Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

Relatedness and Genotype-by-environment Interaction Affect Prediction Accuracies in Genomic Selection: a Study in Cassava

Delphine Ly, Martha Hamblin,* Ismail Rabbi, Gedil Melaku, Moshood Bakare, Hugh G. Gauch, Jr., Richardson Okechukwu, Alfred G.O. Dixon, Peter Kulakow, and Jean-Luc Jannink

D. Ly, Montpellier Supagro, Montpellier, France; M.T. Hamblin, and J-L Jannink, Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA; H.G. Gauch, Jr.,Department of Crop and Soil Sciences, Cornell University, Ithaca, NY, USA; I.Y. Rabbi, M. Gedil,M. Bakare, R. Okechukwu, and P. Kulakow, International Institute for Tropical Agriculture, Ibadan, Nigeria; A.G.O. Dixon, Sierra Leone Agricultural Research Institute, Sierra Leone; and J-L Jannink, USDA-ARS, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA

Received _____. *Corresponding author (mth3@cornell.edu)

AYT, advanced yield trial; BLUE, best linear unbiased estimator; BLUP, best linear unbiased predictor; CV-CR, cross-validation close relatives; CV-GE, cross-validation genotype-by-environment; CV-noCR, cross-validation no close relatives; GxE, genotype-by-environment; GS, genomic selection; PYT, preliminary yield trial; RKHS, Reproducing kernel Hilbert spaces; RR-BLUP, ridge regression best linear unbiased predictor; UYT, uniform yield trial

Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

ABSTRACT

Prior to implementation of genomic selection, evaluation of the potential accuracy of prediction can be obtained by cross validation. In this procedure, a population with both phenotypes and genotypes is split into training and validation sets. The prediction model is fitted using the training set, and its accuracy is calculated on the validation set. The degree of genetic relatedness between the training and validation sets may influence the expected accuracy, as may the genotype-by-environment (GxE) interaction in those sets. We developed a method to assess these effects and tested it in cassava (*Manihot esculenta*). We used historical phenotypic data available from the International Institute of Tropical Agriculture Genetic Gain trial and performed genotyping by sequencing for these clones. We tested cross validation sampling schemes preventing the training and validation sets from sharing a) genetically close clones or b) similar evaluation locations. For 19 traits, plot-basis heritabilities ranged from 0.04 to 0.66. The correlation between predicted and observed phenotypes ranged from 0.15 to 0.47. Across traits, predicting for less related clones decreased accuracy from 0 to 0.07, a small but consistent effect. For 17 traits, predicting for different locations decreased accuracy between 0.01 and 0.18. Genomic selection has potential to accelerate gains in cassava and the existing training population should give a reasonable estimate of future prediction accuracies.

The revolution in sequencing technologies has enabled fast and relatively inexpensive genome information (Metzker, 2010). The increase in DNA-marker information available is considerable, leading to the development of a new approach to marker-assisted selection (MAS) called genomic selection (GS) (Meuwissen et al., 2001; Goddard and Hayes, 2007; Heffner et al., 2009; Hayes et al., 2009a; Lorenz et al., 2011). The GS approach has been developed to use all markers across the genome, instead of only those with larger effects as in traditional MAS, to predict the performance of individuals (Meuwissen et al., 2001; Jannink et al., 2010). With a sufficient accuracy, selection can be done based on the predictions only, for any trait. Thus, GS offers the possibility to accelerate breeding cycles. Because prediction requires from the selection candidates only genotypic data, the selection of the seedling happens at an early stage (Heffner et al., 2009). In addition, selecting on the basis of predicted breeding values of individuals rather than their phenotypic records may also make the choice of the parents more accurate.

Our study focused on cassava, which, unlike other crops, for which GS has been evaluated, is a strongly outcrossing species, characterized by monoecism and protogyny. This outcrossing characteristic is shared with cattle, a species for which GS has been shown to work effectively (VanRaden et al., 2009; Hayes et al., 2009a). Indeed, a cassava GS study using a relatively small training population and relatively low-density markers has reported reasonable prediction accuracies (Oliveira et al., 2012). Nevertheless, GS accuracies in cassava need further empirical testing.

The realized gains of a genomic selection program will depend on the quality of the predictions, which can be assessed by estimating the prediction accuracies. Genomic

selection studies on empirical data generally use cross validation to estimate prediction accuracies (Kohavi, 1995; Goddard and Hayes, 2007; Lorenz et al., 2011). In plants, the use of cross validation studies on inbred cultivars has been useful (Melchinger et al., 2004; Schön et al., 2004; Crossa et al., 2010; Riedelsheimer et al., 2012; Massman et al., 2012; Windhausen et al., 2012). Cross validation is meant to estimate the accuracy with which predictions can be made for selection candidates based on models developed in the training population, by treating a portion of the training population as selection candidates. There are some important differences between cross-validation and the prediction of breeding values in selection candidates, particularly with respect to two factors: the relatedness between individuals and GxE interaction. In particular, a random cross validation might split data into the training and the validation sets so that the information for close relatives or for locations is unrealistically similar in the two sets.

It has been shown that the additive genetic relationship of the training data influences the breeding-value accuracies of the selection candidates (Habier et al., 2007, 2010; Clark et al., 2012). Animals that shared close relationship to the training dataset had highest prediction accuracies (Habier et al., 2007, 2010; Clark et al., 2012; Pszczola et al., 2012; Pérez-Cabal et al., 2012; Cleveland et al., 2012). In a study on US dairy cattle, Pérez-Cabal et al., (2012) emphasized that the type of relatedness between the training and validation sets also influenced the prediction accuracies. Cleveland et al. (2012) have pointed out that using validation approaches that take into account relatedness between populations can correct for potential overestimation of genomic breeding-value accuracies. In the case of cassava, there are a number of factors that may affect the relatedness of clones in the training population.

Because cassava in Africa originated from South America (Jones, 1959), African cassava germplasm experienced a genetic bottleneck (Kawuki et al., 2011). Furthermore, selection, by frequently using specific elite parents in breeding programs, could make cassava clones in Africa relatively genetically similar. In addition, because of the way landrace germplasm has been collected, there are situations when virtually identical clones may be given different names. This relatedness between clones may have a strong impact on the assessment of the efficiency of GS in cassava. With a random k-fold cross validation, the genetic relatedness of individuals between the training and validation sets might be higher than that between this whole population, *i.e.*, the training population, and the individuals of the next generation of the breeding cycle, *i.e.*, the selection candidates.

A second factor that may influence prediction accuracies is GxE interaction. Genotype and environment effects are not independent: a phenotypic response to a change in environments depends on genotype, and vice versa (Comstock and Moll, 1963). With a random k-fold cross validation, the data in training and validation sets are likely to have been evaluated in the same locations. In that case, GxE interaction would generate a common error component between the predictions and the clone estimates based on the observations (Lorenz et al., 2011, 2012; Burgueño et al., 2012). Consequently, GxE interaction may be a confounding factor that upwardly biases the prediction accuracy.

The objectives of our study were to assess the impact of random k-fold cross validation on the overestimation of prediction accuracies attributable 1) to the relatedness of the individuals between the training and the validation sets and 2) to the GxE interaction. The data for the study came from the International Institute for Tropical

Agriculture (IITA) Genetic Gain population, a large collection of historically important clones, maintained at Ibadan, Nigeria.

MATERIALS AND METHODS

Phenotypic Trials

Historical phenotypic evaluation data from several types of trials have been used in the training population. All of these trials were conducted by the cassava breeding program at IITA, Ibadan, Nigeria.

The Genetic Gain is a collection of historically important clones that were selected across four decades, from the 1970s to 2007 (Maziya-Dixon et al., 2007; Okechukwu and Dixon, 2008). A small fraction of the clones are landraces and clones from East Africa with uncertain cloning dates. Most additions to the Genetic Gain population came from clones advanced to multi-environment Uniform Yield Trials. The design of the Genetic Gain trial, the major trial type, usually consists of five-plant plots, with no borders, in a single row. Most, but not all, of the Genetic Gain nurseries are replicated twice. Sometimes the plants are grown at a different density, e.g., 0.5 m apart within rows, if the land area available is limited. The plots are always planted in an incomplete block design with two checks per block.

The second most common trial type, called the Uniform Yield Trial (UYT), contains clones that are at an advanced stage in the breeding process. Compared to the Genetic Gain trial, only 15 to 30 genotypes are evaluated in a single trial because UYTs are formed after several stages of selection. Often genotypes are grouped by particular types of traits, such as multiple pest resistance, high dry matter content or poundability. This

type of trial has larger, bordered plots: generally 6 rows of 6 plants spaced 1m apart, in 4 replications, planted in a randomized complete-block design (RCBD) with two checks. Because of the borders, only 16 plants per plot are harvested. Some variation in plot size for these trials occurred as they were conducted across a period of 12 years. Uniform Yield Trials are almost always multi-location and multi-year trials, most commonly conducted across 2 years and 5 locations.

Two other types of trial, Preliminary Yield Trial (PYT) and Advanced Yield Trial (AYT), represented less than 10% of the observations. Those trials were conducted earlier than the UYTs in the breeding process, so their design was intermediate between the Genetic Gain, and the UYT designs. In the PYT, there are usually 10-plant plots grown in a single 10-meter-long row, in one location with two replications. In recent years, the AYT plot design has been the same as for the PYT, but there are usually 4 replications and one location. The design of both PYT and AYT was an RCBD with two checks.

The data were collected from 2000 to 2011, in 13 locations in Nigeria: Abuja (8.99°, 7.51°), Ibadan (7.40°, 3.90°), Ilorin (8.50°, 4.53°), Ikenne (6.7°, 3.5°), Jos (9.94°, 8.85°), Kano (12°, 8.5°), Mallam Madori (12.3°, 9.7°), Mokwa (9.3°, 5.0°), Ubiaja (6.66°, 6.38°), Onne (4.74°, 7.15°), Shonga (9.14°, 5,1°), Warri (5.52°, 5.75°) and Zaria (10.98°, 7.76°).

Eleven agronomic traits and two morphological traits were measured (Table 1). The agronomic trait “plant stands harvested” (NOHAV) was used only as a covariate in the statistical models for other traits with which it was correlated (see below). Seven of the traits are related to four biotic stresses: cassava mosaic disease (CMD), caused by a virus

from the *Begomovirus* genus that belongs to the Geminiviridae, vectored by the whitefly; cassava bacterial blight (CBB), caused by *Xanthomonas axonopodis* pv. Manihotis; cassava anthracnose disease (CAD), caused by *Colletotrichum gloeosporioides*; and cassava green mite (CGM), *Mononychellus tanajoa*.

Genotyping

DNA was extracted from 645 clones from the 2011 Genetic Gain trial at IITA, using DNeasy Plant Mini Kits (Qiagen), and was quantified using PicoGreen. The genotype data were generated using genotyping by sequencing (GBS; Elshire et al., 2011). Six 95-plex and one 75-plex PstI libraries were constructed and sequenced on Illumina HiSeq, one lane per library.

Single nucleotide polymorphisms (SNPs) were extracted from the raw data by using the TASSEL pipeline (maizegenetics.net) with alignment to the *Manihot esculenta* reference genome (www.phytozome.net). Single nucleotide polymorphisms were filtered by the following criteria: no more than 80% missing data by clone; no more than 50% missing data by SNP; amount of missing data was consistent with read depth; genotype frequencies were consistent with allele frequencies. The final data set consisted of 2069 SNPs scored in 626 clones, with a mean heterozygosity of 0.28 and mean missingness of 17.6%. Because the cassava reference genome is not assembled into chromosomes, it is not possible to show the distribution of SNPs across the genome. However, an overlapping set of GBS SNPs from a PstI library have been genetically mapped, and are well distributed across the 18 linkage groups of the cassava genome (I. Y. Rabbi, M. T. Hamblin, M. Gedil, P. Kulakow,A. S. Ikpan, D. Ly, and J.L. Jannink, unpublished). The

missing genotypic data were imputed using a classification method called Random Forest (Breiman, 2001; Poland et al., 2012).

Statistical Models for Phenotypic Data

Several statistical models were used. The first group of models was used to calculate broad-sense heritabilities on a single plot basis and generate best linear unbiased predictors (BLUPs) for data curation. The second group of models was used to generate best linear unbiased estimators (BLUEs) as an intermediate step to make predictions. The difference between these groups of models was whether they considered the clone effect as random or fixed. For the first group of models, to calculate heritabilities and generate BLUPs, we used mixed models that considered the available clones as a random sample. Mixed models were performed using the *lme4* package in R.

Data were available for 12 years and 13 locations, but not all locations were evaluated every year. Each combination of a particular year and a particular location was considered as an environment. Within each environment, several types of trials were conducted. Within each trial, clones were usually replicated in blocks. Clones measured in only one environment or only one trial were excluded. For most traits, the model was:

$$y_{i,j,k,l} = \mu + \beta_i + t_{j(i)} + r_{k(i,j)} + c_l + \varepsilon_{i,j,k,l} \qquad \begin{cases} i = 1,...,100 \\ l = 1,...,603 \end{cases} \qquad \text{(Model 1)}$$

where $y_{i,j,k,l}$ was the phenotype, μ the overall mean, β_i the fixed effect of the combination of year and location, with *i* varying from 1 to 100 for 100 combinations of year and location, $t_{j(i)}$ the random effect of the trial within an environment with a normal distribution $N(0,\sigma^2_T)$, $r_{k(i,j)}$ the random effect of the replication (or block) within a trial within an environment with a normal distribution $N(0,\sigma^2_R)$, c_l the effect of a clone considered random with a normal distribution $N(0,\sigma^2_C)$, without considering the additive

relationship matrix as the variance-covariance matrix , and with *l* varying from 1 to 603 for 603 clones, and $\varepsilon_{i,j,k,l}$ was the residual considered as random and following a normal distribution $N(0,\sigma^2)$. The assumption of homogeneity of clonal variance derives from assuming all clones were sampled from the same conceptual population of the IITA breeding program. Thus, even though different trials sampled different sets of clones, the variance was assumed consistent across trials. The assumption of homogeneity of error variance is no doubt incorrect (e.g., Edwards and Jannink, 2006), but it is assumed for expediency, as in many studies.

Some traits, such as the number of storage roots, the total fresh weight of harvested foliage and stems, and the total fresh weight of storage roots harvested, depended on the number of harvested plants: the correlation between those traits and the number of plants harvested was higher than 0.6. Because of this dependency, the number of harvested plants was taken into account in the model as a fixed effect. For these traits, the model was:

$$y_{i,j,k,l,m} = \mu + \beta_i + t_{j(i)} + r_{k(i,j)} + c_l + \delta x_{m(i,j,k,l)} + \varepsilon_{i,j,k,l,m} \quad \begin{cases} i = 1,...,100 \\ l = 1,...,603 \end{cases} \quad \text{(Model 2)}$$

with the same notations as above, and where $y_{i,j,k,l,m}$ represents the phenotype and $x_{m(i,j,k,l)}$ the number of plants harvested in plot *m*, δ is a regression coefficient and $\varepsilon_{i,j,k,l,m}$ is the residual considered as random and following a normal distribution $N(0,\sigma^2)$. We estimated the heritability as the ratio of the clonal variance to the sum of the clonal variance and the residuals variance.

Statistical models for Genomic Predictions

This study used a 2-step approach to make the genomic predictions. The first step consisted in generating BLUEs from all the phenotypic observations, so that each clone had a single phenotypic value for each trait. This reduced the computation time in the subsequent prediction step. By using BLUEs instead of BLUPs, there was no shrinkage attributable to the treatment of clones as random effects (Garrick et al., 2009).

The two models described below were used to generate BLUEs from the curated data. As in the models generating BLUPs, the statistical models to make BLUEs depended on whether a given trait was correlated with the number of harvested plants. For traits not correlated to the number of harvested plants, the model was:

$$y_{i,j,k,l} = \mu + \beta_i + t_{j(i)} + r_{k(i,j)} + \chi_l + \varepsilon_{i,j,k,l} \qquad \begin{cases} i = 1,...,93 \\ l = 1,...,580 \end{cases} \qquad \text{(Model 3)}$$

For traits correlated to the number of harvested plants, the model was:

$$y_{i,j,k,l,m} = \mu + \beta_i + t_{j(i)} + r_{k(i,j)} + \chi_l + \delta x_{m(i,j,k,l)} + \varepsilon_{i,j,k,l,m} \qquad \begin{cases} i = 1,...,93 \\ l = 1,...,580 \end{cases} \qquad \text{(Model 4)}$$

In both cases, χ_l is the effect of the clones (which is here fixed, unlike in models 1 and 2); l varies from 1 to 580 because these models are applied to curated clones and i varies from 1 to 93 because these models are applied to curated environments.

The BLUEs and their genotypic data in the training population were used to make genomic predictions of the validation population, using the R package rrBLUP (Endelman, 2011), which considers marker-based relationships as random effect covariates (Endelman, 2011). The statistical model to generate those genomic predictions is a mixed model, described below in matrix notation:

$$y = \mu + Zu + e \qquad \text{(Model 5)}$$

where **y** is the vector of phenotypes, μ is the population mean, **u** is a vector of the genotypic values considered a random effect and following a normal distribution $N(0, K\sigma^2_u)$, where *K* is the realized additive relationship matrix (Endelman, 2011), **Z** is an incidence matrix for **u**, and **e** refers to the vector of random residuals following a normal distribution $N(0, I\sigma^2)$. We measured the prediction accuracy $r(\hat{a}, y)$ as the correlation between the estimated breeding value, accounting only for additive effects (â), and the BLUE (y).

Data Curation

In a first analysis, we sought to identify if any of the 13 locations was particularly different from the others in relative clonal performance. We removed such outlier locations to avoid excessive GxE interaction that would reduce accuracies. The software MATMODEL 3.0 (Gauch and Furnas, 1991) was used to perform the AMMI (Additive Main effect and Multiplicative Interaction) analysis. The AMMI analysis integrates additive components to explain main effects and multiplicative components to account for interactions (Zobel et al., 1988). This analysis was done on the UYT datasets, year by year, from 2006 to 2008, years for which the data were the most balanced. In addition, we developed three curation methods to identify potential clone labeling errors, *i.e.*, cases where the genotypic data did not correspond to phenotypic data from the same clone. Especially in a trial such as Genetic Gain that conserves historical data, one labeling error in the past could propagate across time and reduce the accuracy of genomic predictions.

The three curation methods all used a similar ad hoc approach: 1) regress phenotypic observations on predictors derived from independent phenotypes or from genotypes; 2)

weight the residuals of this regression inversely to trait heritabilities or accuracies and sum their absolute values across traits; 3) identify clones whose total residual scores are extreme relative to the global distribution. Clones that appeared to be outliers in at least two of the methods were removed from subsequent analyses.

In the first curation method, BLUPs for the Genetic Gain dataset were regressed on the BLUPs for the other datasets (UYT/PYT/AYT). The assumption here was that BLUPs of the two types of data should be similar, unless labeling errors had occurred. This method examined phenotypic data only, whereas the following methods also considered the genotypic data.

The second curation method was based on the expectation that genetically similar clones should also be phenotypically similar. For each pair of clones, we examined their relatedness in the relationship matrix A. The A matrix was computed by using the rrBLUP function Amat (Endelman, 2011), which uses the matrix multiplication WW'/c, where $W_{ik} = G_{ik} + 1 - 2p_k$, where G_{ik} is the genotype for the i^{th} individual at the k^{th} marker (coded as -1, 0, and 1 for one homozygote, the heterozygote, and the other homozygote, respectively), and p_k is the frequency of one of the alleles and the normalization constant is $c = 2\sum_k p_k (1 - p_k)$. For each pair of clones, we also calculated the difference between their BLUPs. A linear regression of the BLUP difference on the genetic relationship was performed for each pair of clones, for each trait.

The third curation method evaluated the residuals of predictions of clone effects calculated using RKHS (Reproducing kernel Hilbert spaces) regression estimates (de los Campos et al., 2009) with a Gaussian kernel function (Endelman, 2011). We chose RKHS for this purpose because it fits the training set phenotypes with a high coefficient

of determination (Heslot et al., 2012). We reasoned, therefore, that large residuals from RKHS predictions within the training set could be indicative of problems in the correspondence of phenotypes to genotypes.

Cross Validation Schemes

Accounting for relatedness between training and validation sets: Clones were assigned to clusters based on genotypic data using the k-means clustering algorithm, a method that attempts to minimize the distance between points in a cluster and the center of that cluster. We performed the k-means method using the Hartigan and Wong (1979) algorithm on marker data and generated for each trait N = (number of non-missing individuals for a given trait/5) clusters, from 10 random initial cluster centers.

Two cross validation schemes were designed to evaluate the influence of relatedness on prediction accuracy. The first scheme (CV-noCR) avoided closely related clones, *i.e.*, clones from the same cluster were not allowed in both the training and the validation sets. Thus, CV-noCR (Figure 1b) assigned to the validation population a sample of clusters, such that they represented 20% of the whole population. In contrast, the second cross validation scheme (CV-CR) forced close relatives between the training and validation sets (Figure 1). That is, CV-CR (Figure 1c) always distributed individuals in clusters to both training and validation sets. Note that in CV-CR scheme, clusters of only one individual were never in the validation population and were not predicted. This procedure was repeated five times and we considered for each clone the mean of the five predictions.

The relationship between the training set and the validation set was measured by identifying for each individual of the validation set the 10 most closely related

individuals ("top10") in the training set (Clark et al., 2012). Each validation set was then characterized by the mean relatedness of the top10 individuals. Different training-validation sets were considered until all individuals had been predicted. For a given cross-validation scheme, we considered the a_{top10} statistic, which is the mean relatedness of the top10 individuals in the validation set to those in the training set, averaged across the different training-validation sets, and finally across the five repetitions.

Accounting for the GxE interaction: In the CV-GE cross-validation scheme, the observations were split, by location, into two disjoint sets of 6 locations. BLUEs were calculated separately in each set and according to the models presented above. Five folds of 20% of the BLUEs of one set of locations were used for the validation set so that all clones in this set were predicted. Clones of the training set that also appeared in the validation set were removed from the training set, so that there were no common clones and no common locations represented between the training and the validation sets. This scheme of cross validation was repeated 15 times with different random sets of locations. Because prediction accuracies of CV-GE used smaller training populations built on only half of the observations, we compared their results to cross-validation schemes CV-Random_Half that also split observations randomly into two sets, repeated 5 times.

Estimation of GxE interaction effects

To calculate the relative magnitude of GxE effects, variance components were estimated using ASReml (Gilmour et al., 2002) on data from Genetic Gain trials. The linear model was $y_{i,k,l} = \mu + \beta_i + r_{k(i)} + c_l + \gamma_{i,l} + \epsilon_{i,k,l}$ with terms defined similarly as for Model 1, all fitted as random, with the additional term $\gamma_{i,l}$ accounting for clone by environment interaction. The covariance matrix for the additive clone effects (c_l) was proportional to

the realized relationship matrix (A) calculated in rrBLUP, $c_l \sim N(0, A\sigma_g^2)$. The covariance matrix for the additive clone by environment interaction effects ($\gamma_{i,l}$) was block diagonal, $\gamma_{i,l} \sim N(0, B\sigma_{ge}^2)$: for effects estimated in the same environment B = A, while for effects estimated in different environments B = 0. The ratio $\sigma_{ge}^2/(\sigma_g^2 + \sigma_{ge}^2)$ reveals the magnitude of GxE interaction effects.

RESULTS

Heritabilities of the different traits in the different trials

Heritabilities for each trait, calculated using all the phenotypic data from all the trials, are shown in Table 2 (see Methods). For the two traits related to cassava mosaic disease (CMD), the heritabilities were high (0.66 and 0.63). Cassava mosaic disease is a trait controlled primarily by a few major genes (Lokko et al., 2005). Agronomic traits such as those related to yield and growth had much lower heritabilities, between 0.11 and 0.28. Traits related to diseases other than CMD had heritabilities lower than the ones for CMD; in the case of bacterial blight severity and incidence, heritabilities were below 0.10. Disease traits were sometimes difficult to score accurately because of the uneven spreading of inoculum or the presence or absence of the disease in a particular season or location. The heritabilities of the morphological traits were quite low (0.07 and 0.12). For root number, in particular, this could be partly explained by the high influence of the age of the plant at harvest.

In this study, the heritabilities for dry matter content (DM), shoot weight (SHTWT) and fresh yield (FYLD) were considerably lower (between 0.11 and 0.28) than those reported by Oliveira et al. (2012), which were, respectively, 0.67, 0.83 and 0.76. Furthermore, the heritability values that we obtained were lower than expected for many

other traits. Calculation of heritabilities by trial type (UYT, AYT, PYT, and Genetic Gain) showed that the heritabilities in the Genetic Gain dataset were generally lower than those in the other datasets, decreasing the overall heritability (Figure 2). The Genetic Gain trials used smaller plots and were expected to have a larger environmental variance than the other trial types.

In addition to the trial type, we hypothesized that two other factors might contribute to the low heritabilities. First, there might be some outlier locations that produced GxE interactions. Because we aimed to have a training population whose predictions would be generally useful across locations, we identified and removed locations where clones behaved very differently. The AMMI results for UYT data showed that clone effects for Onne location had a strong negative correlation with clone effects from the other locations in 2006 (Figure 3) and 2007 (not shown), two out of the three years that were examined. Indeed, the score for Onne on the first principal component axis of the PCA of the GxE interaction effects was strongly negative relative to the scores of the other locations. Consequently, the Onne location, which is one of the highest rainfall areas of Nigeria, was removed from subsequent analyses.

Finally, because of the long-term historical nature of the Genetic Gain collection, some labeling errors might have occurred, so that genotype and phenotypic data were incorrectly associated with the same clone ID. We looked for possible labeling errors using three approaches: (i) comparing types of trial (Figure 4A); (ii) regressing BLUP differences of all pairs of clones on their genetic relationship (Figure 4B); and (iii) regressing the predicted genotypic values on the calculated BLUPs (Figure 4C). For each approach, we plotted the distribution of the residuals and arbitrarily identified outliers in

the distribution tail (Figures 4A, 4B, 4C). The three curation methods identified, respectively, 22, 23, and 4 outliers. In total, 23 clones were identified as outliers in at least 2 out of 3 curation methods.

The curation method comparing types of trial could only be used if clones had data in the Genetic Gain population and in at least one of the other datasets. Consequently, some potential outlier clones may not have been identified. We removed the clones identified as outliers for the following analyses.

Cross Validation Schemes to Account for Relatedness

Principal components analysis of the SNP data provided no evidence of genetic structure in the Genetic Gain population (data not shown). However, many pairs of clones were genetically very close, possibly because clones had been renamed by farmers and collected as distinct accessions (Figure 5). To assess the effect of relatedness on prediction accuracies, we tried different methods to identify individuals that were closely related (while the dendrogram provides a visual representation of the problem, it does not provide a basis for assignment of clones to groups). The k-means algorithm created different clusters where many of those clusters contained two clones. These clusters of genetically close individuals were then used to design the training and validation sets in the cross validation schemes.

As explained in the Methods, we generated cross-validation sets according to two different schemes: CV-noCR, avoiding close relatives between training and validation sets, and CV-CR forcing close relatives between those sets. We calculated a statistic, a_{top10}, to measure the relatedness between the training and validation sets. As expected, this statistic was lowest in CV-noCR and highest in CV-CR (Table 2). The variance of

the a_{top10} across individuals within the top10 (see Methods) for each cross-validation scheme and for each trait was about 10^{-3}. The a_{top10} of CV-Random, while intermediate, was most often closer to the a_{top10} of CV-CR than to that of CV-noCR. In several cases, the relatedness statistics for the CV-Random and the CV-CR were equal. This confirmed that the CV-Random scheme created validation sets whose members had close relatives in the training population.

When we compared the prediction accuracies for the different cross-validation schemes, we found that higher accuracies were associated with higher relatedness measurements. The prediction accuracies between the validation schemes CV-Random and CV-CR were very close. Across traits, the CV-noCR scheme showed lower accuracies compared with the other two.

Cross Validation Scheme to Account for GxE Interaction

We tested another cross-validation scheme, CV-GE, to assess the impact of GxE interaction as a confounding factor. We were specifically interested in the GxE interaction explained by locations, because the variability across years cannot be experimentally controlled. Because splitting locations between training and validation sets resulted in a training set containing only about half of the total observations, we compared the CV-GE scheme with a random k-fold scheme that used only a random half of the observations in the training set (CV-Random_Half). As expected, the prediction accuracies decreased for CV-Random_Half relative to CV-Random (Figure 6). In the CV-GE scheme, the training population used the GxE interaction effects of six out of 12 locations (across all 12 years) to predict the six other locations. In 17 of 19 traits evaluated, accuracies for CV-GE (Figure 6: red crosses) were lower than accuracies for

CV-Random_Half (Figure 6: blue triangles). The loss of the GxE interaction effects of the six remaining locations reduced the prediction accuracies, indicating that for most traits there were shared genotype-by-location residuals across training and validation sets, biasing estimated prediction accuracies upward in CV-Random. Furthermore, the a_{top10} values were close to what was obtained using CV-Random or CV-CR. Therefore, the cross validation avoiding GxE interaction did not avoid close relatives between the training and the validation set and only evaluated the impact of GxE interaction.

Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

DISCUSSION

The curation work presented here was motivated by the low heritabilities observed when using all datasets. The curation work eliminated some clones, but did not have much impact on the heritabilities. It appeared that these heritabilities were low because of the low heritabilities in the Genetic Gain data, which might be explained by the field-plot design and by the long-term maintenance of the Genetic Gain collection. Some noise was possibly introduced by phenotyping-protocol variation across time. It is also possible that epigenetic variation between generations of vegetative propagation and/or some somatic mutations might have occurred across time, causing variation in clones' phenotypes (McKey et al., 2010). Indeed, IITA established the Genetic Gain population not for direct breeding purposes but to maintain clones. Rigorous phenotyping of the Genetic Gain population was therefore not a priority, though having this data has been critical to start a genomic selection program. The next steps for implementation of GS will have data from larger and more replicated plot designs than are currently available. As the quality of the phenotype data improves, so should heritabilities and prediction accuracies (Pszczola et al., 2012).

While traits with higher heritabilities tended to show higher prediction accuracies (Table 2), this trend was fairly weak ($r^2 = 0.26$) and there were some striking exceptions. For example, prediction accuracies for harvest index and cassava mosaic disease incidence (CMDI) were similar although heritability for CMDI was much higher. Some discrepancies were presumably caused by differences in genetic architecture and the proximity of SNP markers to QTL. In the case of CMD traits, selection history might have reduced accuracy; CMD has been a target of very strong and directed selection.

Because it is affected primarily by major genes, efforts at resistance-allele introgression could cause two clones with divergent genetic backgrounds to have similar disease resistance (or, conversely, clones with similar background could be divergent in their resistance). These cases would make CMD resistance difficult for GS models to predict.

Prediction accuracies are a correlation between BLUEs, which include both additive and non-additive components, and the estimated breeding values, which only consider the additive component; this causes a systematic underestimation of prediction accuracies. That said, accuracies for most of the traits studied are not currently high enough for GS to be effective in cassava breeding. If the breeding cycle is reduced from 5 years to 2 years, prediction accuracies must be at least 0.4 to match or exceed gain from phenotypic selection. However, as noted above, anticipated improvement in heritabilities will increase accuracy, as will larger training population size and higher marker density.

Testing the CV-noCR and CV-CR cross-validation schemes showed that the more closely the training and the validation sets were related, the higher the prediction accuracies (Table 2). The mean accuracy of CV-Random was quite close to that of CV-CR, showing that the presence of close relatives in the population would bias CV-Random accuracy estimates upward relative to a realistic expectation in a selection program. Accounting for relatedness between the populations might then be a way to correct this potential overestimation (Cleveland et al., 2012). Nevertheless, our study showed that the differences in prediction accuracies between CV-noCR and CV-CR were low. The very closely related pairs of clones did not affect strongly the prediction accuracies, possibly because the number of those pairs was relatively small. We note also that the a_{top10} values did not differ greatly between CV-CR and CV-noCR, suggesting an

overall high level of relatedness among clones in the IITA population. Caution should therefore be exercised in using this population for prediction in other populations that might well not be within its domain of inference.

The differences in prediction accuracies that we observed were much lower than those reported in the Pérez-Cabal et al. (2012) and Cleveland et al. (2012) animal studies. This might be because those studies evaluated relatedness based on pedigree, whereas our study evaluated relatedness based on genotypic information. Studies showed that prediction accuracies increased when using marker information instead of pedigree, if the marker density was high enough (Villanueva et al., 2005; Hayes et al., 2009b). Similarly, high-density markers should be able to generate more distinct training and validation sets than use of pedigree information. However, our study used marker data at a relatively low density, compared with other studies, for example, in dairy cattle or maize (*Zea mays* L.). Therefore, the use of a relatively small number of markers to assess relatedness in the training and validation sets might explain the lower difference in prediction accuracies between CV-noCR and CV-CR compared with the difference assessed in animal studies.

Moreover, the size of the population and its diversity influence the impact of close relatives in the training set. Indeed Pérez-Cabal et al. (2012) and Cleveland et al., (2012) selected the individuals for their studies so that their populations were not only quite large but also quite diverse. Because cassava is propagated vegetatively, farmers do not propagate all clones at the same rate, and in the long term some clones are likely to be lost (McKey et al., 2010). It is not clear whether the genetic bottleneck of African cassava, because of its introduction from the Amazonian region (Jones, 1959), has been more severe than that experienced by domesticated cattle. These factors might cause a

lower diversity compared with animals, and thus reduce the impact of close relatives in the training set on prediction accuracies.

Even if, for cassava breeding programs, the relatedness between the training and the validation sets does not have as much impact as for some animal cases, it may still be worth taking into account that it could affect prediction accuracy. This analysis should consider the level of relatedness expected between training population and selection candidates in the initial generations of the breeding cycle. This contrasts with the relatedness between training and validation sets in the CV-noCR scheme, both of which included individuals from the same generations, *i.e.*, all the years of the Genetic Gain program. A high level of relatedness within the training population might depend on the germplasm used at the beginning and the selection history of the population. The results of our study suggest that it would be interesting in a genomic selection breeding program to compare the a_{top10} assessed by a k-fold random cross validation within the training population, and the actual a_{top10} between the training population and the selection candidates, when the training population contains parents of the validation population. We measured the mean relatedness between parent and offspring for nine clones for which pedigree data were available, and obtained a value of 0.30 ± 0.01. If the relatedness assessed in a k-folds cross validation were higher than the mean relatedness expected between parent and offspring, we might overestimate the prediction accuracy. In our study, the a_{top10} using the CV-noCR scheme (Table 2) was lower than this value estimated for parents and offspring, so the CV-noCR scheme probably did not overestimate prediction accuracy. Note that 0.30 is not the parent-offspring relatedness expected from methods of calculation using pedigree (which would be 0.5). Coefficients

calculated from marker data are not comparable to those calculated from pedigree (e.g., they can be negative; Endelman & Jannink, 2012).

When we analyzed the effect of GxE interaction, we found that the prediction accuracies of CV-GE were generally lower than those obtained by CV-Random_Half (Figure 6). When clones were evaluated in the same locations for both the training and validation sets, GxE interaction, which has been shown to have a significant effect on traits related to yield in Nigerian cassava (Aina et al., 2009), seemed to be a strong confounding factor that leads to overestimation of prediction accuracies. There was considerable variation in the magnitude of this effect for different traits, e.g., accuracy for bacterial blight incidence (MCBBI) dropped almost to zero, while accuracy for DM was unaffected. To test whether the loss of accuracy correlated to the magnitude of GxE interaction effects, we estimated an additive-by-environment interaction effect for each trait (see Methods). The ratio of additive-by-environment variance compared with additive-genetic plus additive-by-environment variance varied from 0.84 for MCBBI to 0.10 for DM, consistent with the hypothesis that traits with a smaller ratio should show a smaller reduction in accuracy in the GxE interaction cross-validation scheme. Across all the traits, the linear correlation between accuracy reduction and magnitude of GxE interaction was significantly positive ($P = 0.04$).

The presence of GxE interaction effects reduces our ability to make predictions when selecting for locations where no evaluations have been done previously. In those cases, the phenotypic observations are likely not to be as correlated to the predictions as the random cross-validation prediction accuracies. When expanding a genomic selection breeding program to new locations, in order to have a better estimation of our prediction

ability, cross-validations schemes should aim at reducing the overestimation caused by GxE interaction by using training and validation sets that do not share common locations. Furthermore, given the impact of the GxE interaction on the prediction accuracies, instead of removing the GxE interaction effect, exploiting it would be a worthwhile goal. It may be worth delineating mega-environments to make predictions within them, thus exploiting the narrow adaptations of genotypes in those mega-environments (Gauch, 1997; Annicchiarico, 2002).

CONCLUSIONS

Prediction accuracies obtained by random cross-validations, used to evaluate the prospects for success of genomic selection, will be overestimated if there are close relatives in the training population. Relatedness should therefore be examined in a genomic selection breeding process to better evaluate prediction accuracies, and should be considered in designing the training population. Genotype-by-environment interactions also contribute to overestimation of prediction accuracies, and should be considered when expanding a breeding program to new experimental sites. Prediction accuracies need improvement if GS is to outperform phenotypic selection on a per-year basis; these improvements are expected as training populations increase in size and are less dependent on historical phenotype data.

AUTHOR CONTRIBUTIONS

DL and JLJ designed the study and interpreted the results. DL performed the statistical analyzes. AGOD, PK, IYR, RO and MB selected and assembled the genetic gain collection, and contributed to phenotypic evaluation. GM oversaw cassava tissue collection and DNA preparation. MTH was responsible for the GBS data and bioinformatic analyses. HGG performed the AMMI analysis. DL wrote the paper with assistance from MTH and JLJ.

ACKNOWLEDGMENTS

This work was supported by the project “Genomic Selection: The next frontier for rapid gains in maize and wheat improvement”, through funds from The Bill & Melinda Gates Foundation. We thank Deniz Akdemir and Jeff Endelman for help with statistical analyses; Lisa Blanchard for preparation of GBS libraries; and Jeff Glaubitz, Rob Elshire, and Qi Sun for help with the GBS bioinformatics pipeline.

REFERENCES

Annicchiarico, P. 2002. Case study□: durum wheat. p. 89–104. *In* Genotype x environment interactions: Challenges and opportunities for plant breeding and cultivar recommendations. Food and Agriculture Organization of the United Nations, Rome.

Breiman, L.E.O. 2001. Random forests. Statistics: 5–32.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci. 52:707-719.

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Select. Evol. 44:4.

Cleveland, M.A., J.M. Hickey, and S. Forni. 2012. A common dataset for genomic analysis of livestock populations. G3 2: 429–35.

Comstock, R.E., and R.H. Moll. 1963. Genotype-environment intersctions. p. 164-196. *In*: W.D. Hanson and H.F. Robinson (ed.) Statistical genetics and plant breeding. NAS-NRC Publ. 982.

Crossa, J., G.D.L. Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–24.

Edwards, J.W., and J.-L. Jannink. 2006. Bayesian modeling of heterogeneous error and genotype by environment interaction variances. Crop Sci. 46:820-833.

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

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255.

Endelman, J.B., and J.-L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. G3 2:1405–1413.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Select. Evol. 41:55.

Gauch, H.G. 1997. Identifying mega-environments and targeting genotypes. Crop Sci. 37:311-326.

Gauch, H.G., and R.E. Furnas. 1991. Statistical analysis of yield trials with MATMODEL. Agron. J. 83:916–920.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. 2002. ASReml user guide release 1.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124:323–330.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–97.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Select. Evol. 42:5.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009a. Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433–443.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91:47–60.

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. Crop Sci. 49:1-12.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9:166–77.

Jones, W. 1959. Manioc in Africa. Stanford University Press, Stanford,CA.

Kawuki, R.S., M. Ferguson, M.T. Labuschagne, L. Herselman, J. Orone, I. Ralimanana, M. Bidiaka, S. Lukombo, M.C. Kanyange, G. Gashaka, G. Mkamilo, J. Gethi, and H. Obiero. 2011. Variation in qualitative and quantitative traits of cassava germplasm from selected national breeding programmes in sub-Saharan Africa. Field Crops Res. 122:151–156.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on

Artificial Intelligence, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. 2:1137-1143.

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

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells, and J. Jannink. 2011. Genomic selection in plant breeding: Knowledge and prospects. Adv. Agron. 110:77-123.

Lorenz, A.J., K.P. Smith, and J.-L. Jannink. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci. 52:1609-1621.

de los Campos, G., D. Gianola, and G.J.M. Rosa. 2009. Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. J. Animal Sci. 87:1883–1887.

Massman, J.M., A. Gordillo, R.E. Lorenzana, and R. Bernardo. 2012. Genomewide predictions from maize single-cross data. Theor. Appl. Genet. 126:13-22.

Maziya-Dixon, B., A.G.O. Dixon, and A.-R. a. Adebowale. 2007. Targeting different end uses of cassava: genotypic variations for cyanogenic potentials and pasting properties. Int. J. Food Sci. Technol. 42:969–976.

McKey, D., M. Elias, B. Pujol, and A. Duputié. 2010. The evolutionary ecology of clonally propagated domesticated plants. New Phytol. 186:318–32.

Melchinger, A.E., H.F. Utz, and C.C. Sch. 2004. QTL analyses of complex traits with cross validation, bootstrapping and other biometric methods. Euphytica 137:1–11.

Metzker, M.L. 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11:31–46.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Okechukwu, R.U., and a. G.O. Dixon. 2008. Genetic gains from 30 years of cassava breeding in Nigeria for storage root yield and disease resistance in elite cassava genotypes. J. Crop Improv. 22:181–208.

Oliveira, E.J., M.D.V. Resende, V. Silva Santos, C.F. Ferreira, G.A.F. Oliveira, M.S. Silva, L.A. Oliveira, and C.I. Aguilar-Vildoso. 2012. Genome-wide selection in cassava. Euphytica 187:263-276.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, S. Dreisigacker, J. Crossa, H. Sanchez-villeda, and M. Sorrells. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5:103-113.

Pszczola, M., T. Strabel, H. a Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400.

Pérez-Cabal, M.A., A.I. Vazquez, D. Gianola, G.J.M. Rosa, and K.A. Weigel. 2012. Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. Front. Genet. 3:27.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer, and A.E. Melchinger. 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. 44:217–20.

Schön, C.C., H.F. Utz, S. Groh, B. Truberg, S. Openshaw, and A.E. Melchinger. 2004. Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485–98.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

Villanueva, B., J. Fernández, M.A. Toro, and J. Ferna. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J. Animal Sci. 83:1747–1752.

Windhausen, V.S., G.N. Atlin, J.M. Hickey, J. Crossa, J.-L. Jannink, M.E. Sorrells, B. Raman, J.E. Cairns, a. Tarekegne, K. Semagn, Y. Beyene, P. Grudloyma, F. Technow, C. Riedelsheimer, and A.E. Melchinger. 2012. Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2:1427–1436.

Zobel, R.W., M.J. Wright, and H.G. Gauch. 1988. Statistical analysis of a yield trial. Agron. J. 80:388–393.

Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

FIGURE LEGENDS

Figure 1: Cross validation (CV) schemes taking into account the effects of relatedness. (a) The big square represents a sample of 25 individuals, where each individual is represented by a little square. Cluster membership is indicated by color. (b) and (c) A red "v" in a square indicates that the corresponding individual is in the validation set, and the remaining individuals are in the training set.

Figure 2: Broad sense heritabilities of the different traits in the different trial types. The trait abbreviations are explained in Table 1.

Figure 3: Identification of outlier locations. AMMI graph of the Main Effects and the first Axis of the Principal Component Analysis (IPCA1) obtained by AMMI Analysis on UYT data in 2006.

Figure 4: Distribution of the sum of the absolute value of the residuals, across traits. A. Comparing the BLUPs between the Genetic Gain and the other trial types, weighted by heritabilities. B: Comparing the phenotypic data to the genotypic data, weighted by heritabilities. C: For predictions using the Gaussian kernel, weighted by accuracies.

Figure 5: Dendrogram of the genotypic data. A hierarchical clustering using the Euclidean distance between the genotypes was used in order to represent a dendrogram of the clones, and allowed capturing visually the strongest pattern to represent in our clustering methods.

Figure 6: Comparison of the prediction accuracies of different cross validation schemes. The trait abbreviations are explained in Table 1. Traits are ranked according to their heritabilities, from lower (on the left) to higher.

Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

1 **Table 1: Description of the traits of interest**

Type of trait	Abbreviation	Name of Trait	Description
agronomic	SPROUT	sprouting	proportion of stakes germinated scored one month after planting
	VIGOR	initial vigor	degree of initial vigor of the establishment scored one month after planting. It is scored from 3 which corresponds to a low vigor to 7 which is high.
	HI	harvest index	ratio of fresh root weight divided by total biomass
	DM	root dry matter content	percentage dry matter storage root. It measures root dry weight as the percentage of 100g of the root tubers
	RTWT	fresh weight of storage root	total fresh weight of storage roots harvested per plot measured in kg
	FYLD	fresh root yield	fresh weight of harvested roots expressed in tons per hectares per plant at harvest
	DYLD	dry yield	dry weight of harvested roots derived by multiplying fresh storage root yield by dry matter content expressed in tons per hectares
	SHTWT	fresh shoot weight	total fresh weight of harvested foliage and stems in kilograms per plot
	TYLD	The top yield	the total fresh weight of harvested foliage and stems expressed in tons per hectare
	RTNO	root number	number of storage roots per plot at harvest
	NOHAV	plant stands	counts the number of plant stand at

		harvested	harvest
morphological measured by visual rating	NKLGT	root neck length	is usually scored on a scale of 0 (=absent or sessile), 3 (=short), 5 (=medium) and 7 (=long)
	ROTNO	rotted storage roots	counts the number of the rotted root per plot at the time of harvest
biotic stresses	CMDS	Cassava mosaic disease Severity	cassava mosaic disease (CMD) severity rated on a scale from 1 (no symptoms) to 5 (extremely severe)
	CMDI	Cassava Mosaic disease Incidence	cassava mosaic disease incidence is the proportion of plants showing CMD symptoms
	CBBS	Cassava bacterial blight Severity	cassava bacterial blight (CBB) severity rated on a scale from 1 (no symptoms) to 5 (extremely severe)
	CBBI	Cassava bacterial blight Incidence	Cassava bacterial blight incidence is the proportion of plants showing CBB symptoms
	CADS	Cassava anthracnose disease Severity	cassava anthracnose disease (CAD) severity rated on a scale from 1 (no symptoms) to 5 (extremely severe)
	CADI	Cassava anthracnose disease Incidence	Cassava anthracnose disease incidence is the proportion of plants showing CAD symptoms
	CGM	Cassava green mite	cassava green mite (CGM) measure the severity rated on a scale from 1 (no symptoms) to 5 (extremely severe)

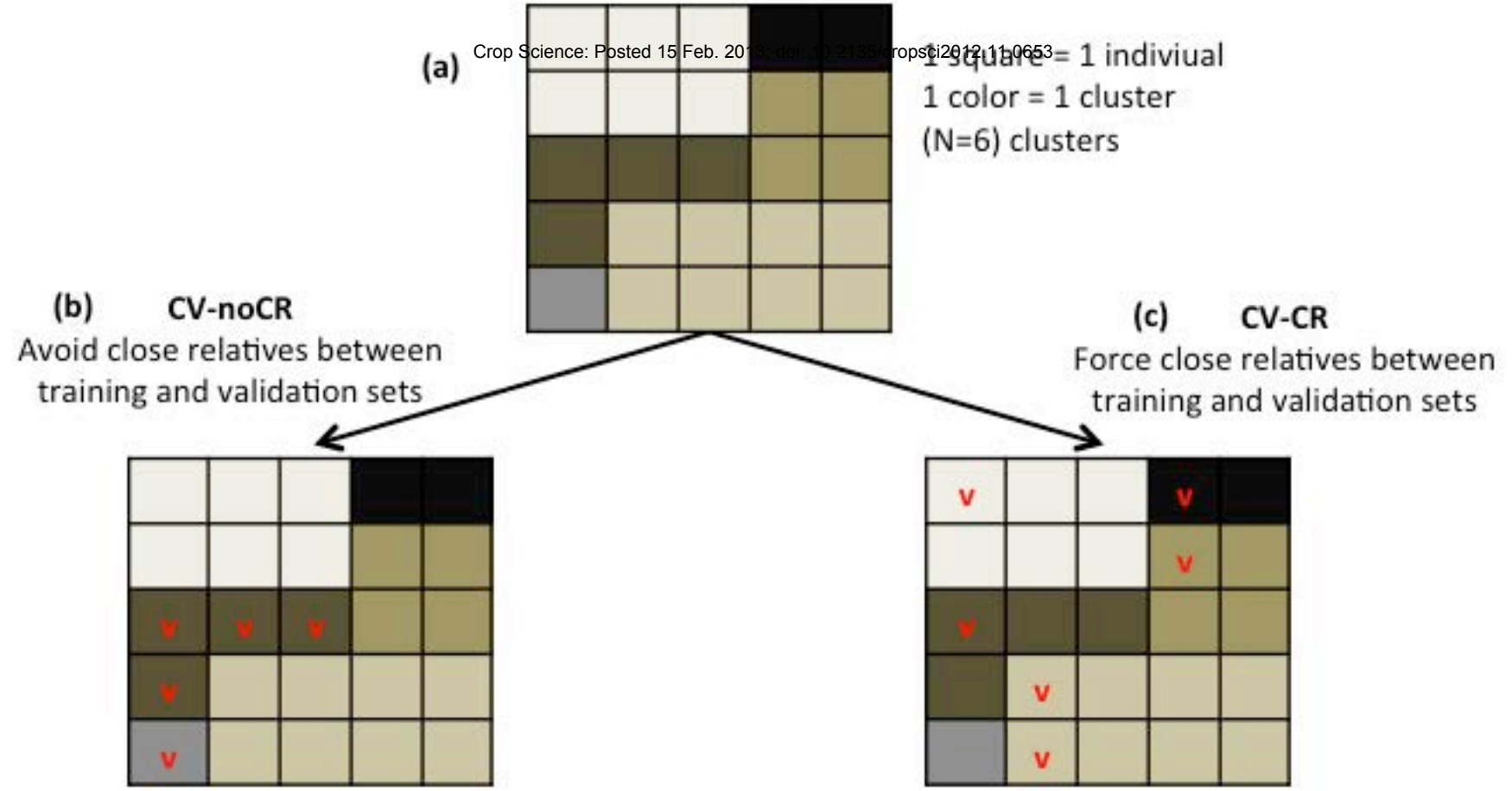
Table 2: Heritabilities of cassava traits of interest, accuracies of prediction and a_{top10} of different cross validation (CV) schemes

Trait	Heritability	CV without Close relatives		CV Random 5-fold		CV with Close relatives	
		Accuracy	a_{top10}	Accuracy	a_{top10}	Accuracy	a_{top10}
MCMDI	0.66	0.417	0.231	0.487	0.260	0.474	0.267
MCMDS	0.63	0.462	0.23	0.503	0.261	0.513	0.266
MCADI	0.38	0.177	0.267	0.184	0.201	0.202	0.285
DM	0.28	0.459	0.229	0.482	0.258	0.477	0.268
SPROUT	0.28	0.259	0.23	0.304	0.260	0.306	0.268
HI	0.27	0.431	0.23	0.483	0.259	0.479	0.268
FYLD	0.26	0.358	0.231	0.407	0.261	0.395	0.267
DYLD	0.21	0.231	0.23	0.304	0.259	0.296	0.268
TYLD	0.2	0.195	0.229	0.251	0.260	0.232	0.267
MCGM	0.18	0.47	0.231	0.308	0.259	0.501	0.267
MCADS	0.17	0.145	0.266	0.177	0.201	0.207	0.284
VIGOR	0.17	0.277	0.23	0.494	0.259	0.299	0.268
RTNO	0.14	0.342	0.228	0.399	0.260	0.384	0.267
RTWT	0.14	0.308	0.23	0.352	0.260	0.349	0.267
NKLGT	0.12	0.202	0.233	0.190	0.258	0.195	0.267
SHTWT	0.11	0.228	0.231	0.299	0.260	0.297	0.266
MCBBS	0.09	0.266	0.229	0.303	0.260	0.316	0.267
ROTNO	0.07	0.188	0.241	0.211	0.252	0.229	0.276

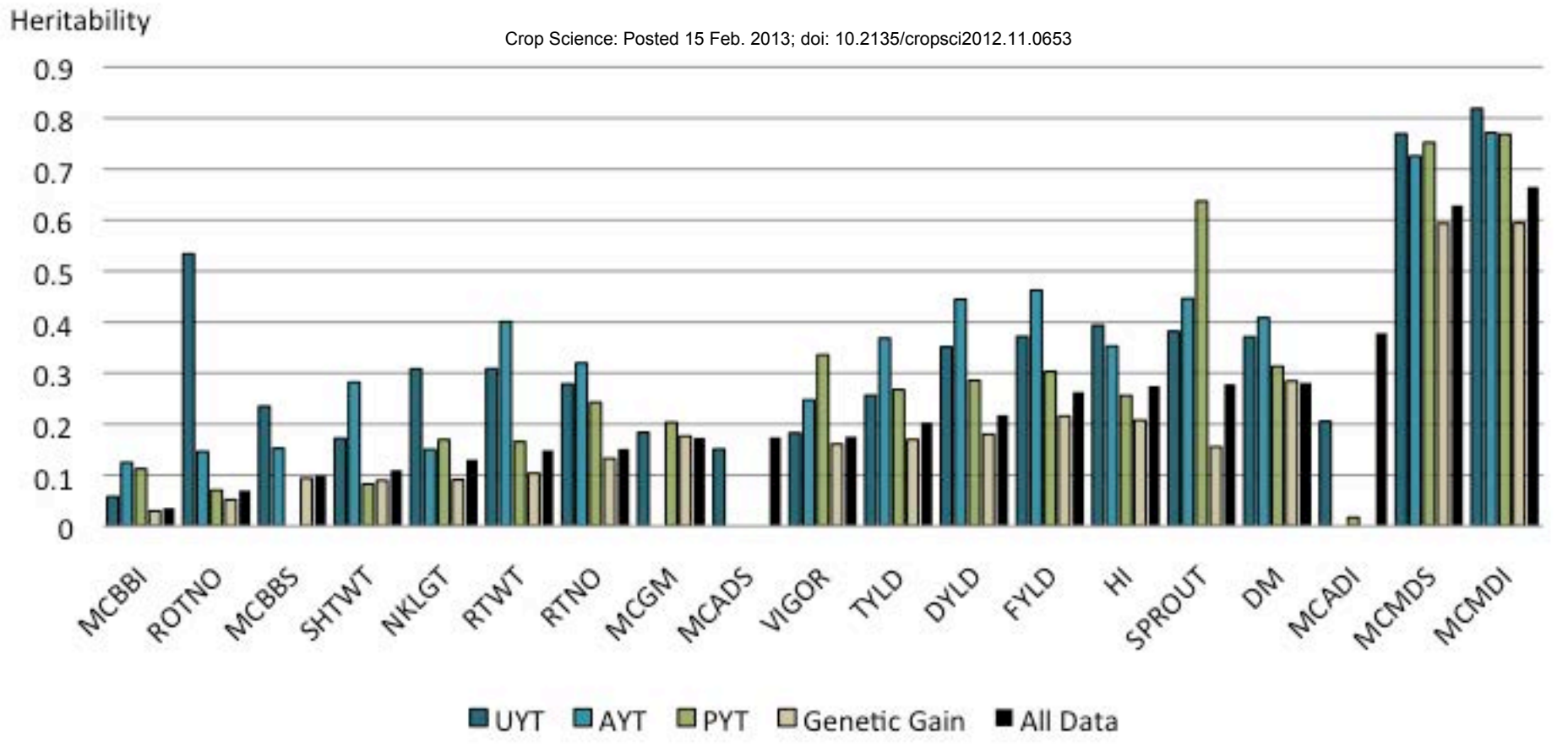
MCBBI	0.04	0.238	0.229	0.255	0.260	0.26	0.266

1

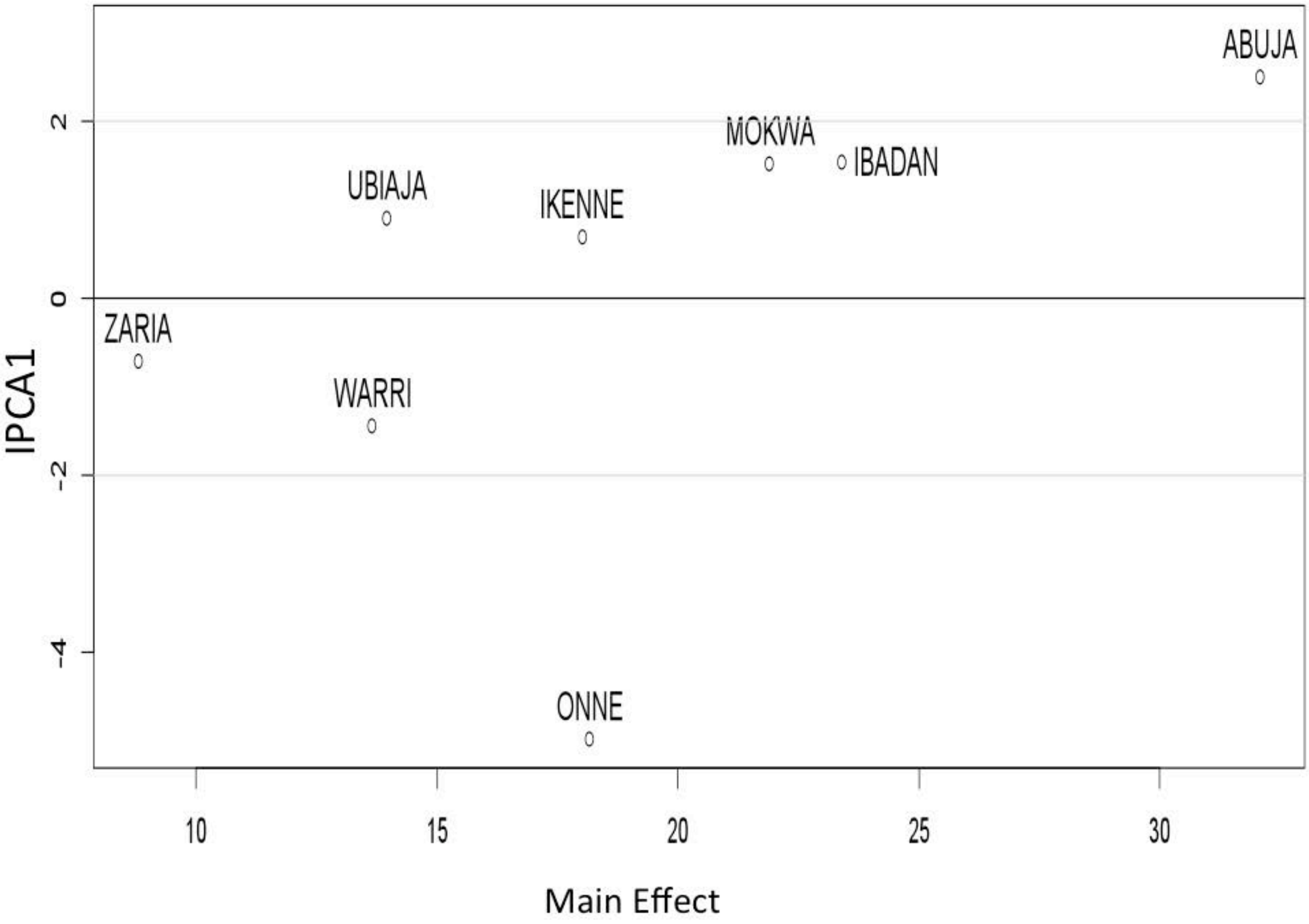
2 For trait abbreviations, refer to Table 1.



Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

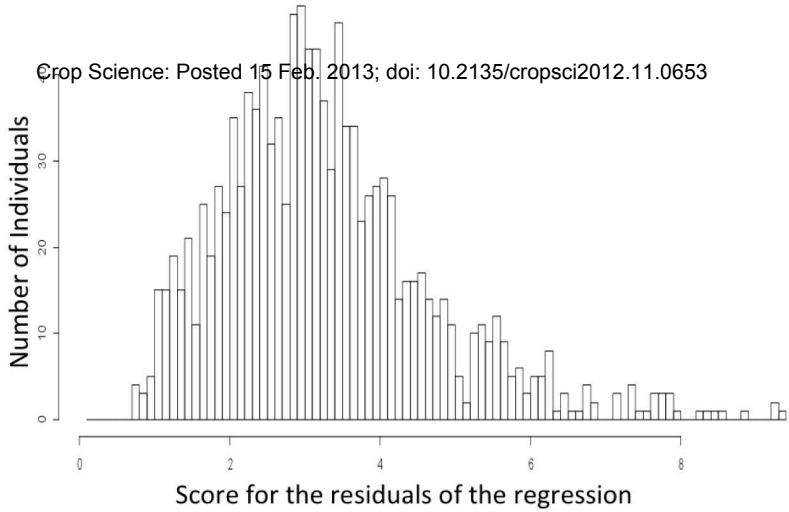


Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

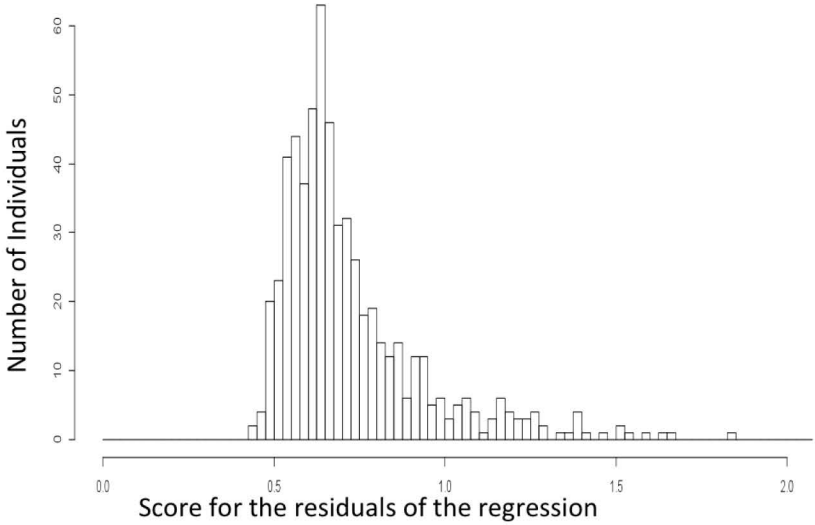


Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653

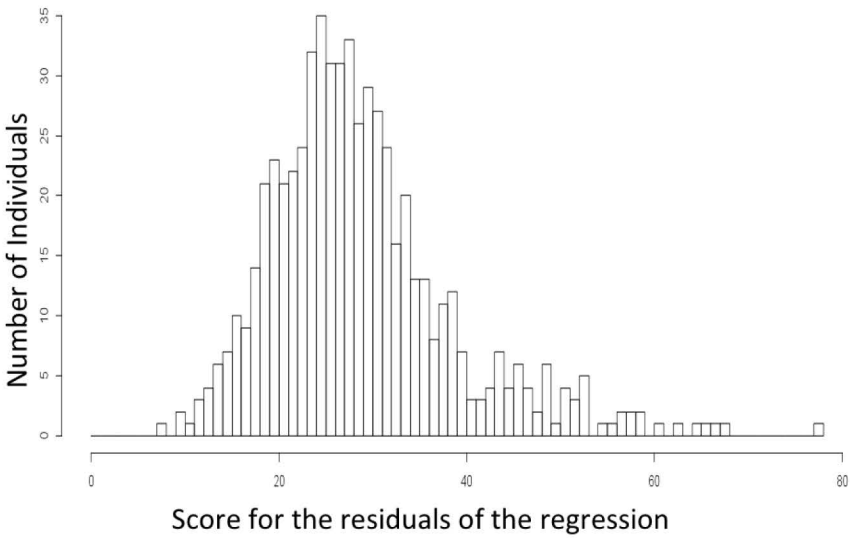
A.



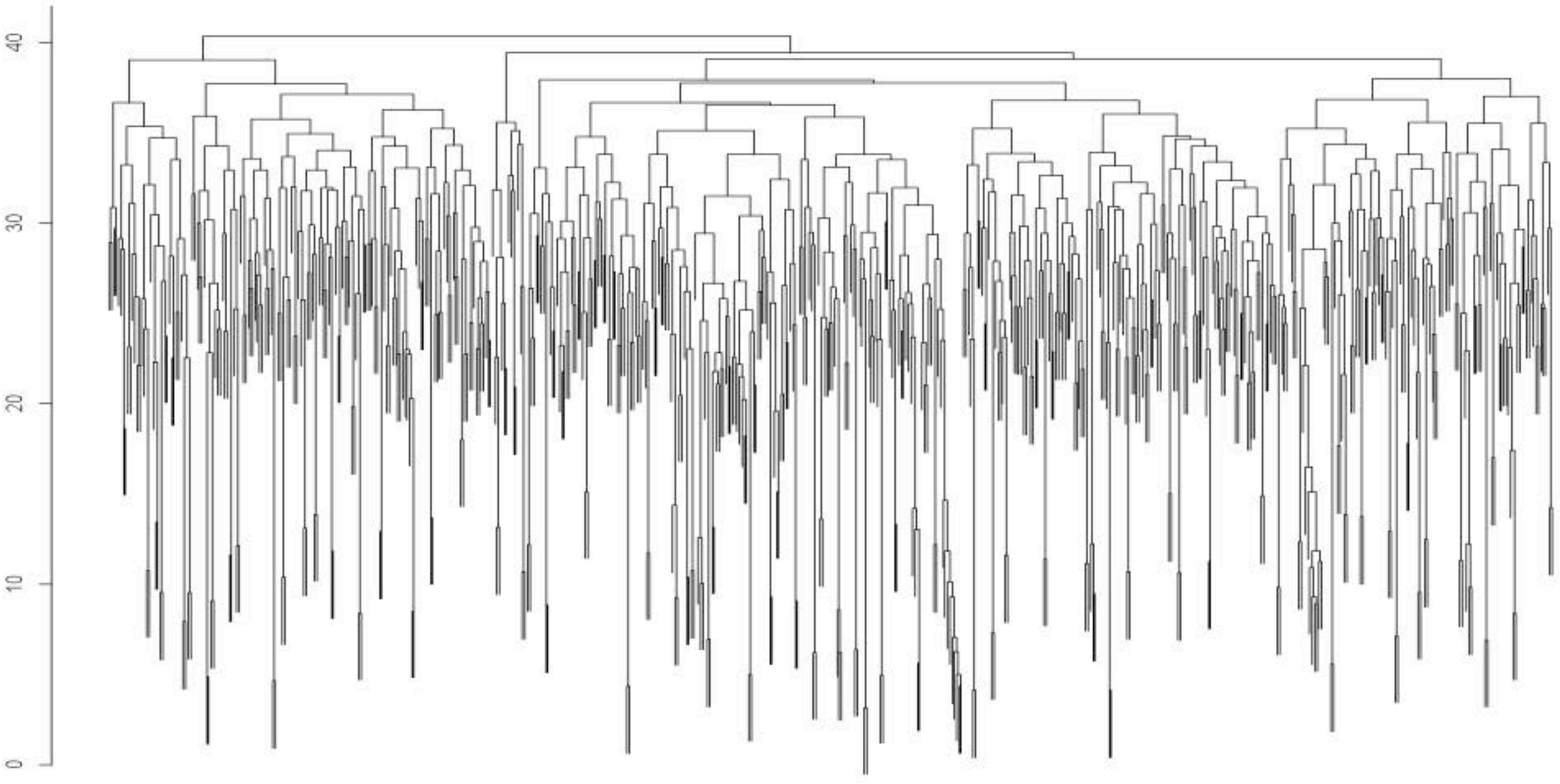
B.



C.



Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653



Crop Science: Posted 15 Feb. 2013; doi: 10.2135/cropsci2012.11.0653