Genome-wide association mapping and genomic prediction unravels CBSD resistance in a *Manihot esculenta* breeding population

Siraj Ismail Kayondo[1,2], Dunia Pino Del Carpio[3], Roberto Lozano[3], Alfred Ozimati[1,3], Marnin Wolfe[3], Yona Baguma[1], Vernon Gracen[2,3], Offei Samuel[2], Morag Ferguson[5], Robert Kawuki[1] and Jean-Luc Jannink[3,4]

1) National Crop Resources Research Institute, NaCRRI, P.O. Box, 7084 Kampala, Uganda,
2) West Africa Center for Crop Improvement, (WACCI), University of Ghana,
3) School of Integrative Plant Sciences, Section on Plant breeding and Genetics, Cornell University, Ithaca, New York,
4) US Department of Agriculture – Agricultural Research Service (USDA-ARS)
5) International Institute for Tropical Agriculture (IITA), Nairobi
Corresponding author: kawukisezirobert@gmail.com

ABSTRACT

Cassava (*Manihot esculenta* Crantz), a key carbohydrate dietary source for millions of people in Africa, faces severe yield loses due to two viral diseases: cassava brown streak disease (CBSD) and cassava mosaic disease (CMD). The completion of the cassava genome sequence and the whole genome marker profiling of clones from African breeding programs (www.nextgencassava.org) provides cassava breeders the opportunity to deploy additional breeding strategies and develop superior varieties with both farmer and industry preferred traits. Here the identification of genomic segments associated with resistance to CBSD foliar symptoms and root necrosis as measured in two breeding panels at different growth stages and locations is reported. Using genome-wide association mapping and genomic prediction models we describe the genetic architecture for CBSD severity and identify loci strongly associated on chromosomes 4 and 11. Moreover, the significantly associated region on chromosome 4 colocalises with a *Manihot glaziovii* introgression segment and the significant SNP markers on chromosome 11 are situated within a cluster of nucleotide-binding site leucine-rich repeat (NBS-LRR) genes previously described in cassava. Overall, predictive accuracy values found in this study varied between CBSD severity traits and across GS models with Random Forest and RKHS showing the highest predictive accuracies for foliar and root CBSD severity scores.

Key words: Genome-wide association studies (GWAS), virus severity, augmented designs, de-regressed best linear unbiased Predictions (drg-BLUPs), NBS-LRR proteins

33

34 **INTRODUCTION**

35 Cassava (*Manihot esculenta* Crantz), is a major source of income and dietary calories for more
36 than 800 million people across the globe especially in Sub Saharan Africa (SSA) and recently,
37 due to the unique starch qualities of the storage roots cassava is also turning into an industrial
38 crop (Pérez *et al.*, 2011). Although cassava is a resilient crop, its production is threatened by
39 viral diseases such as Cassava brown streak virus disease (CBSD), which causes major yield
40 losses to poor farming families (ASARECA:, 2013; Ndunguru *et al.*, 2015; Patil *et al.*, 2015).
41 CBSD is caused by two major strains; *Cassava brown streak virus* (CBSV) and *Ugandan*
42 *cassava brown streak virus* (UCBSV) both CBSVs have successfully colonized the lowland and
43 highland altitudes across East Africa and new strains are emerging (Winter *et al.*, 2010;
44 Ndunguru *et al.*, 2015; Alicai *et al.*, 2016). In Uganda, because of CBSVs and agronomical
45 practices, cassava yields were recorded to be eight times lower than the yield potential for this
46 crop (ASARECA:, 2013).

47 In addition to the uncontrolled exchange of infected cassava stakes among farmers across
48 borders, CBSVs are transmitted by the African whitefly (*Besimia tobaci*) in a semi-persistent
49 manner (Legg, Sseruwagi, *et al.*, 2014; McQuaid *et al.*, 2016). Upon infection, the viruses use
50 the transport system of the plant and cause yellow chlorotic vein patterns along minor veins of
51 leaves in susceptible cassava clones (Ogwok *et al.*, 2010; Maruthi *et al.*, 2016; Anjanappa *et al.*,
52 2016). On the stem, prominent brown elongated lesions commonly referred to as "brown
53 streaks" are formed and in the storage roots, necrotic hard-corky layers are formed in the root
54 cortex of the most susceptible cassava clones (Hillocks *et al.*, 1996; Legg, Somado, *et al.*, 2014;
55 Ndyetabula *et al.*, 2016).

56 Earlier, CBSD resistance breeding initiatives have highlighted the polygenic nature of
57 inheritance in both intraspecific and interspecific cassava hybrids (Nichols, 1947; Hillocks and
58 Jennings, 2003; Munga, 2008; Kulembeka, 2010). In view of the rapid virus evolution and the
59 insufficiency of dependable virus diagnostic tools (Alicai *et al.*, 2016) breeding for durable
60 CBSD resistance, has been the main strategy to control CBSD spread in Eastern Africa. Most of
61 the available elite cassava lines have exhibited some level of sensitivity to CBSVs ranging from

62 mild sensitivity to total susceptibility. Moreover, clones classified as resistant and tolerant show
63 diverse symptom expression, restricted virus accumulation or recovery after clonal propagation
64 (Hillocks and Jennings, 2003; Alicai *et al.*, 2016).
65 Overall, in cassava for many traits the rate of genetic improvement following a traditional
66 breeding pipeline has been slower due to the combination of several biology-related issues such
67 as: poor flowering, length of breeding cycle, limited genetic diversity and slow rate of
68 multiplication of planting materials.
69 Recently, using genotypic and phenotypic information genome wide association mapping
70 (GWAS) has been used to unravel the genetic architecture of cassava mosaic disease (CMD)
71 (Wolfe *et al.*, 2016) and beta carotene content (Esuma *et al.*, 2016). Both studies have been
72 successful in identifying associated loci with traits of interest. In addition, the performance of
73 genomic prediction for different traits was previously evaluated using historical phenotypic and
74 genotyping by sequencing (GBS) datasets from the International Institute of Tropical Agriculture
75 in Nigeria (Elshire *et al.*, 2011; Ly, Hamblin, Rabbi, Melaku, Bakare, Gauch, *et al.*, 2013).
76 Genomic Selection (GS) is a breeding method alternative to marker assisted selection and
77 conventional phenotypic selection which can accelerate genetic gains through the use of
78 phenotypic and genotypic data from a training population (Meuwissen *et al.*, 2001; Jannink *et*
79 *al.*, 2010; Lorenz *et al.*, 2011). The performance of different GS models has been evaluated in
80 various species and in many traits (Resende *et al.*, 2012; Gouy *et al.*, 2013; Heslot *et al.*, 2014;
81 Charmet *et al.*, 2014; Cros *et al.*, 2015). Recently the potential of GS for CMD resistance has
82 been reported with predictive accuracies ranging from 0.53 to 0.58 (Wolfe *et al.*, 2016).
83 In the present study we followed a GWAS approach in combination with genomic prediction to
84 unravel the genetic architecture of CBSD in two Ugandan breeding populations. While one of
85 our main objectives was to assess the current predictive accuracy for CBSD we also aimed to
86 identify the most promising genomic prediction models that can account for CBSD genetic
87 architecture.
88 GWAS identified loci strongly associated with CBSVs resistance to foliar symptoms which co-
89 locate with an introgression block from a cassava wild progenitor, *M. glaziovii* (Bredeson *et al.*,
90 2016) and with root necrosis which were close to a cluster of plant defence response-related

genes annotated in the cassava genome (Lozano *et al.*, 2015). The presence of introgressions segments from the wild progenitors into the elite breeding lines is the result of cassava improvement programs at the Amani Research Station throughout the 1940s and 1950s (Jennings and Iglesias, 2002; Hillocks and Jennings, 2003).

Here we demonstrated with the synergistic implementation of GWAS and GS that GWAS could be used as a prioritization tool to identify markers for genomic prediction for CBSD resistance in cassava.In addition to unravelling the genetics of CBSD resistance these findings may help in the identification of significant causal polymorphisms to guide marker-assisted breeding for CBSD severity that may greatly improve cassava breeding in the face of increasing disease threats to agricultural production.

120

121 **MATERIALS AND METHODS**

122 **Plant material**

123 Phenotypic data was collected from two GWAS panels (Supplementary table 1), GWAS panel 1
124 composed of 429 clones and GWAS panel 2 which was composed of 872 clones.The combined
125 dataset of 1281 cassava clones were developed through three cycles of genetic recombination
126 between cassava introductions and local elite lines by the National root crops breeding program
127 at NaCRRI. These cassava clones have a diverse genetic background whose pedigree could be
128 traced back to introductions from the International Institute of Tropical Agriculture (IITA),
129 International Center for Tropical Agriculture (CIAT) and the Tanzania national cassava breeding
130 program (Supplementary table 1).

131

132 **Phenotyping**

133 The GWAS panel trials were conducted in five locations; Namulonge, Kamuli, Serere, Ngetta
134 and Kasese in Uganda.

135 GWAS panel 1 data was collected in two years across three locations, each trial was designed
136 and laid out as a 6 by 30 alpha-lattice design with two-row plots of five plants each at a spacing
137 of 1 meter by 1 meter. GWAS panel 2 was evaluated in three locations, on each location, five
138 rows of test clones were bordered by two CBSD susceptible clones in order to increase CBSD
139 disease pressure (TME204). Clones from GWAS panel 2 were evaluated as single entries per
140 location being connected by six common checks in an augmented completely randomized block
141 design with 38 blocks per site (Federer *et al.*, 2002; Federer and Crossa, 2012).

142 CBSD severity was scored at 3 (CBSD3S), 6 (CBSD6S), and 9 (CBSD9S) months after planting
143 (MAP) for foliar and 12 MAP (CBSDRS) for root symptoms respectively. The CBSD9S scores
144 were not available for GWAS panel 1.

145 CBSD severity was measured based on a 5-point scale with a score of 1 implying asymptomatic
146 conditions and a score 5 implying over 50% leaf vein clearing under foliar symptoms. However,
147 at 12 MAP a score of 5 implies over 50% of root-core being covered by a necrotic corky layer.
148 (Supplementary Figure 1)

Clones were classified with a score of 5 if pronounced vein clearing at major leaf veins were jointly displayed with brown streaks on the stems and shoot die-back that appeared as a candle-stick. Clones with 31 – 40% leaf vein clearing together with brown steaks at the stems were classified under score 4. A Score of 3 was assigned to clones with 21 – 30% leaf vein clearing with emerging brown streaks on the stems. While a score of 2 was assigned to clones that only displayed 1 – 20% leaf vein clearing without any visible brown streak symptoms on the stems. Plants classified with a score of 1 showed no visible sign of leaf necrosis and brown streaks on the stems. On the other hand, root symptoms were also classified into 5 different categories based on a 5 – point standard scale (Jennings and Iglesias, 2002; Hillocks and Jennings, 2003).

Two-stage genomic analyses

For the two stage analyses, the first stage involved accounting for trial-design using a linear mixed model to obtained de-regressed BLUPs (drgBLUPs) and the second stage involved the use of de-regressed BLUPs in GWAS and Genomic prediction.

For the panel 1 we fitted the model: $= \mathbf{X}\beta + \mathbf{Z_{clone}}c + \mathbf{Z_{range(loc.)}}r + \mathbf{Z_{block(range)}}b + \varepsilon$, using the *lmer* function from the *lme4* R package (Bates et al., 2015). In this model, $\boldsymbol{\beta}$ included a fixed effect for the population mean and location. The incidence matrix $\mathbf{Z_{clone}}$ and the vector c represent a random effect for clone $c \sim \mathrm{N}(0, \mathbf{I}\sigma_c^2)$ and **I** represent the identity matrix. The range variable, which is the row or column along which plots are arrayed, is nested in location-rep and is represented by the incidence matrix $\mathbf{Z_{range(loc.)}}$ and random effects vector $r \sim \mathrm{N}(0, \mathbf{I}\sigma_r^2)$. Block effects were nested in ranges and incorporated as random with incidence matrix $Z_{block(range)}$ and effects vector $b \sim \mathrm{N}(0, \mathbf{I}\sigma_b^2)$. Residuals ε were fit as random, with $\varepsilon \sim \mathrm{N}(0, \mathbf{I}\sigma_\varepsilon^2)$.

For panel 2 we fitted the model $\boldsymbol{y} = \mathbf{X}\beta + \mathbf{Z_{clone}}c + \mathbf{Z_{block}}b + \varepsilon$ Where y was the vector of raw phenotypes, $\boldsymbol{\beta}$ included a fixed effect for the population mean and location with checks included as a covariate. The incidence matrix $\mathbf{Z_{clone}}$ and the vector c are the same as the aforementioned model and the blocks were also modeled with incidence matrix $\mathbf{Z_{block}}$ and **b** represents the random effect for block. The best linear predictors (BLUPs) of the clone effect ($\hat{c}$) were extracted as de-regressed BLUPS following the formula (Garrick *et al.*, 2009):

$$\textbf{deregressed BLUP} = \frac{\textbf{BLUP}}{1 - \frac{\textbf{PEV}}{\sigma_c^2}}$$

177

178 Where PEV is the prediction error variance of the BLUP and σ_c^2 is the clonal variance
179 component.

180

181

182 **DNA preparation and Genotyping by sequencing (GBS)**

183 Total genomic DNA was extracted from young tender leaves of all cassava clones included in
184 the phenotyping trials according to standard procedures using the DNAeasy plant mini extraction
185 kit (QIAGEN, 2012). Genotyping-by-sequencing (GBS)(Elshire *et al.*, 2011) libraries were
186 constructed using the ApeKI restriction enzyme (Hamblin and Rabbi, 2014). Marker genotypes
187 were called using TASSEL GBS pipeline V4 (Glaubitz *et al.*, 2014) after aligning the reads to
188 the Cassava v6 reference genome (Prochnik *et al.*, 2012; Goodstein *et al.*, 2014). Variant Calling
189 Format (VCF) files were generated for each chromosome. Markers with more than 60% missing
190 calls were removed. Genotypes with less than five reads were masked before imputation.
191 Additionally, only biallelic SNP markers were considered for further processing.

192 The marker dataset consisted of a total of 173,647 bi-allelic SNP markers called for 986
193 individuals. This initial dataset was imputed using Beagle 4.1 (Browning and Browning, 2016).
194 After timputation, 63,016 SNPs had an AR2 (Estimated Allelic r-squared) higher than 0.3 and
195 were kept for analysis; from these, 41,530 had a minor allele frequency (MAF) higher than 0.01
196 in our population. Dosage files for this final dataset were generated and used for both GWAS
197 and GS analyses.

198

199 **Genetic correlations and heritability estimates**

200

201 Correlation across CBSD traits was estimated using pairwise correlations for each location using
202 the drgBLUPs values obtained after fitting the aforementioned linear mixed model. Broad sense

203 heritabilities (plot-mean basis) were calculated using the estimated variance components from
204 the first step of the two-step genomic analysis as explained previously.
205 In addition,SNP-based heritabilities were calculated for each GWAS panel by fitting a single-
206 step mixed-effects model, the full models which specified clone as a random effect were fitted
207 using the *emmreml* function from the *EMMREML* R package (Akdemir and Okeke, 2015). The
208 random effect was modeled as having co-variance proportional to the kinship matrix, which was
209 calculated using the *A.mat* function from the *rrBLUP* R package (Endelman, 2011).
210
211
212 **Genome-wide association analysis for CBSV severity**
213 Although pedigree records indicate the two GWAS panels to be closely related a principal
214 component analysis (PCA) was performed in order to characterize these panels and to identify
215 any population stratification between the two GWAS panels. We used the imputed dataset of
216 63,016 SNP markers to calculate the PCs with the function *princomp* in R.
217 With the imputed dataset of 63,016 SNP markers and 986 individuals genome wide association
218 was performed using a mixed linear model association analysis (MLMA) accounting for kinship
219 as implemented in GCTA (v 1.26.0) (Yang *et al.*, 2011) . Specifically, we followed a leave one
220 chromosome out approach, with this approach the chromosome on which the candidate SNP
221 markers are tested gets excluded from the genomic relationship (GRM) calculation. Bonferroni
222 correction (reference) was used to correct for multiple testing with a significance threshold set at
223 5.9. Manhattan plots with transformed -$\log_{10}$(P-value) were generated using R package *qqman*
224 (Turner, 2014).
225
226 **Genomic prediction models**
227 To assess the potential of implementing genomic selection for CBSD, seven genomic prediction
228 models were keenly examined; genomic best linear unbiased prediction (GBLUP), reproducing
229 kernel Hilbert spaces (RKHS), BayesCpi, Bayesian LASSO, BayesA, BayesB and Random
230 forest (RF).

231 **GBLUP**. In this prediction model, the GEBVs are obtained after fitting a linear mixed model
232 where the genomic realized relationship matrix is based on SNP marker dosages. Accordingly,
233 the genomic relationship matrix was constructed using the function *A.mat* in the R package
234 rrBLUP (Endelman, 2011) and follows the formula of VanRaden (2008), method two. GBLUP
235 predictions were made with the function *emmreml* in the R package EMMREML (Akdemir and
236 Okeke, 2015).
237 **Multi-kernel GBLUP**. Because the most significant QTLs for foliar severity 3 and 6 MAP were
238 mapped on chromosomes 4 and 11 (this paper) we followed a multi-kernel approach by fitting
239 three kernels with genomic relationship matrices constructed with SNP markers from
240 chromosomes 4 (G_{chr4}), 11 (G_{chr11}) and SNPs from the other chromosomes ($G_{allchr\text{-}[4,11]}$). Multi-
241 kernel GBLUP predictions were made with the function *emmremlMultiKernel* in the R package
242 EMMREML (Akdemir and Okeke, 2015).
243 **RKHS**. Unlike GBLUP for RKHS we use a Gaussian kernel function: $K_{ij} = \exp\left(-\left(\mathrm{d_{ij}}\theta\right)\right)$,
244 where K_{ij} is the measured relationship between two individuals, d_{ij} is their Euclidean genetic
245 distance based on marker dosages and θ is a tuning ("bandwidth") parameter that determines the
246 rate of decay of correlation among individuals. This function is nonlinear and therefore the
247 kernels used for RKHS can capture non-additive as well as additive genetic variation. To fit a
248 multiple-kernel model with six covariance matrices we used the *emmremlMultiKernel* function
249 in the EMMREML package, with the following bandwidth parameters: 0.0000005, 0.00005,
250 0.0005, 0.005, 0.01, 0.05 (Multi-kernel RKHS) and allowed REML to find optimal weights for
251 each kernel.
252 **Bayesian maker regressions**. We tested four Bayesian prediction models: BayesCpi (Habier *et*
253 *al.*, 2011), the Bayesian LASSO (Park and Casella, 2008), BayesA, and BayesB (Meuwissen *et*
254 *al.*, 2001). The Bayesian models we tested allow for alternative genetic architectures by way of
255 differential shrinkage of marker effects. We performed Bayesian predictions with the R package
256 BGLR (Pérez and De Los Campos, 2014)
257 **Random Forest**. Random forest (RF) is a machine learning method used for regression and
258 classification (Breiman, 2001; Strobl *et al.*, 2009; Charmet and Storlie, 2012). Random forest
259 regression with marker data has been shown to capture epistatic effects and has been successfully

260 used for prediction (Breiman, 2001; Motsinger-Reif *et al.*, 2008; Heslot *et al.*, 2012; Charmet *et*
261 *al.*, 2014; Spindel *et al.*, 2015). We implemented RF using the random Forest package in R
262 (Liaw and Wiener, 2002) with the parameter, *ntree* set to 500 and the number of variables
263 sampled at each split (*mtry*) equal to 300.
264
265 **Introgression Segment Detection**
266 To identify the genome segments in the two GWAS panels, we followed the approach described
267 in Bredeson et al . (Bredeson *et al.*, 2016). We used the *M. glaziovii* diagnostic markers
268 identified in Supplementary Dataset 2 of Bredeson et al. (Bredeson *et al.*, 2016), these ancestry
269 diagnostic (AI) SNPs were identified as being fixed for different alleles in a sample of two pure
270 *M. esculenta* (Albert and CM33064) and two pure *M. glaziovii.*
271 Out of 173,647 SNP in our imputed dataset, 12,502 matched published AI SNPs. For these AI
272 SNPs, we divided each chromosome into non-overlapping windows of 20 SNP. Within each
273 window, for each individual, we calculated the proportion of genotypes that were homozygous
274 (G/G) or heterozygous (G/E) for *M. glaziovii* allele and the proportion that were homozygous for
275 the *M. esculenta* allele (E/E). We assigned G/G, G/E or E/E ancestry to each window, for each
276 individual only when the proportion of the most common genotype in that window was at least
277 twice the proportion of the second most common genotype. We assigned windows a "No Call"
278 status otherwise.
279 We also used this approach on six whole-genome sequenced samples from the cassava HapMap
280 II (Ramu *et al.*, 2016). These included the two "pure cassava" and *M. glaziovii* (S) from
281 Bredeson et al. (Bredeson *et al.*, 2016), plus an additional *M. glaziovii*, and two samples labeled
282 Namikonga. Because these samples came from a different source from most our samples, we
283 could find only 11,686 SNPs that matched both the sites in the rest of our study sample and the
284 list of ancestry informative sites for analysis.
285
286
287
288

289 **Linkage disequilibrium plots**

290 To confirm whether a large haplotype block present on chromosome 4 colocate with a GWAS
291 QTL identified on this chromosome we calculated LD scores of every SNP marker on
292 chromosome 4 in a 1Mb window using GCTA (Yang *et al.*, 2011). Briefly, LD score for a given
293 marker is calculated as the sum of R^2 adjusted between the index marker and all markers within a
294 specified window. The adjusted R^2 is an unbiased measure of LD:

$$R^2_{adj} = R^2 - \frac{(1 - R^2)}{(n-2)}$$

295 Where "n" is the population size and R^2 is the usual estimator of the squared Pearson's
296 correlation (Bulik-Sullivan *et al.*, 2015). The resulting LD scores were then plotted against the
297 GWAS $\log_{10}$ (Pvalue) of every marker on chromosome 4.

298 To highlight the importance of the associated markers on chromosome 11 we calculated pairwise
299 squared Pearson's correlation coefficient (r^2) between the top significant GWAS SNP hit on this
300 chromosome and neighboring markers in a window of 2Mb (1Mb upstream and 1Mb
301 downstream). (plink ref)

302

303 **Candidate gene identification**

304 We used the mlma GCTA output to filter out SNP markers based on -log10 (P-value) values
305 higher than the Bonferroni threshold (~ 5.9). The resulting significant SNP markers were then
306 mapped onto genes using the SNP location and gene description from the
307 M.esculenta_305_v6.1.gene.gff3 available in Phytozome 11(Goodstein et al., 2014) for *Manihot*
308 *esculenta* v6.1 using the intersect function from bedtools (Quinlan and Hall, 2010).

309

310

311

312

313

314

315 **RESULTS**

316 **Phenotypic variability for severity to cassava brown streak virus infection**

317 In the present study field disease scoring was done based on a standard CBSD scoring scale that
318 ranges from 1 to 5 for both foliar and root symptoms (Supplementary figure 1).

319 Datasets for CBSD foliar and root severities of the evaluated germplasm are presented in
320 Supplementary figures 2 and 3, both GWAS panels exhibited differential response to CBSVs at
321 three,six, nine and twelve months as revealed in the great variability of the deregressed BLUPs.
322 Interestingly, clones which displayed an intermediate response were by far more abundant than
323 clones with susceptible or resistance response.

324 Phenotypic correlations for foliar and root severities (CBSD3S, CBSD6S and CBSDRS) within
325 panels and within and across locations are presented in Supplementary figure 4, Supplementary
326 tables 2 and 3 with clear differences in CBSD severity scores.

327 For panel 1, results varied across locations and CBSD severity traits the lowest correlation value
328 was between Ngetta and Kasese (0.09) and the highest between Namulonge and Kasese (0.60)
329 both values correspond root severity scoring (Supplementary table 2A).

330 For panel 2 the results varied across locations and CBSD severity traits with correlation values
331 ranging between -0.08 for CBSD9S (Namulonge-Kamuli) and 0.51 CBSD3S (Kamuli-Serere)
332 (Supplementary table 2B).

333 Within locations across traits the highest correlation values were found in panel 1 for foliar
334 scorings CBSD3S and CBSD6S ($r^2 > 0.5$) (Supplementary table 3A). For panel 2, correlation
335 across traits varied depending on the location, nonetheless correlations across foliar traits were
336 generally higher than those between foliar and root severity (Supplementary table 3B).

337 Heritability estimate values for CBSD3S, CBSD6S and CBDRS were low to intermediate with
338 broad-sense heritability (H^2) estimates spanning a wide range (11% to 73%) for both panels
339 across locations (Table 1). For GWAS panel 2, broad-sense heritability (H^2) estimates ranged
340 between 56% and 63% for CBSD3S and between 60% and 62% for CBSD6S; while for GWAS
341 panel 1 ranged between 11% and 51%.

342 Narrow-sense heritability (h^2), also referred to as SNP heritability, was estimated using the
343 variance components obtained as a result of fitting a one step model using the genetic

344 relationship matrix (GRM) for each panel. For panel 1, the broad- and narrow-sense heritability
345 values were comparable across locations except for the multi-location model. For panel 2, for
346 most locations the broad-sense heritability estimates were larger than the narrow-sense
347 heritability estimates. The high variability observed within and across GWAS panels reflects
348 differences in population composition, field design and environmental effects.
349
350 **Genome wide association mapping for CBSV severity in cassava**
351 The extent of subpopulation structure between the two GWAS panels was examined by PCA,
352 which showed no distinct clusters: clones from both panels had mixed distribution. Overall, the
353 first three PCAs accounted for 60% of the genetic variation observed in the data (Figure 1). The
354 first PC accounted for 30% of the observed variation while the second and third PCs contributed
355 20% and 10% respectively.
356 Genotype-phenotype associations for CBSD severity traits based on the combination of multi-
357 location data and 986 individuals are presented in Figure 2. Additional GWAS analyses
358 performed on each panel individually are presented in Supplementary tables 4 and 5 and
359 Supplementary figures 5-12.
360 We characterized SNP markers with a -log10 (P-value) above the Bonferroni threshold > 5.9 as
361 significant marker–trait associations and further annotated those into candidate genes
362 (Supplementary table 4).
363 For the combined dataset, we identified 83 significant SNP markers associated to CBSD3S; the
364 markers mapped to chromosome 11 with 61 markers located within genes (Supplementary Table
365 4). The QTL on chromosome 11, top hit reference SNP $-\log_{10}$ (P-value) = 9.38, explained 6% of
366 the observed phenotypic variation.
367 On the other hand, for CBSD6S, we identified significant SNPs on chromosome 11,
368 chromosome 4 and chromosome 12. On chromosome 11, 33 SNPs surpassed the Bonferroni
369 threshold with 27 SNP markers located within genes. The QTL on chromosome 11 is located on
370 the same region as the QTL identified for CBSD3S and explained 5% of the observed
371 phenotypic variation (Figure 3A).

372 It suffices to note that although several SNPs on chromosome 11 for CBSD6S exceeded the
373 Bonferroni threshold, six SNPs were in linkage disequilibrium ($r^2 > 0.6$) with the top reference
374 SNP hit. The SNP markers, with an $r^2 > 0.2$ to the reference SNP, were annotated into candidate
375 genes: Manes11G130500, a gene that is known to encode glycine-rich protein.
376 Manes11G130000 gene that encodes Leucine-rich repeat (LRR) containing protein,
377 Manes11G130200 gene that encodes the trigger factor chaperone and *peptidyl-prolyl* trans and
378 Manes11G131100 that encodes a protein kinase (Figure 3B).
379 Since several SNPs on the chromosome 4 QTL region are in high LD, no single locus can be
380 highlighted as candidate gene(s) to be associated with CBSD severity (Figure 4A). The large
381 haplotype on chromosome 4 is an introgression block from the a wild relative of cassava (*M.*
382 *glaziovii*) (Jennings, 1959; Bredeson *et al.*, 2016). We further confirmed the presence and
383 segregation of the introgressed genome segment in both panels using a set of diagnostic markers
384 from *M. glaziovii* (Figure 4B, supplementary figure 13 and 14).
385 The significant QTL on chromosome 12 has been previously identified for CMD resistance in
386 cassava (Wolfe *et al.*, 2016) Accordingly, after correction for CMD scoring in the first step
387 calculation of CBSD deregressed BLUPs, the QTL on chromosome 12 was no longer significant
388 and only QTLs on chromosomes 4 and 11 remained (supplementary Figure 15).
389 For CBSDRS we could not identify SNPs surpassing the Bonferroni correction partly to the
390 complexity of this trait with apparently several small effect genes and low heritability. However,
391 the results of the analysis of CBSDRS multi-location data of panel 1 identified significant
392 regions on chromosomes 5, 11 and 18 (-log10 (P-value) > 6.5), which explained 8, 6 and 10%
393 phenotypic variance respectively.
394
395 **Genome-wide prediction for CBSV severity in cassava**
396 An important objective within this study was to assess the accuracy of prediction in cassava for
397 CBSD-related traits. Using the combined dataset, we compared the performance of seven
398 genomic prediction models with contrasting assumptions on trait genetic architecture. Some
399 model predictions represent genomic estimated breeding values (GEBV) in that they are sums of
400 additive effects of markers, while other model predictions represent genomic estimated total

genetic value (GETGV) because they include non-additive effects. Predictive accuracy for CBSD related traits had mean values across methods of 0.29 (CBSD3S), 0.40 (CBSD6S) and 0.34 (CBSDRS) (Figure 5 and Supplementary table 6).

Predictive accuracies for CBSD3S varied in the range of 0.27 (BayesB and GBLUP) and 0.32 (RF), for CBSD6S we obtained a predictive value of 0.40 for most methods except for RKHS (0.42) and RF (0.41) and for CBSD root severity scores varied from 0.31 (BayesA, B, C and GBLUP) to 0.42 (RF and RKHS). It is clear from the results that higher predictive accuracies were consistently achieved when using Random forest and RKHS for the prediction of both foliar and root CBSD resistance traits. Although for foliar symptoms the increase in predictive accuracy using those methods is modest, for CBSDRS the increase in predictive accuracy was 0.10.

Based on the GWAS results, we identified for CBSD3S, CBSD6S and CBSDRS the strongest marker associations on chromosomes 4 and 11 . Markers from chromosomes 4, 11 and markers on other chromosomes were used independently to construct covariance matrices that were fitted in a multikernel GBLUP model (Supplementary figure 16). For all CBSD traits the mean predictive accuracy values from the single-kernel GBLUP model were similar to the mean total predictive accuracy following the multi-kernel approach (Supplementary table 6).

Differences were found on the contribution of the individual kernels to the total predictive accuracies. For example, the multikernel GBLUP model for CBSD3S had the lowest total predictive accuracy (0.27) with the highest contribution coming from chromosome 11 and the rest of the genome (0.19). In contrast, the multikernel GBLUP model for CBSD6S gave the highest predictive accuracy (0.40) and most of the accuracy came from chromosome 4 (0.29). The multikernel GBLUP approach for CBSDRS had a total predictive accuracy of 0.30 with the rest of the genome (0.29) contributing the most to the total predictive accuracy (Supplementary figure 16).

430 **DISCUSSION**

431 Cassava brown streak disease has been identified as one of the most serious threats to food
432 security (Pennisi, 2010) owing to the significant loses it imparts in cassava wherever it occurs.
433 Host plant resistance, that is obtained through breeding efforts has been so far the most effective
434 approach. However, this is only achievable when the host-pathogen behaviour and interaction is
435 well understood and/or when the genetics of resistance to CBSD are clearly known.
436 In the present study, ~1200 cassava clones from the NaCRRI breeding program in Uganda were
437 evaluated for CBSD severity scores in leaves and root. Specifically, this paper sought to provide
438 fundamental information on the genetics of resistance to CBSD which was previously unknown.
439 From our analyses it was evident that correlation among foliar CBSD severities were higher than
440 correlation between foliar and root severities.
441 Selection of resistant clones has been hampered by the fact that some clones do not show
442 symptoms on leaves or storage roots, while other varieties may only express symptoms on leaves
443 and not on roots and still others do not show symptoms on leaves but instead on roots only
444 (ASARECA:, 2013). Moreover, a lack of correlation between virus load and symptom
445 expression in a field evaluation of selected cassava genotypes has been reported (Kaweesi *et al.*,
446 2014). Previous studies have also reported that 79% plants with above- ground symptoms of
447 CBSD also exhibited root necrosis and 18% of plants had no visible symptoms of CBSV
448 (Hillocks *et al.*, 1996)
449 Recently, efforts to understand CBSD have focused on CBSD resistance population development
450 and preliminary insights into chromosomal regions and genes involved in resistance (Kawuki *et*
451 *al.*, 2016; Anjanappa *et al.*, 2016, 2017). These studies have highlighted the existence of a QTL
452 on chromosome 11 for CBSD root necrosis among cassava clones of Tanzanian origin (Kawuki
453 *et al.*, 2016).
454 In our study, based on foliar CBSD severity scoring using a multi-location dataset we identified
455 significant QTL regions on chromosome 4 and 11, though these associations were not always
456 consistent when the panels were analyzed separately and per location. Overall, these results
457 highlight the advantage of using a large GWAS panel and a multi-location approach were plants
458 are exposed to different disease pressures to identify additional genomic regions.

459 On chromosome 11, a cluster of genes underlies the significant QTL ; candidate genes for further
460 study are: Manes11G131100, Manes11G130500, Manes11G130200 and Manes11G130000.
461 Lozano et al. 2015 previously reported Manes11G130000 when studying the distribution of
462 *NBS-LRR* in cassava. Furthermore, a recent study on early transcriptome response to brown
463 streak virus infection in susceptible and resistant cassava varieties identified Manes.11G130000
464 among the differentially expressed genes in the susceptible line 60444 from the ETH cassava
465 germplasm collection (Anjanappa *et al.*, 2017). The QTL on chromosome 11 is particularly
466 unstable across locations, which may be related to NBS-LRR genes conferring resistance to a
467 particular strain, UCBSV exhibits a lower mutation rate, while CBSV is more aggressive and
468 mutates faster.
469 Throughout the 1940s and 1950s at the Amani Research Station, *Manihot glaziovii* and cassava
470 varieties of Brazilian origin were used for crosses to obtain CBSD resistant varieties (Jennings
471 and Iglesias, 2002). One of the introgression segments from these wild relatives has been
472 reported to be located on chromosome 4, however the level of linkage disequilibrium in that
473 region remains as a major constraint for the identification of the gene or genes that are
474 responsible for CBSD resistance (Bredeson *et al.*, 2016). Current on-going research efforts are
475 focused on dissecting the extent of the effects of wild introgressions on cassava traits (Marnin
476 Wolfe personal communication).
477 One important objective of the present study was to test our ability to predict CBSD severity in
478 cassava, which is, particularly relevant in two situations. First, when the objective is the
479 introduction of germplasm from Latin america and/or from West Africa to East Africa and for
480 early seedling or clonal selection of resistant lines.
481 Thus, using a cross-validation approach, we evaluated the suitability of seven GS models with
482 the expectation that the results may differ due to differences in genetics of foliar and root CBSD
483 severity traits (B. J. Hayes *et al.*, 2009; Grattapaglia *et al.*, 2011).
484 In cassava, previous genomic prediction studies have evaluated the predictive ability of GBLUP
485 using historical phenotypic data from the International Institute of Tropical Agriculture (IITA)
486 and GBS markers and in a small training population with relatively low-density markers (de
487 Oliveira *et al.*, 2012; Ly, Hamblin, Rabbi, Melaku, Bakare, Okechukwu, *et al.*, 2013).

488 Principally, the GS models evaluated have varying underlying assumptions genomic-BLUP
489 (GBLUP) model assume an infinitesimal genetic architecture; Bayesian methods such as BayesA
490 and BayesB relax the assumption of common variance across marker effects (De Los Campos *et*
491 *al.*, 2009; Habier *et al.*, 2011; Legarra *et al.*, 2014), RKHS and random forest methods can model
492 epistatic and other non-additive effects.
493 A first assessment of predictive accuracy of CBSD foliar and root traits in cassava indicate that
494 the use of genomic selection is a promising breeding method for resistance to Cassava brown
495 streak virus. We found moderate to high predictive accuracies for these traits in relation to results
496 from other traits in cassava (Ly, Hamblin, Rabbi, Melaku, Bakare, Okechukwu, *et al.*, 2013)**.**
497 However, predictive accuracy values are lower in comparison to the values reported for cassava
498 mosaic virus (Ly, Hamblin, Rabbi, Melaku, Bakare, Okechukwu, *et al.*, 2013) possibly due the
499 presence of a large effect GWAS QTL (CMD2) for CMD .
500 Although, a priori knowledge of the loci affecting a trait is not needed for GS, we also tested a
501 multiple kernel approach using GWAS results as a reference to construct covariance matrices.
502 GWAS results have been incorporated in genome-wide prediction models to increase predictive
503 accuracy through *de-novo* GWAS or using previously published GWAS results (Zhang *et al.*,
504 2014; Spindel *et al.*, 2016).
505 In our study, to avoid a correlation effect across covariance matrices we partitioned SNP markers
506 into three sets: markers on GWAS QTLs chromosomes (chr 4 and 11) and markers on rest of the
507 genome to built genomic relationship matrices (G_{chr4},G_{chr11},$G_{allchr-[4,11]}$).Remarkably, the
508 predictive accuracy of each kernel modeled the genetic architecture found though GWA
509 analyses. Our GWAS and GS results indicate that resistance to CBSD root necrosis severity is
510 polygenic in nature, which is in accordance to Kawuki et al.'s (2016) results.
511 Our results suggest that non-additive effects are likely to play a role shaping CBSD resistance
512 particularly root necrosis. This conclusion derives from GS results using Random Forest and
513 RKHS, which gave the highest predictive accuracies, and from the observed differences in broad
514 sense and narrow sense heritability values.
515 CBSD is a disease that has devastating consequences in cassava production and poses a risk
516 particularly to countries in Central and West Africa where CBSD is not currently present. Our

517 study provides, through GWAS and genomic prediction, an insight into the genetic regulation of
518 CBSD severity in leaves and roots. Although we were able to identify a candidate NBS-LRR
519 gene on chromosome 11, the function of this gene in CBSD resistance requires further validation
520 and more importantly, there is a risk that this gene might not be a source of durable resistance to
521 CBSVs. Within this context, genomic selection arises as a promising tool that can accelerate
522 breeding, though the average predictive accuracy is lower than CMD, this is highly variable
523 across locations and the breeding panel evaluated. Further work will require screening of large
524 diversity panels in multiple environments, identification of QTLs specific to viral strains and the
525 introgression of genomic regions conferring resistance to CBSD from wild relatives and Latin
526 American accessions.
527
528
529
530

REFERENCES

Akdemir D, Okeke UG (2015). EMMREML: Fitting Mixed Models with Known Covariance Structures.

Alicai T, Ndunguru J, Sseruwagi P, Tairo F, Okao-Okuja G, Nanvubya R, *et al.* (2016). *Characterization by Next Generation Sequencing Reveals the Molecular Mechanisms Driving the Faster Evolutionary rate of Cassava brown streak virus Compared with Ugandan cassava brown streak virus*. Cold Spring Harbor Labs Journals.

Anjanappa RB, Mehta D, Maruthi MN, Kanju E, Gruissem W, Vanderschuren H (2016). Characterization of Brown Streak Virus–Resistant Cassava. **29**: 527–534.

Anjanappa RB, Mehta D, Okoniewski MJ, Szabelska A, Gruissem W, Vanderschuren H (2017). Early transcriptome response to brown streak virus infection in susceptible and resistant cassava varieties. : 1–22.

ASARECA: (2013). *ASARECA Annual Report 2012: Transforming Agriculture for Economic Growth in Eastern and Central Africa*. Entebbe, Uganda.

B. J. Hayes, H. D. Daetwyler, P. Bowman, G. Moser, B. Tier4, R. Crump, *et al.* (2009). Accuracy of Genomic Selection: Comparing Theory and Results. In: Daetwyler HD (ed) *Genome-Wide Evaluation of Populations*, Proc. of Assoc. Advmt. Anim. Breed., pp 352–355.

Bredeson J V, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, *et al.* (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol* **34**: 562–570.

Breiman L (2001). Random Forests. *Mach Learn* **45**: 5–32.

Browning BL, Browning SR (2016). Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**: 116–126.

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, *et al.* (2015). An atlas of genetic correlations across human diseases and traits. *Nat Publ Gr* **47**.

Charmet G, Storlie E (2012). Implementation of genome-wide selection in wheat. *Russ J Genet Appl Res* **2**: 298–303.

Charmet G, Storlie E, Oury FX, Laurent V, Beghin D, Chevarin L, *et al.* (2014). Genome-wide prediction of three important traits in bread wheat. *Mol Breed* **34**: 1843–1852.

Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-gasselin T, *et al.* (2015). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theor Appl Genet* **128**: 397–410.

Elshire RJ, Glaubitz JC, Sun Q, Poland J a, Kawamoto K, Buckler ES, *et al.* (2011). A robust, simple

563 genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**: e19379.
564 Endelman JB (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package
565 rrBLUP. *Plant Genome J* **4**: 250.
566 Esuma W, Herselman L, Labuschagne MT, Ramu P, Lu F, Baguma Y, *et al.* (2016). Genome-wide
567 association mapping of provitamin A carotenoid content in cassava. *Euphytica*.
568 Federer WT, Crossa J (2012). Screening Experimental Designs for Quantitative Trait Loci, Association
569 Mapping, Genotype-by Environment Interaction, and Other Investigations. *Front Physiol* **3**.
570 Federer WT, Nguyen N-K, others (2002). Constructing Augmented Experiment Designs with Gendex.
571 Garrick DJ, Taylor JF, Fernando RL (2009). Deregressing estimated breeding values and weighting
572 information for genomic regression analyses. *Genet Sel Evol* **41**: 55.
573 Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, *et al.* (2014). TASSEL-GBS: a high
574 capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**: e90346.
575 Goodstein D, Batra S, Carlson J, Hayes R, Phillips J, Shu S, *et al.* (2014). Phytozome Comparative Plant
576 Genomics Portal.
577 Gouy M, Rousselle Y, Bastianelli D, Lecomte P, Bonnal · L, Roques D, *et al.* (2013). Experimental
578 assessment of the accuracy of genomic selection in sugarcane. *Theor Appl Genet* **126**: 1–12.
579 Grattapaglia D, Deon M, Resende V, Resende MR, Sansaloni CP, Petroli CD, *et al.* (2011). Genomic
580 Selection for growth traits in Eucalyptus: accuracy within and across breeding populations. *From*
581 *IUFRO Tree Biotechnol Conf BMC Proc* **5**.
582 Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011). Extension of the bayesian alphabet for genomic
583 selection. *BMC Bioinformatics* **12**: 186.
584 Hamblin MT, Rabbi IY (2014). The Effects of Restriction-Enzyme Choice on Properties of Genotyping-
585 by-Sequencing Libraries: A Study in Cassava (). *Crop Sci* **0**: 0.
586 Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014). Integrating environmental covariates and crop
587 modeling into the genomic selection framework to predict genotype by environment interactions.
588 *Theor Appl Genet* **127**: 463–480.
589 Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012). Genomic Selection in Plant Breeding: A
590 Comparison of Models. *Crop Sci* **52**: 146.
591 Hillocks RJ, Jennings DL (2003). Cassava brown streak disease: a review of present knowledge and
592 research needs. *Int J Pest Manag* **49**: 225–234.
593 Hillocks RJ, Raya M, Thresh JM (1996). The association between root necrosis and above-ground
594 symptoms of brown streak virus infection of cassava in southern Tanzania. *Int J Pest Manag* **42**:

285–289.

Jannink J-LL, Lorenz AJ, Iwata H (2010). Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* **9**: 166–177.

Jennings DL (1959). Manihot melanobasis Müll. Arg.—a useful parent for cassava breeding. *Euphytica* **8**: 157–162.

Jennings DL, Iglesias C (2002). Breeding for Crop Improvement. *Cassava Biol Prod Util*: 149–166.

Kaweesi T, Kawuki R, Kyaligonza V, Baguma Y, Tusiime G, Ferguson ME (2014). Field evaluation of selected cassava genotypes for cassava brown streak disease based on symptom expression and virus load. *Virol J* **11**: 216.

Kawuki RSRS, Kaweesi T, Esuma W, Pariyo A, Kayondo IS, Ozimati A, *et al.* (2016). Eleven years of breeding efforts to combat cassava brown streak disease. *Breed Sci* **66**: 560–571.

Kulembeka HP (2010). Genetic linkage mapping of Field Resistance to cassava brown streak Disease in cassava landraces from Tanzania. University of the Free State.

Legarra A, Christensen OF, Aguilar I, Misztal I (2014). Single Step, a general approach for genomic selection. *Livest Sci* **166**: 54–65.

Legg J, Somado EA, Barker I, Beach L, Ceballos H, Cuellar W, *et al.* (2014). A global alliance declaring war on cassava viruses in Africa. *Food Secur* **6**: 231–248.

Legg JP, Sseruwagi P, Boniface S, Okao-Okuja G, Shirima R, Bigirimana S, *et al.* (2014). Spatio-temporal patterns of genetic change amongst populations of cassava Bemisia tabaci whiteflies driving virus pandemics in East and Central Africa. *Virus Res* **186**: 61–75.

Liaw a, Wiener M (2002). Classification and Regression by randomForest. *R news* **2**: 18–22.

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, *et al.* (2011). Genomic Selection in Plant Breeding. Knowledge and Prospects. In: *Advances in Agronomy*, Elsevier Inc Vol 110, pp 77–123.

De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, *et al.* (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182**: 375–385.

Lozano R, Hamblin MT, Prochnik S, Jannink J-L (2015). Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC Genomics* **16**: 1–14.

Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch HG, *et al.* (2013). Relatedness and genotype ?? environment interaction affect prediction accuracies in genomic selection: A study in cassava. *Crop Sci* **53**: 1312–1325.

Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Okechukwu R, *et al.* (2013). Relatedness and

Genotype-by-environment Interaction Affect Prediction Accuracies in Genomic Selection: a Study in Cassava 2. *Crop Sci*.

Maruthi MN, Jeremiah CS, Mohammed IU, Legg JP (2016). Virus-vector relationships and the role of whiteflies, Bemisia tabaci, and farmer practices in the spread of cassava brown streak viruses.

Mbanzibwa DR, Tian YP, Tugume AK, Mukasa SB, Tairo F, Kyamanywa S, *et al.* (2011). Simultaneous virus-specific detection of the two cassava brown streak-associated viruses by RT-PCR reveals wide distribution in East Africa, mixed infections, and infections in Manihot glaziovii. *J Virol Methods* **171**: 394–400.

McQuaid CF, Sseruwagi P, Pariyo A, van den Bosch F (2016). Cassava brown streak disease and the sustainability of a clean seed system. *Plant Pathol* **65**: 299–309.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD (2008). A comparison of analytical methods for genetic association studies. *Genet Epidemiol* **32**: 767–778.

Munga TL (2008). Breeding for Cassava Brown Streak Resistance in Coastal Kenya. University of KwaZulu-Natal Republic of South Africa.

Ndunguru J, Sseruwagi P, Tairo F, Stomeo F, Maina S, Djinkeng A, *et al.* (2015). Analyses of twelve new whole genome sequences of cassava brown streak viruses and ugandan cassava brown streak viruses from East Africa: Diversity, supercomputing and evidence for further speciation. *PLoS One* **10**: e0139321.

Ndyetabula IL, Merumba SM, Jeremiah SC, Kasele S, Mkamilo GS, Kagimbo FM, *et al.* (2016). Analysis of Interactions Between Cassava Brown Streak Disease Symptom Types Facilitates the Determination of Varietal Responses and Yield Losses. *Plant Dis* **100**: 1388–1396.

Nichols RFW (1947). Breeding cassava for virus resistance. *East African Agric J* **12**: 184–94.

Ogwok E, Patil BL, Alicai T, Fauquet CM (2010). Transmission studies with Cassava brown streak Uganda virus (Potyviridae: Ipomovirus) and its interaction with abiotic and biotic factors in Nicotiana benthamiana. *J Virol Methods* **169**: 296–304.

de Oliveira EJ, de Resende MDV, da Silva Santos V, Ferreira CF, Oliveira GAF, da Silva MS, *et al.* (2012). Genome-wide selection in cassava. *Euphytica* **187**: 263–276.

Park T, Casella G (2008). The Bayesian Lasso. *J Am Stat Assoc* **103**: 681–686.

Patil BL, Legg JP, Kanju E, Fauquet CM (2015). Cassava brown streak disease: A threat to food security in Africa. *J Gen Virol* **96**: 956–968.

659 Pennisi E (2010). Armed and dangerous. *Science* **327**: 804–5.
660 Pérez JC, Lenis JI, Calle F, Morante N, Sánchez T, Debouck D, *et al.* (2011). Genetic variability of root
661 peel thickness and its influence in extractable starch from cassava (Manihot esculenta Crantz) roots.
662 *Plant Breed* **130**: 688–693.
663 Pérez P, De Los Campos G (2014). Genome-wide regression and prediction with the BGLR statistical
664 package. *Genetics* **198**: 483–495.
665 Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, *et al.* (2012). The Cassava
666 Genome: Current Progress, Future Directions. *Trop Plant Biol* **5**: 88–94.
667 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007). PLINK: A Tool
668 Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**:
669 559–575.
670 QIAGEN (2012). *DNeasy ® Plant Handbook DNeasy Plant Mini Kit and tissues , or fungi Sample &*
671 *Assay Technologies QIAGEN Sample and Assay Technologies*.
672 Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features.
673 *Bioinforma Appl NOTE* **26**: 841–84210.
674 Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson J V, *et al.* (2016). Cassava HapMap:
675 Managing genetic load in a clonal crop species. *bioRxiv*: 1–15.
676 Rentería ME, Cortes A, Medland SE (2013). Using PLINK for Genome-Wide Association Studies
677 (GWAS) and Data Analysis. In: Gondro C, van der Werf J, Hayes B (eds) Humana Press: Totowa,
678 NJ Vol 1019, pp 193–213.
679 Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, *et al.* (2012). Accelerating the
680 domestication of trees using genomic selection: accuracy of prediction models across ages and
681 environments. *New Phytol* **193**: 617–624.
682 Spindel JE, Begum H, Akdemir D, Collard B, Redoña E, Jannink J, *et al.* (2016). Genome-wide
683 prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice
684 improvement. *Heredity (Edinb)* **116**: 395–408.
685 Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, *et al.* (2015). Genomic Selection and
686 Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training
687 Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic
688 Selection in Elite, Tropical Rice Breeding Lines. *PLoS Genet* **11**: 1–25.
689 Strobl C, Malley J, Tutz G (2009). An Introduction to Recursive Partitioning: Rationale, Application, and
690 Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychol*

691 *Methods* **14**: 323–348.
692 Turner SD (2014). *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots*.
693 VanRaden PM (2008). Efficient methods to compute genomic predictions. *J Dairy Sci* **91**: 4414–23.
694 Winter S, Koerbler M, Stein B, Pietruszka A, Paape M, Butgereitt A (2010). Analysis of cassava brown
695 streak viruses reveals the presence of distinct virus species causing cassava brown streak disease in
696 East Africa. *J Gen Virol* **91**: 1365–1372.
697 Wolfe MD, Rabbi IY, Egesi C, Hamblin M, Kawuki R, Kulakow P, *et al.* (2016). Genome-wide
698 association and prediction reveals the genetic architecture of cassava mosaic disease resistance and
699 prospects for rapid genetic improvement. *Plant Genome* **9**: 1–13.
700 Yang J, Lee SH, Goddard ME, Visscher PM (2011). GCTA: A tool for genome-wide complex trait
701 analysis. *Am J Hum Genet* **88**: 76–82.
702 Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, *et al.* (2014). Improving the accuracy of whole genome
703 prediction for complex traits using the results of genome wide association studies (X Cai, Ed.).
704 *PLoS One* **9**: e93017.
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721

Figures and supplementary legends

Figure 1. Principal components analysis of panel 1 and panel 2 clones. The top two panels and the lower left panel show the distribution of clones in PC1-PC3. The lower right panel shows the variance explained by the first ten principal components. Green color shows the distribution of panel 1 clones and the orange color shows the distribution of panel 2 clones.

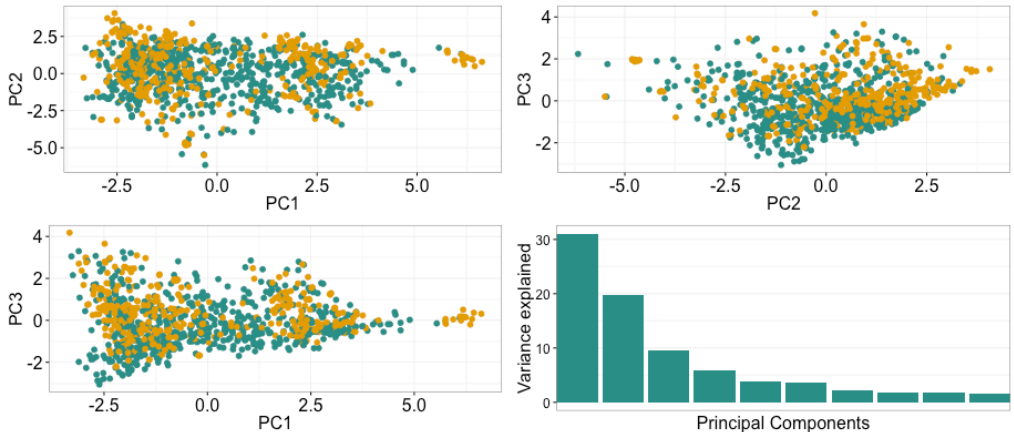
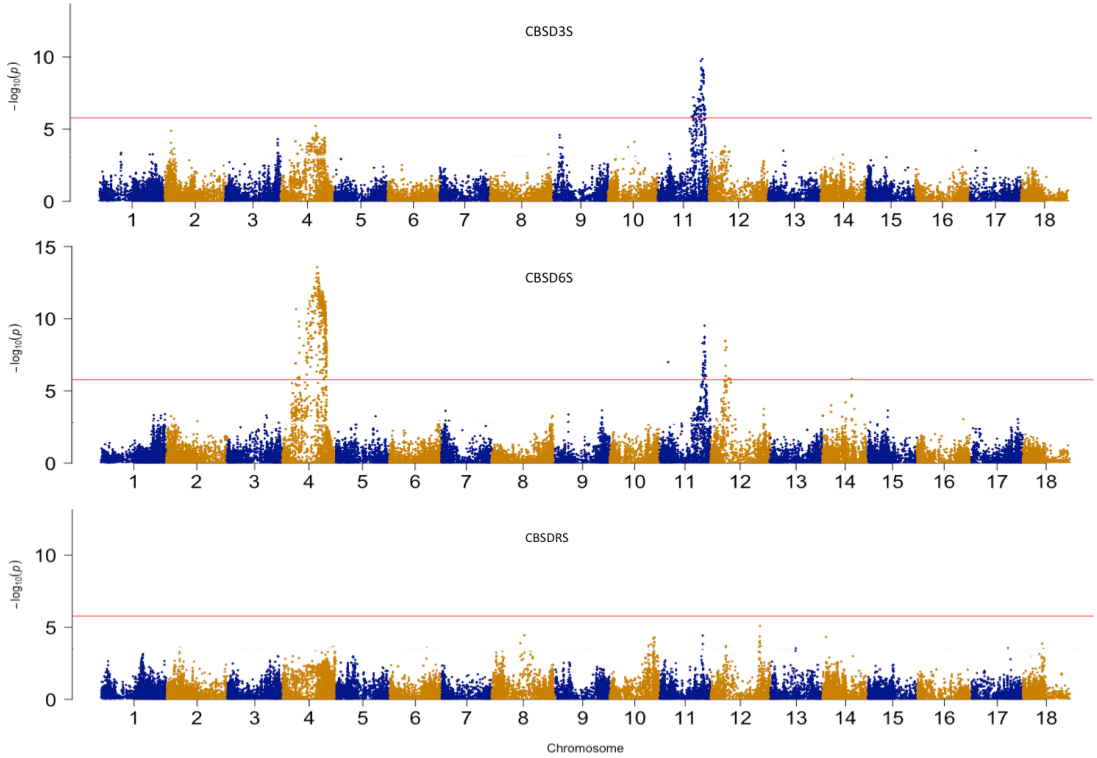
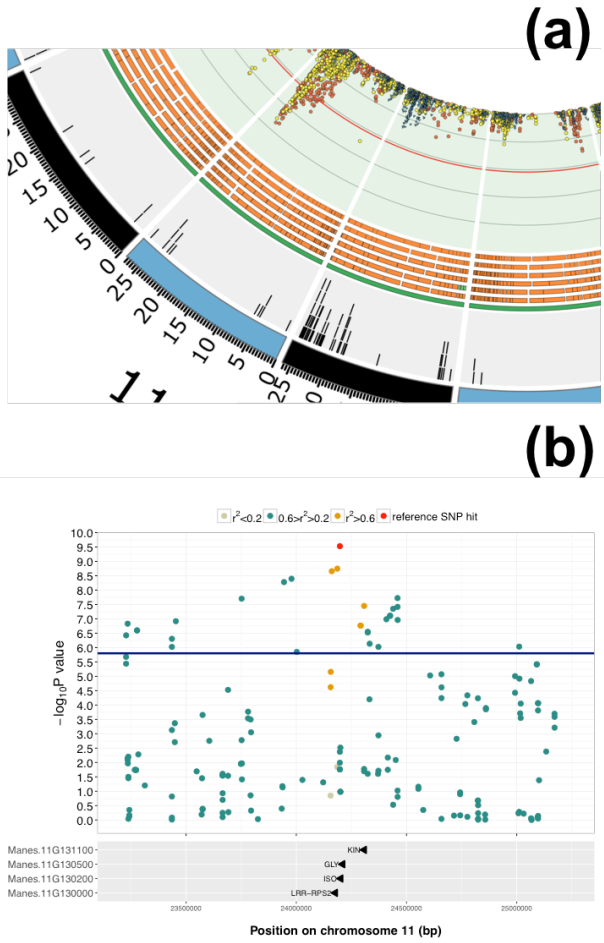


Figure 2. **GWAS results for CBSD severity** .Analysis was performed with a multilocation combined dataset of panels 1 and 2.(a) scoring 3 MAP (b) 6 MAP and (c) root necrosis severity. Red line indicates Bonferroni threshold.

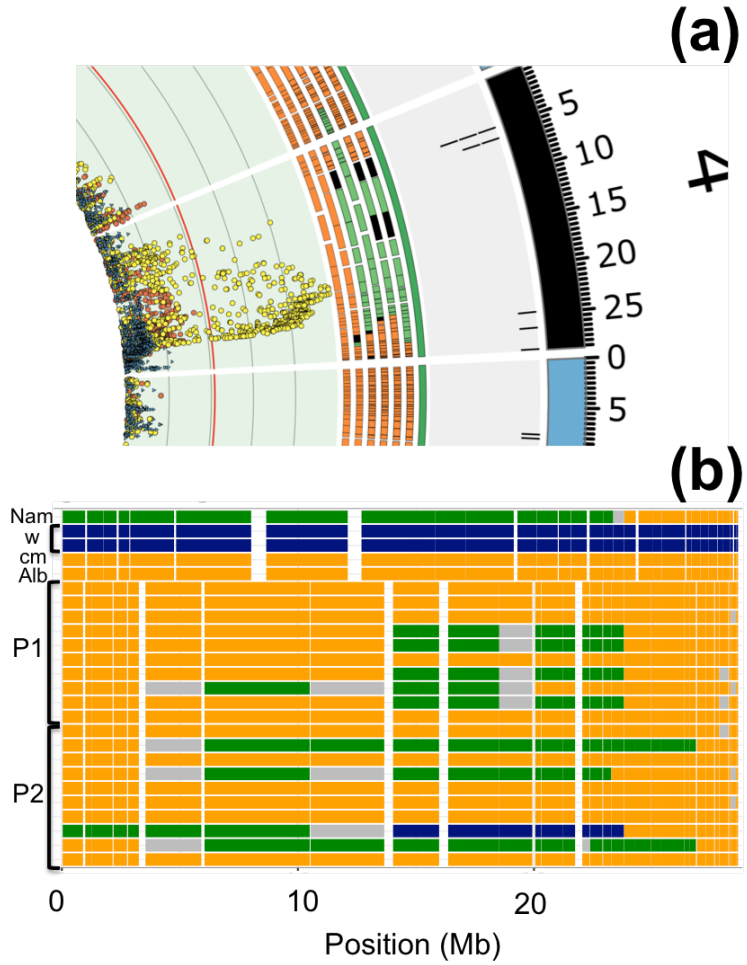


732 **Figure 3**. **Chromosome 11 region with QTL for CBSD severity** (a) 3 MAP (yellow), 6 MAP and root
733 necrosis (blue).Outer ring black lines indicate clusters of NBS-LRR genes (Lozano et al 2015).
734 Intermediate ring indicate regions homozygous (G/G)(blue) or heterozygous (G/E)(green) for *M. glaziovii*
735 allele and the proportion that were homozygous for the *M. esculenta* allele (E/E)(orange) on seven clones.
736 (b) LD association plot, 2 Mb region in chromosome 11, top SNP indicated in red, annotated genes within
737 that region are indicated in the panel below.



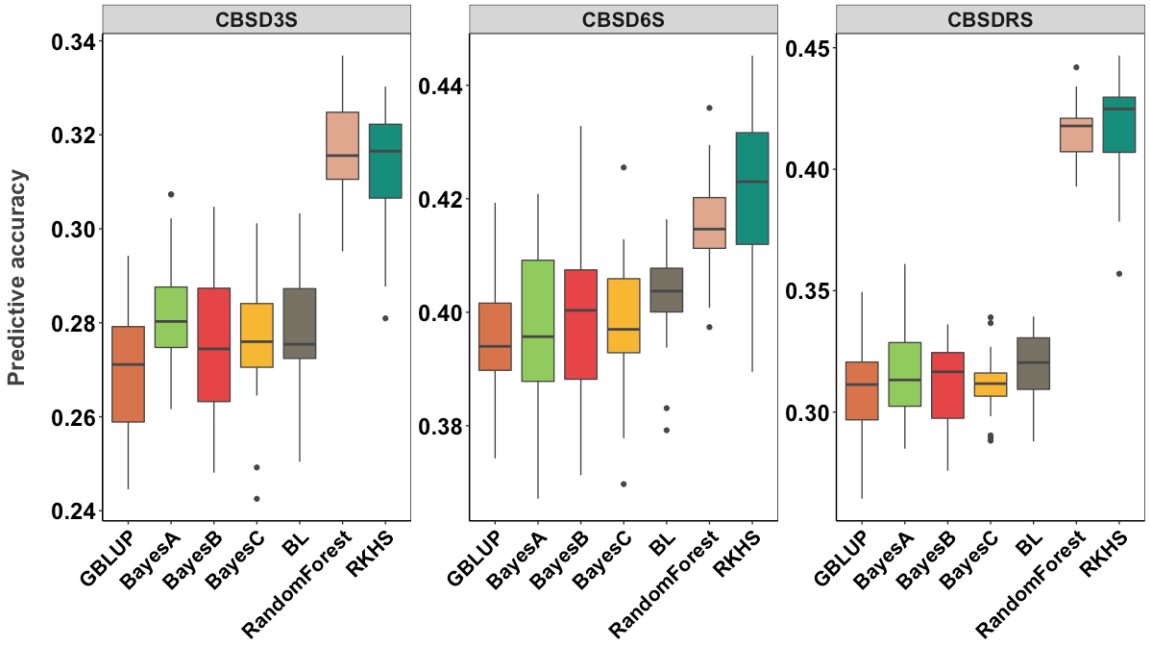
738
739
740
741
742
743
744
745

746 **Figure 4**. **Chromosome 4 region with QTL for CBSD severity with introgression segment** (a) 3 MAP
747 (yellow), 6 MAP and root necrosis (blue). Outer ring black lines indicate clusters of NBS-LRR genes
748 (Lozano et al 2015). Intermediate ring indicate regions homozygous (G/G)(blue) or heterozygous
749 (G/E)(green) for *M. glaziovii* allele and the proportion that were homozygous for the *M. esculenta* allele
750 (E/E)(orange) on seven clones. (b) Introgression region on chromosome 4 (colors description) are the
751 same as the aforementioned),Nam: Namikonga,w: wild *M. glaziovii*, cm: CM330645,Alb:Albert, P1:
752 panel 1 clones and P2 panel 2 clones.
753



754
755
756
757

Figure 5. **Cross validation results for CBSD severity**. 3 MAP (CBSD3S), 6 MAP (CBSD6S) and Root necrosis (CBSDRS). x-axis : predictive accuracy and y-axis : genomic prediction model.



776 **Table 1. Broad sense heritability (H^2) and SNP heritability (h^2) of foliar and root CBSD severity**.
777 Broad-sense heritability (H^2) values were calculated using the variance components obtained from a
778 model fitted using the *lmer* function from the lme4 R package. SNP heritability values were calculated
779 using the variance components obtained obtained from a model fitted using the EMMREML R package.
780 Heritability values estimates were calculated for sets 1 and 2 separately.
781

Trait	H^2	h^2	LOCATION-YEAR	Panel
CBSD3S	0.11	0.32	NAMULONGE	1
CBSD6S	0.31	0.39	NAMULONGE	1
CBSDRS	0.55	0.59	NAMULONGE	1
CBSD3S	0.43	0.48	NGETTA	1
CBSD6S	0.51	0.53	NGETTA	1
CBSDRS	0.73	0.72	NGETTA	1
CBSD3S	0.27	0.29	KASESE	1
CBSD6S	0.21	0.27	KASESE	1
CBSDRS	0.39	0.47	KASESE	1
CBSD3S	0.61	0.17	MULTI LOCATION	1
CBSD6S	0.35	0.31	MULTI LOCATION	1
CBSDRS	0.37	0.34	MULTI LOCATION	1
CBSD3S	0.60	0.37	NAMULONGE	2
CBSD6S	0.60	0.32	NAMULONGE	2
CBSD9S	0.68	0.34	NAMULONGE	2
CBSDRS	0.24	0.53	NAMULONGE	2
CBSD3S	0.63	0.28	SERERE	2
CBSD6S	0.60	0.28	SERERE	2
CBSD9S	0.73	0.34	SERERE	2
CBSDRS	0.15	0.48	SERERE	2
CBSD3S	0.56	0.27	KAMULI	2
CBSD6S	0.62	0.29	KAMULI	2
CBSD9S	0.75	0.34	KAMULI	2
CBSDRS	0.28	0.44	KAMULI	2
CBSD3S	0.42	0.28	MULTI LOCATION	2
CBSD6S	0.47	0.34	MULTI LOCATION	2
CBSD9S	0.56	0.38	MULTI LOCATION	2
CBSDRS	0.25	0.33	MULTI LOCATION	2

782
783
784

785 **Supplementary figure 1**. Cassava brown streak disease symptoms on leaves and roots of
786 sampled plants; Severity Score from 1 (no visible symptoms) to 5 (severely disease plants. **(a)**
787 leaf veins chlorosis severity progresses with severity score, **(b**) dark brown necrotic areas within
788 storage roots severity scale.
789
790 **Supplementary figure 2.**Panel 1 phenotypic distribution of CBSD severity traits.
791 (A) deregressed BLUPs distribution of CBSD 3 months foliar severity, (B) deregressed BLUPs
792 distribution of CBSD 6 months foliar severity, (C) deregressed BLUPs distribution of CBSD 12
793 months root severity
794
795 **Supplementary figure 3.**Panel 2 phenotypic distribution of CBSD severity traits.
796 (A) deregressed BLUPs distribution of CBSD 3 months foliar severity, (B) deregressed BLUPs
797 distribution of CBSD 6 months foliar severity, (C) deregressed BLUPs distribution of CBSD 9
798 months foliar severity, (D) deregressed BLUPs distribution of CBSD 12 months root severity
799
800 **Supplementary figure 4. Correlation plots between de-regressed BLUPs for foliar and root**
801 **symptoms**. De-regressed BLUPs were calculated for different locations in panel 1 and panel 2.
802
803 **Supplementary figure 5**. GWAS results for CBSD severity in panel 1 measure at Kasese.(a)
804 scoring CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity and (c) root necrosis
805 severity. Red line Bonferroni correction. Blue line $\log_{10}$ P-value = 3.8.
806
807 **Supplementary figure 6**. GWAS results for CBSD severity in panel 1 measure at Ngetta.(a)
808 scoring CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity and (c) root necrosis
809 severity . Red line Bonferroni correction. Blue line $\log_{10}$ P-value = 3.8.
810
811 **Supplementary figure 7**. GWAS results for CBSD severity in panel 1 measure at Namulonge.
812 (a) scoring CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity and (c) root
813 necrosis severity. Red line Bonferroni correction. Blue line $\log_{10}$ P-value = 3.8.

814 **Supplementary figure 8**. GWAS results for CBSD severity in with a multilocation dataset of
815 panel 1 (a) scoring CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity and (c)
816 root necrosis severity. Red line Bonferroni correction. Blue line $\log_{10}$ P-value = 3.8.
817
818 **Supplementary figure 9**. GWAS results for CBSD severity in panel 2 measure at Kamuli. (a)
819 scoring CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity (c) 9 CBSD 9
820 months foliar and (c) root necrosis severity. Red line Bonferroni correction. Blue line $\log_{10}$ P-
821 value = 3.8.
822
823 **Supplementary figure 10**. GWAS results for CBSD severity in panel 2 measure at Namulonge.
824 (a) Scoring CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity (c) 9 CBSD 9
825 months foliar and (c) root necrosis severity. Red line Bonferroni correction. Blue line $\log_{10}$ P-
826 value = 3.8.
827
828 **Supplementary figure 11**. GWAS results for CBSD severity in panel 2 at Serere. (a) Scoring
829 CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity (c) 9 CBSD 9 months foliar
830 and (c) root necrosis severity. Red line Bonferroni correction. Blue line $\log_{10}$ P-value = 3.8.
831
832 **Supplementary figure 12**. GWAS results for CBSD severity in with a multilocation dataset of
833 panel 2 (a) scoring CBSD 3 months foliar severity (b) 6 CBSD 6 months foliar severity (c) 9
834 CBSD 9 months foliar and (c) root necrosis severity. Red line Bonferroni correction. Blue line
835 $\log_{10}$ P-value = 3.8.
836
837 **Supplementary figure 13.** local LD in chromosome 4. Plot of the mean LD score for each
838 marker .With a smooth line representing a relative measure of the local LD in chromosome 4.
839 Dots are colored with the -$\log_{10}$ P-value for the association test for CBSD severity six months
840 after planting.
841
842

843 **Supplementary figure 14.** Introgressions segment detection. For each clone of the two GWAS
844 panels we calculated the proportion of genotypes that were homozygous (G/G) or heterozygous
845 (G/E) for *M. glaziovii* allele and the proportion that were homozygous for the *M. esculenta* allele
846 (E/E).
847
848 **Supplementary figure 15. (a)** GWAS results for 6MAP CBSD severity panels 1 and 2 **(b)**
849 GWAS Results after correction including markers in chromosome 12 as a covariate.
850
851 **Supplementary figure 16.** Multi-kernel GBLUP approach by fitting three kernels constructed
852 with non-overlapping SNPs (MAF> 0.01) from chromosomes 4, 11 and SNPs from the other
853 chromosomes. Crossvalidation GS predictive accuracies results for CBSD severity were
854 calculated using the multilocation dataset of the combined panels. Scoring CBSD 3 months
855 foliar severity (CBSD3S),CBSD 6 months foliar severity (CBSD6S) and root necrosis severity
856 (CBSDRS).
857
858 **Supplementary Table 1**.Pedigree information from GWAS panels 1 and 2. Details are shown
859 on the parental lines per clone and selected traits that came from the maternal side.
860
861 **Supplementary table 2**. Correlation values across locations in panel 1 and panel 2. (A)
862 Correlation of deregressed BLUPs across locations within traits in panel 1 measured in three
863 locations.(B) Correlation of deregressed BLUPs across locations within traits in panel 2
864 measured in three locations
865
866 **Supplementary table 3**. Correlation values across locations and traits in panel 1 and panel 2.
867 (A) Correlation of deregressed BLUPs across locations and traits in panel 1 measured in three
868 locations.(B) Correlation of deregressed BLUPs across locations and four traits in panel 2
869 measured in three locations
870

871 **Supplementary table 4**. Panel 1 and 2 and combined panels GWAS results. Gene annotation is
872 only shown for significant SNPs.
873
874 **Supplementary table 5**. Explained variance of phenotypic traits. Details are shown of the
875 reference SNP, the -log10(pval)(score),chromosome and explained variance.
876
877 **Supplementary table 6**.Genomic prediction accuracy values. (A) Cross validation results using
878 7 GS models for CBSD severity prediction of 3 MAP CBSD3S, 6 MAP CBSD6S and Root
879 necrosis (CBSDRS) (B) Multi-kernel GBLUP crossvalidation by fitting three kernels constructed
880 with non-overlapping SNPs (MAF> 0.01) from chromosomes 4, 11 and SNPs from the other
881 chromosomes. RKHS = Reproducing kernel Hilbert spaces regression, Total accuracy is the
882 accuracy obtained by following the GBLUP multikernel approach.
883
884
885
886
887
888
889