New accuracy estimators for genomic selection with application in a cassava (*Manihot esculenta*) breeding program

C.F. Azevedo[1], M.D.V. Resende[1,2], F.F. Silva[3], J.M.S. Viana[4], M.S.F. Valente[4], M.F.R. Resende Jr[5] and E.J. Oliveira[6]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2]Embrapa Floresta, Colombo, PR, Brasil
[3]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[4]Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[5]RAPiD Genomics, Florida Innovation Hub, Gainesville, FL, USA
[6]Embrapa Mandioca e Fruticultura, Cruz das Almas, BA, Brasil

Corresponding author: C.F. Azevedo
E-mail: camila.azevedo@ufv.br

Genet. Mol. Res. 15 (4): gmr.15048838
Received May 31, 2016
Accepted July 4, 2016
Published October 5, 2016
DOI http://dx.doi.org/10.4238/gmr.15048838

ABSTRACT. Genomic selection is the main force driving applied breeding programs and accuracy is the main measure for evaluating its efficiency. The traditional estimator (TE) of experimental accuracy is not fully adequate. This study proposes and evaluates the performance and efficiency of two new accuracy estimators, called regularized estimator (RE) and hybrid estimator (HE), which were applied to a practical cassava breeding program and also to simulated data. The simulation study considered two individual narrow sense heritability levels and two genetic architectures for traits. TE, RE, and HE were

compared under four validation procedures: without validation (WV), independent validation, ten-fold validation through jacknife allowing different markers, and with the same markers selected in each cycle. RE presented accuracies closer to the parametric ones and less biased and more precise ones than TE. HE proved to be very effective in the WV procedure. The estimators were applied to five traits evaluated in a cassava experiment, including 358 clones genotyped for 390 SNPs. Accuracies ranged from 0.67 to 1.12 with TE and from 0.22 to 0.51 with RE. These results indicated that TE overestimated the accuracy and led to one accuracy estimate (1.12) higher than one, which is outside of the parameter space. Use of RE turned the accuracy into the parameter space. Cassava breeding programs can be more realistically implemented using the new estimators proposed in this study, providing less risky practical inferences.

Key words: Genomic prediction; Accuracy estimator; Cross-validation; Cassava breeding

INTRODUCTION

Genome wide selection (GWS; Meuwissen et al., 2001) is a technique of great importance in plant and animal breeding, allowing efficiency in genetic evaluation and prediction of genetic gains. It is based on genomic values predicted by phenotypes and a large number (n) of molecular markers widely distributed in the genome. Genomic breeding values (GBV) of N individuals are predicted by appropriate functional models, which estimate the effect of each marker on phenotypes, allowing early identification of the genetically superior individuals. However, genomic prediction poses statistical challenges such as estimability, owing to the high dimensionality problem (N << n case), and multicollinearity between the covariates, since the molecular markers are highly correlated. These challenges require the use of statistical methods to consider the regularization of the estimation process and/or the selection of covariates (Gianola et al., 2003).

To address these drawbacks, many statistical methods have been proposed for genomic selection, such as penalized estimation methods (Meuwissen et al., 2001; Van Raden, 2008), Bayesian estimation (Meuwissen et al., 2001; de Los Campos et al., 2009; Habier et al., 2011; Legarra et al., 2011), implicit regression methods (Gianola et al., 2006; Gianola et al., 2009; de Los Campos et al., 2009), and methods with dimensional reduction (Solberg et al., 2009; Azevedo et al., 2014; Azevedo et al., 2015a), among others such methods for genome-wide association studies (Hayes, 2013) and methods for additive-dominance genomic models (Azevedo et al., 2015b). These methods were typically evaluated based on their prediction accuracy.

Accuracy is the main measure for evaluating the efficiency of the prediction of GBV. A traditional estimator for the experimental accuracy of GWS was introduced by Legarra et al. (2008) and Hayes et al. (2009a). This estimator is the ratio between the predictive ability and the square root of the trait heritability. However, in some circumstances, this estimator has inconsistencies, such as the fact that the higher the trait heritability, the lower the accuracy, and can lead to estimates outside the parameter space (higher than 1). Ould Estaghvirou et al. (2013) conducted a comparative study among alternative accuracy estimators for GWS, but such estimators only differed from the estimator of Legarra et al. (2008) and Hayes et

al. (2009a) in the estimation of trait heritability. In fact, to date, there are no other proposed estimators for the experimental (after obtaining data) accuracy of genomic selection, but only for the expected (before obtaining data) accuracy (Resende et al., 2008; Goddard, 2009; Hayes et al., 2009b; Daetwyler et al., 2008, 2010; Goddard et al., 2011).

Thus, we propose an estimator for experimental accuracy, called the regularized estimator (RE), which is given by the multiplication of the traditional estimator by the square root of the molecular heritability. This adjustment corrects the accuracy estimator so that it occurs within the parameter space (from 0 to 1) and produces less biased and more precise estimates. Furthermore, the predictive ability combined with the proportion of genetic variance explained by markers yields a hybrid estimator (HE), which uses both the experimental information and the theoretical expectation.

In addition, the estimation of accuracy is associated with the validation form and the marker selection, because when thousands of effects are estimated, there is a risk of over-parameterization, i.e., experimental errors in the data explaining marker effects (Meuwissen, 2007). Thus, it is necessary to study and assess the behavior of the accuracy estimators in different forms of validation and under markers selection. It is noteworthy that recent methodologies for GWS have been evaluated with simulation studies (Piccoli et al., 2014; Bhering et al., 2015).

Cassava (*Manihot esculenta*) is a crop of great economic importance and the efficiency of its breeding programs should be investigated. Cassava genetic resources have been evaluated for the prediction of breeding values of traits of interest (Nassar, 2007; Graciano-Ribeiro et al., 2009); genomic selection may be of great value for this crop. Thus, the objective of this study was to propose and evaluate the performance and efficiency of the two new estimators (called regularized and hybrid) for the accuracy of GWS, which includes the predictive ability and both genomic heritability and pedigree-based heritability, considering different validation forms and the selection of markers. The estimators were also evaluated for cassava because of its importance in plant breeding programs and to elucidate the importance of the estimation of accuracy during practical genomic selection.

MATERIAL AND METHODS

Experimental material

Genomic selection was conducted on five traits evaluated in cassava (*M. esculenta*). The experiment was arranged in a randomized complete block design with three replications and 10 plants per plot, including 358 cassava accessions belonging to the germplasm collection of the Embrapa Cassava which were genotyped for 390 SNP markers. The trial was established at Cruz das Almas, Brazil, under the guidelines of Embrapa. The traits evaluated were shoot weight (SW), fresh root yield, amylose content, dry matter content, and shoot yield. The RR-BLUP method (described below) was used.

Simulated datasets

The simulated data set was described by Azevedo et al. (2015b). A total of 2000 equidistant SNP markers separated by 0.1 cM across ten chromosomes were simulated. One hundred of the markers were genes (QTL). A total of 1000 individuals from 20 full-sib families were genotyped and phenotyped.

Scenarios

There were four scenarios studied: 1) heritability of 0.22 in a trait controlled by genes with small effects; 2) heritability of 0.37 in a trait controlled by genes with small effects; 3) heritability of 0.20 in a trait controlled by many genes with small effects and some with major effects; and 4) heritability of 0.32 in a trait controlled by many genes with small effects and some with major effects. These 4 scenarios were analyzed considering four forms of validation and three accuracy estimators. Each scenario was simulated 10 times. Accuracies and genomic heritabilities were calculated for each replicate of the simulation and averaged across replicates. For validation forms based on folds, the accuracies and genomic heritabilities were also averaged across validation cycles in each replicate of the population.

Traditional accuracy estimator (TE)

This estimator is traditionally applied in evaluating the efficiency of the GWS method in several breeding programs of different organisms (Resende Jr et al., 2012; Resende et al., 2012, 2014; Azevedo et al., 2015a). Considering the phenotypic model given by $y = \mu + g + e$, where μ is the overall mean, g is a genetic effect, and e is an residual effect, an accuracy estimator for genomic selection ($r_{\hat{y}g}$) was proposed by Legarra et al. (2008) and Hayes et al. (2009a) and is given by $r_{\hat{y}g} = (r_{\hat{y}y}/h)$ or $r_{\hat{y}g} = (r_{\hat{y}y}/r_{\hat{a}a})$, respectively, where $r_{\hat{y}y}$ is the predictive ability of GWS represented by the correlation between the phenotype y and the predicted genomic breeding values $\hat{y}$, $r_{\hat{a}a}$ is the accuracy of the pedigree-based evaluation (a and â are the true additive genetic value and estimated additive genetic value, respectively), and h^2 is the heritability of the trait given by $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$, where σ_g^2 is the genetic variance and σ_e^2 is the residual variance. It can be seen that h is itself the accuracy ($r_{\hat{a}a}$) of the individual phenotypic selection.

Considering that the quantities $\hat{g}_M$ and g stand for the predicted genomic breeding value and true breeding value (with variance σ_g^2), respectively. The algebraic proof of this estimator can be made considering the corrected phenotypic values ($y = g + e$) and predicted genomic breeding values ($\hat{y}$) in the validation population and under the assumption that $\hat{y}$ only captures genetic effects ($\hat{y} = \hat{g}_M$), i.e., it does not explain any environmental effect (e). This assumption holds only when the validation is perfect (totally independent) and simultaneously LD (linkage disequilibrium) between the marker and QTL is complete. Thus, the covariance of these two variables is:

$$Cov(\hat{y}, y) = Cov(\hat{g}_M, g + e) = Cov(\hat{g}_M, g) = \sigma_g^2 \qquad \text{(Equation 1)}$$

This leads to $Cov(\hat{g}_M, g) = \sigma_g^2$, indicating that $\hat{g}_M$ captures all of g. The variances of $\hat{y}$ and y are $Var(\hat{y}) = \sigma_{\hat{y}}^2 = Var(\hat{g}_M) = \sigma_{g_M}^2$ and $Var(y) = \sigma_y^2 = \sigma_g^2 + \sigma_e^2 = \sigma_g^2/h^2$, respectively.

Thus, the predictive ability equals the correlation value between $\hat{y}$ and y, and is given by:

$$r_{\hat{y}y} = Cor(\hat{y}, y) = Cov(\hat{y}, y)/(\sigma_{\hat{y}}\,\sigma_y) = \sigma_g^2/(\sigma_{\hat{y}}\,\sigma_y) = \sigma_g^2/(\sigma_{\hat{y}}\,(\sigma_g/h)) = r_{\hat{y}g}h. \qquad \text{(Equation 2)}$$

Therefore, the accuracy estimator is equal to $r_{\hat{y}g} = r_{\hat{y}y}/h$.

However, this estimator has some inconsistencies. In $r_{\hat{y}g} = r_{\hat{y}y}/h$, given $r_{\hat{y}y}$, if the square root of the trait heritability h is larger than $r_{\hat{y}g}$ is smaller, a fact that is not consistent with reality, because, theoretically, the higher the heritability of the trait, the greater the accuracy of selection. In addition, this estimator produces $r_{\hat{y}g}$ outside (greater than 1) the parameter space from 0 to 1. Conceptually, it does not make sense that $r_{\hat{y}g}$ tends to infinity when h approaches zero. In practice, this can lead to the situation where a population without genetic variability (heritability of zero) is able to show accuracy of reasonable magnitude with GWS, which is inconsistent with the genomic selection theory.

RE

The proposed regularized estimator is given by $r_{g_M g} = r_{\hat{y}y}\frac{h_M}{h}$ or $r_{g_M g} = r_{\hat{y}g}h_M$, where h_M is the square root of the genomic or molecular heritability. This estimator multiplies the proximity (distance) between y and $\hat{y}$ given by $r_{\hat{y}y}$ by the proximity (distance) between h_M and h (given by $\frac{h_M}{h}$), producing the proximity (distance) between g_M and g. In this case, multiplying $r_{\hat{y}y}$ by $\frac{h_M}{h}$ decreases the accuracy, i. e., penalizes $r_{\hat{y}y}$, which makes sense biologically.

Under the assumption that $\hat{y}$ captures both genetic and environmental effects, i.e., $\hat{y} = \hat{g}_M + \hat{e}$, as the LD is imperfect ($Cov(\hat{g}_M, g) = \sigma^2_{g_M}$) and/or validation is not performed or is inaccurate, which leads to covariance of errors ($Cov(\hat{e}, e) = \sigma^2_e$). In this case, we have:

$$\begin{aligned} Cov(\hat{y}, y) &= Cov(\hat{g}_M + \hat{e}, g + e) = Cov(\hat{g}_M, g) + Cov(\hat{e}, e) = Var(\hat{g}_M) + Var(e) \\ &= \sigma^2_{g_M} + \sigma^2_e = Var(y_M) \end{aligned} \quad \text{(Equation 3)}$$

The variances of the corrected phenotypic values y and predicted values $\hat{y}$ are equal to $Var(y) = \sigma^2_y = \sigma^2_g + \sigma^2_e = \sigma^2_g / h^2$ and $Var(\hat{y}) = \sigma^2_{y_M} = \sigma^2_{g_M} / h^2_M$, respectively.

Thus, the predictive ability equals the correlation between y and $\hat{y}$ given by:

$$\begin{aligned} r_{\hat{y}y} &= Cor(\hat{y}, y) = Cov(\hat{y}, y) / (\sigma_{\hat{y}}\, \sigma_y) = \sigma^2_{y_M} / (\sigma_{y_M}\, \sigma_y) = \sigma^2_{y_M} \Big/ \left(\frac{\sigma_{g_M}}{h_M}\frac{\sigma_g}{h}\right) = \\ &= \frac{\sigma^2_{y_M} h_M h}{\sigma_{g_M}\sigma_g} = \frac{\sigma_{y_M} h \sigma_{g_M}}{\sigma_{g_M}\sigma_g}\frac{\sigma_{g_M}}{\sigma_{g_M}} = \frac{\sigma_{y_M} h \sigma^2_{g_M}}{\sigma_{g_M}\sigma_g\sigma_{g_M}} = \frac{\sigma_{y_M} h r_{g_M g}}{\sigma_{g_M}} = \frac{h r_{g_M g}}{h_M} \end{aligned} \quad \text{(Equation 4)}$$

and, therefore, the accuracy estimator is equal to $r_{g_M g} = r_{\hat{y}y}\frac{h_M}{h}$. The regularized estimator $r_{g_M g} = r_{\hat{y}y}\frac{h_M}{h}$ has a good statistical property, producing $r_{g_M g}$ in the parameter space as the product of two fraction, since $\frac{h_M}{h}$ is always smaller or equal to 1. In addition, the lower the $\frac{h_M}{h}$, the lower the $r_{g_M g}$, which is consistent with reality. Since the smaller $\frac{h_M}{h}$, the lower the proportion of the variance explained by the markers and therefore less accurate. Thus, the inconsistencies in the traditional formula are corrected. This estimator is conservative (produces smaller accuracies than the traditional formula) and is a function of three parameters, not two, also including genomic heritability. The ratio of the heritabilities takes into account the efficiency of markers in capturing QTLs, i.e., it considers the degree of imperfection in LD.

It is important to emphasize that the genomic or molecular heritability (h^2_M) must be equal to or less than the heritability of the trait (h^2) as h^2_M is the fraction of h^2 that is captured by markers, and the maximum limit of the squared accuracy of GWS is h^2_M (de Los Campos and Sorensen, 2013; de Los Campos et al., 2014). The molecular heritability is given by $h^2_M = \sigma^2_{g_M} / (\sigma^2_{g_M} + \sigma^2_e)$, where $\sigma^2_{g_M} = \sum_{i=1}^{n} 2p_i q_i \sigma^2_{m_i}$ is the additive genomic variance, $\sigma^2_{m_i}$, and p_i and q_i are the variance and allele frequency of marker i, respectively (Gianola et al., 2009).

HE

The HE combines the experimental predictive ability and the theoretical expectation of $r^2_{mq} = \sigma^2_{g_M} / \sigma^2_g$, which also is a regularized estimator. Assuming $\sigma^2_{y_M} \approx \sigma^2_y$, we have $h^2_{mq} / h^2 \approx \sigma^2_{g_M} / \sigma^2_g$. Given that $r^2_{mq} = \sigma^2_{g_M} / \sigma^2_g$ is the proportion of g explained by markers, we have $h_M / h = r_{mq}$. Thus, using the RE, the accuracy is given by $r_{g_M g} = r_{\hat{y}y} \frac{h_M}{h} = r_{\hat{y}y} r_{mq} = r^*_{g_M g}$, where $r^2_{mq} = \frac{n}{n + M_e}$, n is the number of markers and $M_e = 2N_eL$, N_e is the effective population size, and L is the genome size in Morgans (Goddard et al., 2011; Meuwissen et al., 2011). It is noteworthy that r^2_{mq} changes with the number of selected markers. The HE provides an accuracy estimate without requiring the estimation of h^2_M and h^2. In addition, it penalizes the selection of a very small number of markers.

A deterministic formula for the predictive ability is derived and presented in Appendix 1. It demonstrates the relationship between the predictive ability and the genomic heritability.

Parametric accuracy

Parametric accuracies under the additive model were computed using the formula of Resende et al. (2008); Goddard et al. (2011); Grattapaglia and Resende (2011); Resende et al. (2014): $r_{\hat{g}g} = \sqrt{\frac{r^2_{mq}(Nr^2_{mq}h^2 / n_{QTL})}{1 + Nr^2_{mq}h^2 / n_{QTL}}}$, where n_{QTL} is the number of QTL, N is number of genotyped and phenotyped individuals, h^2 is the trait heritability and r^2_{mq} proportion of genetic variance explained by markers.

Supervised RR-BLUP

In the context of genomic selection, according to Meuwissen et al. (2001), the basic linear model is as follows:

$$y = 1\mu + Xm + e \qquad \text{(Equation 5)}$$

where y is the vector of phenotypes (N X 1, where N is the number of genotyped and phenotyped individuals), 1 is a vector with all entries equal to 1 (N X 1), μ is the average of the trait, m is the vector of additive genetic marker effects (n X 1, where n is the number of markers) with incidence matrix X (N X n), and $m \sim N(0, I\sigma^2_m)$ where σ^2_m is the variance of markers, e is the model of the residual vector with $e \sim N(0, I\sigma^2_e)$, and σ^2_e is the residual variance. In this study, we evaluated the accuracies under the RR-BLUP (Meuwissen et al., 2001) and RR-BLUP_B (a marker selection procedure by Resende Jr et al., 2012) methods.

The RR-BLUP method was applied to each data set considering the total number of markers. After this preliminary analysis, markers groups of size 50, 100, 150, and so on up to 2,000 were selected based on the highest magnitudes of their effects.

Estimation and validation populations

The estimation population is used to estimate the effects of the markers while the validation population is used to analyze the efficiency of the estimated effects in recovering individual genomic values in an independent sample of the population.

An estimation-validation approach was cross-validated by the jacknife procedure. According to this method, the original data set with 1000 individuals was divided into 10 training data sets of 100 individuals each (ten-folds procedure or TF) (Efron, 1982).

The genomic heritability was estimated as the average of heritabilities obtained in each of the cycles. The estimation of accuracy was done in two ways: first as the average of the predictive abilities obtained in each of the cycles allowing for different markers to be selected (TFD scheme) in each cycle; and second, restricting the selected markers to be the same (TFS scheme) in each cycle, after identifying them through the average effects across ten cycles. Under the supervised RR-BLUP method, the markers with the greatest magnitudes of effects were selected. However, the chosen sets of markers were alternatively variable in each cycle (form I of the jacknife validation or TFD) or unique for all validation cycles (form II of the jacknife validation or TFS).

In this study, the following forms of validation were considered, i) physically distinct (2 different subpopulations), called independent validation (IV); ii) jacknife procedure considering forms I (TFD) and II (TFS); and iii) without validation (WV) with the population used for estimation and validation at the same time.

The data were simulated using the RealBreeding software (Viana, 2011). All computational analyses were implemented in the R software (R Development Core Team, 2010) using the *rrBLUP* package and *mixed.solve* function.

RESULTS

Results on the average accuracy using the RE, the TE, and the HE for independent validation (900 and 100 individuals in the estimation and validation populations, respectively) are presented in Figure 1. For IV, at high heritability (0.37 and 0.32), RE and TE performed similarly in terms of the distance between the estimated accuracy and the parametric accuracy, but RE underestimated and TE overestimated the accuracy, thus favoring RE. For IV at low heritability (0.22 and 0.20), TE showed a smaller distance from the parametric accuracy but it still overestimated accuracy, while RE underestimated it. HE was poorer for IV, underestimating the accuracy.

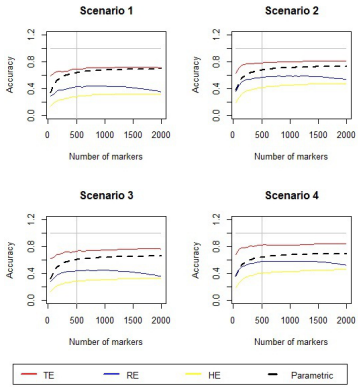

Figure 1. Average prediction accuracy by the regularized estimator (RE - blue), the traditional estimator (TE - red), and the hybrid estimator (HE - yellow) for independent validation. The gray line indicates the maximum accuracy (1); 500 refers to the number of markers which maximizes the estimated accuracy. The scenarios are defined as Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; and Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

Results of the average accuracy for the RE, TE, and HE by the jacknife procedure (with k = 100) considering the I and II validation forms (TFD and TFS, respectively) are presented in Figure 2.

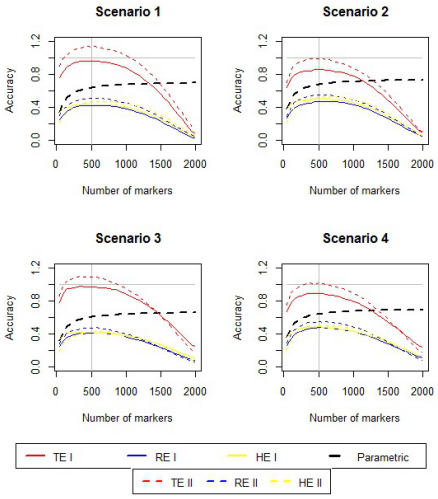

Figure 2. Average prediction accuracy by the regularized estimator (RE - blue), the traditional estimator (TE - red), and the hybrid estimator (HE - yellow) for the jacknife procedure considering forms I and II (TFD and TFS, respectively). The gray line indicates the maximum accuracy (1); 500 refers to the number of markers which maximizes the estimated accuracy. The scenarios are defined as Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; and Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

The behaviors of the genomic heritability (h_M^2) estimates through cross-validation were similar to the parametric accuracy curve by IV, reaching a maximum and then staying constant (Figure 3). The predictive ability with cross-validation showed a different behavior, decreasing with an increase in the numbers of markers (Figure 3).

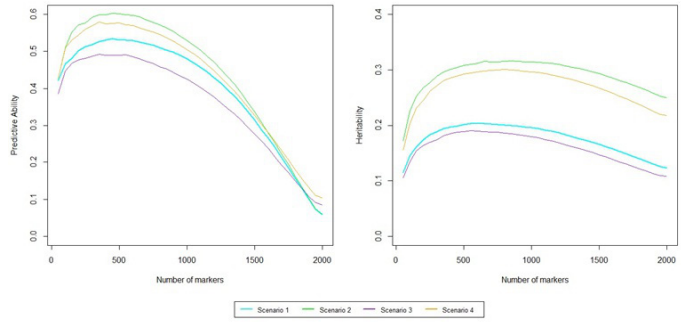

Figure 3. Behavior of the predictive ability and genomic heritability across selected groups of SNPs with cross-validation. The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

The results of the average accuracy for the RE, the TE, and HE without validation (WV) are presented in Figure 4.

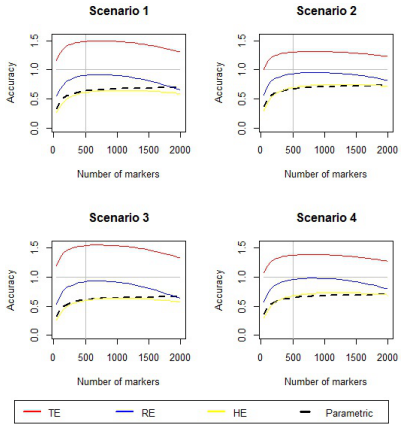

Figure 4. Average prediction accuracy by the regularized estimator (RE - blue), the traditional estimator (TE - red), and the hybrid estimator (HE - yellow) without validation. The gray line indicates the maximum accuracy (1); 500 refers to the number of markers which maximizes the estimated accuracy. The scenarios are defined as Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; and Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

The standard deviations of the accuracy estimates with RE, TE, and HE are presented in Figures 5 and 6 for independent validation and forms I (TFD) and II (TFS) of the jacknife procedure, respectively. For independent validation the best (the ones with smaller standard deviation) estimators for the accuracy were HE, RE, and TE, in this order, for all scenarios. For the jacknife procedures the same tendencies were observed. In this case, the RE and HE methods were by far more precise than the TE method.

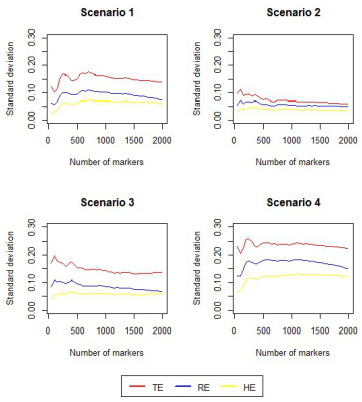

Figure 5. Standard deviation of the accuracy by the regularized estimator (RE - blue), the traditional estimator (TE - red), and the hybrid estimator (HE - yellow) for independent validation. The scenarios are defined as Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; and Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

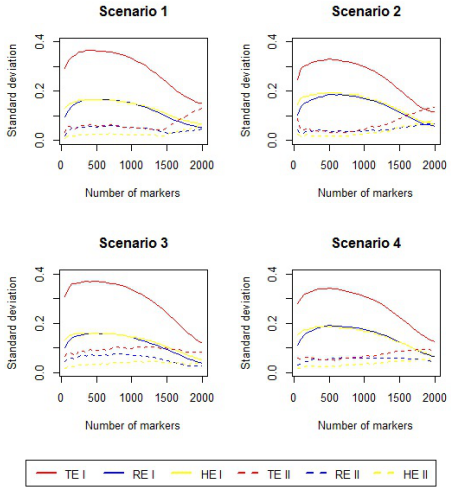

Figure 6. Standard deviation of by the regularized estimator (RE - blue), the traditional estimator (TE - red), and the hybrid estimator (HE - yellow) for the jacknife procedure considering forms I and II (TFD and TFS, respectively). The scenarios are defined as Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; and Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

To elucidate the importance of the new estimators for evaluating the quality of prediction in genomic selection, the results for an applied breeding program of cassava are presented in Table 1. Accuracies ranged from 0.67 to 1.12 with the TE estimator, which turned out 0.22 and 0.51 with the RE estimator. These results revealed that TE overestimated the accuracy and led to one accuracy estimate (1.12) higher than one, which is outside of the parameter space. Under the same conditions, the RE estimator produced a value of 0.22, which is in the parameter space. The estimates with TE have the highest possible accuracy and those with RE have the lowest possible accuracy. Such inferences provide greater reliability in the practical evaluation of the efficiency of genomic selection in cassava improvement. For example, for the SW trait, the most likely accuracy is between 0.51 and 0.83, being more likely close to 0.51. Besides being more precise, the new estimator provides more conservative inferences, which are less risky from a practical standpoint. Overall, the cassava breeding program can be more realistically implemented using the new estimators proposed in this study.

Table 1. Estimates of genomic heritability (h_M^2), trait heritability (h^2), predictive ability ($r_{\hat{y}y}$), and accuracy by the traditional estimator ($r_{\hat{y}g}$- TE) and by regularized estimator (r_{g_Mg}- RE) for the traits shoot weight (SW), fresh root yield (FRY), dry matter content (DMC), shoot yield (SY), and amylose content (AC) in cassava (*Manihot esculenta*).

	SW	FRY	DMC	SY	AC
h_M^2	0.37	0.29	0.26	0.32	0.04
h^2	0.37	0.35	0.30	0.38	0.05
$r_{\hat{y}y}$	0.51	0.40	0.37	0.39	0.25
$r_{\hat{y}g}$ (TE)	0.83	0.68	0.68	0.67	1.12
r_{g_Mg} (RE)	0.51	0.37	0.35	0.38	0.22

DISCUSSION

The behavior of the parametric accuracy curve, along with the number of markers with the greatest effects in the analysis, show that an asymptote is reached around 500 markers and, subsequently, a constant value is maintained up to the total number of 2000 markers. For IV (Figure 1), estimators of the accuracies approximately reach the maximum value at 500 selected markers. This result indicates that, on average, 5 markers are enough to capture one QTL. After that point, the accuracy tends to stabilize. This occurred for all estimators, all scenarios in IV, and also for the WV case. These results were similar to the parametric case.

The results showed notably greater distances from the parametric values when considering the IV (Figure 1) than jacknife procedures (Figure 2) for all estimators. In addition, the IV was superior to the jacknife procedures. This is consistent with the results of Wray et al. (2013a, b). The IV was superior to the jacknife procedures, approaching better parametric accuracy with or without marker selection. With RE, the TFS was better than the TFD validation scheme.

For the key value of 500 markers, RE produces accuracy values very close to parametric ones in scenarios 2 and 4 (Figure 1) and all scenarios in Figure 2. In these situations, the TE estimator overestimated the accuracies. In addition, TE exceeded the parameter space (values greater than 1) in some scenarios (Figure 2). This is not permissible for a good estimator. For RE and HE, this never occurred. In the comparative study of Ould Estaghvirou et al. (2013), selection of markers was not considered, but out of space parameter estimates were still obtained. This result has also been observed in genomic selection analysis using real data (Resende et al., 2012).

For TFD and TFS validations when (number of markers equal to 500) the maximum accuracy is reached, RE and HE were approximately coincident and superior to TE in terms of both distance (smaller bias) and direction of the estimation (underestimation) in all four scenarios. In this case (number of markers equal to 500), the parametric values were 0.64, 0.68, 0.61, and 0.65; the estimates for RE were 0.51, 0.55, 0.48, and 0.55; and the estimates for TE were 1.13, 0.99, 1.09, and 1.02, for the four scenarios, respectively. This shows that TE should not be used with cross-validation.

For TFD and TFS, estimators of the accuracies approximately reach the maximum value at 500 selected markers. Following this, they tend to decrease (Figure 2). The decay in accuracy with increasing number of markers is consistent with Fernando et al. (2007) also working with RR-BLUP and two thousand markers. As in Figure 3, Wray et al. (2013a,b) also found that the predictive ability ($r_{\hat{y}y}$) decreases with increasing number (n) of independent markers and consequently with increasing of the ratio $\frac{n}{N}$ (N = number of individuals). This result is due to the fact that increasing n causes greater variance of the estimated genetic relationship coefficients (Wray et al., 2013a,b). This decrease in predictive ability was also observed in association with cross-validation for a few traits (Resende et al., 2012).

For WV, the HE very closely matched the parametric accuracy curve while RE and TE overestimated it. However, when all markers were used, RE also had accuracies close to the parametric. In general, compared to TE, the regularized estimator presented accuracies closer to the parametric ones, mainly when selecting markers. It was also less biased and more precise, with smaller standard deviations than the traditional estimator. In all scenarios, RE and HE are conservative with respect to estimation of accuracy. In theory, RE assumes that validation was poor and HE does not need validation. Thus, without validation, the best results were for HE (coincident with parametric value) and RE, in this order. Without marker selection, RE without validation perfectly matched the parametric accuracy. In contrast, TE

was the worst estimator and exceeded the parameter space in the entire curve, proving to be completely inadequate without validation or when it validation is poor. Even in such cases, the use of an adequate accuracy estimator (other than TE) can guarantee a valid estimate.

TE can be used only with IV and, even in this case, it overestimates accuracy and is less precise. The HE proved to be very effective in the absence of validation and in the jacknife procedures, but it is not recommended for IV. The regularized estimator revealed that not only the predictive ability of GWS methods matters, but also their capacity of precisely estimating the genomic heritability (Azevedo et al., 2015b). Thus, not only the predictive ability and bias should be used in comparing methods but also the genomic heritability and accuracy produced by the new method should be considered. The following inferences can be made according to the accuracy estimator and kind of validation: i) for the best accuracy, HE without validation should be used; ii) for the highest possible accuracy, TE with independent validation should be used; and iii) for the lowest possible accuracy, RE with independent validation should be used.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors thank the following Brazilian funding organizations: CAPES and CNPq. Acknowledgments to FAPEMIG for financial resources for publication costs.

REFERENCES

Azevedo CF, Silva FF, de Resende MD, Lopes MS, et al. (2014). Supervised independent component analysis as an alternative method for genomic selection in pigs. *J. Anim. Breed. Genet.* 131: 452-461. http://dx.doi.org/10.1111/jbg.12104

Azevedo CF, Nascimento M, Silva FF; Resende MDV, et al. (2015a). Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. *Genet. Mol. Res.* 14: 1 2217-12227.

Azevedo CF, de Resende MD, E Silva FF, Viana JMS, et al. (2015b). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC Genet.* 16: 105. http://dx.doi.org/10.1186/s12863-015-0264-2

Bhering LL, Junqueira VS, Peixoto LA, Cruz CD, et al. (2015). Comparison of methods used to identify superior individuals in genomic selection in plant breeding. *Genet. Mol. Res.* 14: 10888-10896. http://dx.doi.org/10.4238/2015.September.9.26

Daetwyler HD, Villanueva B and Woolliams JA (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3: e3395. http://dx.doi.org/10.1371/journal.pone.0003395

Daetwyler HD, Pong-Wong R, Villanueva B and Woolliams JA (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185: 1021-1031. http://dx.doi.org/10.1534/genetics.110.116855

de los Campos G and Sorensen DA (2013). A commentary on Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14: 894-894. http://dx.doi.org/10.1038/nrg3457-c1

de Los Campos G, Gianola D and Rosa GJM (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J. Anim. Sci.* 87: 1883-1887. http://dx.doi.org/10.2527/jas.2008-1259

de Los Campos G, Sorensen DA and Gianola D (2014). Genomic heritability: what is it? Proceedings of the 10th WCGALP, Vancouver, Canada.

Efron B (1982). The jacknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia.

Ould Estaghvirou SB, Ogutu JO, Schulz-Streeck T, Knaak C, et al. (2013). Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC Genomics* 14: 860. http://dx.doi.org/10.1186/1471-2164-14-860

Fernando RL, Habier D, Stricker C, Dekkers JMC, et al. (2007). Genomic selection. *Acta Agric. Scand.* 57: 192-195.
Gianola D, Perez-Enciso M and Toro MA (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* 163: 347-365.
Gianola D, Fernando RL and Stella A (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173: 1761-1776. http://dx.doi.org/10.1534/genetics.105.049510
Gianola D, de los Campos G, Hill WG, Manfredi E, et al. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* 183: 347-363. http://dx.doi.org/10.1534/genetics.109.103952
Goddard M (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136: 245-257. http://dx.doi.org/10.1007/s10709-008-9308-0
Goddard ME, Hayes BJ and Meuwissen THE (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* 128: 409-421. http://dx.doi.org/10.1111/j.1439-0388.2011.00964.x
Graciano-Ribeiro D, Hashimoto DYC, Nogueira LC, Teodoro D, et al. (2009). Internal phloem in an interspecific hybrid of cassava, an indicator of breeding value for drought resistance. *Genet. Mol. Res.* 8: 1139-1146. http://dx.doi.org/10.4238/vol8-3gmr629
Grattapaglia D and Resende MDV (2011). Genomic selection in forest tree breeding. *Tree Genet. Genomes* 7: 241-255. http://dx.doi.org/10.1007/s11295-010-0328-4
Habier D, Fernando RL, Kizilkaya K and Garrick DJ (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12: 186. http://dx.doi.org/10.1186/1471-2105-12-186
Hayes BJ (2013). Overview of statistical methods for genome-wide association studies (GWAS). In: Genome-wide association studies and genomic prediction (Gondro C, Van Der Werf J and Hayes B, eds.). Humana Press, New York, 149-169.
Hayes BJ, Bowman PJ, Chamberlain AJ and Goddard ME (2009a). Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92: 433-443. http://dx.doi.org/10.3168/jds.2008-1646
Hayes BJ, Visscher PM and Goddard ME (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91: 47-60. http://dx.doi.org/10.1017/S0016672308009981
Legarra A, Robert-Granié C, Manfredi E and Elsen JM (2008). Performance of genomic selection in mice. *Genetics* 180: 611-618. http://dx.doi.org/10.1534/genetics.108.088575
Legarra A, Robert-Granié C, Croiseau P, Guillaume F, et al. (2011). Improved Lasso for genomic selection. *Genet. Res.* 93: 77-87. http://dx.doi.org/10.1017/S0016672310000534
Meuwissen T (2007). Genomic selection: marker assisted selection on a genome wide scale. *J. Anim. Breed. Genet.* 124: 321-322. http://dx.doi.org/10.1111/j.1439-0388.2007.00708.x
Meuwissen TH, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819-1829.
Meuwissen TH, Luan T and Woolliams JA (2011). The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *J. Anim. Breed. Genet.* 128: 429-439. http://dx.doi.org/10.1111/j.1439-0388.2011.00966.x
Nassar NMA (2007). Cassava genetic resources and their utilization for breeding of the crop. *Genet. Mol. Res.* 6: 1151-1168.
Piccoli ML, Braccini J, Cardoso FF, Sargolzaei M, et al. (2014). Accuracy of genome-wide imputation in Braford and Hereford beef cattle. *BMC Genet.* 15: 157. http://dx.doi.org/10.1186/s12863-014-0157-9
R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <http://www.R-project.org>.
Resende MDV, Lopes PS, Silva RL and Pires IE (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesq. Florest. Brasil.* 56: 63-78.
Resende MDV, Resende MFR, Jr., Sansaloni CP, Petroli CD, et al. (2012). Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol.* 194: 116-128. http://dx.doi.org/10.1111/j.1469-8137.2011.04038.x
Resende Jr MFR, Muñoz P, Resende MDV, Garrick DJ, et al. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). *Genetics* 190: 1503-1510. http://dx.doi.org/10.1534/genetics.111.137026
Resende RMS, Casler M and Resende MDV (2014). Genomic selection in forage breeding: accuracy and methods. *Crop Sci.* 54: 143-156. http://dx.doi.org/10.2135/cropsci2013.05.0353
Solberg TR, Sonesson AK, Woolliams JA and Meuwissen THE (2009). Reducing dimensionality for prediction of genome-wide breeding values. *Genet. Sel. Evol.* 41: 29. http://dx.doi.org/10.1186/1297-9686-41-29
VanRaden PM (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414-4423. http://dx.doi.org/10.3168/jds.2007-0980
Viana JMS (2011). Programa para análises de dados moleculares e quantitativos RealBreeding. UFV, Viçosa.
Wray NR, Yang J, Hayes BJ, Price AL, et al. (2013a). Author reply to A commentary on Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14: 894-894. http://dx.doi.org/10.1038/nrg3457-c2
Wray NR, Yang J, Hayes BJ, Price AL, et al. (2013b). Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14: 507-515. http://dx.doi.org/10.1038/nrg3457

Appendix 1

Deterministic formula for the predictive ability $r_{\hat{y}y}$, molecular heritability h_M^2, and accuracy $r_{\hat{g}g}$ of GWS.

From the expression $r_{mq}^2 = \frac{h_M^2}{h^2}$, we get $h_M^2 = r_{mq}^2 h^2$ (1)

From the expression for the expected squared accuracy

$$r_{g\hat{g}}^2 = \frac{r_{mq}^2 (N r_{mq}^2 h^2 / n_{QTL})}{[1 + N r_{mq}^2 h^2 / n_{QTL}]} \quad (2)$$

substituting n_{QTL} for M_e, we have $r_{g\hat{g}}^2 = \frac{r_{mq}^2 (N r_{mq}^2 h^2 / M_e)}{[1 + N r_{mq}^2 h^2 / M_e]}$ (3)

As $h_M^2 = r_{mq}^2 h^2$ (1), it follows that the squared experimental accuracy is $r_{g\hat{g}}^2 = \frac{r_{mq}^2 (N h_M^2 / M_e)}{[1 + N h_M^2 / M_e]}$ (4)

From (4) the squared accuracy of the RE, we get the squared predictive ability as:

$$r_{\hat{y}y}^2 = r_{g\hat{g}}^2 \frac{h^2}{h_M^2} = \frac{r_{mq}^2 (N h_M^2 / M_e)}{[1 + N h_M^2 / M_e]} \frac{h^2}{h_M^2} = \frac{r_{mq}^2 (N h^2 / M_e)}{[1 + N h_M^2 / M_e]} = \frac{(N h_M^2 / M_e)}{[1 + N h_M^2 / M_e]} = \frac{N h_M^2}{N h_M^2 + M_e} \quad (5)$$

which simplifies to $r_{\hat{y}y}^2 = \frac{N h_M^2}{N h_M^2 + M_e} = \frac{h_M^2}{h_M^2 + \frac{M_e}{N}} = \frac{1}{1 + \frac{M_e}{N h_M^2}}$ (6)

showing that it depends mainly on h_M^2 and N.

From (6) we get back to the squared experimental accuracy as $r_{g\hat{g}}^2 = r_{mq}^2 r_{\hat{y}y}^2 = \frac{h_M^2}{h^2} r_{\hat{y}y}^2 = \frac{h_M^2}{h^2 + \frac{M_e h^2}{N h_M^2}}$ (7)

From the formula for the squared predictive ability via the RE, the genomic heritability is given by $h_M^2 = \frac{r_{\hat{y}y}^2 M_e}{(1 - r_{\hat{y}y}^2) N}$ (8)