Biometry, Modeling, and Statistics | Research Article

SCIENTIA AGRICOLA

DOI: http://dx.doi.org/10.1590/1678-992X-2017-0369

ISSN 1678-992X

Triple categorical regression for genomic selection: application to cassava breeding

Leísa Pires Lima[1], Camila Ferreira Azevedo[1]*, Marcos Deon Vilela de Resende[2], Fabyano Fonseca e Silva[3], José Marcelo Soriano Viana[4], Eder Jorge de Oliveira[5]

[1]Universidade Federal de Viçosa – Depto. de Estatística, Av. Peter Henry Rolfs, s/n – 36570-900 – Viçosa, MG – Brasil.
[2]Embrapa Floresta, Est. da Ribeira, km 111 – 83411-000 – Colombo, PR – Brasil.
[3]Universidade Federal de Viçosa – Depto. de Zootecnia.
[4]Universidade Federal de Viçosa – Depto. de Biologia Geral.
[5]Embrapa Mandioca e Fruticultura, R. Embrapa, s/n, C.P. 007 – 44380-000 – Cruz das Almas, BA – Brasil.
*Corresponding author <camila.azevedo@ufv.br>

Edited by: Marcin Kozak

Received October 24, 2017
Accepted May 14, 2018

ABSTRACT: Genome-wide selection (GWS) is currently a technique of great importance in plant breeding, since it improves efficiency of genetic evaluations by increasing genetic gains. The process is based on genomic estimated breeding values (GEBVs) obtained through phenotypic and dense marker genomic information. In this context, GEBVs of N individuals are calculated through appropriate models, which estimate the effect of each marker on phenotypes, allowing the early identification of genetically superior individuals. However, GWS leads to statistical challenges, due to high dimensionality and multicollinearity problems. These challenges require the use of statistical methods to approach the regularization of the estimation process. Therefore, we aimed to propose a method denominated as triple categorical regression (TCR) and compare it with the genomic best linear unbiased predictor (G-BLUP) and Bayesian least absolute shrinkage and selection operator (BLASSO) methods that have been widely applied to GWS. The methods were evaluated in simulated populations considering four different scenarios. Additionally, a modification of the G-BLUP method was proposed based on the TCR-estimated (TCR/G-BLUP) results. All methods were applied to real data of cassava (*Manihot esculenta*) with to increase efficiency of a current breeding program. The methods were compared through independent validation and efficiency measures, such as prediction accuracy, bias, and recovered genomic heritability. The TCR method was suitable to estimate variance components and heritability, and the TCR/G-BLUP method provided efficient GEBV predictions. Thus, the proposed methods provide new insights for GWS.

Keywords: G-BLUP, BLASSO, genomic prediction, genomic heritability

Introduction

Genome-wide selection (GWS) is based on dense marker maps covering the entire genome, where all genes of a quantitative trait are expected to be in linkage disequilibrium (LD) with these markers. Thus, GWS is used to explain the entire genetic variation of a quantitative trait and predict its individual genetic merit (Meuwissen et al., 2001). The practical application of this genomic information is a challenge, since the main problem is estimation of a large number of marker effects (n) from a limited number of phenotyped and genotyped individuals (N). Additionally, multicollinearity between markers is a relevant issue to be overcome in GWS modeling (Gianola et al., 2003). A feasible solution is to treat the markers as random effects under genomic best linear unbiased predictor (G-BLUP) (Goddard, 2009; Van Raden, 2008; Whittaker et al., 2000) or Bayesian frameworks (De los Campos et al., 2009b). These methods estimate simultaneously n effects based on N observations ($n >> N$), but the smaller the n/N ratio, the more accurate the estimation process is. Resende et al. (2014) introduced a new and simple GWS method named triple categorical regression (TCR); nevertheless, it has not been implemented and evaluated yet. This method returns phenotypes in the three categories (MM, Mm, and mm) of marker genotypes ($3/N << n/N$) to capture the genetic effects in a locus with genotype categories BB, Bb, and bb, where B is the favorable allele. This is consistent with the philosophy of the infinitesimal genetic model and G-BLUP.

Breeding programs for cassava (*Manihot esculenta*) have increased intensely (Graciano-Ribeiro et al., 2009; Nassar, 2007) and have recently used genomic information (Azevedo et al., 2016; Oliveira et al., 2012). Thus, studies to improve the statistical methods of genomic selection would have a positive impact on genetic improvement of cassava. First, the efficiency of the TCR method was evaluated and then compared with the G-BLUP and Bayesian least absolute shrinkage and selection operator (BLASSO) methods using simulated populations based on four different scenarios (two heritability levels and two dominance status). In addition, we proposed a new method denominated as TCR/G-BLUP, which estimates heritability through TCR and uses it in the G-BLUP method. The efficiency of all the methods was also tested using real data on six traits of cassava.

Materials and Methods

Simulated datasets

Data were generated as described by Azevedo et al. (2015) and simulated using Real Breeding software (Viana, 2011; Viana et al., 2016b). It was generated 5,000 individuals from the crossing of two populations with linkage equilibrium. This resultant population was subjected to five generations of random mating without mutation, selection, or migration. Thus, the resulting composite population presented both Hardy-Weinberg equilibrium and linkage disequilibrium (LD). The LD value (Δ) in a composite population is given by

$$\Delta_{ab} = \left(\frac{1-2\theta_{ab}}{4}\right)(p_a^1 - p_a^2)(p_b^1 - p_b^2) ,$$

where a and b are two single nucleotide polymorphisms (SNPs), two quantitative trait loci (QTLs), or one SNP and one QTL, θ is the frequency of recombinant gametes, and p_1 and p_2 are the allele frequencies in the parental populations 1 and 2, respectively (Viana, 2004). Consequently, the LD value depends on the allele frequencies in the parental populations.

It was generated 1,000 individuals from 20 full-sib families (each one with 50 individuals) with diploid genomes of 200 centimorgans (cM) (L = 2 morgans) in length and with 2,000 equidistant SNP markers separated by 0.10 cM among the ten chromosomes. One hundred QTLs were distributed in the genome, where according to the expression presented by Goddard et al. (2011), 95 % of the expected proportion of the genetic variation has to be explained by markers. This value shows that the genome was sufficiently saturated by markers.

This simulation provides a typical small effective population size (N_e = 39.22) and a large LD in the breeding populations. The phenotypic traits were simulated taking into account the infinitesimal model or polygenic inheritance, that is, traits controlled by genes with a small effect. In other words, each of the 100 QTLs had one additive effect of small magnitude on the phenotype. The additive and dominance effects were considered independently and were both normally distributed with mean zero and genetic variance, allowing the desired level of heritability. The genotypic values of homozygotes were obtained taking into account Gmax = 100(m + a) as the maximum value and Gmin = 100(m – a) as the minimum value, where a is the genotypic value of the homozygote and m is the mean of the genotypic values. To obtain the phenotypic value, a random deviation was added to the genotypic value considering the normal distribution $N(0, \sigma_e^2)$, where the variance σ_e^2 was defined according to two levels of heritability in the narrow sense at approximately 0.30 and 0.50. According to Azevedo et al. (2015), these heritability levels are chosen to represent a trait with low and moderate heritability in cases where the genomic selection is expected to be higher than the phenotypic selection. The minor allele frequency was smaller than 5 % for all loci.

Scenarios

Four different scenarios were simulated and used in the analyses: two heritability levels (0.30 and 0.50, associated with the restricted-sense heritability values of 0.20 and 0.35, respectively) and dominance (absence and complete). The description of the scenarios is presented in Table 1. These four scenarios were analyzed using the TCR, G-BLUP, and BLASSO methods. Each scenario was simulated ten times, where nine replicates were used as training populations and one replicate was used as the validation population. Estimates based on each of the nine replicates were validated to calculate accuracy, bias, and genomic heritability. Thus, these measures were calculated in each repetition of the simulation and the mean was generated.

Table 1 – Scenarios with the respective averages of the additive heritability (h_a^2) due to dominance (h_d^2) and total heritability (h_g^2), genetic architectures (traits controlled by genes of small effect; polygenic inheritance), and dominance status (absence of dominance and complete dominance).

Scenario	Dominance status	h_a^2	h_d^2	h_g^2
Scenario 1	Absence	0.22	-	0.22
Scenario 2	Absence	0.33	-	0.33
Scenario 3	Complete	0.21	0.10	0.31
Scenario 4	Complete	0.35	0.17	0.52

Real data

Genomic selection was performed for six traits evaluated in cassava (*Manihot esculenta*). The experiment was carried out under a randomized block design with three replicates (10 plants per plot), using 358 accessions of cassava. The accessions were genotyped for 390 SNP markers. The experiment was established in Cruz das Almas, Brazil (12°48′38″ S and 39°6′26″ W; 220 m above sea level) in 2010 and 2011. We evaluated shoot weight, total root productivity, percentage amylose content of the starch fraction, starch content, hydrogen cyanide, and starch yield. Further details of the experiment are described at Oliveira et al. (2012).

Triple categorical regression

Using the TCR procedure, for estimation, the population was initially divided into two subpopulations: one with individuals or families above the general average (subpopulation 1, with a higher phenotypic mean u_1 value) and another with individuals or families below the general average (subpopulation 2, with a lower phenotypic mean u_2 value). The difference between these values ($u_1 - u_2$) is attributed to the higher frequency (p) of favorable alleles (and lower frequency of unfavorable alleles) in subpopulation 1 in relation to those in subpopulation 2. Thus, $u_1 - u_2$ is explained by $\Delta p = p_1 - p_2$, where Δp is the difference in allele frequencies p_1 and p_2 between these two subpopulations. Δp values were calculated for each marker. Those with positive signals were allocated as favorable (type B), that is, their latent additive genetic effects or allelic substitution effects (α_i) were taken as positive. Likewise, those with negative Δp signs had their α_i value assigned as negative. The encoding of the incidence matrix (W) was reconfigured. The marker genotypes consisting of 0 (mm), 1 (Mm), and 2 (MM) are compatible with a gene genotype given by 0 (bb), 1 (Bb), and 2 (BB), where allocation of BB or bb is dictated by α_i signal. Obviously, the correct allocation of BB or bb is probabilistic. On average (statistical expectation), there is correctness in most loci and most errors are found in markers with very small effects (tending to zero). The approach does not demand an iterative computational method and only uses the concept of genetic distance (Δp_i signal) associated to both subpopulations.

The complete algorithm of the method is described below:

i) Subdivide the training population into two according to phenotype and corrected for controllable environmental effects, as described by De los Campos et al. (2013). Consider the following model:

$$y_{RAW} = X_1 f + X_2 r + e$$

where y_{RAW} is the total phenotype vector without correction, f is the vector of fixed environmental effects with incidence matrix X_2, r is the vector of random environmental effects with a matrix of incidence Z , and e is the random residual vector assumed as $e \sim N(0, I\sigma_e^2)$, where σ_e^2 is the residual variance and I is an identity matrix. The corrected phenotype is given by

$$\hat{y} = y_{RAW} - X_1\hat{f} + X_2\hat{r}$$

where $\hat{f}$ and $\hat{r}$ are estimated and predicted via mixed models.

ii) Calculate the value of Δp_i.

iii) For the marker genotypes consisting of 0 (mm), 1 (Mm), and 2 (MM), change zero by 2 and 2 by zero in each marker column with a negative Δp_i signal, creating the gene genotypes to be used in the next step.

iv) For the gene genotypes given by 0 (bb), 1 (Bb) and 2 (BB), determine the quantity (n_{BB}) of code 2 in the line corresponding to each individual j of the marker file, and do the same for codes 1 and zero, obtaining n_{Bb} and n_{bb}, respectively.

v) Fit the TCR, defined as follows:

$$y = 1\mu + \beta_{BB} n_{BB} I_{(BB)} + \beta_{Bb} n_{Bb} I_{(Bb)} + \beta_{bb} n_{bb} I_{(bb)} + e$$

where $I_{(BB)}$, $I_{(Bb)}$ and $I_{(bb)}$ are indicator variables. If the analyzed category is BB, then $I_{(BB)} = 1$ and $I_{(Bb)} = I_{(bb)} = 0$. Similarly, the same can be defined for the other genotype categories. Regression coefficients ($\hat{\beta}$) are estimated using the least-squares method thus providing the global genetic value of each genotype category as follows:

$\hat{\beta}_{BB} = Cov(y, n_{BB}) / Var(n_{BB})$; $\hat{\beta}_{Bb} = Cov(y, n_{Bb}) / Var(n_{Bb})$; and $\hat{\beta}_{bb} = Cov(y, n_{bb}) / Var(n_{bb})$

vi) Obtain the genotypic values ($\hat{u}_{BB}$, $\hat{u}_{Bb}$, and $\hat{u}_{bb}$) according to the genotype category of markers by calculating the sum of all loci in each individual of the validation population as follows:

$\hat{u}_{BB} = \hat{\beta}_{BB}\, n_{BB} = 2\alpha_B + \delta_{BB}$ (total genotypic value of category BB in the n_{BB} locus), $\hat{u}_{Bb} = \hat{\beta}_{Bb}\, n_{Bb} = \alpha_B + \alpha_b + \delta_{Bb}$ (total genotypic value of category Bb at the n_{Bb} locus), and $\hat{u}_{bb} = \hat{\beta}_{bb}\, n_{bb} = 2\alpha_b + \delta_{bb}$ (total genotypic value of category bb in the n_{bb} locus), where α_B is the additive genetic effect of genotype B, α_b is the additive genetic effect of genotype b, $\delta_{BB} = -2(1 - p)^2 d$, $\delta_{bb} = -2p^2 d$, and $\alpha_i = \alpha_B - \alpha_b$, where d the genotypic value for one heterozygote (Falconer, 1989).

vii) Allocate the total genotypic values of each individual in a vector.

viii) Compute the genetic variances, as detailed in Table 2.

ix) Estimate the heritability given by $\hat{h}_a^2 = \hat{\sigma}_{u_a}^2 / \sigma_y^2$ and $\hat{h}_d^2 = \hat{\sigma}_{u_d}^2 / \sigma_y^2$, where σ_y^2 is the variance between individual phenotypic values.

Compositions of genotypes in terms of frequencies, additive effects, and dominance and variances are shown in Table 2 (Falconer, 1989). This information was used to compose genetic variance estimators by the TCR method.

Estimators of genetic effects

Since genotypic values a and $-a$ (Table 2) of BB and bb are related to the additive effects, the sum $\hat{\mu}_a = f(\hat{\alpha}) = \hat{u}_{BB} + \hat{u}_{bb}$ provides an estimate of additive effects of the individual. These can be used for computation of selective accuracy and prediction bias.

Since d is the genotypic value of heterozygote Bb (Table 2) and is related to the dominance effect, it is assumed that $\hat{\mu}_d = \hat{u}_{Bb}$, thus, providing an estimate of dominance effects of the individual. This allows to compute selective accuracy and bias in predicting dominance effects. With p tending to q ($p \approx q \approx 0.50$), the quantity $\hat{\mu}_d = \hat{u}_{Bb} - \hat{u}_{BB} + \hat{u}_{bb}$ is also defined as an estimator of these effects.

Table 2 – Allele frequencies (freq), genotypic values (GVs), parametric (theoretical) genetic additive effects (u_a), dominance effects (u_d), and variances obtained from the triple categorical regression method.

Genotype	freq	GV	u_a	u_d
BB	p^2	a	$2\alpha_B = 2q\alpha$	$\delta_{BB} = -2q^2d$
Bb	$2pq$	d	$\alpha_B + \alpha_b = (q - p)\alpha$	$\delta_{Bb} = 2pqd$
bb	q^2	$-a$	$2\alpha_b = -2\,p\alpha$	$\delta_{bb} = -2p^2d$
Genotype	freq		$\sigma_{u_a}^2$	$\sigma_{u_d}^2$
BB	p^2		$p^2(2\alpha_B)^2 = p^2(2q\alpha)^2$	$p^2(-2q^2d)^2$
Bb	$2pq$		$2pq(\alpha_B + \alpha_b)^2 = 2pq[(q - p)\alpha]^2$	$2pq(2pqd)^2$
bb	q^2		$q^2(2\alpha_b)^2 = q^2(-2p\alpha)^2$	$q^2(-2p^2d)^2$
Sum			$\sigma_{u_a}^2 = 2pq\alpha^2$	$\sigma_{u_d}^2 = (2pqd)^2$

p is the allele frequency of B; $q = 1 - p$ is the allele frequency of b; a and d are the genotypic values for one homozygote and heterozygote, respectively; α_B and α_b are the additive genetic effects of genotypes B and b, respectively; and α is the allelic substitution effect.

Estimators of genetic variances

Additive variance

According to Table 2, $\sigma^2_{u_a} = 2pq\alpha^2$ and $\hat{\mu}_a = f(\hat{\alpha}) = \hat{u}_{BB} + \hat{u}_{bb}$; thus, $\sigma^2_{u_a} = 2pqf(\hat{\alpha}) = 2pqVar(\hat{u}_{BB} + \hat{u}_{bb})$ is an estimator for the additive genetic variance, where $\hat{\alpha}$ is an intrinsic estimator for the allelic substitution effect on the loci.

Variance of dominance

Based on Table 2, $\sigma^2_{u_d} = (2pqd)^2$. The contrast 2 $\hat{u}_{Bb} - \hat{u}_{BB} + \hat{u}_{bb}$ provides an estimate of d, and, therefore, $\sigma^2_{u_d} = (2pq)^2Var(2\hat{u}_{Bb} - \hat{u}_{BB} + \hat{u}_{bb})$ is an estimator for the genetic variance of dominance. With $p \approx q \approx 0.50$, the quantity $\sigma^2_{u_d} = (2pq)^2 4Var(\hat{u}_{Bb} - \hat{u}_{BB} + \hat{u}_{bb})$ is also defined as an estimator of $\sigma^2_{u_d}$.

Total genotypic variance

The variance of the summation $\hat{u}_{Bb} + \hat{u}_{BB} + \hat{u}_{bb}$ provides information on the total genotypic variance as a function of p, d, and α. Thus, $Var(\hat{u}_{Bb} + \hat{u}_{BB} + \hat{u}_{bb}) = f(p, d, \alpha)$ and the additive and dominance genetic variances can be extracted from $f(p, d, \alpha)$ through $\sigma^2_{u_a} = 2pqVar(\hat{u}_{Bb} + \hat{u}_{BB} + \hat{u}_{bb})$ and $\sigma^2_{u_d} = (2pq)^2Var(\hat{u}_{Bb} + \hat{u}_{BB} + \hat{u}_{bb})$, respectively. Thus, the total genotypic variance is given by $\sigma^2_{u_g} = [2pq + (2pq)^2]Var(\hat{u}_{Bb} + \hat{u}_{BB} + \hat{u}_{bb})$.

G-BLUP method

The G-BLUP method is based on the following linear mixed model given by

$$y = 1\mu + Za + Zd + e$$

where y is a vector of phenotypes ($N \times 1$, where N is the number of genotypes and phenotypes of individuals); μ is the general mean and 1 is the vector with dimension ($N \times 1$), whose elements are equal to 1; a is the vector of additive genomic values of individuals ($N \times 1$) with incidence matrix Z ($N \times N$), given the variance structure $a \sim N(0, G_a\sigma^2_a)$, where σ^2_a is the additive variance and G_a (N × N) is the additive genomic relationship matrix; d is the vector of dominance genomic values of individuals (N × 1) with incidence matrix Z (N × N), given the variance structure $d \sim N(0, G_d\sigma^2_d)$, where σ^2_d is the dominance variance and G_d ($N \times N$) is the dominance genomic relationship matrix; and e is the random residual vector, with $e \sim N(0, I\sigma^2_e)$, where σ^2_e is the residual variance and I an identity matrix.

According to Vitezica et al. (2013), the genomic relationship matrices for the additive and dominance effects (G_a and G_d) are given, respectively, by

$$G_a = \frac{WW'}{\sum_{i=1}^{n}(2p_iq_i)} \text{ and } G_d = \frac{SS'}{\sum_{i=1}^{n}(2p_iq_i)^2}$$

where p_i and q_i are the allele frequencies of locus i, W is an incidence matrix for the allelic substitution vectors of markers (α), and S is the incidence matrix for the effect of vectors due to marker dominance (δ). According to Da et al. (2014), Resende et al. (2014), Van Raden (2008), Vitezica et al. (2013), and Wang and Da (2014), the elements of W and S are given by

$$W = \begin{cases} If\ MM,\ then\ 2 - 2p \rightarrow 2q \\ If\ Mm,\ then\ 1 - 2p \rightarrow q - p \\ If\ mm,\ then\ 0 - 2p \rightarrow -2p \end{cases} \text{ and}$$

$$S = \begin{cases} If\ MM,\ then\ 0 \rightarrow 2q^2 \\ If\ Mm,\ then\ 1 \rightarrow 2pq \\ If\ mm,\ then\ 0 \rightarrow -2p^2 \end{cases}$$

TCR/G-BLUP method

In order to make the G-BLUP genomic values more accurate, an improvement to the method was proposed using the estimated heritability provided by TCR, characterizing the TCR/G-BLUP method. In this method, the strategy to determine the TCR-estimated heritability in the mixed model equations of G-BLUP was adopted.

BLASSO method

The BLASSO (Park and Casella, 2008) regression for genomic selection was proposed by De los Campos et al. (2009b). BLASSO includes a common variance term for the genetic and residual effects of markers. Therefore, the basic linear model is used to predict the effects of markers, $y = 1\mu + Wm_a + Sm_d + e$, where y, 1, W, S, and e were defined previously, m_a is the vector of additive genetic effects of markers, and m_d is the vector of genetic dominance effects of markers.

The BLASSO method is a penalized Bayesian regression procedure whose general estimator is given by

$$\hat{m} = argmin_m\{(\hat{y} - Wm_a - Sm_d)'(\hat{y} - Wm_a - Sm_d) + \lambda_a\sum_{i=1}^{n}|m_{a_i}| + \lambda_d\}$$

where λ_a and λ_d are regularization parameters and $\hat{m} = [\hat{m}_a\ \hat{m}_d]'$. The BLASSO method was implemented in the Bayesian Generalized Linear Regression (BGLR) package (De los Campos et al., 2009b; Pérez et al., 2010) of the R software package, using 100,000 Markov chain Monte Carlo iterations, with a burn-in and thin of 20,000 and ten iterations, respectively.

Computer resources

The computational codes of all methods were implemented in R software (R Core Team, 2016). The G-BLUP method was performed with the Ridge Regression and Other Kernels for Genomic Selection (rrBLUP) package with the mixed.solve function. The BLASSO method was implemented through the BGLR package with the BLR function. The algorithm used for development of the TCR method is available at http://www.ppestbio.ufv.br/?page_id=1811.

Comparison of the methods

The methods were compared by means of independent validations in which the first nine replicates were assumed as training populations and used to estimate the effects of SNP markers on the phenotype. The tenth repetition was assumed as a validation population and used to predict the GEBVs by estimating the effects of the markers obtained in the training population. The measurements to predict efficiency used were accuracy ($r_{\hat{a}a}$ and $r_{\hat{d}d}$), recovered prediction bias ($b_{y\hat{a}}$ and $b_{y\hat{d}}$), and additive and dominance genomic heritability (h^2_{aM} and h^2_{dM}) of the estimates based on each of the four simulated scenarios.

Accuracy is defined as the correlation between the GEBVs and the parametric genetic values. Prediction bias is defined as the regression coefficient between phenotype and the GEBV, where it is understood that the GEBVs were overestimated for regression coefficients < 1 and underestimated for regression coefficients > 1. The recovered additive molecular heritability is given by

$$h^2_{aM} = \frac{\sigma^2_{a_M}}{\sigma^2_{a_M} + \sigma^2_{d_M} + \sigma^2_e},$$

where $\sigma^2_{a_M} = \sum_{i=1}^{n} 2p_i q_i m_i^2$ is the additive genomic variance, where m_i^2 is the square of the ith marker with allele frequencies equal to p_i and q_i. Molecular heritability due to dominance is given by

$$h^2_{dM} = \frac{\sigma^2_{d_M}}{\sigma^2_{a_M} + \sigma^2_{d_M} + \sigma^2_e},$$

where $\sigma^2_{d_M} = \sum_{i=1}^{n} (2p_i q_i d_i)^2$, where d_i is the genotypic value of the heterozygote. The relative efficiency is given by the ratio between accuracies from the methods compared. All these measurements were obtained for each replicate in each scenario and the general results were reported as average values.

Efficiencies of the G-BLUP and TCR/G-BLUP methods were compared using six traits evaluated in *Manihot esculenta*. The experiment was set up in a randomized block design, with three replicates and ten plants per plot, and phenotypes were corrected for block effects. Of the 358 individuals, 50 were randomly separated to compose the validation population. Efficiency measurements were the predictive ability ($r_{\hat{y}y}$), consisting of the correlation between the estimated genomic values and the phenotypic values of the validation population, and prediction bias.

Results and Discussion

Simulated data

Table 3 shows the average results of accuracy, prediction bias, and heritability obtained by the TCR, G-BLUP, and BLASSO methods, associated with the predicted additive genomic values that considered the absence of dominance and complete dominance. For the additive effects, the TCR method outperformed the G-BLUP and BLASSO methods in terms of heritability estimation (providing estimates very close to the parametric heritability), except for scenario 1. For scenario 2, all three methods did not report suitable estimations of heritability. In addition, the TCR method provided estimates of non-biased additive genomic values, since the regression coefficients were close to the unit. The unbiased property is important when selection involves individuals of many generations using effects of estimated markers in a single generation. On the other hand, the TCR method provided lower accuracy than the G-BLUP and BLASSO methods did, with the BLASSO method standing out in terms of prediction accuracy. In GWS studies for genetic improvement of table grapes, Viana et al. (2016a) also reported the superiority of the BLASSO method over the ridge regression BLUP method (reparametrization of G-BLUP method) for its efficiency in predicting additive genomic values.

Table 3 – Additive heritability (h^2_{aM}), accuracy ($r_{\hat{a}a}$), and bias ($b_{y\hat{a}}$), with respective standard deviations, of the additive genomic values estimated by the triple categorical regression (TCR), genomic best linear unbiased predictor (G-BLUP), and Bayesian least absolute shrinkage and selection operator (BLASSO) methods, considering the additive-dominance model on simulated data.

Model	Scenario	Method	h^2_{aM}	$r_{\hat{a}a}$	$b_{y\hat{a}}$
Additive	Scenario 1	TCR	0.31 ± 0.03	0.65 ± 0.02	1.09 ± 0.01
		G-BLUP	0.27 ± 0.04	0.64 ± 0.03	1.48 ± 0.04
		BLASSO	0.28 ± 0.03	0.76 ± 0.02	1.03 ± 0.06
Additive	Scenario 2	TCR	0.47 ± 0.04	0.69 ± 0.02	0.77 ± 0.01
		G-BLUP	0.50 ± 0.04	0.79 ± 0.02	1.30 ± 0.02
		BLASSO	0.50 ± 0.05	0.82 ± 0.01	1.00 ± 0.08
Additive-dominance	Scenario 3	TCR	0.23 ± 0.03	0.57 ± 0.05	1.09 ± 0.01
		G-BLUP	0.15 ± 0.05	0.63 ± 0.03	1.25 ± 0.35
		BLASSO	0.17 ± 0.09	0.63 ± 0.03	1.44 ± 0.65
Additive-dominance	Scenario 4	TCR	0.35 ± 0.04	0.62 ± 0.02	1.09 ± 0.01
		G-BLUP	0.27 ± 0.03	0.70 ± 0.02	1.17 ± 0.13
		BLASSO	0.18 ± 0.05	0.69 ± 0.03	1.69 ± 0.45

Scenarios with traits controlled by genes of small effects: Scenario 1 (h^2_a = 0.22), Scenario 2 (h^2_a = 0.33), Scenario 3 (h^2_a = 0.21 and h^2_d = 0.10), and Scenario 4 (h^2_a = 0.35 and h^2_d = 0.17).

The average results of accuracy, prediction bias, and heritability obtained through the TCR, G-BLUP, and BLASSO methods, associated with the predicted dominance genomic values, are presented in Table 4. For the dominance effects, the TCR method presented, on average, heritability estimates that were coincident with the parametric heritability. The G-BLUP and BLASSO methods underestimated heritability and showed biased values. The TCR method also provided higher accuracy than the G-BLUP and BLASSO methods did (about 0.40 in the TCR method; from 0.31 to 0.40 in G-BLUP; and from 0.29 to 0.35 in BLASSO) and was able to better extract the ratio between dominance variance and additive variance. In addition, the TCR method presented higher accuracy values for the effects due to dominance and ratio between variances close to parametric values (ratio between dominance and additive variances of around 0.50). Thus, for the dominance effects, the TCR method showed superiority for all criteria considered.

The results of the simulation study revealed suitability of the TCR method estimators. According to De los Campos et al. (2009a), the ability to estimate heritability accurately may be a more sensitive criterion for discriminating and evaluating statistical methods in GWS. This greater sensitivity is because heritability is a more complex parameter than the simple correlation coefficients used to estimate accuracy of predictions. Furthermore, according to Azevedo et al. (2015) and Makowsky et al. (2011), the heritability recovered can be considered a measurement of quality of GWS model fitting.

The results in Tables 3 and 4 were obtained from simulated data, whose mean value of $2pq$ is equal to 0.49 and thus has a mean p approximately equal to mean q of 0.5, which fits well with broad-based populations, such as compounds and F_2 populations. In this case, the TCR method fits better and is a recommended alternative. These results are valid for Fisher's infinitesimal genetic model, which does not admit genes of larger effects as those in the simulated scenarios.

Table 4 – Heritability due to dominance (h^2_{dM}), accuracy ($r_{\hat{d}d}$), and bias ($b_{y\hat{d}}$), with their respective standard deviations, of the genomic values according to the dominance estimated by the triple categorical regression (TCR), genomic best linear unbiased predictor (G-BLUP), and Bayesian least absolute shrinkage and selection operator (BLASSO) methods, and the ratio between the heritability values according to the dominance and additive values (h^2_{dM} / h^2_{aM}), considering the additive-dominance model on simulated data.

Scenario	Method	h^2_{dM}	$r_{\hat{d}d}$	$b_{y\hat{d}}$	h^2_{dM} / h^2_{aM}
Scenario 3	TCR	0.10 ± 0.01	0.40 ±0.02	0.90 ± 0.14	0.43
	G-BLUP	0.13 ± 0.06	0.31 ± 0.04	0.70 ± 0.30	0.87
	BLASSO	0.13 ± 0.02	0.29 ± 0.05	3.20 ± 5.34	0.76
Scenario 4	TCR	0.17 ± 0.02	0.40 ± 0.02	0.96 ± 0.12	0.49
	G-BLUP	0.20 ± 0.02	0.40 ± 0.04	0.74 ± 0.22	0.74
	BLASSO	0.29 ± 0.03	0.35 ± 0.03	0.46 ± 0.08	1.61

Scenarios with traits controlled by genes of small effects: Scenario 3 (h^2_a = 0.21 and h^2_d = 0.10), and Scenario 4 (h^2_a = 0.35 and h^2_d = 0.17).

Table 5 shows the average results of additive heritability and heritability due to dominance associated with the heritability estimation through total genotype variation. Heritability, in the broad sense, estimated directly by the estimator of total genotypic variance, is an important genetic parameter for plants that undergo vegetative propagation (e.g., cassava, eucalyptus, and sugarcane) or self-fertilization in which the genotype is inherited integrally by the offspring. On the other hand, heritability in the restricted sense, estimated directly by TCR, is an important genetic parameter, especially when the interest is the prediction gain due to selection for sexual propagation (Falconer, 1989). The results in Table 5 show that it is possible to obtain the additive (restricted sense) heritability using the total genotypic variance estimator and the dominance status.

Real data

The average results of predictive ability, prediction bias, and heritability obtained by the TCR and G-BLUP methods, associated with the predicted additive genotypic values of six traits evaluated in *Manihot esculenta*, are presented in Table 6. The TCR method provided unbiased estimates and smaller values of predictive ability than the G-BLUP method did. Because TCR reported better heritability and G-BLUP showed higher accuracy (simulated data) or predictive ability (real data), the strategy to establish the TCR-estimated heritability in the mixed-model equations of G-BLUP was adopted, generating the TCR/G-BLUP method. This approach increased the predictive ability and reduced the bias of the G-BLUP method and is therefore recommended for practical uses.

According to Oliveira et al. (2012), cassava cultivation is of great importance to Brazil, as it is one of the most relevant commodities for food security. Thus, the

Table 5 – Additive heritability (h^2_{aM}), heritability due to dominance (h^2_{dM}), and heritability in the broad sense (h^2_{gM}), with respective standard deviations, estimated by the triple categorical regression method, considering the genotypic variances and the additive-dominant model on simulated data.

Scenario	Direct estimator	h^2_{aM}	h^2_{dM}	h^2_{gM}
Scenario 3	σ^2_{aM} and σ^2_{dM}	0.23 ± 0.03	0.10 ± 0.01	0.33
	σ^2_{gM}	0.24 ± 0.03	0.12 ± 0.01	0.36
Scenario 4	σ^2_{aM} and σ^2_{dM}	0.35 ± 0.04	0.17 ± 0.02	0.52
	σ^2_{gM}	0.37 ± 0.04	0.18 ± 0.02	0.55

Scenarios with traits controlled by genes of small effects: Scenario 3 (h^2_a = 0.21 and h^2_d = 0.10) and Scenario 4: (h^2_a = 0.35 and h^2_d = 0.17).

Table 6 – Additive heritability (h^2_{aM}), predictive ability ($r_{\hat{a}y}$), and bias ($b_{y\hat{a}}$) of the additive genomic values estimated by the triple categorical regression (TCR), genomic best linear unbiased predictor (G-BLUP), and TCR/G-BLUP models on real cassava data.

Variable	Method	h^2_{aM}	$r_{\hat{a}y}$	$b_{y\hat{a}}$
SW	TCR	0.36	0.44	1.25
	G-BLUP	0.17	0.60	1.68
	TCR/G-BLUP	0.36	0.65	1.54
TRP	TCR	0.28	0.44	1.44
	G-BLUP	0.15	0.57	1.87
	TCR/G-BLUP	0.28	0.64	1.64
AC	TCR	0.12	0.30	1.45
	G-BLUP	0.07	0.50	3.77
	TCR/G-BLUP	0.12	0.56	3.02
SC	TCR	0.20	0.40	1.52
	G-BLUP	0.23	0.66	2.10
	TCR/G-BLUP	0.20	0.65	2.21
HCN	TCR	0.47	0.46	1.13
	G-BLUP	0.50	0.83	1.26
	TCR/G-BLUP	0.47	0.83	1.28
SY	TCR	0.30	0.45	1.39
	G-BLUP	0.15	0.57	1.80
	TCR/G-BLUP	0.30	0.64	1.56

Traits evaluated were shoot weight (SW); total root productivity (TRP); amylose content (AC); starch content (SC); hydrogen cyanide (HCN); and starch productivity (SY).

prospects of using GWS for cassava traits are crucial, since the estimation of genomic values of the individuals allows the selection of genetically superior plants in the seedling phase, increasing selection gain per unit of time. The estimates of heritability obtained by the TCR/G-BLUP method based on cassava traits were similar to the values found by Azevedo et al. (2016) and Oliveira et al. (2012).

Conclusions

Based on simulated data, the TCR method outperformed the G-BLUP and BLASSO methods, showing heritability estimates close to the parametric value. Moreover, compared with the other methods, the TCR method presented greater accuracy and less bias in the prediction of the genomic values due to dominance. However, for the additive genomic values, it was less accurate. Based on real data and considering the additive model, TCR was less accurate in terms of prediction ability. However, when combined with G-BLUP, it was more accurate. The TCR/G-BLUP method was superior to the G-BLUP method, with increased predictive ability and lower bias production, for the traits evaluated in cassava.

Acknowledgements

The authors thank the following Brazilian funding organizations: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação Arthur Bernardes (FUNARBE).

Authors' Contributions

Conceptualization: Lima, L.P.; Azevedo, C.F.; Resende, M.D.V.; Silva, F.F. Data acquisition: Viana, J.M.S.; Oliveira, E.J. Data Analysis: Lima, L.P.; Azevedo, C.F.; Resende, M.D.V. Design of methodology: Lima, L.P.; Azevedo, C.F.; Resende, M.D.V.; Silva, F.F. Software development: Lima, L.P.; Azevedo, C.F.; Resende, M.D.V.; Viana, J.M.S. Writing and Editing: Lima, L.P.; Azevedo, C.F.; Resende, M.D.V.; Silva, F.F.; Oliveira, E.J.

References

Azevedo, C.F.; Resende, M.D.V.; Silva, F.F.; Viana, J.M.S.; Valente, M.S.F.; Resende Junior, M.F.R.; Muñoz, P. 2015. Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genetics 16: 105.

Azevedo, C.F.; Resende, M.D.V.; Silva, F.F.; Viana, J.M.S.; Valente, M.S.F.; Resende Junior, M.F.R.; Oliveira, E.J. 2016. New accuracy estimators for genomic selection with application in a cassava (*Manihot esculenta*) breeding program. Genetics and Molecular Research 15: 4.

Da, Y.; Wang, C.; Wang, S.; Hu, G. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS ONE 9: e87666.

De los Campos, G.; Gianola, D.; Rosa, G.J.M. 2009a. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. Journal of Animal Science 876: 1883-1887.

De los Campos, G.; Hickey, J.M.; Pong-Wong, R.; Daetwyler, H.D.; Calus, M.P.L. 2013. Whole genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327-345.

De los Campos, G.; Naya, H.; Gianola, D.; Crossa, J.; Legarra, A.; Manfredi, E.; Cotes, J.M. 2009b. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375-385.

Falconer, D.S. 1989. Introduction to Quantitative Genetics. Longman Scientific & Technical, New York, NY, USA.

Gianola, D.; Perez-Enciso, M.; Toro, M.A. 2003. On marker-assisted prediction of genetic value: beyond the ridge. Genetics 163: 347-365.

Goddard, M.E. 2009. Genomic selection: prediction of accuracy and maximization of long term response. Genetics 136: 345-357.

Goddard, M.E.; Hayes, B.J.; Meuwissen, T.H.E. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of Animal Breeding and Genetics 128: 409-421.

Graciano-Ribeiro, D.; Hashimoto, D.Y.C.; Nogueira, L.C.; Teodoro, D.; Miranda, S.F.; Nassar, N.M.A. 2009. Internal phloem in an interspecific hybrid of cassava, an indicator of breeding value for drought resistance. Genetics and Molecular Research 8: 1139-1146.

Makowsky, R.; Pajewski, N.M.; Klimentidis, Y.C.; Vazquez, A.I.; Duarte, C.W.; Allison, D.B.; De los Campos, G. 2011. Beyond missing heritability: prediction of complex traits. PLoS Genetics 7: e1002051.

Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Nassar, N.M. 2007. Cassava genetic resources and their utilization for breeding of the crop. Genetics and Molecular Research 6: 1151-1168.

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

Park, T.; Casella, G. 2008. The Bayesian lasso. Journal of the American Statistical Association 103: 681-686.

Pérez, P.; De los Campos, G.; Crossa, J.; Gianola, D. 2010. Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome 3: 106-116.

Resende, M.D.V.; Silva, F.F.; Azevedo, C.F. 2014. Mathematical, Biometric and Computational Statistics: Mixed, Multivariate, Categorical and Generalized Models (REML/BLUP), Bayesian Inference, Random Regression, Genomic Selection, QTL-GWAS, Spatial and Temporal Statistics, Competition, Survival. = Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. Suprema Editora, Visconde do Rio Branco, MG, Brazil. (in Portuguese).

Van Raden, P.M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414-4423.

Viana, A.P.; Resende, M.D.V.; Riaz, S.; Walker, M.A. 2016a. Genome selection in fruit breeding: application to table grapes. Scientia Agricola 73: 142-149.

Viana, J.M.S. 2004. Quantitative genetics theory for non-inbred populations in linkage disequilibrium. Genetics and Molecular Biology 27: 594-601.

Viana, J.M.S. 2011. Program for Molecular and Quantitative Data Analysis. = Programa para Análises de Dados Moleculares e Quantitativos: Real Breeding, UFV, Viçosa, MG, Brazil (in Portuguese).

Viana, J.M.S.; Piepho, H.P.; Silva, F.F. 2016b. Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations. Scientia Agricola 73: 243-251.

Vitezica, Z.G.; Varona, L.; Legarra, A. 2013. On the additive and dominance variance and covariance of individuals within the genomic selection scope. Genetics 195: 1223-1230.

Wang, C.; Da, Y. 2014. Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. PLoS ONE 9: e114484.

Whittaker, J.C.; Thompson, R.; Denham, M.C. 2000. Marker-assisted selection using ridge regression. Genetics Research 75: 249-252.