

1 **Prospects for genomic selection in cassava breeding.**

2

3 **Running head: Genomic selection in cassava**

4 Authors:

5 Marnin D. Wolfe\*, Dunia Pino Del Carpio\*, Olumide Alabi, Chiedozie Egesi, Lydia  
6 C. Ezenwaka, Ugochukwu N. Ikeogu, Robert S. Kawuki, Ismail S. Kayondo, Peter  
7 Kulakow, Roberto Lozano, Ismail Y. Rabbi, Esuma Williams, Alfred A. Ozimati,  
8 Jean-Luc Jannink

9

10 Marnin D. Wolfe, Dunia Pino Del Carpio, Ugochukwu N. Ikeogu, Roberto Lozano, Alfred A. Ozimati  
11 and Jean-Luc Jannink, Section on Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA;  
12 Olumide Alabi, Ismail Y. Rabbi and Peter Kulakow, International Institute for Tropical Agriculture  
13 (IITA), Ibadan, Oyo, Nigeria; Lydia C. Ezenwaka, Ugochukwu N. Ikeogu and Chiedozie Egesi,  
14 National Root Crops Research Institute (NRCRI), Umudike, Umuahia, Nigeria; Robert S. Kawuki,  
15 Ismail S. Kayondo, Esuma Williams and Alfred A. Ozimati, National Crops Resources Research  
16 Institute (NaCRRI), Namulonge, Uganda; Jean-Luc Jannink, USDA-ARS, R.W. Holley Center for  
17 Agriculture and Health, Ithaca, NY, USA.

18

19 Marnin D. Wolfe and Dunia Pino Del Carpio contributed equally to this work.

20 \*Corresponding authors Marnin Wolfe (wolfemd@gmail.com) and Dunia Pino Del  
21 Carpio (dpd64@cornell.edu).

22

23

24 Received \_\_\_\_\_

25

26

27 Abbreviations:

28 genomic selection (GS); genotype-by-sequencing (GBS); International Institute of  
29 Tropical Agriculture (IITA); National Root Crops Research Institute (NRCRI);  
30 National Crops Resources Research Institute (NaCRRI); genomic estimated breeding  
31 values (GEBVs); training population (TP); fresh root weight (RTWT), root number  
32 (RTNO); fresh shoot weight (SHTWT); harvest index (HI); dry matter (DM) content;  
33 cassava mosaic disease (CMD); mean CMD severity (MCMDS); early vigor  
34 (VIGOR).

35

36

37 **Key words:** genomic selection, genomic prediction, cassava, training population  
38 optimization

39

40

41

42 **ABSTRACT**

43

44 Cassava (*Manihot esculenta* Crantz) is a clonally propagated staple food crop  
45 in the tropics. Genomic selection (GS) reduces selection cycle times by the prediction  
46 of breeding value for selection of unevaluated lines based on genome-wide marker  
47 data. GS has been implemented at three breeding programs in sub-Saharan Africa.  
48 Initial studies provided promising estimates of predictive abilities in single  
49 populations using standard prediction models and scenarios. In the present study we  
50 expand on previous analyses by assessing the accuracy of seven prediction models for  
51 seven traits in three prediction scenarios: (1) cross-validation within each population,  
52 (2) cross-population prediction and (3) cross-generation prediction. We also evaluated  
53 the impact of increasing training population size by phenotyping progenies selected  
54 either at random or using a genetic algorithm. Cross-validation results were mostly  
55 consistent across breeding programs, with non-additive models like RKHS predicting  
56 an average of 10% more accurately. Accuracy was generally associated with  
57 heritability. Cross-population prediction accuracy was generally low (mean 0.18  
58 across traits and models) but prediction of cassava mosaic disease severity increased  
59 up to 57% in one Nigerian population, when combining data from another related  
60 population. Accuracy across-generation was poorer than within (cross-validation) as  
61 expected, but indicated that accuracy should be sufficient for rapid-cycling GS on  
62 several traits. Selection of prediction model made some difference across generations,  
63 but increasing training population (TP) size was more important. In some cases, using  
64 a genetic algorithm, selecting one third of progeny could achieve accuracy equivalent  
65 to phenotyping all progeny. Based on the datasets analyzed in this study, it was  
66 apparent that the size of a training population (TP) has a significant impact on  
67 prediction accuracy for most traits. We are still in the early stages of GS in this crop,  
68 but results are promising, at least for some traits. The TPs need to continue to grow  
69 and quality phenotyping is more critical than ever. General guidelines for successful  
70 GS are emerging. Phenotyping can be done on fewer individuals, cleverly selected,  
71 making for trials that are more focused on the quality of the data collected.

## 72 INTRODUCTION

73

74 Cassava (*Manihot esculenta* Crantz), a root crop with origins in the Amazon basin  
75 (Olsen and Schaal, 1999), provides staple food for more than 500 million people  
76 worldwide (Howeler et al., 2013). It is widely cultivated in Sub-Saharan Africa where  
77 the storage roots serve as primary source of carbohydrates and can be processed into a  
78 wide variety of products such as Fufu, Lafun, Gari, Abacha, Tapioca and starch  
79 (Chukwuemeka, 2007; Bamidele et al., 2015).

80 Cassava is a diploid ( $2n=36$ ) and highly heterozygous non-inbred crop that is  
81 propagated vegetatively by farmers using stem cuttings, though most genotypes do  
82 flower and can be used to produce botanical seeds from either self or cross-  
83 pollination. Among the most important traits targeted for improvement are storage  
84 root yield, dry matter content, starch content, tolerance to postharvest physiological  
85 deterioration, carotenoids content and resistance to pests/diseases (Esuma et al.,  
86 2016).

87 Development and implementation of breeding strategies in cassava represent a  
88 challenge due to the crop's heterozygous nature and long breeding cycle. A traditional  
89 cassava-breeding program relies on phenotypic characterization of mature plants that  
90 have been clonally propagated. Typically, cycles of selection take three to six years  
91 from seedling germination to multi-location yield trials and additional years are  
92 required for evaluation of promising genotypes before variety release (Figure 1).

93 Marker-assisted selection (MAS) has been effective in cassava for the  
94 selection of promising genotypes for resistance to cassava mosaic disease (Okogbenin  
95 et al., 2007; Ceballos et al., 2015; Parkes et al., 2015). However, the use of MAS is  
96 limited primarily to monogenic traits, which makes this method infeasible for  
97 complex traits (Dekkers and Hospital, 2002; Heffner et al., 2009a).

98 With the advent of next generation sequencing technologies, it is now  
99 affordable to profile single nucleotide polymorphic (SNP) markers genome-wide  
100 (Barabaschi et al., 2015), which can support the use of genomic selection (GS), a  
101 breeding method that uses such markers to predict breeding values of unevaluated  
102 individuals (Meuwissen et al., 2001). GS can optimize and accelerate pipelines for  
103 population improvement, variety development and release (Heffner et al., 2009b) with  
104 reduction in breeding time due to selection of parental genotypes with superior  
105 breeding values at seedling stage based on genotypes alone.

106 In general, GS models differ with respect to the assumptions they make about  
107 genetic architecture. While random-regression (RRBLUP) and genomic-BLUP  
108 (GBLUP) models assume an infinitesimal genetic architecture (nearly equal and small  
109 contribution of all genomic regions to the phenotype), Bayesian methods are available  
110 that alter that assumption (Gianola et al., 2009; Legarra et al., 2011; Habier et al.,  
111 2011). Evaluation of different GS models using non-simulated data indicates that  
112 prediction accuracy varies across species and traits (Heslot et al., 2012; Resende et al.,  
113 2012; Gouy et al., 2013; Charmet et al., 2014; Rutkoski et al., 2014; Cros et al.,  
114 2015).

115 Previous studies in cassava have estimated genetic parameters and evaluated  
116 prediction accuracy applying the GBLUP model with small training sets and low-  
117 density markers (Oliveira et al., 2012, 2014). Historical phenotypic data from the  
118 International Institute of Tropical Agriculture (IITA) combined with markers obtained  
119 from genotyping-by-sequencing (GBS) showed promising results for cassava  
120 breeding using genomic selection (Ly et al., 2013). In that study, the predictive ability  
121 (accuracy) measured as the correlation between predictive values and the phenotypic  
122 value ranged from 0.15 to 0.47 across traits (Ly et al., 2013).

123 There are ongoing efforts under the Next Generation Cassava Breeding  
124 (NextGen Cassava) project ([www.nextgencassava.org](http://www.nextgencassava.org)) to increase the rate of genetic  
125 improvement in cassava and unlock the full potential of cassava production. The  
126 project is currently in the early stages of implementing genomic selection at three  
127 African research institutes: the National Crops Resources Research Institute  
128 (NaCRRI) in Uganda, the National Root Crops Research Institute (NRCRI) and the  
129 IITA, both in Nigeria.

130 In the present study, we evaluated the potential of genomic selection as a  
131 breeding tool to increase rates of genetic gain in datasets associated with all three  
132 NextGen Cassava breeding programs. We assessed predictive ability by cross-  
133 validation within training population datasets for seven traits: dry matter (DM)  
134 content, fresh root weight (RTWT), root number (RTNO), shoot weight (SHTWT),  
135 harvest index (HI), severity of cassava mosaic disease (MCMDS) and plant vigor  
136 (VIGOR). We compared the performance of seven GS models for these traits in each  
137 of the breeding programs.

138 One important topic in genomic selection concerns the feasibility of prediction  
139 across generations and across training populations from different breeding

140 populations or programs. To maximize the rate of gain achievable by GS, prediction  
141 models will need to accurately rank unevaluated progenies rather than genotypes  
142 contemporary to the training population. It is well known that recombination and  
143 divergence associated with recurrent selection reduces the accuracy of across-  
144 generation prediction, making this kind of prediction a major challenge for genomic  
145 selection. Accuracies in these scenarios have not been previously estimated in  
146 cassava. Therefore, we tested accuracy of across-generation prediction using the IITA  
147 training population and two successive cycles of progenies that have been  
148 phenotyped. Similarly, given that previous results indicated only a small genetic  
149 differentiation among clones from different populations (Wolfe et al., 2016a), we  
150 tested whether combining information from different populations could increase  
151 prediction accuracy in the smaller populations.

152 Finally, in a typical scenario a GS program will phenotype all selected  
153 materials and a subset of the unselected material in order to update the training model.  
154 We further investigated the impact of phenotyping different size subsets of materials  
155 for TP update. We compared random subset selections to selections based on a  
156 training population optimization algorithm (Akdemir et al. 2015).

157 This study is a starting point for successful application of genomic selection in  
158 African cassava. Similar to other studies, factors such as trait heritability, prediction  
159 model and training population composition play an important role. For example, traits  
160 with higher heritability like DM are considered to be more likely to respond to  
161 selection and lead to larger genetic gain over cycles of selection (Kawano et al 1998,  
162 Ceballos et al 2015). Our results will serve to guide implementation strategies for GS  
163 in cassava breeding programs.

164

165

## 166 **MATERIALS & METHODS**

### 167 **Germplasm**

168 In this study, we analyzed data from the genomic selection programs at three African  
169 cassava breeding institutions: NaCRRI, NRCRI and IITA. Germplasm from NaCRRI  
170 included 411 clones descended from crosses among accessions from East Africa,  
171 West Africa and South America. The collection from NRCRI was made up of 899  
172 clones, 211 of them being in common with the IITA breeding germplasm. The  
173 remaining 688 clones were materials derived either in part or directly from the  
174 International Center for Tropical Agriculture (CIAT) in Cali, Columbia. Wolfe et al.  
175 (2016a) shows details of origins and pedigrees of the NaCRRI and NRCRI clones  
176 used in this study.

177 The primary IITA germplasm we have analyzed is also known as the Genetic  
178 Gain (GG) collection, which comprises 709 elite and historically important breeding  
179 clones and a few landraces that have been collected starting in the 1970's. These  
180 materials have also been previously described in Okechukwu and Dixon (2008), Ly et  
181 al. (2013) and Wolfe et al. (2016a).

182 In addition, two generations of GS progenies were analyzed (Figure 2). The  
183 first, GS cycle 1 (C1) comprised 2,890 clones, from 166 full-sib (FS) families with 85  
184 parents from the GG collection. Because successful crossing is a challenge in cassava,  
185 and in order to obtain the full set of desired matings among parents of C1, crossing  
186 blocks were planted in two successive years (2013 and 2014). In 2013, 79 parents  
187 produced 2,322 seedlings (135 FS families). In 2014, 17 parents, of which, 11 were  
188 re-used from the previous year and six were new parents from the GG collection, gave  
189 rise to an additional 568 seedlings (31 new FS families). C1 families have a mean size  
190 of 17.4 siblings (median 15, range 2 to 78).

191 Finally, in 2014, a crossing block was planted with 89 selected C1 parents and  
192 generated 1648 GS cycle 2 (C2) seedlings in 242 FS families. Cycle 2 families had a  
193 mean size of 6.8 individuals (median 6, range 1 to 20).

194

195

## 196 **Phenotyped traits**

197 In total, seven traits were analyzed in this study. Plant vigor (VIGOR) was recorded  
198 as 3 (low), 5 (medium) and 7 (high), one month after planting (1 MAP) at IITA and  
199 NRCRI and three MAP at NaCRRI. We used the across-season average cassava  
200 mosaic disease severity score (MCMDS) for our analyses. MCMDS is the mean of  
201 measurements taken at 1, 3 and 6 MAP, on a scale of 1 (no symptoms) to 5 (severe  
202 symptoms). DM was expressed as a percentage of dry root weight relative to fresh  
203 root weight (RTWT). At IITA, DM was measured by drying 100 g of fresh roots in an  
204 oven whereas at NRCRI and NaCRRI, the specific gravity method (Kawano et al.,  
205 1987) was used. RTWT and SHTWT were expressed in kilograms per plot, whereas  
206 HI was the proportion of total biomass per plot that is RTWT. Meanwhile, RTNO was  
207 the number of fresh roots harvested per plot.

208 The phenotyping trials analyzed in this study have been described in part in  
209 previous publications (Wolfe et al. 2016a; Wolfe et al. 2016b). However, complete  
210 details on the phenotyping trial design particular to this study are provided in  
211 **Supplementary Methods**. All phenotyping trials were conducted between 2013 and  
212 2015. NaCRRI clones were evaluated in three locations with different agro-ecological  
213 conditions in Uganda: Namulonge, Kasese and Ngetta. NRCRI clones were tested in  
214 three locations in Nigeria: Kano, Otobi and Umudike. Meanwhile, IITA clones were  
215 evaluated in four locations within Nigeria: Ibadan, Ikenne, Ubiaja and Mokwa.

216

## 217 **Two-stage genomic analyses**

218 Except where noted, a two-step approach was used to evaluate genomic prediction in  
219 this study. This approach was used to correct for the heterogeneity in experimental  
220 designs and increase computational efficiency. The first stage involved accounting for  
221 trial-design related variables using a linear mixed model.

222 For NaCRRI we fitted the model:  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{clone}\mathbf{c} + \mathbf{Z}_{range(loc.year)}\mathbf{r} +$   
223  $\mathbf{Z}_{block(range)}\mathbf{b} + \boldsymbol{\varepsilon}$ , where  $\boldsymbol{\beta}$  included a fixed effect for the population mean, the  
224 location-year combination and for plot-basis traits (RTWT, RTNO and SHTWT), the  
225 number of plants harvested per plot was included as a covariate; vector  $\mathbf{c}$  and  
226 corresponding incidence matrix  $\mathbf{Z}_{clone}$  represented a random effect for clone where  
227  $\mathbf{c} \sim N(0, \mathbf{I}\sigma_c^2)$ ;  $\mathbf{I}$  represented the identity matrix, while the range variable was nested in  
228 location-year-replication and was represented by the incidence matrix  $\mathbf{Z}_{range(loc.year)}$

229 and random effects vector  $r \sim N(0, \mathbf{I}\sigma_r^2)$ . Ranges were equivalent to a row or column  
230 along which plots were arrayed. Blocks were also modeled, with a block being a  
231 subset of a range. Block effects were nested in ranges and were incorporated as  
232 random with incidence matrix  $\mathbf{Z}_{\text{block}(\text{range})}$  effects vector  $b \sim N(0, \mathbf{I}\sigma_b^2)$ . Finally, the  
233 residuals  $\varepsilon$  were random, with  $\varepsilon \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$ .

234 The model for NRCRI was:  
235  $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_{\text{clone}}c + \mathbf{Z}_{\text{set}(\text{loc.year})}s + \mathbf{Z}_{\text{rep}(\text{set})}r + \mathbf{Z}_{\text{block}(\text{rep})}b + \varepsilon$ . Here,  $\mathbf{Z}_{\text{set}}$  was the  
236 incidence matrix corresponding to the random effect for the planting group (see  
237 above), which was nested in location-year, with  $s \sim N(0, \mathbf{I}\sigma_s^2)$ . Replication effects  
238 were nested in sets and treated as random with incidence matrix  $\mathbf{Z}_{\text{rep}(\text{set})}$  and effects  
239 vector  $r \sim N(0, \mathbf{I}\sigma_r^2)$ . Blocks were nested in replications, treated as random and  
240 represented by design matrix  $\mathbf{Z}_{\text{block}(\text{rep})}$  and effects vector  $b \sim N(0, \mathbf{I}\sigma_b^2)$ . The fixed  
241 effects for NRCRI included were the same as for NaCRRI, with the addition of a term  
242 for trial (i.e. TP1 and TP2; see above).

243 For IITA, data from all trials described above were fitted together using the  
244 following model:  $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_{\text{clone}}c + \mathbf{Z}_{\text{range}(\text{loc.year})}r + \varepsilon$ . The range effect was fit  
245 as random. The fixed effects were the same as those described for NaCRRI except  
246 the proportion of harvested plants (out of the total originally planted) was used instead  
247 of the number harvested as a cofactor. This was done to correct for differences in plot  
248 sizes.

249 BLUP ( $\hat{c}$ ) for the clone effect, which represents an estimate of the total genetic  
250 value (EGV) for each individual, was extracted. EGVs were de-regressed by dividing  
251 by their reliability  $(1 - \frac{\text{PEV}}{\sigma_c^2})$ , where PEV is the prediction error variances of the  
252 BLUP. The mixed models above were solved using the *lmer* function of *lme4* package  
253 (Bates et al., 2014) in R.

254 For downstream genomic evaluations, we used the de-regressed EGVs and  
255 weighted error variances according to Garrick et al. (2009), using one divided by the  
256 square root of  $\frac{1-H^2}{0.1 + \frac{1-r^2}{r^2}H^2}$ , where  $H^2$  is the proportion of the total variance explained by  
257 the clonal variance component,  $\sigma_c^2$ .

258

259

## 260 **Genotyping data**

261 Cassava collections described above were genotyped using GBS (Elshire et al. 2011)  
262 with the *ApeKI* restriction enzyme recommended by Hamblin and Rabbi (2014).  
263 SNPs were called using the TASSEL 5.0 GBS pipeline v2 (Glaubitz et al., 2014) and  
264 aligned to cassava reference genome, v6.1 (<http://phytozome.jgi.doe.gov>; ICGMC,  
265 2015). Genotype calls were only allowed when a minimum of two reads were present,  
266 otherwise the genotype was imputed (see below). Furthermore, the GBS data was  
267 filtered such that clones with >80% missing and markers with >60% missing  
268 genotype calls were removed. Markers with extreme deviation from Hardy-Weinberg  
269 equilibrium ( $X^2 > 20$ ) were also removed. Only biallelic SNP markers were  
270 considered for further analyses. We used a combination of custom scripts and  
271 common variant call file (VCF) (Danecek et al., 2011) manipulation tools to  
272 accomplish the above pipeline. Finally, imputation was conducted with Beagle v4.0  
273 (Browning & Browning, 2009). A total of 155,871 markers were obtained following  
274 the procedures described above. For genomic prediction in a given population/dataset,  
275 we further filtered out SNPs with a minor allele frequency (MAF) less than 0.01.

276

## 277 **Assessment of prediction accuracy by cross-validation**

278 In order to obtain unbiased estimates of prediction accuracy, we used a *k*-fold cross  
279 validation scheme (Kohavi, 1995). In brief, each breeding program dataset (NR, UG  
280 and GG) was split randomly into *k* = 5 fold mutually exclusive training and validation  
281 sets. The training set composed by four out of five of the folds was used to estimate  
282 marker effects for predictions. The estimated marker effects were used to predict the  
283 breeding value of validation set individuals. The process of fold assignment and  
284 genomic prediction was repeated 25 times for each model. For each repeat,  
285 predictions were accumulated from each individual when it was in the validation fold.  
286 Prediction accuracy was then calculated as the Pearson correlation (*cor* function in R)  
287 between the EGV and the accumulated predicted values for that repeat.

288

## 289 **Genomic prediction methods**

290 In this study, we compared the accuracy of genomic prediction using seven methods  
291 that are briefly described below. These methods differ in their assumptions about  
292 genetic architecture and whether the prediction being made represents a genome  
293 estimated breeding value (GEBV, which includes additive effects) or a genome

294 estimated total genetic value (GETGV, which includes additive plus non-additive  
295 effects). Prediction models were compared using several prediction scenarios  
296 (described in detail below), including 25 replications of 5-fold cross-validation, cross-  
297 generation and cross-population prediction.

298

299 **GBLUP.** Prediction with genomic BLUP (GBLUP) involves fitting a linear mixed  
300 model of the following form:  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \boldsymbol{\varepsilon}$ . Here,  $\mathbf{y}$  is a vector of the  
301 phenotype,  $\boldsymbol{\beta}$  is a vector of fixed, non-genetic effects with design matrix  $\mathbf{X}$ . The  
302 vector  $\mathbf{g}$  is a random effect, the best linear unbiased prediction (BLUP), which  
303 represents the GEBV for each individual.  $\mathbf{Z}$  is a design matrix pointing observations  
304 to genotype identities and  $\boldsymbol{\varepsilon}$  is a vector of residuals. The GEBV is obtained by  
305 assuming  $\mathbf{g} \sim N(0, \mathbf{K}\sigma_g^2)$ , where  $\sigma_g^2$  is the additive genetic variance and  $\mathbf{K}$  is the  
306 square, symmetric genomic realized relation matrix based on SNP marker dosages.  
307 The genomic relationship matrix used was constructed using the function *A.mat* in the  
308 R package rrBLUP (Endelman, 2011) and follows the formula of VanRaden (2008),  
309 method two. GBLUP predictions were made with the function *emmreml* in the R  
310 package EMMREML (Akdemir and Okeke, 2015).

311

312 **RKHS.** We made predictions using reproducing kernel Hilbert spaces (RKHS). The  
313 genomic relationship matrix used in the GBLUP model described above can be  
314 considered as a parametric (additive genetic) kernel function and exists as a special  
315 case of RKHS (Gianola and van Kaam, 2008; Morota and Gianola, 2014). For RKHS  
316 predictions, we used a mixed model of the same form as for GBLUP above. Unlike  
317 for GBLUP, we used a Gaussian kernel function:  $K_{ij} = \exp(-d_{ij}\theta)$ . Here,  $K_{ij}$  was  
318 the measured relationship between two individuals,  $d_{ij}$  was their euclidean genetic  
319 distance based on marker dosages and  $\theta$  was a tuning (sometimes called a  
320 “bandwidth”) parameter that determines the rate of decay of correlation among  
321 individuals. Because this is a nonlinear function, the kernels we used for RKHS could  
322 capture non-additive as well as additive genetic variation. Thus, the BLUPs from  
323 RKHS models represent GETGVs rather than GEBVs.

324 Because the optimal  $\theta$  must be determined, a range of values was tested in two  
325 ways. First, we did cross-validation with the following  $\theta$  values and selected the one  
326 with the best accuracy: 0.0000005, 0.000005, 0.00005, 0.0001, 0.0005, 0.001, 0.004,

327 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1 (Single kernel RKHS). Second, we used  
328 the *emmremlMultiKernel* function in the EMMREML package to fit a multiple-kernel  
329 model with six covariance matrices, with the following bandwidth parameters and  
330 allowed REML to find optimal weights for each: 0.0000005, 0.00005, 0.0005, 0.005,  
331 0.01, 0.05 (Multi-kernel RKHS).

332

333 **Bayesian Marker Regressions.** We tested four well-established Bayesian prediction  
334 models: BayesCpi (Habier et al., 2011), the Bayesian LASSO (BL; Park and Casella,  
335 2008), BayesA, and BayesB (Meuwissen et al., 2001). In ridge-regression (equivalent  
336 to GBLUP), marker effects are all shrunk by the same amount, because we assume  
337 they are all drawn from a normal distribution with the same variance. Further, all  
338 markers have nonzero effect and most have small effects, essentially assuming that  
339 the genetic architecture of the trait is infinitesimal. In contrast, the Bayesian models  
340 we tested allow for alternative genetic architectures by inducing differential shrinkage  
341 of marker effects. For BayesA and BL, all markers have nonzero effect but marker  
342 variances are drawn from scaled-t and double-exponential distributions respectively,  
343 which are both distributions with thicker tails and greater density at zero. BayesB and  
344 BayesCpi are variable selection models, because the marker variances come from a  
345 two-component mixture of a point mass at zero and either a scaled-t distribution  
346 (BayesB) or a normal distribution (BayesCpi). Fitting BayesB and BayesCpi begins  
347 by estimating a parameter  $\pi$ , the proportion of markers with nonzero effect. We  
348 performed Bayesian predictions with the R package BGLR (Pérez and De Los  
349 Campos, 2014). Following Heslot et al. (2012) and others, we ran BGLR for 10,000  
350 iterations, discarded the first 1000 iterations as burn-in and thinned to every 5<sup>th</sup>  
351 sample. Marker dosages were mean-centered for each training population before  
352 analysis. Convergence was confirmed visually in initial test runs using the *coda*  
353 package in R (Plummer et al., 2006).

354

355 **Random Forest.** Random forest (RF) is a machine learning method used widely in  
356 regression and classification (Breiman, 2001; Strobl et al., 2009). The use of RF  
357 regression with marker data has been shown to capture epistatic effects and has been  
358 successfully used for prediction of GETGV (Breiman, 2001; Motsinger-Reif et al.,  
359 2008; Michaelson et al., 2010; Heslot et al., 2012; Charmet et al., 2014; Sarkar et al.,  
360 2015; Spindel et al., 2015). In prediction, a random forest is a collection of  $r$

361 regression trees grown on a subset of the original dataset that is bootstrapped over  
362 observations and randomly sampled over predictors. Averaging the prediction over  
363 trees for validation observations then aggregates information. We used RF with the  
364 parameter, *n*tree set to 500 and the number of variables sampled at each split (*m*try)  
365 equal to 300. We implemented RF using the randomForest package in R (Liaw and  
366 Wiener, 2002). As in the Bayesian regressions, marker dosages were mean-centered  
367 before RF analysis.

368

### 369 **Comparison of models based on similarity of rankings**

370 In order to test for GS model similarities among breeding programs we clustered the  
371 GEBV output on a breeding program basis. GEBVs from each model were scaled and  
372 centered on a column basis, using the *scale* function in R, and were then used to  
373 construct a matrix of Euclidean distances between models. Distance matrices were  
374 used as an input for hierarchical clustering using the Ward criterion implemented in  
375 the *hclust* R function (Heslot et al., 2012).

376

### 377 **Across-generation genomic predictions**

378 Because nearly all of the IITA germplasm from C1 and C2 had been clonally  
379 evaluated, we were able to test the prospects for prediction of unevaluated progeny.  
380 We predicted all traits using all methods in four scenarios: GG predicts C1, GG  
381 predicts C2, C1 predicts C2, GG+C1 predicts C2. Unlike in the other predictions  
382 presented in this study, cross-generation predictions were done in a single step (raw  
383 phenotype and genomic data fit simultaneously). The exception was for RF, where  
384 correction for location and blocking factors is not supported. For RF prediction, we  
385 used the same de-regressed EGVs as for cross-validation. The software and  
386 parameters used were the same as already described. The design model is the same as  
387 described for IITA above.

388

### 389 **Training population update**

390 We evaluated the impact on cross-generation prediction accuracy of phenotyping  
391 different size subsets of the un-selected C1 (materials selected for crossing in each  
392 cycle were phenotyped, but unselected materials were not phenotyped in all cases).  
393 We selected subsets of C1 using two methods: randomly and with a genetic algorithm  
394 implemented in the R package STPGA (Akdemir et al., 2015).

395 STPGA uses an approximation of the mean prediction error variance (PEV)  
396 expected for a given set of training individuals in combination with a given set of test  
397 genotypes as a criterion (which does not require phenotype data) for selecting the  
398 “optimal” training set. The genetic algorithm implemented by STPGA is used to  
399 rapidly find the training set that minimized the selection criterion (mean PEV of the  
400 test set; Akdemir et al., 2015). In order to speed up computation, STPGA uses  
401 principal components rather than raw SNP markers as genetic predictors.

402 Parents selected for further recombination were cloned into a crossing block.  
403 This is the point at which additional, un-selected seedlings must be chosen for  
404 phenotyping in order to incorporate their data in the prediction of the eventual  
405 progeny that are produced. Since the next generation of progenies had not yet been  
406 produced, we targeted STPGA on the parents of C2 (PofC2). Figure 3 provides a  
407 schematic of genomic selection with training population update and optimization  
408 using STPGA. We constructed a genomic relationship matrix with only C1 (including  
409 the PofC2). We did PCA on the kinship matrix and took the first 100 principal  
410 components as genomic predictors. We ran 1000 iterations of the genetic algorithm 10  
411 times at each sample size. Sample sizes ranged from 200 to 2400 at increments of  
412 400 (Supplementary Table 1). Predictions at each sample size were then made with  
413 each of 10 random and 10 optimized training sets using GBLUP in two scenarios:  
414 either just the sample of C1 was used to train the model or the sample of C1 plus all  
415 of the GG were used.

416

#### 417 **Across-population genomic predictions**

418 We predicted all traits using all methods in three scenarios: GG (IITA Genetic Gain)  
419 +NR (NRCRI) predicts UG (NaCRRI), GG+UG predicts NR, NR+UG predicts GG  
420 (Supplementary Table 2A). Across-population predictions were made using the  
421 prediction models described above and were done following the two-step approach as  
422 also described above.

423 We selected optimized subsets of the combined datasets with a genetic  
424 algorithm implemented in the R package STPGA (Akdemir et al., 2015). Random  
425 subsets of the same size as the optimized subsets (300, 600, 900 and 1200) were  
426 selected for comparison between predictive accuracies. Predictions at each sample  
427 size were then made with each of 10 random and 10 optimized training sets using  
428 GBLUP.

## 429 RESULTS

430 After quality control and keeping only markers with >1% MAF, the datasets had  
431 between 70,010 and 78,212 SNP markers (Table 1). Principal component analysis  
432 (PCA) of the genomic relationship matrix indicated some genetic differentiation  
433 between Nigerian populations (GG and NR) and the Ugandan training population  
434 (UG; Figure S1a). In contrast, there was little differentiation between the NRCRI and  
435 IITA GG datasets, even when comparing only the non-overlapping clones. We also  
436 calculated  $F_{ST}$  between populations as implemented in *vcftools* (Danecek et al., 2011).  
437 In agreement with results from PCA,  $F_{ST}$  between GG and NR was only 0.008, but  
438 was 0.019 and 0.021 between the Ugandan and the Nigerian populations, GG and NR,  
439 respectively. There was a similar amount of genetic differentiation between the IITA  
440 C2 progeny and its grandparental GG population ( $F_{ST} = 0.02$ ) as there was between  
441 GG and UG (Table 1, Figure S1b).

442 The mean inbreeding coefficient ( $F$ ), as measured by the mean of the diagonal  
443 of the genomic relationship matrix, was similar for all populations, ranging from  
444 0.933 in GG to 0.965 in C1. The mean rate of heterozygous loci was also similar  
445 between populations, ranging from 0.15 to 0.17. There was no notable decrease in  
446 heterozygosity or increase in inbreeding coefficient from GG to C1 or from C1 to C2  
447 (Table 1; Figure S2).

448 In general, broad-sense heritability ( $H^2$ ) was highest in the C1 (mean 0.46  
449 across traits), lowest for NRCRI (mean 0.13) and similar for the IITA GG and  
450 NaCRRRI TPs. Averaging across populations,  $H^2$  was highest for MCMDS (0.57)  
451 followed by HI (0.43) and DM (0.39). However,  $H^2$  was generally low for yield  
452 components (Table 1).

453

### 454 Prediction within breeding populations

455 We tested seven genomic prediction models that differ by the extent and the kind of  
456 shrinkage, which is relevant to model different genetic architectures, and by their  
457 ability to capture non-additive effects (Figures S3-5).

458 Overall, breeding populations exhibited differences in the cross-validated  
459 prediction accuracies between methods and across traits. For NRCRI ( $n = 899$ ), the  
460 mean predictive accuracy values across methods ranged between -0.02 for plant vigor  
461 and 0.27 for HI. For NaCRRRI ( $n = 411$ ), the mean predictive accuracy values ranged

462 between 0.23 for shoot weight and 0.46 for HI. Meanwhile, the predictive accuracy  
463 values for GG (n = 709) ranged between 0.22 for plant vigor and 0.66 for DM.

464 In the NRCRI population, methods that capture non-additive effects like  
465 RKHS and random forest had the highest predictive accuracy values for all traits,  
466 except plant vigor. The trait with the highest predictive accuracy was root weight  
467 (Random forest (0.34)) and the lowest predictive accuracy was found for vigor  
468 (MultiKernel RKHS (-0.03)).

469 In the NaCRRRI population, RKHS Multikernel showed highest predictive  
470 accuracies for all traits except for CMD, for which BayesB showed the highest value  $r$   
471 = 0.50. In this population CMD had the overall highest predictive accuracy across  
472 traits while shoot weight exhibited the lowest predictive accuracy (Bayesian LASSO,  
473  $r=0.18$ ).

474 In the IITA GG population, Bayesian approaches performed better for vigor,  
475 CMD, shoot weight and DM, while RKHS method showed higher predictive  
476 accuracies for HI and for yield related traits such as root number and root weight.  
477 Meanwhile, RF gave a better predictive accuracy when used to estimate GEBVs.

478 Some trait-dataset combinations exhibited better predictive accuracies than  
479 others. For example, NaCRRRI population had better predictive accuracies for yield  
480 components like HI, root weight and root number while the highest predictive values  
481 for CMD and DM were obtained in the GG population.

482 Similar to Heslot et al. (2012), we compared the cross-validated GEBV  
483 following a clustering approach. Results in Figure S6 show the hierarchical cluster  
484 trees from the combined results of the three breeding populations. Differences in  
485 clustering of methods are observed across datasets (Figure 4). In the NRCRI data, we  
486 found two groups of clustering GS methods. With BayesB, BayesC and GBLUP in  
487 one group and the rest on the other group. In the NaCRRRI and IITA populations, non-  
488 parametric methods such as RKHS and Random Forest clustered together as well as  
489 the BayesA with Bayesian LASSO and GBLUP cluster with BayesC or BayesB.

490

#### 491 **Across-population prediction**

492 Previous studies have reported close relatedness between the clones in the NextGen  
493 training populations (Wolfe et al., 2016). One important question within this project is  
494 whether or not datasets from different breeding programs can be combined in a  
495 training set to increase predictive accuracy. The application of any prediction model

496 with the combined dataset would then benefit from an increase in the training  
497 population size with an outlook of using such models by other cassava breeding  
498 programs in Africa. With that in mind, we used combined datasets of GG+NR,  
499 GG+UG and UG+NR to predict the population that was not included in the training  
500 set UG, NR and GG respectively.

501 When predicting the traits in the UG dataset, with the combined GG+NR full  
502 set, Bayesian models gave better predictive accuracies for MCMDS, RTNO and DM.  
503 Random Forest gave better predictive accuracies for HI and RKHS for root weight  
504 and shoot weight (Table S2a).

505 The average predictive accuracy with the combined GG+NR full set as  
506 training set using the GBLUP model was consistently lower for all the traits when  
507 compared to the average GBLUP cross validation results (Table S2a). Furthermore,  
508 the subsets selected by STPGA to predict the NaCRRRI (UG) validation set gave, for  
509 all traits and all subset sizes, lower predictive accuracies than the GBLUP cross-  
510 validation model (Table 3; Figure S7; Table S2b).

511 For plant vigor, MCMDS and HI, the optimized STPGA subsets gave higher  
512 predictive accuracies than the combined GG+NR full training dataset. With few  
513 exceptions (MCMDS, SHTWT and DM) the optimized STPGA datasets gave better  
514 prediction accuracies than the same size random sets. As the optimized STPGA  
515 dataset increased in size, the predictive accuracy did not increase, except for root  
516 number where the highest predictive accuracy was found when the training population  
517 size was 1200.

518 When combined GG+UG full training dataset was used to predict the NRCRI  
519 training population, Random Forest and RKHS prediction models performed better  
520 for root weight, shoot weight, root number and plant vigor. Bayesian models gave  
521 better predictive accuracies for MCMDS and DM. For plant vigor, MCMDS and DM,  
522 the combined UG+GG full dataset gave better predictive accuracies than the GBLUP  
523 cross validation model (Figure S8; Table S2b). For prediction of the NRCRI training  
524 population, the optimized STPGA selected datasets gave better predictive accuracies  
525 for plant vigor, root weight, root number and shoot weight than the combined UG+  
526 GG full training dataset.

527 To predict the NRCRI training population for all traits except root number (at  
528 n=900 and n=1200) and CMD (n=900), the optimized datasets gave higher predictive  
529 accuracies than the random datasets. For plant vigor, CMD resistance and DM the

530 selection of optimized datasets with STPGA gave better predictive accuracies than the  
531 GBLUP cross validation model.

532 Among the STPGA datasets, the highest predictive accuracy was not always  
533 the result of an increase in training population size. For CMD resistance, the highest  
534 predictive accuracy was found, with the same value than the highest optimized size,  
535 for the smallest optimized dataset.

536 Predictive accuracy results of traits in the GG dataset using the full training set  
537 (UG+NR) varied across methods. Whereas Bayesian methods gave better predictive  
538 accuracy values for MCMD and plant vigor, RKHS performed better for DM, HI, root  
539 number and shoot weight. The combined (UG+NR) full training dataset for prediction  
540 of the GG population gave lower predictive accuracies than the GBLUP cross-  
541 validation model for all the traits. GBLUP cross-validation model also gave better  
542 predictive accuracies for all the traits than the random and optimized STPGA datasets.  
543 The optimized STPGA datasets gave better predictive accuracies compared to the  
544 random sets for all the traits except for plant vigor and for DM (optimized dataset n =  
545 900) (Figure S9; Table S2b). For all traits except MCMDS and DM, the optimized  
546 STPGA subsets gave higher predictive accuracies than the combined UG+NR full  
547 training dataset.

548 For all the cross population results, we tested if the optimized STPGA sets  
549 would do better than random with a binomial test, assuming independence of the  
550 comparisons. We compared how many times the prediction accuracy of STPGA was  
551 greater than random for all traits. We found that for the prediction of the NR and UG  
552 sets, the STPGA optimized sets perform better than the random sets. On the contrary,  
553 when applying the same comparison of the STPGA sets with the prediction with full  
554 sets, the latter had significantly higher number of full set greater than STPGA  
555 predictive accuracy results.

556 Additionally, we tested if there was differential enrichment in the optimized  
557 STPGA training set of any of the populations relative to the source sets. We found a  
558 significant enrichment of the GG population ( $p < 0.001$ ) in the STPGA of different  
559 sizes, for the prediction of NR set using GG+UG. Similarly, we found a significant  
560 enrichment of the NR population ( $p < 0.001$ ), in the STPGA of different sizes, for the  
561 prediction of the GG set using the UG-NR. On the contrary, we found no significant  
562 enrichment of any population in the STPGA optimized sets for the prediction of the  
563 UG population.

564

### 565 **Across-generation prediction**

566 One major area where analysis was needed concerned prediction across generations.  
567 Selections can be done at the seedling stage if GEBV can be predicted based on the  
568 previous generations and training data. Because nearly all of the IITA germplasm  
569 from C1 and C2 were clonally evaluated, we were able to use these data to assess the  
570 accuracy of genomic prediction on unevaluated genotypes of the next generation. In  
571 general, the accuracy of prediction across generation was greatest when predicting C2  
572 as evidenced by averaging across prediction models and traits for predictions trained  
573 either with C1 (mean  $0.19 \pm$  standard error  $0.02$ ) or GG+C1 ( $0.19 \pm 0.02$ ). The  
574 accuracy was lower on average when predicting C2 with GG ( $0.11 \pm 0.01$ ) compared  
575 to predicting C1 with GG ( $0.17 \pm 0.02$ ). Accuracy was lowest for both VIGOR and  
576 RTWT ( $0.06 \pm 0.005$ ) and highest for MCMDS ( $0.32 \pm 0.03$ ) and DM ( $0.38 \pm 0.01$ ).  
577 Most prediction models performed similarly as evidenced by the averaged accuracy  
578 across traits and training-test combinations with RF performing worst ( $0.08 \pm 0.01$ )  
579 and BayesA and BayesB performing best (both  $0.20 \pm 0.03$ ). For MCMDS, we found  
580 that prediction accuracy was greatest using BayesA and BayesB (Figure 5, Figure  
581 S10, Table S3).

582

### 583 **Training population update**

584 The first 100 PCs of the C1 kinship matrix were used as predictors for STPGA and  
585 explained 97.7% of the genetic variance. In all cases the genetic algorithm converged  
586 within the 1000 iterations run (Figure S11).

587         Given the constraints of breeding programs described above, it was necessary  
588 to select samples of C1 that were optimized for predicting the parents of C2 (PofC2),  
589 rather than the C2 themselves. Despite targeting the PofC2, we used selected training  
590 sets to predict C2, thus simulating the addition of phenotypes to the training set.  
591 Because of this, we compared the accuracy of subsets of C1 predicting C2 to accuracy  
592 predicting the PofC2. As the number sampled increased from 200 to 2,400, averaging  
593 across traits and methods for subset selection (STPGA and Random), accuracy  
594 increased by 120 and 105% when predicting C2 and PofC2, respectively. Accuracy  
595 increase was smaller when including the 709 GG clones in the prediction, increasing  
596 only by 43 and 36% respectively when predicting C2 and PofC2 (Supplementary  
597 Table 4).

598 STPGA consistently selected training datasets with lower expected mean PEV  
599 on the test set compared to random and across training set sizes (Figure S12). Further,  
600 using STPGA to select clones for phenotyping gave an average 13% better accuracy  
601 (average accuracy of 0.242 vs. 0.214, two-tailed  $t=6.29$ ,  $df=4458$ ,  $p<0.0001$ )  
602 compared to random sampling. Broken down by validation set, STPGA was  
603 significantly better than random predicting PofC2 ( $t=9.8$ ,  $df=2147$ ,  $p<0.0001$ ), but not  
604 significantly better for predicting C2 ( $t=1.41$ ,  $df=2227$ ,  $p=0.16$ ).

605 We compared these accuracies with that of the full set of C1 (or GG+C1) *and*  
606 to the cross-validation accuracy within the test set (C1 for prediction of PofC2, C2 for  
607 predictions of C2). When predicting C2, which was our primary goal, subsets were  
608 almost always inferior to the full set, with the exceptions of the middle sizes for  
609 RTWT, but the advantage was very small (Figure 6, Figure S13). However, STPGA-  
610 selected subsets tended to have better accuracy than the full set, especially for yield  
611 components when predicting the PofC2, which were the genotypes targeted by the  
612 optimization algorithm (Figure 7, Figure S14).

613 The correlation between the selection criterion, PEV<sub>mean</sub>, used by STPGA  
614 and the training set size is strong for all traits (range -0.57 to -0.61). Aside from  
615 simply increasing the TP size, we wanted to assess the extent to which the PEV<sub>mean</sub>  
616 could be used as a predictor of the achievable accuracy. Regression of prediction  
617 accuracies for each sample (regardless of whether it was selected randomly or by  
618 STPGA) on PEV<sub>mean</sub> explains between 8% (RTNO) and 46% (DM) of the variance  
619 in accuracy. Multiple regression including PEV<sub>mean</sub> and training set size (N<sub>train</sub>) as  
620 predictors showed PEV to be the more significant predictor (across all traits). In fact,  
621 N<sub>train</sub> was not a significant explanatory variable for RTWT or RTNO (Table S5).

622

623

624

## 625 **DISCUSSION**

626           The Next Generation Cassava Breeding Project ([www.nextgencassava.org](http://www.nextgencassava.org))  
627 aims to assess the potential of genomic selection in cassava to reduce the length of the  
628 breeding cycle and increase the number of crosses and selection per unit time. The  
629 project is implementing genomic selection in three breeding programs from Nigeria  
630 and Uganda, with genotypic and phenotypic data from training populations and two  
631 cycles of selection available on a database dedicated to cassava  
632 ([www.cassavabase.org](http://www.cassavabase.org)).

633           Using a cross-validation scheme, we contrasted the performance of GBLUP,  
634 RKHS (Single-kernel and Multi-kernel), BayesA, BayesB, BayesCpi, Bayesian  
635 LASSO and Random Forest for yield components (RTWT, RTNO, SHTWT, HI, DM)  
636 and CMD resistance data from the breeding programs.

637           In general, the performance of predictive models is known to be conditional  
638 on the genetic architecture of the trait under consideration (Daetwyler et al., 2010; Su  
639 et al., 2014). While non-additive models including RF and RKHS capture dominance  
640 and epistasis effects, GBLUP is more suitable for prediction when traits are  
641 determined by an infinite number of unlinked and non-epistatic loci, with small effect.

642           Not surprisingly, heritability varied between populations, conceivably as a  
643 consequence of the differences in the number and design of field trials between  
644 breeding programs. For most traits, it is not possible to determine exactly the reason  
645 for differences in heritability. However, for DM, we can hypothesize that differences  
646 in phenotyping protocols between programs (specific gravity method at NRCRI and  
647 NaCRRRI versus oven drying at IITA) could account for differences. We note the  
648 estimate of zero heritability for RTWT, RTNO and SHTWT in the IITA C2 and  
649 acknowledge this is likely to account for the quality of cross-generation prediction of  
650 that dataset.

651           Cross-validation results were mostly consistent across breeding programs and  
652 the superiority of one prediction method over the others was trait-dependent. RF and  
653 RKHS usually predicted phenotypes more accurately for yield-related traits, which  
654 are known to have a significant amount of non-additive genetic variation (Wolfe et al  
655 2016b). Similar findings have been made in wheat, for grain yield, an additive and  
656 epistatic trait, in which RKHS, radial basis function neural networks (RBFNN), and  
657 Bayesian regularized neural networks (BRNN) models clearly had a better predictive

658 ability than additive models like BL, Bayesian ridge-regression, BayesA, and BayesB  
659 (Perez-Rodriguez et al., 2013).

660 While cross validation results within breeding programs are encouraging for  
661 the use of genomic selection, across breeding program prediction values were fairly  
662 low. Mean  $F_{ST}$  values lower than 0.05 indicated that the three breeding populations  
663 share genetic material. Despite this, our results indicate that the prospect for sharing  
664 data across Africa to assist in genomic selection is limited to certain traits (most  
665 notably MCMDS) and populations. Indeed, obtaining a larger training set by  
666 combining training population did not always lead to higher prediction accuracies  
667 compared to what could already be achieved within that population as evidenced by  
668 cross-validation.

669 In animal models, prediction with multi-breed populations has also been  
670 shown to be poor with most of the observed accuracy due to population structure  
671 (Daetwyler et al., 2012). An alternative kernel function has been proposed to estimate  
672 the covariance between individuals based on markers, which can improve fit to the  
673 data to account for genetic heterogeneity of breeding populations (Heslot and Jannink,  
674 2015).

675 Conceivably, in our study the addition of individuals from different breeding  
676 programs was detrimental due to the inconsistent heritability for most traits. Another  
677 possibility is genotype-by-environment (GxE) interaction. The impact of GxE  
678 interaction on predictive accuracy has been reported in wheat when the same  
679 population was evaluated in different environments (Crossa et al., 2010; Endelman,  
680 2011). Similarly, in cassava using historical data from the IITA's GG population,  
681 prediction across locations led to a decrease in accuracy (Ly et al., 2013).

682 Using the training sets selected based on optimized algorithm gave better  
683 predictive ability than randomly assigned samples with a decrease in accuracy when  
684 compared with GBLUP cross-validation results. Although in previous studies  
685 predictive accuracies with full sets were lower than optimized subsets (Rutkoski et al.,  
686 2015), in our study we found the contrary, indicating that a larger training set was  
687 more advantageous. Combining data from different experiments and populations for  
688 across population prediction remains promising for traits like CMD where GWAS  
689 results indicate a stable large-effect QTL throughout the tested breeding populations  
690 (Wolfe et al., 2016).

691           When predicting unevaluated progenies from the next generation (cross  
692 generation), our results indicated, in our judgment, that accuracy should be sufficient  
693 for DM, MCMDS and to a lesser extent HI. Although accuracy is stable across the  
694 generations tested for DM using most models, for MCMDS to be successful, we  
695 recommend using a Bayesian shrinkage model such as BayesA or BayesB. The  
696 advantage of these models for CMD resistance over GBLUP likely comes because of  
697 the major known QTL segregating in the population (Rabbi et al., 2014; Wolfe et al.,  
698 2016a) and the ability of these two models to allow differential contribution of  
699 markers near the QTL to the prediction. One disadvantage of BayesB, in particular, is  
700 that the known polygenic background resistance for CMD may become de-  
701 emphasized, in favor of heavy selection on the major effect gene(s) (Hahn et al.,  
702 1980; Legg and Thresh, 2000; Akano et al., 2002; Rabbi et al., 2014; Wolfe et al.,  
703 2016).

704           We noted that RF and RKHS performed poorly across generations; this is a  
705 result that makes sense given that the predictability of epistatic and dominant  
706 interactions declines with recombination (Lynch and Walsh, 1998).

707           Based on the datasets analyzed in this study, it was apparent that the size of a  
708 training population had a significant impact on prediction accuracy for most traits.  
709 Thus, breeding programs will benefit from phenotyping the maximum possible  
710 amount. In agreement with the results in other crops (Rincent et al., 2012; Akdemir et  
711 al., 2015; Isidro et al., 2015), our results do indicate that optimization algorithms like  
712 STPGA can provide at least a small advantage over random selections of materials for  
713 phenotyping.

714           Each breeding program will need to determine the amount of phenotyping vs.  
715 genotyping to do in order to maximize prediction accuracy and selection gain based  
716 on the cost and availability of land, labor and genotyping. An analysis in barley by  
717 Endelman et al. (2014) provides a good example of the potential complexity of these  
718 decisions. The authors show, as we do, that larger number of phenotyped individuals  
719 is always beneficial, and that it is usually beneficial to focus on evaluating new lines  
720 at the expense of additional phenotyping of old lines. However, if genotyping costs  
721 are high, the cost-benefit balance shifts towards more evaluation of existing lines  
722 (Endelman et al., 2014). Endelman et al.'s (2014) study focused on prediction in  
723 biparental populations. Although this is likely to apply to cassava breeding

724 populations, we stress the necessity of doing such an analysis for each breeding  
725 application separately.

726 An important result is that STPGA was able to find subsets that were better than  
727 the full set for predicting the parents of C2 (PofC2). PofC2 are members of C1 and  
728 were the individuals targeted with STPGA. One possible interpretation is that the  
729 benefit comes from phenotyping contemporaries. If that were true, we could make a  
730 significant difference in accuracy by phenotyping a subset of clones from the current  
731 generation before predicting GEBV for the entire set of selection candidates. To do  
732 this without lengthening the selection and recombination cycle, harvested stems  
733 would need to be stored long enough for phenotypic data to be curated, predictions  
734 and selections to be conducted and STPGA to be run. Methods to store cassava stakes  
735 for up to 30 days are available, indicating such a scheme could be possible  
736 (Sungthongw et al., 2016). Even without improved stem cutting storage, this could be  
737 done while only lengthening the selection and recombination cycle to perhaps 1.5-2  
738 years, which would still be significantly faster than conventional cassava breeding.

739 A related possibility is to place annual selection pressure on traits that are  
740 predictable across generation (e.g. MCMDS, HI and DM). Predictions of total genetic  
741 value for yield traits for selection of clones that will be tested as potential varieties  
742 could then be done after clonal evaluation data become available on at least a subset  
743 of contemporary genotypes. Further trials will be necessary to determine whether  
744 there is an advantage to this type of strategy.

745 The primary promise genomic selection offers to cassava breeding is the  
746 ability to select and recombine germplasm more frequently and thus hopefully speed  
747 the rate of population improvement while combining a myriad of quality, disease and  
748 yield related traits into a single genotype that can be released as a variety. The  
749 applicability of results from the different prediction models in cassava is then  
750 dependent on whether the goal is the prediction of breeding value of progeny or the  
751 selection of advanced lines for testing as varieties.

752 We are still in the early stages of GS in this crop, but results are promising, at  
753 least for some traits. The TPs need to continue to grow and quality phenotyping is  
754 more critical than ever. However, general guidelines for successful GS are emerging.  
755 Phenotyping can be done on fewer individuals, cleverly selected, making for trials  
756 that are more focused on the quality of the data collected.

757

## 758 **ACKNOWLEDGEMENTS**

759 We acknowledge the Bill & Melinda Gates Foundation and UKaid (Grant  
760 1048542; <http://www.gatesfoundation.org>) and support from the CGIAR  
761 Research Program on Roots, Tubers and Bananas (<http://www.rtb.cgiar.org>).  
762 We give special thanks to A. G. O. Dixon for his development of many of the  
763 breeding lines and historical data we analyzed. Thanks also to A. I. Smith and  
764 technical teams at IITA, NRCRI and NaCRRI for collection of phenotypic data and  
765 to A. Agbona, P. Peteti, A. Ogbonna, E. Uba and R. Mukisa for data curation.

766

## 767 **CONFLICTS OF INTEREST**

768 No conflicts.

769

## 770 **REFERENCES**

- 771 (ICGMC), I.C.G.M.C. 2015. High-Resolution Linkage Map and Chromosome-Scale  
772 Genome Assembly for Cassava (*Manihot esculenta* Crantz) from Ten  
773 Populations. *G3* 5(1): 133–144 Available at  
774 <http://g3journal.org/cgi/doi/10.1534/g3.114.015008>.  
775 Akano, O., O. Dixon, C. Mba, E. Barrera, and M. Fregene. 2002. Genetic mapping of  
776 a dominant gene conferring resistance to cassava mosaic disease. *Theor. Appl.*  
777 *Genet.* 105(4): 521–525 Available at  
778 <http://www.ncbi.nlm.nih.gov/pubmed/12582500> (verified 29 October 2013).  
779 Akdemir, D., and U.G. Okeke. 2015. EMMREML: Fitting Mixed Models with  
780 Known Covariance Structures.  
781 Akdemir, D., J.I. Sanchez, and J.-L. Jannink. 2015. Optimization of genomic selection  
782 training populations with a genetic algorithm. *Genet. Sel. Evol.* 47(1):  
783 38 Available at <http://www.gsejournal.org/content/47/1/38>.  
784 Bamidele, O.P., M.B. Fasogbon, D.A. Oladiran, and E.O. Akande. 2015. Nutritional  
785 composition of *fufu* analog flour produced from Cassava root (*Manihot*  
786 *esculenta*) and Cocoyam (*Colocasia esculenta*) tuber. *Food Sci. Nutr.* 3(6):  
787 597–603 Available at <http://doi.wiley.com/10.1002/fsn3.250>.  
788 Barabaschi, D., A. Tondelli, F. Desiderio, A. Volante, P. Vaccino, G. Vale, and L.  
789 Cattivelli. 2015. Next generation breeding. *Plant Sci.* 242: 3–13 Available at  
790 <http://dx.doi.org/10.1016/j.plantsci.2015.07.010>.  
791 Bates, D., M. Maechler, B. Bolker, and S. Walker. 2014. Fitting Linear Mixed-Effects  
792 Models Using lme4. *J. Stat. Softw.* 67(1): 1–48 Available at  
793 <http://arxiv.org/abs/1406.5823>.  
794 Breiman, L. 2001. Random forests. *Mach. Learn.* 45(1): 5–32.

- 795 Ceballos, H., R.S. Kawuki, V.E. Gracen, G.C. Yench, and C.H. Hershey. 2015.  
796 Conventional breeding, marker-assisted selection, genomic selection and  
797 inbreeding in clonally propagated crops: a case study for cassava. *Theor. Appl.*  
798 *Genet.* Available at <http://link.springer.com/10.1007/s00122-015-2555-4>.
- 799 Charmet, G., E. Storlie, F.X. Oury, V. Laurent, D. Beghin, L. Chevarin, A. Lapiere,  
800 M.R. Perretant, B. Rolland, E. Heumez, L. Duchalais, E. Goudemand, J. Bordes,  
801 and O. Robert. 2014. Genome-wide prediction of three important traits in bread  
802 wheat. *Mol. Breed.* 34(4): 1843–1852.
- 803 Chukwuemeka, O.C. 2007. Effect of process modification on the physio-chemical and  
804 sensory quality of fufu-flour and dough. *Africa J. Biotechnol.* 6(August): 1949–  
805 1953.
- 806 Cros, D., M. Denis, L. Sánchez, B. Cochard, A. Flori, T. Durand-Gasselin, B. Nouy,  
807 A. Omoré, V. Pomiès, V. Riou, E. Suryana, and J.M. Bouvet. 2015. Genomic  
808 selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis*  
809 *guineensis* Jacq.). *Theor. Appl. Genet.* 128(3): 397–410.
- 810 Crossa, J., G. d. I. Campos, P. Perez, D. Gianola, J. Burgueno, J.L. Araus, D.  
811 Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J.  
812 Braun. 2010. Prediction of Genetic Values of Quantitative Traits in Plant  
813 Breeding Using Pedigree and Molecular Markers. *Genetics* 186(2): 713–  
814 724 Available at <http://www.genetics.org/cgi/doi/10.1534/genetics.110.118521>.
- 815 Daetwyler, H.D., K.E. Kemper, J.H. van der Werf, and B.J. Hayes. 2012.  
816 Components of the accuracy of genomic prediction in a multi-breed sheep  
817 population. *J. Anim. Sci.* 90: 3375–3384 Available at  
818 [file:///C:/Users/juan/Downloads/Conocimientos sobre tuberculosis en agentes](file:///C:/Users/juan/Downloads/Conocimientos sobre tuberculosis en agentes comunitarios de salud en Tacna, Perú.pdf)  
819 [comunitarios de salud en Tacna, Perú.pdf](file:///C:/Users/juan/Downloads/Conocimientos sobre tuberculosis en agentes comunitarios de salud en Tacna, Perú.pdf).
- 820 Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The  
821 impact of genetic architecture on genome-wide evaluation methods. *Genetics*  
822 185(3): 1021–1031.
- 823 Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E.  
824 Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, and R. Durbin.  
825 2011. The variant call format and VCFtools. *Bioinformatics* 27(15): 2156–2158.
- 826 Dekkers, J.C.M., and F. Hospital. 2002. The use of molecular genetics in the  
827 improvement of agricultural populations. *Nat. Rev. Genet.* 3(1): 22–32 Available  
828 at <http://dx.doi.org/10.1038/nrg701>.
- 829 Elshire, R.J., J.C. Glaubitz, Q. Sun, J. a Poland, K. Kawamoto, E.S. Buckler, and S.E.  
830 Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for  
831 high diversity species. *PLoS One* 6(5): e19379 Available at  
832 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract)  
833 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087801&tool=pmcentrez&rendertype=abstract) (verified 21 May 2013).
- 834 Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection  
835 with R Package rrBLUP. *Plant Genome J.* 4(3): 250 Available at  
836 <https://www.crops.org/publications/tpg/abstracts/4/3/250> (verified 22 July 2014).
- 837 Endelman, J.B., G.N. Atlin, Y. Beyene, K. Semagn, X. Zhang, M.E. Sorrells, and J.L.  
838 Jannink. 2014. Optimal design of preliminary yield trials with genome-wide  
839 markers. *Crop Sci.* 54(1): 48–59.
- 840 Esuma, W., L. Herselman, M.T. Labuschagne, P. Ramu, F. Lu, Y. Baguma, E.S.  
841 Buckler, and R.S. Kawuki. 2016. Genome-wide association mapping of  
842 provitamin A carotenoid content in cassava. *Euphytica* Available at  
843 <http://link.springer.com/10.1007/s10681-016-1772-5>.
- 844 Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding

- 845 values and weighting information for genomic regression analyses. *Genet. Sel.*  
846 *Evol.* 41: 55.
- 847 Gianola, D., and J.B.C.H.M. van Kaam. 2008. Reproducing kernel hilbert spaces  
848 regression methods for genomic assisted prediction of quantitative traits.  
849 *Genetics* 178(4): 2289–303 Available at  
850 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2323816&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2323816&tool=pmcentrez&rendertype=abstract)  
851 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2323816&tool=pmcentrez&rendertype=abstract) (verified 4 August 2014).
- 852 Gianola, D., G. de los Campos, W.G. Hill, E. Manfredi, and R. Fernando. 2009.  
853 Additive genetic variability and the Bayesian alphabet. *Genetics* 183(1): 347–  
854 63 Available at  
855 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2746159&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2746159&tool=pmcentrez&rendertype=abstract)  
856 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2746159&tool=pmcentrez&rendertype=abstract) (verified 15 July 2014).
- 857 Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S.  
858 Buckler. 2014. TASSEL-GBS: a high capacity genotyping by sequencing  
859 analysis pipeline. *PLoS One* 9(2): e90346 Available at  
860 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3938676&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3938676&tool=pmcentrez&rendertype=abstract)  
861 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3938676&tool=pmcentrez&rendertype=abstract) (verified 10 July 2014).
- 862 Gouy, M., Y. Rousselle, D. Bastianelli, P. Lecomte, L. Bonnal, D. Roques, J.C. Efile,  
863 S. Rocher, J. Daugrois, L. Toubi, S. Nabeneza, C. Hervouet, H. Telismart, M.  
864 Denis, A. Thong-Chane, J.C. Glaszmann, J.Y. Hoarau, S. Nibouche, and L.  
865 Costet. 2013. Experimental assessment of the accuracy of genomic selection in  
866 sugarcane. *Theor. Appl. Genet.* 126: 2575–2586.
- 867 Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the  
868 bayesian alphabet for genomic selection. *BMC Bioinformatics* 12(1):  
869 186 Available at  
870 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3144464&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3144464&tool=pmcentrez&rendertype=abstract)  
871 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3144464&tool=pmcentrez&rendertype=abstract) (verified 4 August 2014).
- 872 Hahn, S., A. Howland, and E. Terry. 1980. CORRELATED RESISTANCE OF  
873 CASSAVA TO MOSAIC AND BACTERIAL BLIGHT DISEASES. *Euphytica*  
874 29: 305–311 Available at <http://link.springer.com/article/10.1007/BF00025127>  
875 (verified 30 July 2014).
- 876 Hamblin, M.T., and I.Y. Rabbi. 2014. The Effects of Restriction-Enzyme Choice on  
877 Properties of Genotyping-by-Sequencing Libraries: A Study in Cassava (). *Crop*  
878 *Sci.* 54(6): 2603 Available at  
879 <https://www.crops.org/publications/cs/abstracts/54/6/2603> (verified 1 December  
880 2014).
- 881 Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009a. Genomic Selection for Crop  
882 Improvement. *Crop Sci.* 49(1): 1 Available at  
883 <https://www.crops.org/publications/cs/abstracts/49/1/1> (verified 28 May 2013).
- 884 Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009b. Genomic Selection for Crop  
885 Improvement. *Crop Sci.* 49(1): 1 Available at  
886 <https://www.crops.org/publications/cs/abstracts/49/1/1> (verified 19 September  
887 2013).
- 888 Heslot, N., and J.-L. Jannink. 2015. An alternative covariance estimator to investigate  
889 genetic heterogeneity in populations. *Genet. Sel. Evol.* 47(1): 93 Available at  
890 <http://www.gsejournal.org/content/47/1/93>.
- 891 Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic selection in  
892 plant breeding: A comparison of models. *Crop Sci.* 52(February): 146–  
893 160 Available at <https://www.crops.org/publications/cs/abstracts/52/1/146>.
- 894 Howeler, R., N. Litaladio, and G. Thomas. 2013. Save and Grow: Cassava. A Guide

- 895 to Sustainable Production Intensification.
- 896 Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot, and M.E. Sorrells. 2015.
- 897 Training set optimization under population structure in genomic selection. *Theor.*
- 898 *Appl. Genet.* 128: 145–158.
- 899 Kawano, K., W.M.G. Fukuda, and U. Cempukdee. 1987. Genetic and Environmental
- 900 Effects on Dry Matter Content of Cassava Root1. *Crop Sci.* 27(1): 69Available
- 901 at <https://dl.sciencesocieties.org/publications/cs/abstracts/27/1/CS0270010069>.
- 902 Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation
- 903 and Model Selection. *Int. Jt. Conf. Artif. Intell.* 14(12): 1137–1143.
- 904 Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz. 2011.
- 905 Improved Lasso for genomic selection. *Genet. Res. (Camb).* 93(1): 77–87.
- 906 Legg, J.P., and J.M. Thresh. 2000. Cassava mosaic virus disease in East Africa: a
- 907 dynamic disease in a changing environment. *Virus Res.* 71(1–2): 135–
- 908 49Available at <http://www.ncbi.nlm.nih.gov/pubmed/11137168>.
- 909 Liaw, a, and M. Wiener. 2002. Classification and Regression by randomForest. *R*
- 910 *news* 2(December): 18–22.
- 911 Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch, R. Okechukwu,
- 912 A.G.O. Dixon, P. Kulakow, and J.-L. Jannink. 2013. Relatedness and Genotype
- 913 × Environment Interaction Affect Prediction Accuracies in Genomic Selection:
- 914 A Study in Cassava. *Crop Sci.* 53(4): 1312Available at
- 915 <https://www.crops.org/publications/cs/abstracts/53/4/1312> (verified 20
- 916 September 2013).
- 917 Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits.
- 918 Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic
- 919 value using genome-wide dense marker maps. *Genetics* 157(4): 1819–
- 920 29Available at
- 921 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461589&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461589&tool=pmcentrez&rendertype=abstract)
- 922 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461589&tool=pmcentrez&rendertype=abstract).
- 923 Michaelson, J.J., R. Alberts, K. Schughart, and A. Beyer. 2010. Data-driven
- 924 assessment of eQTL mapping methods. *BMC Genomics* 11: 502Available at
- 925 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2996998&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2996998&tool=pmcentrez&rendertype=abstract)
- 926 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2996998&tool=pmcentrez&rendertype=abstract).
- 927 Morota, G., and D. Gianola. 2014. Kernel-based whole-genome prediction of
- 928 complex traits: a review. *Front. Genet.* 5(October): 1–13Available at
- 929 [http://www.frontiersin.org/Statistical\\_Genetics\\_and\\_Methodology/10.3389/fgen](http://www.frontiersin.org/Statistical_Genetics_and_Methodology/10.3389/fgen)
- 930 [e.2014.00363/abstract](http://www.frontiersin.org/Statistical_Genetics_and_Methodology/10.3389/fgen) (verified 22 October 2014).
- 931 Motsinger-Reif, A.A., S.M. Dudek, L.W. Hahn, and M.D. Ritchie. 2008. Comparison
- 932 of approaches for machine-learning optimization of neural networks for
- 933 detecting gene-gene interactions in genetic epidemiology. *Genet. Epidemiol.*
- 934 32(4): 325–340.
- 935 Okechukwu, R.U., and a. G.O. Dixon. 2008. Genetic Gains from 30 Years of
- 936 Cassava Breeding in Nigeria for Storage Root Yield and Disease Resistance in
- 937 Elite Cassava Genotypes. *J. Crop Improv.* 22(2): 181–208Available at
- 938 <http://www.tandfonline.com/doi/abs/10.1080/15427520802212506> (verified 18
- 939 July 2014).
- 940 Okogbenin, E., M. Porto, and C. Egesi. 2007. Marker-assisted introgression of
- 941 resistance to cassava mosaic disease into Latin American germplasm for the
- 942 genetic improvement of cassava in Africa. *Crop Sci.* 47: 1895–1904Available at
- 943 <https://dl.sciencesocieties.org/publications/cs/abstracts/47/5/1895> (verified 4
- 944 November 2013).

- 945 Oliveira, E.J., M.D.V. Resende, V. Silva Santos, C.F. Ferreira, G.A.F. Oliveira, M.S.  
946 Silva, L.A. Oliveira, and C.I. Aguilar-Vildoso. 2012. Genome-wide selection in  
947 cassava. *Euphytica* 187(2): 263–276 Available at  
948 <http://link.springer.com/10.1007/s10681-012-0722-0> (verified 19 September  
949 2013).
- 950 Oliveira, E.J., F. a Santana, L. a Oliveira, and V.S. Santos. 2014. Genetic parameters  
951 and prediction of genotypic values for root quality traits in cassava using  
952 REML/BLUP. *Genet. Mol. Res.* 13(3): 6683–700 Available at  
953 <http://www.ncbi.nlm.nih.gov/pubmed/25177949>.
- 954 Olsen, K.M., and B. a Schaal. 1999. Evidence on the origin of cassava:  
955 phylogeography of *Manihot esculenta*. *Proc. Natl. Acad. Sci. U. S. A.* 96(10):  
956 5586–91 Available at  
957 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21904&tool=pmcent](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21904&tool=pmcentrez&rendertype=abstract)  
958 [rez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21904&tool=pmcentrez&rendertype=abstract).
- 959 Park, T., and G. Casella. 2008. The Bayesian Lasso. *J. Am. Stat. Assoc.* 103(482):  
960 681–686.
- 961 Parkes, E., M. Fregene, A. Dixon, E. Okogbenin, B. Boakye-Peprah, and M.T.  
962 Labuschagne. 2015. Developing Cassava Mosaic Disease resistant cassava  
963 varieties in Ghana using a marker assisted selection approach. *Euphytica* 203:  
964 549–556 Available at <http://dx.doi.org/10.1007/s10681-014-1262-6>.
- 965 Perez-Rodriguez, P., D. Gianola, J.M. Gonzalez-Camacho, J. Crossa, Y. Manes, and  
966 S. Dreisigacker. 2013. Comparison Between Linear and Non-parametric  
967 Regression Models for Genome-Enabled Prediction in Wheat. *G3 Genes|*  
968 *Genomes| Genet.* 2(12): 1595–1605 Available at  
969 <http://g3journal.org/cgi/doi/10.1534/g3.112.003665>.
- 970 Pérez, P., and G. De Los Campos. 2014. Genome-wide regression and prediction with  
971 the BGLR statistical package. *Genetics* 198(2): 483–495.
- 972 Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: convergence  
973 diagnosis and output analysis for MCMC. *R News* 6(March): 7–11 Available at  
974 [http://cran.r-project.org/doc/Rnews/Rnews\\_2006-1.pdf#page=7](http://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf#page=7).
- 975 Rabbi, I.Y., M.T. Hamblin, P.L. Kumar, M. a Gedil, A.S. Ikpan, J.-L. Jannink, and P.  
976 a Kulakow. 2014. High-resolution mapping of resistance to cassava mosaic  
977 geminiviruses in cassava using genotyping-by-sequencing and its implications  
978 for breeding. *Virus Res.* Available at  
979 <http://www.ncbi.nlm.nih.gov/pubmed/24389096> (verified 9 June 2014).
- 980 Resende, M.F.R., P. Munoz, M.D. V. Resende, D.J. Garrick, R.L. Fernando, J.M.  
981 Davis, E.J. Jokela, T. a. Martin, G.F. Peter, and M. Kirst. 2012. Accuracy of  
982 Genomic Selection Methods in a Standard Data Set of Loblolly Pine (*Pinus taeda*  
983 L.). *Genetics* 190(4): 1503–1510.
- 984 Rincént, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V.M. Rodríguez,  
985 J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C.-C. Schoen, N. Meyer, C.  
986 Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A.  
987 Charcosset, and L. Moreau. 2012. Maximizing the reliability of genomic  
988 selection by optimizing the calibration set of reference individuals: comparison  
989 of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*  
990 192(2): 715–28 Available at  
991 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3454892&tool=pmce](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3454892&tool=pmcentrez&rendertype=abstract)  
992 [ntrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3454892&tool=pmcentrez&rendertype=abstract) (verified 8 November 2013).
- 993 Rutkoski, J.E., J.A. Poland, R.P. Singh, J. Huerta-espino, S. Bhavani, H. Barbier,  
994 M.N. Rouse, J. Jannink, and M.E. Sorrells. 2014. Genomic selection for

- 995 quantitative adult plant stem rust resistance in wheat. *Plant Genome J.*  
996 02.006(May): 1–44.
- 997 Rutkoski, J., R. Singh, and J. Huerta-Espino. 2015. Efficient use of historical data for  
998 genomic selection: a case study of stem rust resistance in wheat. *Plant Genome*:  
999 1–45.
- 1000 Sarkar, R.K., A.R. Rao, P.K. Meher, T. Nepolean, and T. Mohapatra. 2015.  
1001 Evaluation of random forest regression for prediction of breeding value from  
1002 genomewide SNPs. *J. Genet.* 94(2): 187–192.
- 1003 Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, E. Redona, G. Atlin, J.L.  
1004 Jannink, and S.R. McCouch. 2015. Genomic selection and association mapping  
1005 in rice (*Oryza sativa*): effect of trait genetic architecture, training population  
1006 composition, marker number and statistical model on accuracy of rice genomic  
1007 selection in elite, tropical rice breeding lines. *PLoS Genet.* 11(2): e1004982.
- 1008 Strobl, C., J. Malley, and G. Tutz. 2009. An Introduction to Recursive Partitioning:  
1009 Rationale, Application and Characteristics of Classification and Regression  
1010 Trees, Bagging and Random Forests. *Psychol Methods* 14(4): 323–348.
- 1011 Su, G., O.F. Christensen, L. Janss, and M.S. Lund. 2014. Comparison of genomic  
1012 predictions using genomic relationship matrices built with different weighting  
1013 factors to account for locus-specific variances. *J. Dairy Sci.* 97(10): 6547–  
1014 59 Available at  
1015 <http://www.sciencedirect.com/science/article/pii/S0022030214005591>.
- 1016 Sungthongw, K., A. Promkhambu, A. Laoken, and A. Polthanee. 2016. Effects of  
1017 Methods and Duration Storage on Cassava Stake Characteristics. *Asian J. Plant*  
1018 *Sci.* 15(3): 86–91 Available at  
1019 <http://www.scialert.net/abstract/?doi=ajps.2016.86.91>.
- 1020 VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy*  
1021 *Sci.* 91(11): 4414–23 Available at  
1022 <http://www.ncbi.nlm.nih.gov/pubmed/18946147> (verified 18 October 2013).
- 1023 Wolfe, M.D., I.Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki, P. Kulakow, R. Lozano,  
1024 D.P. del Carpio, P. Ramu, and J.-L. Jannink. 2016. Genome-wide association  
1025 and prediction reveals the genetic architecture of cassava mosaic disease  
1026 resistance and prospects for rapid genetic improvement. *Plant Genome* 9(2): 1–  
1027 13 Available at <https://dl.sciencesocieties.org/publications/tpg/first-look>.
- 1028  
1029  
1030

1031 **FIGURE LEGENDS**

1032

1033 **Figure 1. Schematic of a conventional cassava breeding cycle.** Arrows between  
1034 trials indicate the selection of materials for further phenotyping trials. Red arrows  
1035 indicate the selection of materials as parents for crossing.

1036

1037 **Figure 2. Schematic of IITA Genomic Selection 2012-2015.** Three generations of  
1038 IITA genomic selection program are illustrated here. From the genetic gain (GG)  
1039 population, 85 parents were selected and crosses over two years (“TMS13F” in 2012-  
1040 2013 and “TMS14F” in 2013-2014) gave rise to 2890 Cycle 1 (C1) progeny.  
1041 Predictions based on data from the GG were used to select 89 parents from among C1  
1042 in 2013, giving rise to 1648 Cycle 2 (C2) progeny in 2014. The GG have been  
1043 clonally evaluated in 2013-2014 and 2014-2015. The “TMS13” C1 were evaluated in  
1044 2013-2014 and 2014-2015. The “TMS14” C1 were evaluated with the C2 in 2014-  
1045 2015.

1046

1047 **Figure 3. Schematic of genomic selection with training population optimization**  
1048 **by STPGA.** Selection is initially made among available, genotyped candidates based  
1049 upon genomic prediction with available phenotype data. Selected parents are grown  
1050 and mated in a crossing block. Resulting Cycle 1 (C1) seeds are subsequently  
1051 collected and grown in a nursery. C1 seedlings are genotyped by GBS and selections  
1052 are made based on genomic prediction alone. Selected parents of C2 are cloned into a  
1053 crossing nursery. STPGA is used to select the optimal additional C1 seedlings to plant  
1054 in a clonal evaluation trial. Because C2 seedlings do not yet exist, STPGA is instead  
1055 used to select the optimal C1 seedlings to predict the selected parents of C2.  
1056 Phenotypes from C1 clonal evaluation are added to the existing genomic prediction  
1057 training dataset. The updated training model is used to predict breeding values of C2  
1058 seedlings when GBS data become available and the selections of parents of C3 is  
1059 made. Subsequent cycles proceed based on this procedure.

1060

1061 **Figure 4. Hierarchical clustering of genomic prediction models based on cross-**  
1062 **validated genomic estimated breeding values (GEBVs).** Height on the y-axis refers  
1063 to the value of the dissimilarity criterion. (A) Clustering of prediction models in the  
1064 NRCRI population. (B) Clustering of prediction models in the NaCRRRI population.  
1065 (C) Clustering of prediction models in Genetic Gain (GG) population.  
1066 GBLUP, genomic best linear unbiased predictor; BL, Bayesian Lasso; RF, random  
1067 forest; RKHS, reproducing kernel Hilbert spaces multi-kernel model.

1068

1069 **Figure 5. Boxplot of cross-generation prediction accuracies.** Seven genomic  
1070 prediction methods were tested for seven traits (panels). For each model – trait  
1071 combination, four predictions were made: GG predicts C1, GG predicts C2, C1  
1072 predicts C2, GG+C1 predicts C2. Boxes show range of accuracies across these four  
1073 prediction scenarios. All data are from the IITA Genomic Selection program.  
1074 GG=Genetic Gain. C1 = Cycle 1. C2 = Cycle 2.

1075

1076 **Figure 6. The relationship between training set size and accuracy predicting**  
1077 **IITA Cycle 2 (across-generation).** The accuracy of prediction for seven traits  
1078 (panels) with the IITA Genetic Gain (GG) population training data plus data from  
1079 different size subsets (x-axis) of their progeny, Cycle 1 (C1) is shown. Subsets of a  
1080 given size were selected either at random or using the genetic algorithm implemented

1081 in the R package STPGA. Ten random and ten STPGA-selected subsets were made at  
1082 each training set size. Error bars are the standard error around the mean for the ten  
1083 samples. Horizontal black lines show the mean cross-validation accuracy for the C2  
1084 (validation set; solid line) and the accuracy of the full set of GG+C1 predicting C2  
1085 (dashed line).

1086

1087 **Figure 7. The relationship between training set size and accuracy predicting the**  
1088 **parents of Cycle 2 (from Cycle 1, within-generation).** The accuracy of prediction  
1089 for seven traits (panels) with the IITA Genetic Gain (GG) population training data  
1090 plus data from different size subsets (x-axis) of their progeny, Cycle 1 (C1) is shown.  
1091 Subsets of a given size were selected either at random or using the genetic algorithm  
1092 implemented in the R package STPGA. Ten random and ten STPGA-selected subsets  
1093 were made at each training set size. Error bars are the standard error around the mean  
1094 for the ten samples. Horizontal black lines show the mean cross-validation accuracy  
1095 for the C1 (validation set; solid line) and the accuracy of the full set of GG+C1  
1096 predicting the parents of C2 (dashed line).

1097

1098 TABLES

**Table 1.** Summary and comparison of phenotype and genotype datasets analyzed in this study.

Trait	Broad-sense Heritabilities						
	All IITA	IITA			NRCRI	NaCRRI	
		GG	C1	C2			
VIGOR		0.25	0.25	0.31	0.19	0.06	0.15
MCMDS		0.69	0.60	0.86	0.25	0.44	0.62
DM		0.49	0.59	0.62	0.51	0.01	0.14
HI		0.57	0.36	0.62	0.55	0.12	0.36
RTWT		0.31	0.10	0.36	0.00	0.10	0.27
RTNO		0.24	0.09	0.26	0.00	0.06	0.22
SHTWT		0.22	0.14	0.21	0.00	0.13	0.25
N Clones		5247	709	2890	1648	899	411
Raw data points		8501	2924	3875	1702	2391	7662
<b>Genetic Diversity Statistics</b>							
	Mean Inbreeding Coeff*	0.933	0.965	0.949	0.946	0.954	
	Std Dev. Kinship Coeff**	0.080	0.089	0.092	0.080	0.118	
	MAF>1%	76137	73096	70010	78212	75923	
	Median(MAF)	0.009	0.0067	0.0047	0.01	0.01	
	Mean(Heterozygosity)***	0.16	0.15	0.17	0.15	0.15	
	Max(Heterozygosity)	0.29	0.27	0.28	0.26	0.24	
	Min(Heterozygosity)	0.07	0.07	0.10	0.07	0.08	
	Mean(MAF)	0.056	0.054	0.056	0.055	0.054	
<b>Mean-Fst between Datasets</b>							
	<b>Populations Compared</b>	<b>F<sub>ST</sub></b>		<b>Populations Compared</b>	<b>F<sub>ST</sub></b>		
	GG vs. NR	0.008		GG vs. C1	0.010		
	GG vs. UG	0.019		GG vs. C2	0.020		
	NR vs. UG	0.021		C1 vs. C2	0.014		

\*Mean of the diagonal of the genomic relationship matrix

\*\*Off-diagonal of the genomic relationship matrix

\*\*\*Heterozygosity per individual per dataset

IITA = International Institute of Tropical Agriculture; GG = IITA Genetic Gain; C1 = IITA Cycle 1; C2 = IITA Cycle 2; NR = National Root Crops Research Institute; UG = National Crops Resources Research Institute

1100 **Table 2.** Summary of cross-validated predictive accuracies by prediction model, trait  
 1101 and breeding program. Highest predictive accuracy across methods within a trait and  
 1102 within breeding program is indicated in bold.

1103 The asterisk (\*) indicates highest predictive accuracy within a trait across breeding  
 1104 programs

1105

1106

Trait	Program	BayesA	BayesB	BayesC	BL	GBLUP	MultiKernel-RKHS	RandomForest	mean
DM	NRCRI	0.12	0.12	0.11	0.12	0.10	<b>0.18</b>	0.15	0.13
	NaCRRI	0.29	0.29	0.30	0.29	0.30	0.33	<b>0.34</b>	0.31
	GG	0.67	0.67	0.67	<b>0.68*</b>	0.67	0.67	0.63	0.66
Harvest index	NRCRI	0.27	0.26	0.27	0.24	0.27	0.30	<b>0.31</b>	0.27
	NaCRRI	0.46	0.45	0.45	0.45	0.45	<b>0.48*</b>	0.47	0.46
	GG	0.37	0.39	0.39	0.40	0.39	<b>0.41</b>	0.39	0.39
Root weight	NRCRI	0.23	0.22	0.23	0.24	0.22	0.32	<b>0.34</b>	0.26
	NaCRRI	0.31	0.30	0.30	0.29	0.31	<b>0.37*</b>	0.35	0.31
	GG	0.31	0.31	0.33	0.33	0.32	0.33	<b>0.34</b>	0.33
Root number	NRCRI	0.19	0.18	0.18	0.19	0.18	<b>0.21</b>	0.20	0.19
	NaCRRI	0.35	0.34	0.34	0.30	0.35	<b>0.39*</b>	0.36	0.34
	GG	0.33	0.33	0.34	0.35	0.35	0.34	<b>0.35</b>	0.34
Shoot weight	NRCRI	0.18	0.19	0.19	0.19	0.17	<b>0.25</b>	0.24	0.20
	NaCRRI	0.21	0.22	0.22	0.18	0.24	<b>0.26</b>	0.25	0.23
	GG	0.31	0.32	0.32	<b>0.33*</b>	0.32	<b>0.33*</b>	0.29	0.31
Cassava mosaic disease	NRCRI	0.23	0.22	0.20	0.21	0.19	0.24	<b>0.29</b>	0.23
	NaCRRI	0.50	<b>0.50</b>	0.42	0.41	0.40	0.45	0.48	0.45
	GG	0.58	<b>0.60*</b>	0.57	0.56	0.56	0.57	<b>0.60*</b>	0.57
Plant vigor	NRCRI	-0.03	<b>-0.02</b>	<b>-0.02</b>	-0.03	<b>-0.02</b>	-0.03	-0.03	-0.02
	NaCRRI	0.35	0.34	0.34	0.34	0.35	<b>0.38*</b>	<b>0.38*</b>	0.34
	GG	0.23	0.23	0.24	<b>0.24</b>	0.23	0.22	0.18	0.22
mean		0.31	0.31	0.30	0.30	0.30	0.33	0.33	

1107

1108 IITA = International Institute of Tropical Agriculture; GG = IITA Genetic Gain;

1109 NRCRI = National Root Crops Research Institute; NaCRRI = National Crops

1110 Resources Research Institute

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

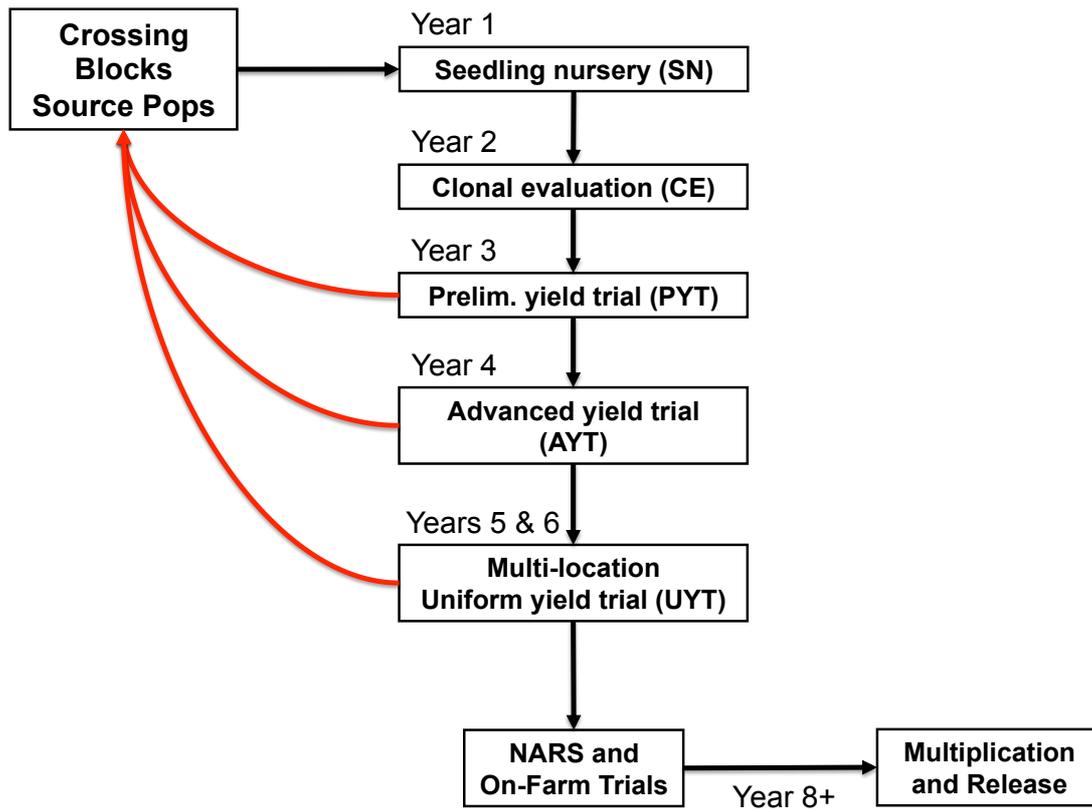
1123

1124 **Table 3.** Summary of mean GBLUP cross-validated predictive accuracies cross  
 1125 populations. Four subset selection methods (random vs. STPGA) and the full set were  
 1126 considered. Highest predictive accuracy across subsets and the full set is indicated in  
 1127 bold, CVGBLUP:crossvalidation GBLUP within the test population.  
 1128 NR:NRCRI,UG:NaCRRI,GG: Genetic gain IITA.  
 1129  
 1130

Train	Test	Trait	300		600		900		1200		FULL	CVGBLUP
			STPGA	Random	STPGA	Random	STPGA	Random	STPGA	Random		
NR+GG	UG	VIGOR	0.199	0.083	0.182	0.102	<b>0.221</b>	0.152	0.200	0.174	0.193	0.353
NR+GG	UG	MCMD5	<b>0.293</b>	0.224	0.284	0.264	0.262	0.279	0.284	0.291	0.285	0.404
NR+GG	UG	DM	0.272	0.209	0.282	0.227	0.258	0.254	0.252	0.272	<b>0.284</b>	0.296
NR+GG	UG	HI	<b>0.294</b>	0.176	0.278	0.230	0.266	0.215	0.228	0.214	0.206	0.454
NR+GG	UG	RTWT	0.155	0.072	0.165	0.124	0.181	0.156	0.179	0.174	<b>0.193</b>	0.314
NR+GG	UG	RTNO	0.149	0.068	0.171	0.151	0.175	0.167	0.195	0.190	<b>0.206</b>	0.348
NR+GG	UG	SHTWT	-0.014	0.059	0.042	<b>0.075</b>	0.027	0.066	0.037	0.071	<b>0.075</b>	0.244
UG+NR	GG	VIGOR	-0.011	0.054	0.032	0.049	0.050	<b>0.061</b>	--	--	0.060	0.231
UG+NR	GG	MCMD5	0.374	0.325	0.377	0.341	0.372	0.374	--	--	<b>0.382</b>	0.558
UG+NR	GG	DM	0.216	0.173	0.221	0.212	0.235	0.238	--	--	<b>0.244</b>	0.666
UG+NR	GG	HI	<b>0.261</b>	0.210	0.252	0.204	0.222	0.213	--	--	0.215	0.386
UG+NR	GG	RTWT	0.079	0.077	<b>0.095</b>	0.073	0.084	0.061	--	--	0.063	0.320
UG+NR	GG	RTNO	<b>0.132</b>	0.096	0.130	0.110	0.113	0.097	--	--	0.099	0.345
UG+NR	GG	SHTWT	0.154	0.110	<b>0.163</b>	0.160	0.145	0.156	--	--	0.162	0.321
GG+UG	NR	VIGOR	<b>0.054</b>	-0.003	0.029	0.003	0.039	0.014	0.017	0.011	0.016	-0.024
GG+UG	NR	MCMD5	0.193	0.138	0.186	0.154	0.189	<b>0.190</b>	0.193	0.188	<b>0.213</b>	0.188
GG+UG	NR	DM	0.116	0.110	0.151	0.142	0.166	0.155	0.168	0.167	<b>0.184</b>	0.104
GG+UG	NR	HI	0.149	0.122	0.157	0.145	0.151	0.151	0.164	0.155	<b>0.181</b>	0.271
GG+UG	NR	RTWT	0.080	0.070	<b>0.120</b>	0.048	0.099	0.058	0.096	0.071	0.082	0.220
GG+UG	NR	RTNO	0.074	0.064	<b>0.066</b>	0.051	0.041	0.054	0.040	0.053	0.053	0.180
GG+UG	NR	SHTWT	0.094	0.089	0.107	0.088	0.107	0.099	0.112	0.106	<b>0.119</b>	0.169

1131  
 1132  
 1133  
 1134  
 1135  
 1136  
 1137  
 1138  
 1139  
 1140  
 1141  
 1142  
 1143  
 1144  
 1145  
 1146  
 1147

1148 **Figure 1.**

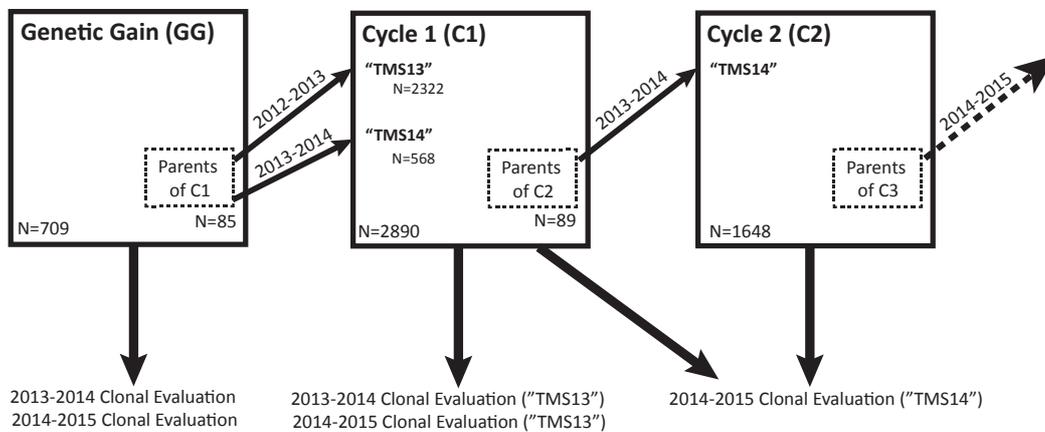


1149

1150

1151

**Figure 2.**



1152

1153

1154

1155

1156

1157

1158

1159

1160

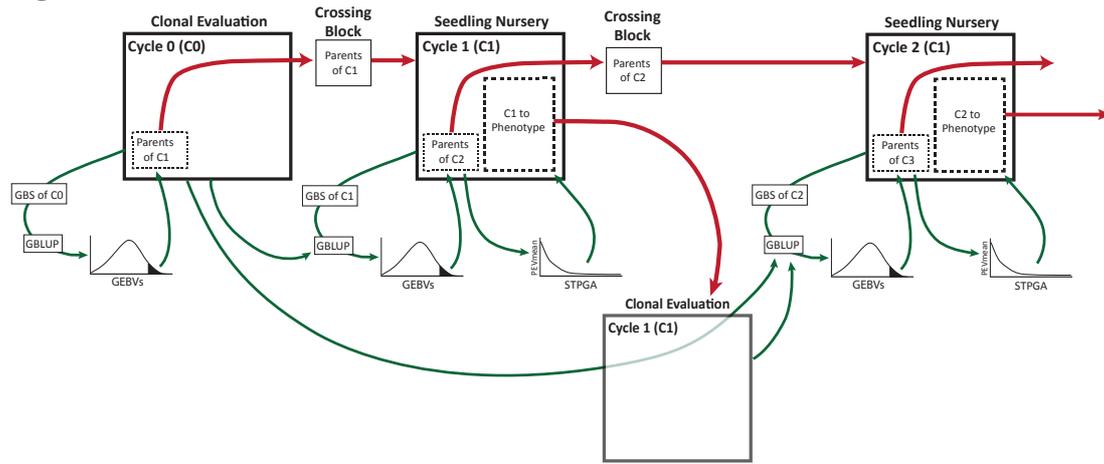
1161

1162

1163

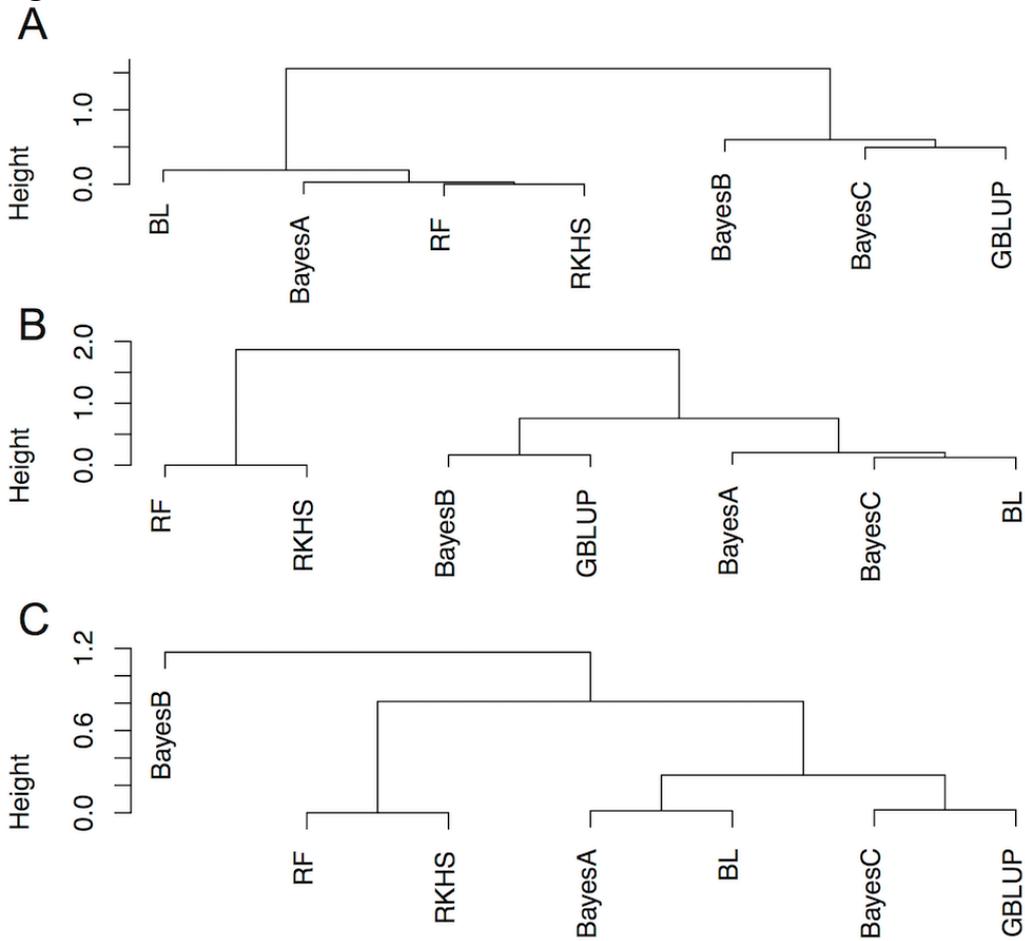
1164

1165 **Figure 3.**



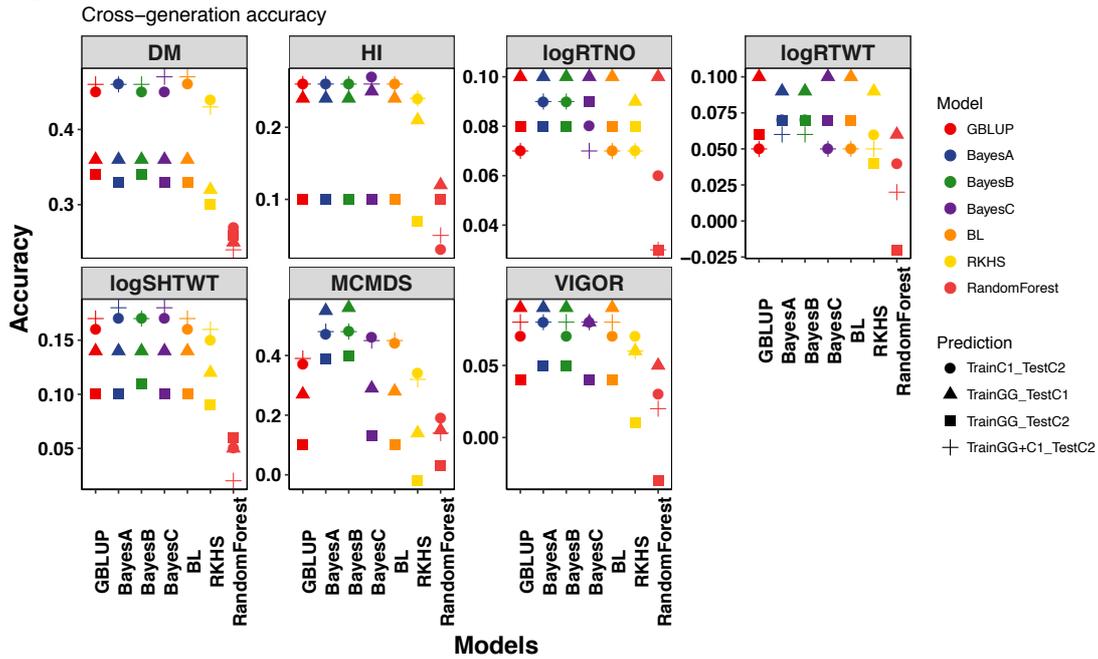
1166  
1167  
1168

**Figure 4.**



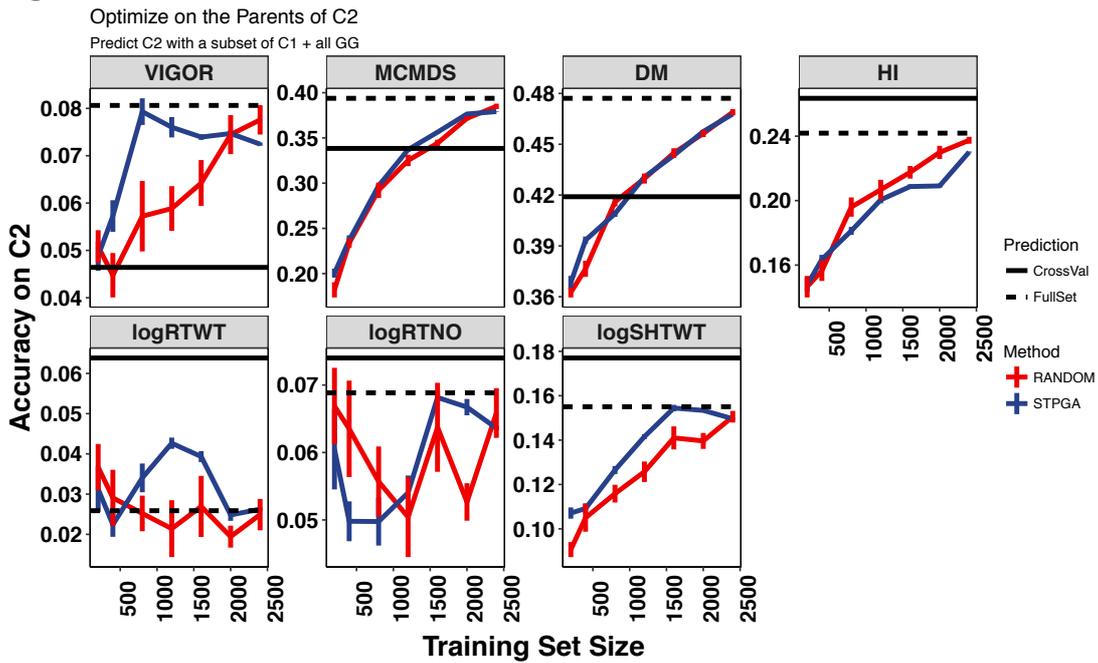
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177

1178 **Figure 5.**



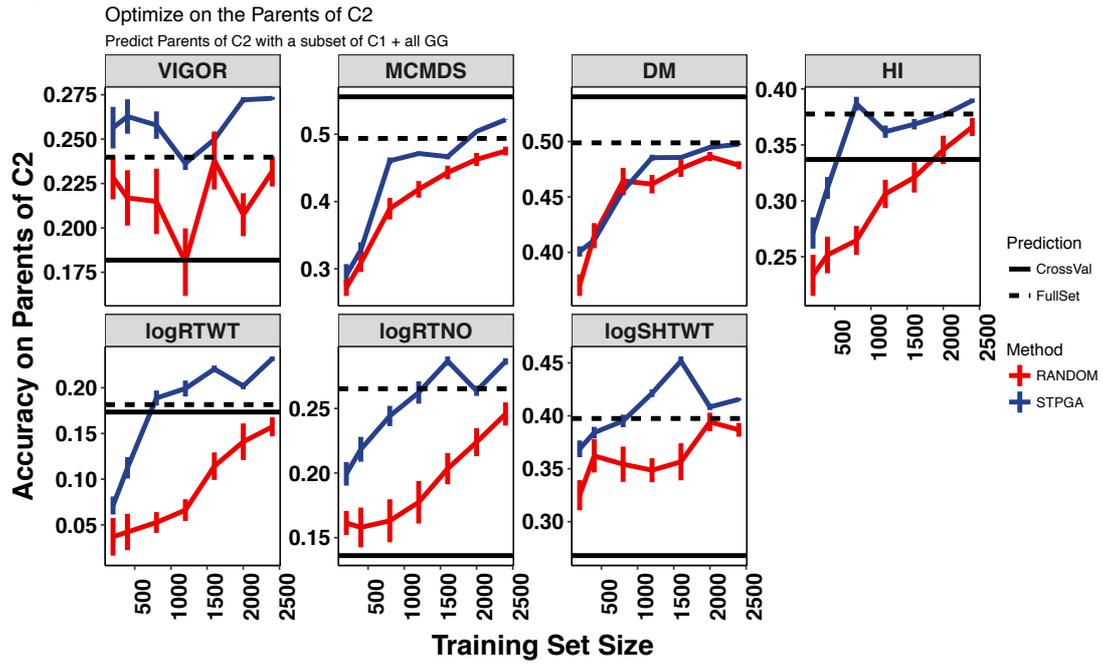
1179  
1180  
1181

**Figure 6.**



1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193

1194 **Figure 7.**



1195