p. 100.

Jansen, J., de Jong, A.G., van Ooijen, J.W., 2001. Constructing dense genetic linkage maps. Theor. Appl. Genet. 102, 1113–1122.

Johnson, R., 1984. A critical analysis of durable resistance. Annu. Rev. Phytopath. 22, 309–330.

Kawuki, R.S., Pariyo, A., Amuge, T., Nuwamanya, E., Ssemakula, G., Tumwesigye, S., Bua, A., Baguma, Y., Omongo, C., Alicai, T., Orone, J., 2011. A breeding scheme for local adoption of cassava (*Manihot esculenta* Crantz). J. Plant Breed. Crop Sci. 3, 120–130.

Legg, J.P., Fauquet, C.M., 2004. Cassava mosaic geminiviruses in Africa. Plant Mol. Biol. 56, 585–599.

Legg, J.P., Ogwal, S., 1998. Changes in the incidence of African cassava mosaic geminivirus and the abundance of its whitefly vector along south–north transects in Uganda. J. Appl. Entomol. 122, 169–178.

Legg, J.P., Owor, B., Sseruwagi, P., Ndunguru, J., 2006. Cassava mosaic virus disease in East and Central Africa: epidemiology and management of a regional pandemic. Adv. Virus Res. 67, 355–418.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

Lokko, Y., Danquah, E.Y., Offei, S.K., Dixon, A.G.O., Gedil, M.A., 2005. Molecular markers associated with a new source of resistance to the cassava mosaic disease. Afr. J. Biotechnol. 4, 873–881.

Lokko, Y., Dixon, A., Offei, S., Danquah, E., Fregene, M., 2006. Assessment of genetic diversity among African cassava *Manihot esculenta* Grantz accessions resistant to the cassava mosaic virus disease using SSR markers. Genet. Resour. Crop Evol. 53, 1441–1453.

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H.G., et al., 2013. Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: a study in Cassava. Crop Sci. 53, 1312–1325.

McDonald, B.A., Linde, C.C., 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annu. Rev. Phytopathol. 40, 349–379.

Michelmore, R.W., Meyers, B.C., 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 8, 1113–1130.

Monde, G., Walangululu, J., Bragard, C., 2012. Screening cassava for resistance to cassava mosaic disease by grafting and whitefly inoculation. Arch. Phytopathol. Pfl. 45, 2189–2201.

Nichols, R.F.W., 1947. Breeding cassava for virus resistance. East Afr. Agric. J. 12, 184–194.

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

Okogbenin, E., Porto, M.C.M., Egesi, C., Mba, C., Espinosa, E., Santos, L.G., et al., 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.

Otim-Nape, G.W., Thresh, J.M., 1998. The current pandemic of cassava mosaic virus disease in Uganda. In: Jones, G. (Ed.), The Epidemiology of Plant Diseases. Kluwer, Dordrecht, Germany, pp. 423–443.

Rabbi, I.Y., Hamblin, M., Gedil, M., Kulakow, P., Ferguson, M., Ikpan, A.S., Ly, D., Jannink, J-L. Genetic mapping using genotyping-by-sequencing in the clonally-propagated cassava. Crop Sci. (in press).

Sserubombwe, W.S., Briddon, R.W., Baguma, Y.K., Ssemakula, G.N., Bull, S.E., Bua, A., Alicai, T., Omongo, C., Otim-Nape, G.W., Stanley, J., 2008. Diversity of begomoviruses associated with mosaic disease of cultivated cassava (*Manihot esculenta* Crantz) and its wild relative (*Manihot glaziovii* Müll. Arg.) in Uganda. J. Gen. Virol. 89, 1759–1769.

Stuthman, D.D., Leonard, K.J., Miller-Garvin, J., 2007. Breeding crops for durable resistances. In: Sparks, D.L. (Ed.), Advances in Agronomy, vol. 95. Elsevier, Amsterdam, pp. 319–367.

Van Ooijen, J.W., 2006. JoinMap 4. Software for the Calculation of Genetic Linkage Maps in Experimental Populations. Kyazma BV, Wageningen, Netherlands.

Vazquez, A.I., Bates, D.M., Rosa, G.J.M., Gianola, D., Weigel, K.A., 2010. Technical note: an R package for fitting generalized linear mixed models in animal breeding. J. Anim. Sci. 88, 497–504.