

# Regularized quantile regression applied to genome-enabled prediction of quantitative traits

M. Nascimento<sup>1</sup>, F.F. e Silva<sup>2</sup>, M.D.V. de Resende<sup>1,3</sup>, C.D. Cruz<sup>4</sup>,  
A.C.C. Nascimento<sup>1</sup>, J.M.S. Viana<sup>4</sup>, C.F. Azevedo<sup>1</sup> and L.M.A. Barroso<sup>1</sup>

<sup>1</sup>Departamento de Estatística, Universidade Federal de Viçosa,  
Viçosa, MG, Brasil

<sup>2</sup>Departamento de Zootecnia, Universidade Federal de Viçosa,  
Viçosa, MG, Brasil

<sup>3</sup>Embrapa Florestas, Colombo, PR, Brasil

<sup>4</sup>Departamento de Biologia Geral, Universidade Federal de Viçosa,  
Viçosa, MG, Brasil

Corresponding author: M. Nascimento  
E-mail: moysesnascim@gmail.com

Genet. Mol. Res. 16 (1): gmr16019538

Received November 16, 2016

Accepted December 19, 2016

Published March 22, 2017

DOI <http://dx.doi.org/10.4238/gmr16019538>

Copyright © 2017 The Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution ShareAlike (CC BY-SA) 4.0 License.

**ABSTRACT.** Genomic selection (GS) is a variant of marker-assisted selection, in which genetic markers covering the whole genome predict individual genetic merits for breeding. GS increases the accuracy of breeding values (BV) prediction. Although a variety of statistical models have been proposed to estimate BV in GS, few methodologies have examined statistical challenges based on non-normal phenotypic distributions, e.g., skewed distributions. Traditional GS models estimate changes in the phenotype distribution mean, i.e., the function is defined for the expected value of trait-conditional on markers,  $E(Y|X)$ . We proposed an approach based on regularized quantile regression (RQR) for GS to improve the estimation of marker effects

and the consequent genomic estimated BV (GEBV). The RQR model is based on conditional quantiles,  $Q_{\tau}(Y|X)$ , enabling models that fit all portions of a trait probability distribution. This allows RQR to choose one quantile function that “best” represents the relationship between the dependent and independent variables. Data were simulated for 1000 individuals. The genome included 1500 markers; most had a small effect and only a few markers with a sizable effect were simulated. We evaluated three scenarios according to symmetrical, positively, and negatively skewed distributions. Analyses were performed using Bayesian LASSO (BLASSO) and RQR considering three quantiles (0.25, 0.50, and 0.75). The use of RQR to estimate GEBV was efficient; the RQR method achieved better results than BLASSO, at least for one quantile model fit for all evaluated scenarios. The gains in relation to BLASSO were 86.28 and 55.70% for positively and negatively skewed distributions, respectively.

**Key words:** Genomic selection; Regularized regression; SNP effects; Statistics; Simulation

## INTRODUCTION

Genomic selection (GS), proposed by Meuwissen et al. (2001), is a variant of marker-assisted selection (MAS), in which genetic markers covering the whole genome are used to predict individual genetic merits for animal and plant breeding. According to Goddard and Hayes (2007), GS potentially increases genetic variance, and because the markers are assumed to be in linkage disequilibrium (LD) with the quantitative trait loci (QTL), the number of effects per QTL to be estimated is small.

GS has been successfully used in breeding. Among several studies on animal breeding, Christensen et al. (2012) evaluated the accuracy of genomic estimated breeding values (GEBV) using 60,000 single nucleotide polymorphisms (SNPs) for feed conversion in a population of Duroc pigs. They concluded that GEBVs were significantly more accurate than pedigree-only estimated breeding values. In plant breeding, Beyene et al. (2014) aimed to estimate genetic gains in grain yield from GS in eight bi-parental maize populations under managed drought stress environments. The authors demonstrated that GS is more effective than pedigree-based conventional phenotypic selection for increasing genetic gains in grain yield under drought stress in tropical maize.

However, owing to the growing number of markers and interest in longitudinal (Crispim et al., 2015) and non-normal (Filardi de Campos et al., 2015) traits, almost 15 years after this seminal publication, statistical issues such as multicollinearity, high dimensionality, and non-normal phenotypic distributions still need to be overcome in GS.

Although several proposed statistical methods have focused on tackling issues with regards to multicollinearity and dimensionality (Riedelsheimer et al., 2012) in GS (RR-BLUP, Bayesian alphabet, PLS, kernel-based regressions), few studies have examined statistical challenges based on non-normal phenotypic distributions, e.g., skewed distributions.

Skewed distributions exist for various traits in plant and animal breeding, such as flowering time (Maurer et al., 2015), order of parity (Varona et al., 2008), and hormone

concentrations (Mathur et al., 2012; Filardi de Campos et al., 2015). Muranty et al. (2015) verified that accuracy was strongly affected by phenotypic distribution. Specifically, traits that showed poor results often had skewed phenotypic distributions or low heritability.

The traditional GS approach focuses on estimating changes in the mean of the phenotype distribution, i.e., the function is defined for the expected value of trait-conditional on markers,  $E(Y|X)$  (Cade and Noon, 2003). However, the mean is not adequate for situations in which the distribution presents some degree of skewness. Mosteller and Tukey (1977) showed that it is possible to fit regression models to other parts of the distribution of the response variable, but this is rarely done, and therefore most regression analyses give an incomplete picture of the relationships between variables. An interesting solution that addresses these issues, which has not yet been explored in GS, is the use of quantile regression (QR) (Koenker and Bassett, 1978). Unlike traditional methods that are based on conditional expectations, QR is based on conditional quantiles,  $Q_{\tau}(Y|X)$ , i.e., it is possible to fit models for all portions of a probability distribution of the trait enabling a more informative study of the relationship between variables. QR facilitates the selection of a quantile function that “best” represents the relationship between the dependent (phenotype) and independent (marker) variables to resolve the skewness issue.

Therefore, when we combined the properties of traditional GS methods (e.g., shrinkage estimation theory) with desirable QR characteristics, we generated a novel and powerful prediction method that can resolve issues related to dimensionality, multicollinearity, and skewed phenotypic distribution. This method is called the regularized quantile regression (RQR) method because the shrinkage (or penalty) parameter regularizes the variance of marker effects to produce a direct variable selection scheme.

The aims of this study were: 1) to propose and evaluate the RQR method for genome selection studies to improve GEBV prediction; 2) to improve the accuracy of the GEBV by choosing the “best” quantile function to represent the relationship between the phenotype and markers; and 3) to compare the accuracies obtained for RQR and Bayesian LASSO (BLASSO) (de los Campos et al., 2009) for predicting genetic merit using a simulated data set with different degrees of skewness. Finally, we explored future applications of the RQR method in the field of genetics and breeding.

## MATERIAL AND METHODS

### Simulated data

To assess the proposed approach, data were simulated using an additive model of the form:

$$y_i = \mu + \sum_{j=1}^{1500} x_{ij}g_j + e_i, \quad (i=1, \dots, 1000), \quad (j=1, \dots, 1500) \quad (\text{Equation 1})$$

where  $y_i$  is the phenotypic value of the  $i^{\text{th}}$  individual;  $\mu$  is the general mean;  $x_{ij}$  is the genotype of the  $j^{\text{th}}$  marker for the  $i^{\text{th}}$  individual;  $g_j$  is the additive genetic effect of the marker at  $j^{\text{th}}$  locus; and  $e_i$  is the independent and identically distributed residuals with different probability distributions that are assumed to have different degrees of skewness. Under the biallelic marker assumption

( $x_{ij} = 0, 1, \text{ and } 2$  for aa, Aa, and AA, respectively), three distinct scenarios were considered in the phenotype ( $y_i$ ) simulations: symmetrical, positively, and negatively skewed distributions. We assumed that most markers (1497) had a small effect and that only a few (3) had a sizable effect. Specifically, marker effects were sampled from the following mixture model:

$$\mathbf{g}_j = \begin{cases} N(1.5, k), \text{ if } j=1; \\ N(2.5, k), \text{ if } j=750; \\ N(1.5, k), \text{ if } j=1500; \\ N(0, k), \text{ if } j=1, \dots, 1500; j \neq 1, 750 \text{ and } 1500 \end{cases} \quad (\text{Equation 2})$$

where  $k = 1000^{-1}$ . For each trait, the degree of dominance is equal to 1.

Insertion of positive and negative skewness was accomplished by adding and subtracting, respectively, the residual term generated from an exponential distribution with parameters equal to 0.50, i.e.,  $e_i \sim \exp(0.50)$ . Thus, the phenotypic values were given by:

$$y_{PS_i} = y_i + \exp(0.50) \text{ and } y_{NS_i} = y_i - \exp(0.50) \quad (\text{Equation 3})$$

where  $y_{PS_i}$  and  $y_{NS_i}$  are the  $i^{\text{th}}$  values of the phenotypes assuming positive (PS) and negative (NS) skewness distributions, respectively. Independent and identically distributed residuals were simulated as normal distributions [ $e_i \sim N(0, \sigma_e^2)$ ] to include situations where the phenotype distributions were symmetrical. In all simulations, it was assumed that the minimum value of minimum allele frequency for all markers was equal to 0.05. A heritability of 0.25 was used, independent of the scenario. The simulation process was replicated 30 times to assess the performance of the RQR.

### Regularized quantile regression (RQR) for GWAS

Unlike traditional methods, which are based on conditional expectations, RQR is based on conditional quantiles (Li and Zhu, 2008). The use of RQR facilitates the assessment of functional relationships according to the quantile that “best” represents the phenotype distribution. For example, a distribution with negative skewness, when defined by higher quantiles (e.g., 0.75) may better describe the relationship between GEBV and the effect markers, thereby increasing prediction accuracy. However, if the phenotype distribution presents positive skewness, the functional relationship can be better described by lower quantiles (e.g., 0.25). In general, the fit of the regularized model with the quantile function  $\tau$  was obtained through estimates of regression coefficients of the model (1), which addressed the following optimization problem:

$$\hat{\mathbf{g}} = \arg \min_{\mathbf{g}} \left\{ \sum_{i=1}^n \rho_{\tau} \left( y_i - \mu - \sum_j x_{ij} g_j \right) + \lambda \sum_{j=1}^p |g_j| \right\} \quad (\text{Equation 4})$$

where  $\sum_{j=1}^p |g_j|$  is the sum of the absolute values of the regression coefficients,  $\lambda$  is the parameter that controls the strength of regularization, and  $\rho_\tau(\cdot)$  is a function that weighs the observations for estimation of the quantile function (Koenker and Bassett, 1978). It was defined by:

$$\rho_\tau\left(y_i - \mu + \sum_j x_{ij} g_j\right) = \begin{cases} \tau \cdot \left(y_i - \mu - \sum_j x_{ij} g_j\right), & \text{if } y_i - \mu + \sum_j x_{ij} g_j > 0; \\ 1 - \tau \cdot \left(y_i - \mu - \sum_j x_{ij} g_j\right), & \text{otherwise} \end{cases} \quad (\text{Equation 5})$$

where  $\tau \in [0,1]$  indicates the quantile of interest. As with BLASSO, owing to the constraints imposed in the estimation process, the RQR method can be used in situations where the number of markers is superior to the total number of observations. Through coefficient shrinkage, this methodology combines variable selection and regularization via shortening of regression coefficients. The GEBV predictions from the RQR were obtained from the expression:

$$GEBV(\tau) = \hat{y}_i(\tau) = \sum_j x_{ij} \hat{g}_j(\tau) \quad (\text{Equation 6})$$

where  $GEBV(\tau)$  is the genomic breeding value obtained for the quantile of interest ( $\tau$ ).

### Comparison of methodologies under a GWS approach

The simulation process was replicated 30 times to address the performance of the RQR. This facilitated the calculation of the mean squared error (MSE) on the  $j^{\text{th}}$  marker effect, which was given by:

$$MSE_j = \frac{1}{30} \sum_{r=1}^{30} \left(\hat{g}_{jr} - g_j\right)^2 \quad (\text{Equation 7})$$

where  $\hat{g}_{jr}$  is the estimated effect of the  $j$  marker at the  $r^{\text{th}}$  repetition and  $g_j$  corresponds to the real effect of the  $j^{\text{th}}$  marker in each simulation. The MSE values of RQRs in different quantiles ( $\tau = 0.25$  - RQR<sub>0.25</sub>,  $\tau = 0.50$  - RQR<sub>0.50</sub>, and  $\tau = 0.75$  - RQR<sub>0.75</sub>) were compared with those calculated from BLASSO.

To assess the predictive performance of both methods, Pearson's correlation coefficients ( $r_{\hat{y}^{ygi}}$ ) were calculated between true breeding values and their respective predicted values using the two evaluated methods ( $\hat{y}_{BLASSO}$  and  $\hat{y}_{RQR(\tau)}$ ). These values were obtained by 30-fold cross-validation process. Thus, each of the first 29 repetitions was defined as a training population, and was used for SNP marker estimation. The last repetition, population 30, was considered the validation population, and was used to evaluate the correlation between the genetic values predicted by training population estimates and true simulated genetic values.

After obtaining the GEBVs through the fit models (RQR<sub>0.25</sub>, RQR<sub>0.50</sub>, RQR<sub>0.75</sub>, and BLASSO), aimed at assessing the agreement between the methodologies, the Spearman's correlation and Cohen's kappa (Cohen, 1960) coefficients were calculated.

The Cohen's kappa coefficient was used to calculate the percentage of individuals in common between the first 100 individuals ranked according to the values of GEBV and the true genetic values obtained from the simulation according to model 1 (genetic breeding value: GBV). The same procedure was performed for the last 100 individuals, with the intention of representing different selection strategies. The Cohen's kappa coefficient was given by:

$$C = \frac{NC - C_{Random}}{1 - C_{Random}}, \quad (\text{Equation 8})$$

where NC is the agreement observed among raters, and  $C_{Random}$  is the hypothetical probability of random agreement. Because correlation was assessed between the first and last 100 individuals, the value of  $C_{Random}$  was given by:  $C_{Random} = 100/1000 = 0.10$ .

### Computational features

The simulation process was carried out using the `simulateSNPglm` function of the `scime` package (Schwender and Fritsch, 2015), and was implemented in R software (R Development Core Team, 2016). The BLASSO and RQR model fittings were carried out using the `BLR` and `rq` functions in the `BLR` (de los Campos et al., 2009; Pérez et al., 2010) and `quantreg` (Koenker, 2015) packages, respectively.

The BLASSO method was implemented using 100,000 Markov chain Monte Carlo iterations, with burn-in and thin values at 10,000 and 10 iterations, respectively. To ensure the methods were directly comparable, the shrinkage strengths (the values of  $\lambda$ ) used in the RQR method were the same as those estimated using the BLASSO method.

The R codes for simulation and implementation are freely accessible at: XXXX.

## RESULTS

The MSE values of the RQR ( $RQR_{0.25}$ ,  $RQR_{0.50}$ , and  $RQR_{0.75}$ ) and BLASSO fit models ranged from 0.003 to 0.791 (Table 1). Overall, the RQR fit models showed better results than those obtained from BLASSO in all evaluated scenarios, because a reduction in MSE values was observed when the quantile models  $RQR_{0.75}$  and  $RQR_{0.25}$  were fitted for negatively and positively skewed phenotypic distributions, respectively (Table 1). For instance, the MSE values for  $RQR_{0.75}$  and  $RQR_{0.25}$  fit models were closer to zero (0.003 and 0.005) than those obtained from BLASSO (0.016 and 0.016) for negatively and positively skewed phenotypic distributions, respectively.

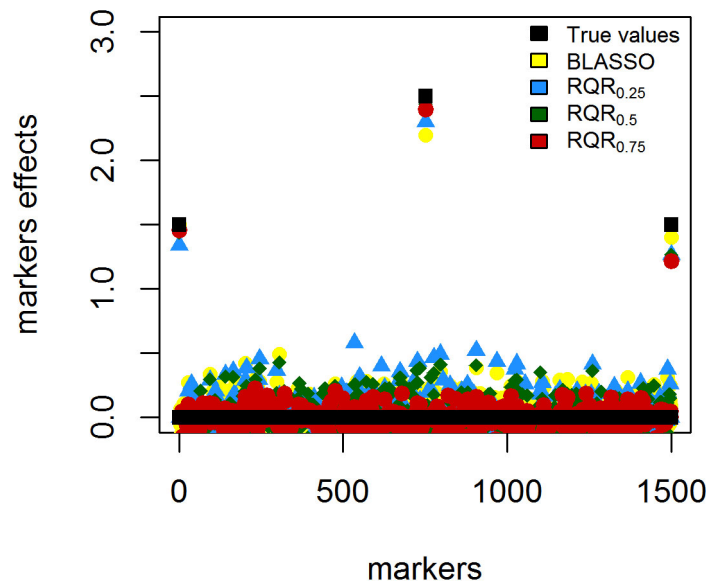
Estimates of accuracy for all scenarios range from 0.421 to 0.785 and are presented in Table 1. According to the accuracy values, the RQR fit models produced better results than those obtained from BLASSO. Specifically, when skewness was present in phenotypic distributions, the  $RQR_{0.75}$  and  $RQR_{0.25}$  fit models for positive and negative skewness resulted in higher (0.696 and 0.785) accuracy compared with the BLASSO method (0.447 and 0.421, respectively). The gains in relation to BLASSO were 86.28 and 55.70% for positive and negative skewness, respectively (Table 1). When the distribution of phenotypic values was symmetrical, both methods demonstrated comparable performance, with lower variations in accuracy compared with skewed distributions. This was expected because the mean and median have equal values on symmetrical distributions.

**Table 1.** Mean square error (MSE) estimates for marker effects and accuracy ( $r_{y,yg}$ ) obtained using BLASSO and  $RQR_{\tau}$  ( $RQR_{0.25}$ ,  $RQR_{0.50}$ , and  $RQR_{0.75}$ ) fit models with different degrees of skewness.

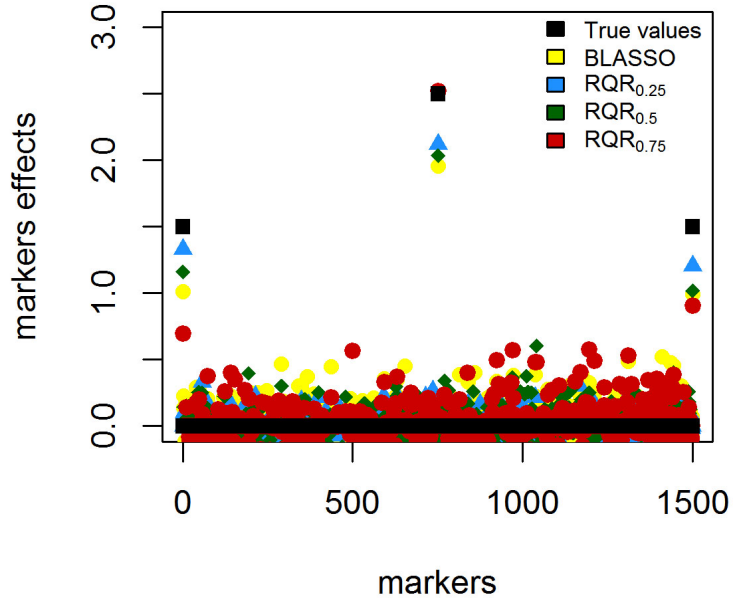
Phenotype distribution	Methods	MSE				$r_{y,yg}$
		( $g_j, j = 1, \dots, 1500$ )	$g_1$	$g_{750}$	$g_{1500}$	
Negative skewness (-1.532 <sup>ns*</sup> )	BLASSO	0.016	0.182	0.380	0.256	0.421
	$RQR_{0.25}$	0.013	0.240	0.554	0.273	0.476
	$RQR_{0.50}$	0.009	0.118	0.254	0.151	0.596
	$RQR_{0.75}$	0.003	0.093	0.139	0.123	0.785
Symmetrical (0.070 <sup>ns</sup> )	BLASSO	0.016	0.244	0.301	0.215	0.483
	$RQR_{0.25}$	0.011	0.303	0.302	0.218	0.535
	$RQR_{0.50}$	0.009	0.184	0.254	0.153	0.518
	$RQR_{0.75}$	0.007	0.398	0.479	0.326	0.588
Positive skewness (1.587 <sup>ns*</sup> )	BLASSO	0.016	0.231	0.251	0.285	0.447
	$RQR_{0.25}$	0.005	0.103	0.103	0.151	0.696
	$RQR_{0.50}$	0.009	0.141	0.097	0.148	0.590
	$RQR_{0.75}$	0.013	0.749	0.791	0.781	0.448

Results are averages of 30 replicates;  $\tau$  indicates the quantile of interest ( $\tau = 0.25, 0.50,$  and  $0.75$ ); <sup>ns</sup>non-significant; <sup>\*</sup>significant, as assessed by the D’Agostino test for skewness at 5% probability.

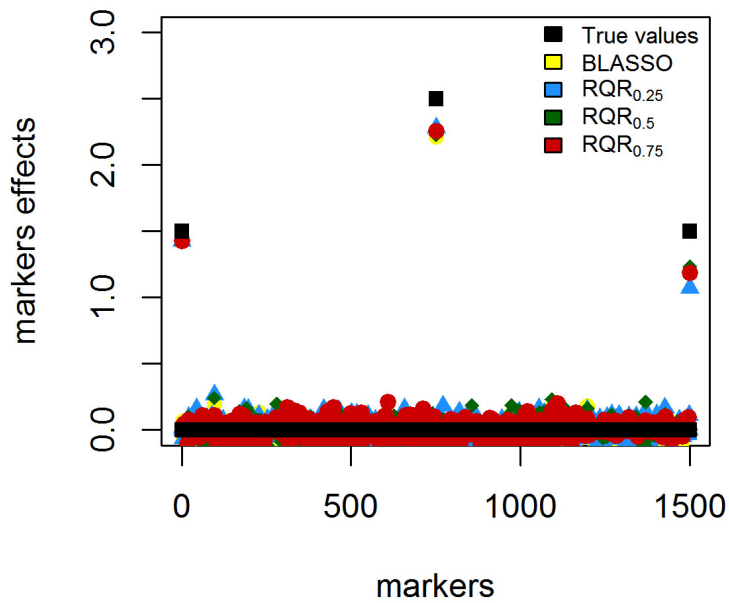
Figures 1, 2, and 3 provided the MSE in terms of bias and prediction variance error (PEV). When the distribution of phenotypic values showed negative skewness, we found that the  $RQR_{0.75}$  fit model yielded estimates with the least bias and PEV, because the estimated values were closer to the true values than those obtained using the other fit models ( $RQR_{0.25}$ ,  $RQR_{0.50}$ , and BLASSO) (Figure 1). In contrast, when considering phenotypic values from negatively skewed distribution, we found that the  $RQR_{0.25}$  fit model produced estimates with the least bias and PEV (Figure 2). In the symmetrical distribution scenario, the fit models produced similar results in term of bias and PEV (Figure 3).



**Figure 1.** True values and estimates of marker effects obtained using BLASSO and  $RQR_{\tau}$  ( $RQR_{0.25}$ ,  $RQR_{0.50}$ , and  $RQR_{0.75}$ ) fit models for negatively skewed phenotypic distribution.



**Figure 2.** True values and estimates of marker effects obtained using BLASSO and  $RQR_{\tau}$  ( $RQR_{0.25}$ ,  $RQR_{0.50}$ , and  $RQR_{0.75}$ ) fit models for positively skewed phenotypic distribution.



**Figure 3.** True values and estimates of marker effects obtained using BLASSO and  $RQR_{\tau}$  ( $RQR_{0.25}$ ,  $RQR_{0.50}$ , and  $RQR_{0.75}$ ) fit models for symmetrical phenotypic distribution.

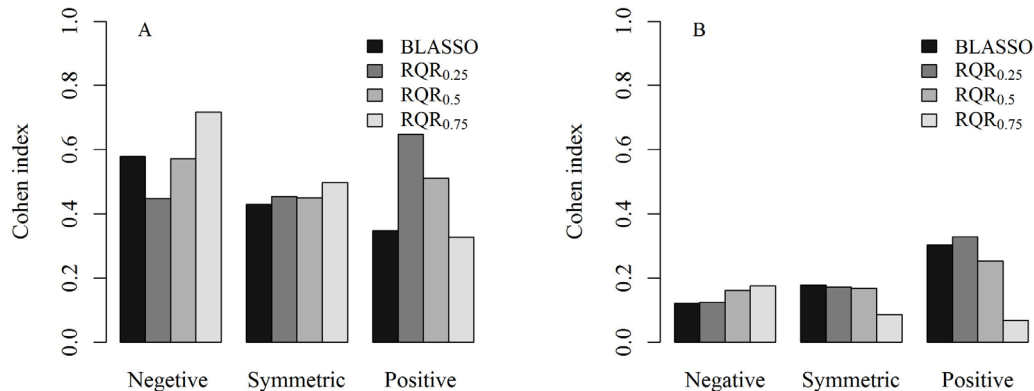


For the estimates between GEBVs obtained from the fit models ( $RQR_{0.25}$ ,  $RQR_{0.50}$ ,  $RQR_{0.75}$  and BLASSO), Spearman's correlations were positive and strong, with correlations ranging from 0.606 to 0.892 (Table 2). As expected, the higher estimates of correlation, 0.883, 0.887, and 0.892, were observed between the GEBVs obtained for  $RQR_{0.50}$  and BLASSO fit models for, respectively, negatively skewed, symmetrical, and positively skewed phenotypic distributions.

**Table 2.** Estimates of Spearman's correlation between genomic estimated breeding values (GEBV) obtained from Bayesian Lasso (BLASSO) and regularized quantile regression (RQR) fit models ( $RQR_{0.25}$ ,  $RQR_{0.50}$ , and  $RQR_{0.75}$ ) under different degrees of skewness.

Models		Scenarios		
		Negatively skewed	Symmetrical	Positively skewed
BLASSO	$RQR_{0.25}$	0.841	0.883	0.606
	$RQR_{0.50}$	0.789	0.887	0.614
	$RQR_{0.75}$	0.807	0.892	0.604

In addition to these results, after ranking the individuals according to the GEBVs obtained from either BLASSO or RQR, the percentage of selected individuals in common was calculated by the Cohen's kappa coefficient (0-1) (Cohen, 1960), based on GBVs for the first and last 100 scenarios (Figure 4A and B).



**Figure 4.** Cohen's kappa coefficient for agreement between genetic and estimated values. **A.** Percent top 100 +; **B.** percent top 100 -.

According to Landis and Koch (1977), a Cohen's kappa coefficient of  $C \geq 0.5$  may be considered a good estimate. We found that when the distributions of simulated phenotypes were symmetrical, the values of Cohen's kappa coefficient were between 0.50 and 0.54 in both methods. This indicates that genomic selection for traits that are symmetrically distributed can be accurately evaluated using both methods (Figure 4A). However, with respect to the 100 individuals with higher GBVs, the RQR method showed a higher percentage of common selected individuals for both positively and negatively skewed distributions than the BLASSO method (82.8 and 24.1%, respectively) (Figure 4A).

Similar results were observed for the 100 individuals with lower genetic values (Figure 4B). Specifically, the improvements using the RQR method were 10.0 and 41.6% with positive and negative skewness, respectively, compared with the BLASSO method (Figure 4B).

## DISCUSSION

In this study, we proposed the RQR method as a novel computational tool for GS that can improve marker estimation and GEBV prediction. We compared the proposed method with the traditional BLASSO method under various simulated scenarios (symmetrical, positively, and negatively skewed distributions). Each scenario considered by the additive model in simulation involved 1,000 individuals with complete information on 1,500 SNP markers. The fit models were  $RQR_{0.25}$ ,  $RQR_{0.50}$ ,  $RQR_{0.75}$ , and BLASSO. We investigated predictive performance and the relationship with GEBV for those approaches.

We found that the RQR method efficiently estimated GEBV; for at least one quantile model fit, assessments using MSE and accuracy confirmed that the RQR method showed better results than BLASSO when negative and positive skewness were present in phenotype distributions. These results are reasonable because, unlike traditional methods based only on conditional expectations  $[E(Y|X)]$ , RQR is an extension of a linear model for estimating functional relationships between variables in all portions of the distribution of a response variable. Moreover, according to Briollais and Durrieu (2014), when the conditional distributions of  $Y$  are non-normal (for instance, skewed), the mean might not be the best summary. The RQR method facilitates discovery of the “best” model for representation of the relationship between the dependent (phenotype) and independent (marker effect) variables, increasing the predictive performance of the model. Furthermore, RQR presented a higher match percentage with the 100 individuals evaluated at higher and lower GBV.

The potential of quantile regression (QR) has been confirmed by many studies. Pourhoseingholi et al. (2009) used linear and quantile regression models to analyze the predictors of duration of stay in hospital for patients with gastrointestinal cancers. The authors observed that if the duration data showed higher skewness, quantile regression led to better interpretation and richer inference. Briollais and Durrieu (2014) pointed out some aspects of the use of QR in genome-wide association studies (GWAS). Because the genetic variants identified through GWAS have small effect sizes and explain a small proportion of trait heritability, Wang et al. (2011), with the aim of enriching the genetic signal, estimated effect sizes related to particular quantiles of the trait distribution, in particular the extreme tails. QR allows direct estimation at the extremes, and specific sampling is not needed.

Altogether, because traits with negative (e.g., time until the occurrence of specific events) and positive (e.g., substance concentration) skewness in phenotypic distributions can be found in plant and animal breeding, the QR model presents an interesting and promising approach. For instance, in these cases a functional relationship defined to higher quantiles (e.g., 0.75) or lower quantiles (e.g., 0.25) can improve GS studies.

However, more studies using different genetic architectures and data set sizes (individuals and markers) are needed to confirm the efficiency of RQR. Furthermore, owing to an infinite number of quantiles that can exist in RQR, finding the “best” one to explain the functional relationship is still a challenge. Another issue is the estimation of the penalty parameter (shrinkage), which in the RQR fit is defined a priori. In this study, to compare RQR with BLASSO methodologies, the shrinkage parameters used in the RQR fit were the same as those used in BLASSO. Indeed, specific penalty values can be accessed exclusively for RQR by cross-validation (Fonseca Silva et al., 2011) or via Bayesian inference (Alhamzawi et al., 2012). These approaches may further improve the performance of RQR prediction because they present different powers of penalization over other methodologies.

## Conflicts of interest

The authors declare no conflict of interest.

## ACKNOWLEDGMENTS

We are grateful for financial support from Fundação de Apoio à Universidade Federal de Viçosa (FUNARBE), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG - #PPM-00518-15 and #APQ-00825-14), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq - Universal- #446176/2014-1).

## REFERENCES

- Alhamzawi R, Yu K and Benoit DF (2012). Bayesian adaptive LASSO quantile regression. *Stat. Model.* 12: 279-297. <http://dx.doi.org/10.1177/1471082X1101200304>
- Beyene Y, Semagn K, Mugo S, Tarekegne A, et al. (2014). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* 55: 154-163. <http://dx.doi.org/10.2135/cropsci2014.07.0460>
- Briollais L and Durrieu G (2014). Application of quantile regression to recent genetic and -omic studies. *Hum. Genet.* 133: 951-966. <http://dx.doi.org/10.1007/s00439-014-1440-6>
- Cade BS and Noon BR (2003). A gentle introduction to quantile regression for ecologists. *Front. Ecol. Environ* 1: 412-420. [http://dx.doi.org/10.1890/1540-9295\(2003\)001\[0412:AGITQR\]2.0.CO;2](http://dx.doi.org/10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;2)
- Christensen OF, Madsen P, Nielsen B, Ostersen T, et al. (2012). Single-step methods for genomic evaluation in pigs. *Animal* 6: 1565-1571. <http://dx.doi.org/10.1017/S1751731112000742>
- Cohen J (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20: 37-46. <http://dx.doi.org/10.1177/001316446002000104>
- Crispim AC, Kelly MJ, Guimarães SE, Fonseca e Silva F, et al. (2015). Multi-trait GWAS and new candidate genes annotation for growth curve parameters in Brahman cattle. *PLoS One* 10: e0139906. <http://dx.doi.org/10.1371/journal.pone.0139906>
- de los Campos G, Naya H, Gianola D, Crossa J, et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182: 375-385. <http://dx.doi.org/10.1534/genetics.109.101501>
- Filardi de Campos C, Soares Lopes M, Fonseca e Silva F, Veroneze R, et al. (2015). Genomic selection for boar taint compounds and carcass traits in a commercial pig population. *Livest. Sci.* 174: 10-17. <http://dx.doi.org/10.1016/j.livsci.2015.01.018>
- Fonseca Silva F, Varona L, de Resende MDV, Bueno Filho JSS, et al. (2011). A note on accuracy of Bayesian LASSO regression in GWS. *Livest. Sci.* 142: 310-314. <http://dx.doi.org/10.1016/j.livsci.2011.09.010>
- Goddard ME and Hayes BJ (2007). Genomic selection. *J. Anim. Breed. Genet.* 124: 323-330. <http://dx.doi.org/10.1111/j.1439-0388.2007.00702.x>
- Koenker R (2015). quantreg: Quantile regression. R package version 4.91. <http://CRAN.R-project.org/package=quantreg>.
- Koenker R and Bassett G (1978). Regression quantiles. *Econometrica* 46: 33-50. <http://dx.doi.org/10.2307/1913643>
- Landis JR and Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174. <http://dx.doi.org/10.2307/2529310>
- Li YJ and Zhu J (2008). L<sub>1</sub>-norm quantile regression. *J. Comput. Graph. Stat.* 17: 1-23. <http://dx.doi.org/10.1198/106186008X289155>
- Mathur PK, ten Napel J, Bloemhof S, Heres L, et al. (2012). A human nose scoring system for boar taint and its relationship with androstenone and skatole. *Meat Sci.* 91: 414-422. <http://dx.doi.org/10.1016/j.meatsci.2012.02.025>
- Maurer A, Draba V, Jiang Y, Schnaithmann F, et al. (2015). Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16: 290. <http://dx.doi.org/10.1186/s12864-015-1459-7>
- Meuwissen TH, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819-1829.
- Mosteller F and Tukey JW (1977). Data analysis and regression. Addison-Wesley, Reading.

- Muranty H, Troggio M, Sadok IB, Rifaï MA, et al. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. *Hortic Res* 2: 15060. <http://dx.doi.org/10.1038/hortres.2015.60>
- Pérez P, de Los Campos G, Crossa J and 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. <http://dx.doi.org/10.3835/plantgenome2010.04.0005>
- Pourhoseingholi MA, Vahedi M, Pourhoseingholi A, Moghimi-Dehkordi B, et al. (2009). Comparing linear regression and quantile regression to analyze the associated factors of length of hospitalization in patients with gastrointestinal tract cancers. *Ital. J. Public Health* 6: 136-139.
- R Development Core Team (2016). R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org>.
- Riedelsheimer C, Technow F and Melchinger AE (2012). Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics* 13: 452. <http://dx.doi.org/10.1186/1471-2164-13-452>
- Schwender H and Fritsch A (2015). *scrim*: Analysis of high-dimensional categorical data such as SNP data. R package version 1.2.8. [<http://CRAN.Rproject.org/package=scrim>].
- Varona L, Ibañez-Escriche N, Quintanilla R, Noguera JL, et al. (2008). Bayesian analysis of quantitative traits using skewed distributions. *Genet. Res.* 90: 179-190. <http://dx.doi.org/10.1017/S0016672308009233>
- Wang K, Li WD, Zhang CK, Wang Z, et al. (2011). A genome-wide association study on obesity and obesity-related traits. *PLoS One* 6: e18939. <http://dx.doi.org/10.1371/journal.pone.0018939>