agronomy

MDPI

Article

Quantile Regression Applied to Genome-Enabled Prediction of Traits Related to Flowering Time in the Common Bean

Ana Carolina Nascimento [1], Moyses Nascimento [1,*], Camila Azevedo [1], Fabyano Silva [2], Leiri Barili [3], Naine Vale [3], José Eustáquio Carneiro [3], Cosme Cruz [4], Pedro Crescencio Carneiro [4] and Nick Serão [5]

1 Department of Statistics, Federal University of Viçosa, Viçosa 36570-977, Brazil; ana.campana@ufv.br (A.C.N.); camila.azevedo@ufv.br (C.A.)
2 Department of Animal Science, Federal University of Viçosa, Viçosa 36570-977, Brazil; fabyano.fonseca@ufv.br
3 Department of Plant Sciences, Federal University of Viçosa, Viçosa 36570-977, Brazil; leyridaiana@hotmail.com (L.B.); nainemartinsdovale@hotmail.com (N.V.); eustaquiocarneiro@yahoo.com.br (J.E.C.)
4 Department of General Biology, Federal University of Viçosa, Viçosa 36570-977, Brazil; cruzcd@ufv.br (C.C.); pcscarneiro@gmail.com (P.C.C.)
5 Department of Animal Science, Iowa State University, Ames, IA 50011, USA; serao@iastate.edu
* Correspondence: moysesnascim@ufv.br; Tel.: +55-31-3612-2408

Received: 1 November 2019; Accepted: 21 November 2019; Published: 23 November 2019

Abstract: Genomic selection (GS) aims to incorporate molecular information directly into the prediction of individual genetic merit. Regularized quantile regression (RQR) can be used to fit models for all portions of a probability distribution of the trait, enabling the conditional quantile that "best" represents the functional relationship between dependent and independent variables to be chosen. The objective of this study was to predict the individual genetic merits of the traits associated with flowering time (DFF—days to first flower; DTF—days to flower) in the common bean using RQR and to compare the predictive abilities obtained from Random Regression Best Linear Unbiased Predictor (RR-BLUP), Bayesian LASSO (BLASSO), BayesB, and RQR for predicting the genetic merit. GS was performed using 80 genotypes of common beans genotyped for 380 single nucleotide polymorphism (SNP) markers. Considering the "best" RQR fit models (RQR0.3 for DFF, and RQR0.2 for DTF), the gains in predictive ability in relation to BLASSO, BayesB, and RR-BLUP were 18.75%, 22.58%, and 15.15% for DFF, respectively, and 15.20%, 24.65%, and 12.55% for DTF, respectively. The potential cultivars selected, considering the RQR "best" models, were among the 5% of cultivars with the lowest genomic estimated breeding value (GEBV) for the DFF and DTF traits—the IAC Imperador, IPR Colibri, Capixaba Precoce, and IPR Andorinha were included in the list of early cycle cultivars.

Keywords: *Phaseolus vulgaris* L.; linear model; conditional quantile; genome-enabled prediction

1. Introduction

Meuwissen et al. [1] introduced genomic selection (GS) as a means of incorporating molecular information directly into the prediction of individual genetic merit. GS has been successfully used in breeding to increase genetic gain per generation through early selection [2] and to improve prediction accuracy [3].

However, some statistical issues, for example, longitudinal [4] and non-normal [5] traits, still pose a challenge for GS. Although several statistical methods have been proposed for GS as solutions for multicollinearity and dimensionality, we believe there are limited reports in the literature that generalize these methods to non-normal traits. Non-normal distributions can be found for some traits in the fields of plant and animal breeding, for example, traits that measure the time until the occurrence of specific events (such as flowering and parity [6,7]) and hormone concentrations [8].

Another issue is related to the residual heteroscedastic variance. This tends to be neglected by existing methods, which focus on the mean of the conditional distribution, E(X|Y). Generally, in the presence of heteroscedastic variance, which is frequently observed in high dimensional data sets such as those found in GS studies, the sets of relevant covariates may differ when the different segments of conditional distribution are considered [9].

Quantile regression (QR) is a method that can be used to address these issues [10]. QR allows models to be fitted to all portions of the probability distribution of the trait, enabling a more complete picture of the conditional distribution than a single estimate of the center [11]. This feature allows QR to examine all of the conditional distributions, in order to investigate skewness and heteroscedasticity. From a GS viewpoint, we can choose the "best" conditional quantile to represent the relationship between the dependent and independent variables, thus increasing the accuracy of the genomic estimated breeding value (GEBV) prediction of individual genetic merits [12]. However, because of the high dimensionality commonly found in GS studies, a variation of QR, denoted by regularized quantile regression (RQR) [13] should be considered. RQR uses an L1-norm penalty for simultaneously controlling the variance of the fitted coefficients and performing automatic variable selection.

Among several breeding programs, the common bean (*Phaseolus vulgaris* L.) fits well in low-input agricultural systems, which are commonly practiced in African and Latin American countries. Moreover, the common bean is a key commodity for improving food security [14]. Besides productivity, traits associated with the flowering time such as days to flowering (DTF) and days to first flower (DFF) are important in the selection of the common bean. The identification of cultivars with an early cycle (shorter time from planting to harvest) allows for the planning of harvests in periods of less rain, the reduction of water consumption by irrigated crops, and earlier freeing of the area for crop succession [15]. In addition, cultivars with an early cycle are exposed to the risk of plague and disease for a shorter period of time.

In this context, we aimed to (1) predict the individual genetic merits of the traits associated with the flowering time (DFF and DTF) in the common bean using RQR, and (2) to compare the predictive abilities obtained for RQR, Random Regression Best Linear Unbiased Predictor (RR-BLUP), BayesB [1], and Bayesian LASSO (BLASSO) [16] for predicting genetic merit.

2. Materials and Methods

2.1. Phenotypic and Genotypic Data

The phenotypic data were obtained from two experimental stations at the Federal University of Viçosa, Minas Gerais State (MG), Brazil. One is located in Viçosa, MG (latitude 20º 45′ 14″ S, longitude 42º 52′ 55″ W, altitude 648m asl), and the other is in Coimbra, MG (latitude 20º 51′ 24″ S, longitude 42º 48′ 10″ W, altitude 720 m asl). Two phenological traits, DFF and DTF, were measured in eighty common bean cultivars, which were studied between 1970 and 2013 by research institutions in Brazil (Brazilian agricultural research corporation—Embrapa, Campinas Agronomic Institute—IAC, Federal University of Viçosa—UFV, Paraná Agronomic Institute—IAPAR, Agricultural Research Company of Minas Gerais—Epamig, Federal University of Lavras—UFLA, State Foundation for Agricultural Research—Fepagro, Rural Extension and Agricultural Research Enterprise—Epagri, and FT seeds).

The cultivars were planted at each location in the dry-summer (February) and winter (July) seasons of 2013 following a randomized complete block design with three replicates. The experimental plots consisted of four 3 m long rows, spaced 0.5 m apart with 15 seeds sown per meter. Fertilization was

applied according to the results of the soil analyses in order to ensure ideal conditions for development and production. Insect pests, invasive plants, and weeds were controlled as needed, according to the official recommendations for the common bean [17]. DFF was measured as the number of days from planting until at least one plant presented a flower. DTF was measured as the number of days from planting until at least 50% of the plants in a plot (replicate) had at least one open flower. These experiments were developed by the bean breeding program at the Federal University of Viçosa, Brazil. More details, such as the list of cultivars and research institutions are described in the literature [17,18].

The DNA samples were genotyped using the Vera Code1 BeadXpress platform (Illumina) at the Biotechnology Laboratory of Embrapa (Goiania, GO, Brazil). A set of 384 single nucleotide polymorphism (SNP) markers was selected as the oligo pool assay (OPA) SNP marker panel. The genotype call was performed using GenomeStudio software version 2011.1 (Illumina, San Diego, CA, USA), with call rate values ranging from 0.96 to 0.99, and GenTrain ≥ 0.46 for SNP clustering.

2.2. Phenotypic Data Analyses

The following statistical model of the phenological traits (DFF and DTF) was fit to the phenotypic data set as follows:

$$y_{ijk} = \mu + g_i + a_j + ga_{ij} + b_{k(j)} + \varepsilon_{ijk} \quad (1)$$

where y_{ijk} is the observed phenotype (DFF and DTF); μ is the overall mean; g_i is the random effect of the genotype (cultivar), i = 1 to 80, assumed to follow a normal distribution, with a mean of 0 and a variance of σ^2_g, with i = 1 to 80; a_j is the fixed effect of environment j, with j = 1 to 4; ga_{ij} is the random effect of the interaction of genotype i with environment j, assumed to follow a normal distribution with a mean of 0 and variance of σ^2_{ga}; $b_{k(j)}$ is the random effect of block (replicate) k (k = 1, 2, and 3) within environment j, assumed to follow a normal distribution with a mean of 0 and variance of σ^2_b; and ε_{ijk} is the experimental error for genotype i in block k of environment j, assumed to follow a normal distribution with a mean of 0 and variance of σ^2_ε. Nascimento et al. [19] used the same data set to evaluate models with and without the interaction between genotypes and environments through a likelihood ratio test (LRT), and showed that the full model had better fit.

The corresponding bivariate model is as follows:

$$\begin{bmatrix} \boldsymbol{y}_{DFF} \\ \boldsymbol{y}_{DTF} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{DFF} & 0 \\ 0 & \boldsymbol{X}_{DTF} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_{DFF} \\ \boldsymbol{b}_{DTF} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Zg}_{DFF} & 0 \\ 0 & \boldsymbol{Zg}_{DTF} \end{bmatrix} \begin{bmatrix} \boldsymbol{g}_{DFF} \\ \boldsymbol{g}_{DTF} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Zga}_{DFF} & 0 \\ 0 & \boldsymbol{Zga}_{DTF} \end{bmatrix} \begin{bmatrix} \boldsymbol{ga}_{DFF} \\ \boldsymbol{ga}_{DTF} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Zr}_{DFF} & 0 \\ 0 & \boldsymbol{Zr}_{DTF} \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_{DFF} \\ \boldsymbol{r}_{DTF} \end{bmatrix} + \begin{bmatrix} \varepsilon_{DFF} \\ \varepsilon_{DTF} \end{bmatrix} \quad (2)$$

where $\boldsymbol{y}_{DFF}$ and $\boldsymbol{y}_{DTF}$ denote the vectors of observed DFF and DTF, respectively; $\boldsymbol{X}_{DFF}$ and $\boldsymbol{X}_{DTF}$ denote the design matrices of the fixed effects for DFF and DTF, respectively; $\boldsymbol{b}_{DFF}$ and $\boldsymbol{b}_{DTF}$ denote the vectors of the fixed effects associated with $\boldsymbol{X}_{DFF}$ and $\boldsymbol{X}_{DTF}$, respectively; $\boldsymbol{Zg}_{DFF}$ and $\boldsymbol{Zg}_{DTF}$ denote the design matrices of the random effect of the genotype for DFF and DTF, respectively; $\boldsymbol{g}_{DFF}$ and $\boldsymbol{g}_{DTF}$ denote the vectors of the random effects with $\boldsymbol{Zg}_{DFF}$ and $\boldsymbol{Zg}_{DTF}$; respectively; $\boldsymbol{Zga}_{DFF}$ and $\boldsymbol{Zga}_{DTF}$ denote the design matrices of the random effect of the interaction of the genotype with the environment for DFF and DTF, respectively; $\boldsymbol{ga}_{DFF}$ and $\boldsymbol{ga}_{DTF}$ denote the vectors of the random effects with $\boldsymbol{Zga}_{DFF}$ and $\boldsymbol{Zga}_{DTF}$, respectively; $\boldsymbol{Zr}_{DFF}$ and $\boldsymbol{Zr}_{DTF}$ denote the design matrices of the random effect of the block (replicates) for DFF and DTF, respectively; $\boldsymbol{r}_{DFF}$ and $\boldsymbol{r}_{DTF}$ denote the vectors of the random effects with $\boldsymbol{Zr}_{DFF}$ and $\boldsymbol{Zr}_{DTF}$; respectively; and ε_{DFF} and ε_{DTF} denote the vectors of the random errors associated with $\boldsymbol{y}_{DFF}$ and $\boldsymbol{y}_{DTF}$, respectively. Assuming random effects distributed as a multivariate normal with the mean equal to zero and a covariance matrix, is as follows:

$$\begin{bmatrix} \boldsymbol{g}_{DFF} \\ \boldsymbol{g}_{DTF} \\ \boldsymbol{ga}_{DFF} \\ \boldsymbol{ga}_{DTF} \\ \boldsymbol{r}_{DFF} \\ \boldsymbol{r}_{DTF} \\ \boldsymbol{\varepsilon}_{DFF} \\ \boldsymbol{\varepsilon}_{DTF} \end{bmatrix} =$$

$$\begin{bmatrix} I\sigma^2_{g,DFF} & I\sigma_{g,DFF;DTF} & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 \\ I\sigma_{g,DFF;DTF} & I\sigma^2_{g,DTF} & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 \\ I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma^2_{ga,DFF} & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 \\ I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma^2_{ga,DTF} & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 \\ I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma^2_{r,DFF} & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 \\ I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma^2_{r,DTF} & I\sigma_{g,}00000 & I\sigma_{g,}00000 \\ I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma^2_{\varepsilon,DFF} & I\sigma_{\varepsilon,DFF;DTF} \\ I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{g,}00000 & I\sigma_{\varepsilon,DFF;DTF} & I\sigma^2_{\varepsilon,DTF} \end{bmatrix} \quad (3)$$

where $\sigma^2_{g,DFF}$ and $\sigma^2_{g,DFF}$ denote the random genotypes variance for DFF and DTF, respectively; $I\sigma_{g,DFF;DTF}$ denote the random genotype covariance between DFF and DTF; $\sigma^2_{ga,DFF}$ and $I\sigma^2_{ga,DTF}$ denote the random genotype by the environment variance for DFF and DFF, respectively; $I\sigma^2_{r,DFF}$ and $I\sigma^2_{r,DTF}$ denote the effect of the block (replicates) variance for DFF and DFF, respectively; $I\sigma^2_{\varepsilon,DFF}$ and $I\sigma^2_{\varepsilon,DTF}$ denote the random error variance for DFF and DTF, respectively; $I\sigma_{\varepsilon,DFF;DTF}$ denotes the random error covariance between DFF and DTF; and I denotes the identity matrix. The genetic parameters (broad sense heritability and correlations) were also estimated for the flowering traits using components of variance estimated by the bivariate model. The analyses were carried out in ASReml 3.0 [19].

2.3. Genomic Prediction Models

Prior to performing the genomic predictions, the adjusted phenotypes (Y^*_i) were obtained as the sum of the random effects (genotypes and error) from the selected model for analyzing the phenotypic data. Specifically, the adjusted phenotypes were obtained to correct for non-genetic sources of variation, for example, blocks, environments, and the interaction between the genotypes and environments (i.e., the combination of two locations (Viçosa and Coimbra, MG) and two seasons (dry and winter)).

The basic genomic model is represented by the following:

$$Y^*_i = \mu + \sum_{m=1}^{384} X_{im}\beta_m + e_i \quad (4)$$

where Y^*_i is the adjusted phenotype of the genotype, which is obtained as the sum of the random effects (genotypes and error); μ is the population mean; X_{im} is the incidence of the mth SNP in the ith adjusted phenotype of the genotype; and e_i is the random error term associated with Y^*_i.

Genomic prediction was performed using regularized quantile regression (RQR) [13]. RQR allows the GEBV at different portions of a probability distribution of the traits to be obtained. This method consists of obtaining parameter effects at level τ (θ_τ) that solve the following optimization problem:

$$\hat{\theta}_\tau = \text{argmin}_{\hat{\theta}_\tau}\left[\sum_{i=1}^{80}\rho_\tau(e_i) + \lambda\sum_{m=1}^{384}\left|\beta_{m\tau}\right|\right] \quad (5)$$

where τ indicates the quantile of interest, $\sum_{m=1}^{384}\left|\beta_{m\tau}\right|$ is the sum of the absolute values of the regression coefficients, and λ is the parameter that controls the strength of regularization. In the current study,

we evaluated five quantiles—(τ): 0.2, 0.3, 0.5, 0.7, and 0.8. The parameter $\rho_\tau(\cdot)$ is denoted as a check function [10] and is defined by the following:

$$\rho_\tau(e_i) = \begin{cases} \tau \cdot e_i, & \text{if } e_i \geq 0, \\ -(1-\tau) \cdot e_i, & \text{if } e_i < 0. \end{cases} \quad (6)$$

where $\tau \in]0,1[$ indicates the quantile of interest. The constrained optimization problem was solved using an algorithm that computes the exact solution for the model parameters [10].

GEBVs were obtained by $GEBV_\tau = \hat{u}_\tau = X\hat{\beta}_\tau$, where τ represents the τth quantile of interest. To define the "best" RQR fit, a grid of pair values given by combinations of τ ($\tau = 0.2,\ 0.3,\ 0.5,\ 0.7,\ 0.8$) and λ (from 0 until λ estimated by BLASSO, varying by 1), was considered. The predictive ability was used as a criterion to define the optimal pair. The GEBVs were also obtained using RR-BLUP, BLASSO, and BayesB.

In order to assess the predictive ability of all of the fitted models, the Pearson's correlation between the predicted values and the phenotypes were calculated using a four-fold cross-validation (CV) random process. This process was repeated randomly 10 times. The folds were the same for each fitted model.

The predictive abilities (PAs) and standard errors (SEs) were estimated from 10 estimates of the predictive ability.

After obtaining the GEBVs through the fit models, the Cohen's kappa [20] coefficient, and the Spearman's and Kendall's τ correlations were calculated to assess the agreement between the methods. The Cohen's kappa coefficient was used to calculate the percentage of individuals in common between the better 10% of individuals ranked according to the GEBVs. The Cohen's kappa coefficient is given by $C = NC - C_{Random}/1 - C_{Random}$, where NC is the relative observed agreement among raters, and C_{Random} is the hypothetical probability of the random agreement. The Spearman's correlation was calculated between the GEBV obtained by the different genomic selection models.

The RQR, RR-BLUP, BLASSO, and BayesB model fittings were carried out using the rq (for RQR) and BGLR (RR-BLUP, BLASSO, and BayesB) functions in the quantreg [21] and BGLR [22] packages, respectively, of R software [23]. The Bayesian methods were implemented using 200,000 Markov chain Monte Carlo iterations, with burn-in and thin values at 10,000 and five iterations, respectively. The convergence of the Markov chains was checked with the Geweke's Diagnostic [24].

3. Results

A summary of the descriptive statistics including the means, standard deviations (SD), ranges, and skewness for the phenological traits is presented in Table 1. The average days from planting until at least one plant presents one flower (DFF) was 34.74. After an average of 42 days, more than 50% of the plants in a plot presented at least one open flower (DTF; Table 1). The skewness coefficient shows that the phenotypes have left-skewed distributions.

Table 1. Means, standard deviations (SD), ranges, and skewness for days to first flower (DFF) and days to flowering (DTF), measured in 80 common bean genotypes.

Trait	Mean (SD)	Minimum	Maximum	Skewness
DFF (days)	34.74 (5.52)	20.00	49.00	−1.08
DTF (days)	42.30 (5.56)	27.00	55.00	−1.32

3.1. Model Selection and Genetic Parameters

The full model (model with the interaction effect) presented lower AIC (AIC_{DFF}: 2206, and AIC_{DTF}: 2318) and BIC (BIC_{DFF} = 2225, and BIC_{DTF} = 2337) values compared with those obtained from the reduced model (model without the interaction effect; AIC_{DFF} = 2233, and AIC_{DTF} = 2334;

BIC_{DFF} = 2248, and BIC_{DTF} = 2348). The LRTs (LRT_{DFF} = 29.8, and LRT_{DTF} = 17.26) between the full and reduced models were significantly different ($p < 0.01$) for both traits.

The estimates of heritability for DFF and DTF were moderate to high, with 0.58 ± 0.06 and 0.49 ± 0.08, respectively. For the estimates of the correlations between DFF with DTF, the genetic correlations were positive and strong, with a correlation of 0.98 ± 0.01 and a phenotypic correlation of 0.68 ± 0.04.

3.2. Prediction Accuracy of Traits

The estimated predictive abilities for two traits (DFF and DTF) ranged from −0.06 (0.04) to 0.38 (0.02) and are presented in Table 2. For DFF and DTF, the highest accuracy values were 0.38 and 0.34, obtained from $RQR_{0.3}$ and $RQR_{0.2}$, respectively (Table 2).

Table 2. Predictive ability for days to first flower (DFF) and to flowering (DTF) measured in 80 common bean genotypes, using a four-fold cross validation repeated 10 times for all of the fitted models (Bayesian LASSO (BLASSO), BayesB, Random Regression Best Linear Unbiased Predictor (RR-BLUP), and regularized quantile regression (RQR)).

Trait		Method							
		BLASSO	**BayesB**	**RR-BLUP**	**$RQR_{0.2}$**	**$RQR_{0.3}$**	**$RQR_{0.5}$**	**$RQR_{0.7}$**	**$RQR_{0.8}$**
DFF	PA [1]	0.32	0.31	0.33	0.37	0.38	0.25	−0.02	−0.05
	SE [2]	0.03	0.02	0.03	0.03	0.02	0.04	0.03	0.06
DTF	PA	0.29	0.27	0.30	0.34	0.32	0.10	0.01	−0.06
	SE	0.03	0.02	0.03	0.02	0.01	0.04	0.03	0.04

[1] Predictive ability. [2] Standard error of parameters estimates.

Considering the "best" RQR fit models ($RQR_{0.3}$ for DFF, and $RQR_{0.2}$ for DTF), the gains in predictive ability in relation to BLASSO, BayesB, and RR-BLUP were 18.75% (the ratio between the predictive ability obtained from "best" RQR model fit and the other methods), 22.58%, and 15.15% for DFF, respectively, and 15.20%, 24.65%, and 12.55% for DTF, respectively.

Estimates of the Spearman's correlation (lower triangle) and Cohen's kappa concordance coefficient (upper triangle) between the GEBVs obtained by the "best" quantile fit models ($RQR_{0.3}$ for DFF, and $RQR_{0.2}$ for DTF), and the BLASSO, BayesB, and RR-BLUP for the two traits are shown in Table 3.

Table 3. Estimates of Spearman's correlation (lower triangle), Kendall's τ rank correlation coefficient (lower triangle—in parenthesis), and Cohen's Kappa concordance coefficient[2] (upper triangle) between the genomic estimated breeding value (GEBV) values, obtained considering four different genomic selection models (BLASSO, RR-BLUP, BayesB, and RQR[1]) for days to first flower (DFF) and to flowering (DTF), measured in 80 common bean genotypes.

Traits	Method	BLASSO	BayesB	RR-BLUP	RQR [1]
DFF	BLASSO		1.00	1.00	0.63
	BayesB	0.99 (0.90)		1.00	0.63
	RR-BLUP	0.99 (0.99)	0.98 (0.89)		0.63
	RQR [1]	0.63 (0.36)	0.69 (0.41)	0.62 (0.32)	
DTF	BLASSO		1.00	1.00	0.63
	BayesB	0.99 (0.91)		1.00	0.63
	RR-BLUP	1.00 (0.99)	0.99 (0.91)		0.63
	RQR [1]	0.67 (0.38)	0.73 (0.43)	0.65 (0.38)	

[1] The "best" RQR fit models ($RQR_{0.3}$ for DFF and $RQR_{0.2}$ for DTF) considering the predictive ability. [2] Based on the lowest 10% GEBVs.

Overall, the Spearman's correlations presented high positive values. The lowest Spearman's correlation was estimated between the RR-BLUP and $RQR_{0.3}$ (0.62) models for DFF. The highest Spearman's correlations were estimated for BLASSO with BayesB (0.99) and RR-BLUP (0.99) for the DFF trait, and between the BLASSO and RR-BLUP (1.00) models for the DTF trait (Table 3). The Kendall's τ rank correlation coefficient estimates vary from moderate to high positive values. For instance, the Kendall's τ rank correlations between the RQR and BayesB, RR-BLUP, and BLASSO present moderate values (0.32–0.43) for both traits.

After ranking the individuals according to the GEBVs, the percentage of selected individuals in common was calculated using Cohen's kappa coefficient and based on the lowest 10% GEBVs. The lowest GEBVs represent those cultivars with an early cycle (shorter time from planting to harvest).

The lowest Cohen's kappa coefficient value was observed between RR-BLUP and $RQR_{0.2}$ (0.25) for the DTF trait. The highest values were observed between BayesB with RR-BLUP (1.00) and BLASSO (1.00) for DFF and DTF, respectively (Table 3, upper triangle).

The computational time to fit a model considering all of the four-fold processes varied from 0.28–264.08 s (Table 4). The Bayesian methods (BLASSO and BayesB) required a higher computational time compared with the RR-BLUP (0.28) and RQR (178.38).

Table 4. Average computational cost (seconds).

Method	Computational Time
BLASSO	264.08
BayesB	258.90
RR-BLUP	0.28
RQR [1]	178.38 [1]

[1] Time related to the fit of several models considering different quantile and shrinkage values.

4. Discussion

In this study, we predicted the individual genetic merits of the traits associated with the flowering time (DFF and DTF) in the common bean using RQR models. Using 80 genotypes of common beans genotyped for 384 markers, we compared the predictive ability of RQR to that obtained by BLASSO, BayesB, and RR-BLUP. The predictive ability of these methods was assessed using a four-fold cross-validation (CV) repeated 10 times. The Spearman's correlation and Cohen's kappa concordance (based on the lowest 10% GEBVs) coefficients between the GEBV values estimated by the four different genomic selection models used in this work were also estimated. However, first, the genetic parameters were estimated for DFF and DTF.

The heritability estimates for DFF (0.58) and DTF (0.49) were consistent with those reported in the literature. Specifically, the heritability estimate for DFF was within the range of estimates for the different crosses in beans (0.29–0.75 [25]; 0.59 [26]). In addition, for DTF, the estimate was close to that reported by the authors of [27], namely, 0.54. The estimates of the genetic and phenotypic correlations between the flowering traits were positive and high, which are similar to those reported in [28] for the phenological traits.

Overall, RQR outperformed the traditional methods for evaluating the two traits of DFF and DTF. Indeed, the better results are reasonable since the RQR allows for estimating the functional relationships between the variables for all portions of the probability distribution of the trait. The ability to choose the "best" relationship between the phenotype and markers increases the predictive performance of the model. According to the authors of [11] and [12], when the conditional distributions of Y are non-normal (for instance, skewed), the mean might not be the best way to describe the functional relationship between the variables.

The skewness coefficient for both traits (DFF and DTF) indicates a negative-skewed phenotypic distribution, and the "best" quantile fit models were $RQR_{0.3}$ for DFF, and $RQR_{0.2}$ for DTF. Therefore, our results indicate that to improve predictive ability, an effective strategy is to evaluate all of the

phenotypic distributions so as to choose the "best" quantile fit model. In addition, the heterogeneous variance that is frequently observed in high dimensional data sets suggests that a single slope is not able to characterize changes over the probability distribution, therefore indicating that RQR is a good tool to deal with those situations [29].

Overall, the models presented moderate to high values of Spearman's correlations (Table 3). On the other hand, the Kendall's τ rank correlation, which is a statistic that measures the ordinal association between two measured quantities, presented moderate correlations between RQR and the traditional genomic selection methodologies. Additionally, based on Cohen's kappa coefficient, the classification agreement between the RQR and non-quantile regression models (BLASSO, BayesB, and RR-BLUP) also showed moderate values (0.63) [30], suggesting differences in the classifications obtained by these models.

Among the 5% of cultivars with the lowest GEBVs for the DFF and DTF traits, the IAC Imperador, IPR Colibri, Capixaba Precoce, and IPR Andorinha were included in the list obtained by RQR "best" models. These are considered as early cycle cultivars [31–33], indicating the model ability to select. On the other hand, considering the GEBV obtained by BLASSO, BayesB, and RR-BLUP, the cultivar IPR Andorinha was replaced by BRSMG Madrepérola, which is not characterized as an early cycle cultivar [34].

Altogether, these results show that the use of RQR to predict the individual genetic merits of flowering time-related traits in the common bean (DTF and DFF) is worthy of interest. RQR showed similar or higher estimates of predictive ability compared with traditional methods (RR-BLUP, BLASSO, and BayesB), and was able to find cultivars with an early cycle. Moreover, RQR allows for the fitting of the regression models to other parts of the distribution of the trait, enabling a more informative study of the relationship between variables. This approach can be useful for representing different selection strategies (individuals with higher or lower values of the trait), or to give more information about potential cultivars for selection.

Overall, the computational time does not present any problems as the higher value was less than five minutes. The higher computational cost observed in the Bayesian methods is related to the estimation process, which is based on Markov Chain Monte Carlo (MCMC) algorithms. The RQR requires more computational time compared with RR-BLUP. The higher time observed in the RQR fitting is related to the necessity of evaluating several quantile and shrinkage values to choose the "best" model.

The potential of quantile regression (QR) has been confirmed by many studies. Briollais and Durrieu [11] pointed out some aspects of the use of QR in genome-wide association studies (GWAS). According to these authors, QR allows for direct estimation at the extremes, and specific sampling is not needed. Extreme sampling is used to enrich the genetic signal, where the main idea is to sample individuals with extreme phenotypes in the hope that rare causal variants will be enriched [35]. Recently, Nascimento et al. [18] used QR to identify genomic regions for phenological traits in the common bean. Unlike the traditional single-SNP GWAS model, the QR methodology was able to find SNP-trait associations considering one extreme quantile ($\tau = 0.1$). Barroso et al. [29] successfully used RQR for the SNP marker effect estimation of pig growth curves, as well as to identify the chromosome regions of the most relevant markers and to estimate the genetic individual weight trajectory over time under different quantiles. However, to the best of our knowledge, reports in the literature about the use of QR for GS are limited. Therefore, here, we introduce the QR for plant breeders, bringing new insights for GS studies.

Finally, QR uses all of the data set in the estimation process, which is different to using a subsample of data, which can result in smaller sample sizes for each regression [36] and introduces sample selection bias [37].

5. Conclusions

The regularized quantile regression (RQR) method was able to predict the individual genetic merits of the traits associated with the flowering time (DTF and DFF) in the common bean. In addition, considering the estimates of predictive ability, RQR presented similar or better results compared with those obtained for RR-BLUP, BLASSO, and BayesB. Moreover, RQR was able to find early cycle cultivars.

Author Contributions: Conceptualization, A.C.N., M.N., and N.S.; formal analysis, A.C.N. and M.N.; investigation, L.B.; methodology, A.C.N., M.N., and N.S.; software, A.C.N., M.N., C.A., and F.S.; writing (original draft), A.C.N., M.N., and N.S.; writing (review and editing), A.C.N., M.N., C.A., F.S., L.B., N.V., J.E.C., C.C., P.C.C., and N.S.

Funding: This research was funded by CAPES, CNPq, FAPEMIG, and FUNARBE.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **2001**, *157*, 1819–1829. [PubMed]
2. Resende, M.F.R., Jr.; Muñoz, P.; Resende, M.D.V.; Garrick, D.J.; Fernando, R.L.; Davis, J.M.; Jokela, E.J.; Martin, T.A.; Peter, G.F.; Kirst, M. Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (*Pinus taeda* L.). *Genetics* **2012**, *190*, 1503–1510. [CrossRef] [PubMed]
3. Crossa, J.; Beyene, Y.; Kassa, S.; Pérez, P.; Hickey, J.M.; Chen, C.; de los Campos, G.; Burgueño, J.; Windhausen, V.S.; Buckler, E.; et al. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3-Genes Genomes Genet.* **2013**, *3*, 1903–1926. [CrossRef]
4. Crispim, A.C.; Kelly, M.J.; Guimarães, S.E.; Silva, F.F.; Fortes, M.R.S.; Wenceslau, R.R.; Moore, S. Multi-trait GWAS and new candidate genes annotation for growth curve parameters in Brahman cattle. *PLoS ONE* **2015**, *10*, e0139906. [CrossRef] [PubMed]
5. Campos, C.F.; Soares, M.L.; Silva, F.F.; Veroneze, R.; Knol, E.F.; Lopes, P.S.; Guimarães, S.E.F. Genomic selection for boar taint compounds and carcass traits in a commercial pig population. *Livest. Sci.* **2015**, *174*, 10–17. [CrossRef]
6. Maurer, A.; Draba, V.; Jiang, Y.; Schnaithmann, F.; Sharma, R.; Schumann, R.; Kilian, B.; Reif, J.C.; Pillen, K. Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genom.* **2015**, *16*, 290. [CrossRef]
7. Varona, L.; Ibañez-Escriche, N.; Quintanilla, R.; Noguera, J.L.; Casellas, J. Bayesian analysis of quantitative traits using skewed distributions. *Genet. Res.* **2008**, *90*, 179–190. [CrossRef]
8. Mathur, P.K.; Ten Napel, J.; Bloemhof, S.; Heres, L.; Knol, E.F.; Mulder, H.A. A human nose scoring system for boar taint and its relationship with androstenone and skatole. *Meat Sci.* **2012**, *91*, 414–422. [CrossRef]
9. Wang, L.; Wu, Y.; Li, R. Quantile regression for analyzing heterogeneity un Ultra-high dimension. *J. Am. Stat. Assoc.* **2012**, *107*, 214–222. [CrossRef]
10. Koenker, R.; Bassett, G.W. Regression quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
11. Briollais, L.; Durrieu, G. Application of quantile regression to recent genetic and omic studies. *Hum. Genet.* **2014**, *133*, 951–966. [CrossRef] [PubMed]
12. Nascimento, M.; Silva, F.F.; Resende, M.D.V.; Cruz, C.D.; Nascimento, A.C.C.; Viana, J.M.S.; Azevedo, C.F.; Barroso, L.M.A. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. *Genet. Mol. Res.* **2017**, *16*, gmr16019538. [CrossRef] [PubMed]
13. Li, Y.; Zhu, J. L_1-norm quantile regression. *J. Comp. Graph. Stat.* **2008**, *17*, 1–23. [CrossRef]
14. Kamfwa, K.; Cichy, K.A.; Kelly, J.D. Genome-Wide Association Study of Agronomic Traits in Common Bean. *Plant Genome* **2015**, *8*, 1–12. [CrossRef]
15. Buratto, J.S.; Cirino, V.M.; Fonseca Junior, N.S.; Prete, C.E.C.; Faria, R.T. Agronomic performance and grain yield in early common bean genotypes in Paraná state. *Semina Ciênc Agrár* **2007**, *28*, 373–380. [CrossRef]
16. De los Campos, G.; Naya, H.; Gianola, D.; Crossa, J.; Legarra, A.; Manfredi, E.; Weigel, K.; Cotes, J.M. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. *Genetics* **2009**, *182*, 375–385. [CrossRef]

17. Barili, L.D.; Vale, N.M.; Prado, A.L.; Carneiro, J.E.S.; Silva, F.F.; Nascimento, M. Genotype-environment interaction in common bean cultivars with carioca grain cultivated in Brazil in the last 40 years. *Crop Breed. Appl. Biotechnol.* **2015**, *15*, 244–250. [CrossRef]
18. Nascimento, M.; Nascimento, A.C.C.; Silva, F.F.; Barili, L.D.; Vale, N.M.; Carneiro, J.E.S.; Carneiro, P.C.; Cruz, C.D.; Serao, N.V.L. Quantile regression for genome-wide association study of flowering time-related traits in common bean. *PLoS ONE* **2018**, *3*, e0190303. [CrossRef]
19. Gilmour, A.R.; Gogel, B.J.; Cullis, B.R.; Thompson, R. *ASReml User Guide Release, 3.0*; VSN International Ltd.: Hemel Hempstead, UK, 2009.
20. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
21. Koenker, R. Quantreg: Quantile Regression. Available online: https://CRAN.R-project.org/package=quantreg (accessed on 20 April 2018).
22. De los Campos, G.; Rodriguez, P.P. BGLR: Bayesian Generalized Linear Regression. Available online: https://cran.r-project.org/web/packages/BGLR/index.html (accessed on 1 March 2018).
23. R Core Team. *A Language and Environment for Statistical Computing*; Foundation for Statistical Computing: Vienna, Austria, 2017.
24. Geweke, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics*; Bernardo, L.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Oxford University: New York, NY, USA, 1992; pp. 625–631.
25. Cerna, J.; Beaver, J.S. Inheritance of early maturity of indeterminate dry bean. *Crop Sci.* **1990**, *30*, 1215–1218. [CrossRef]
26. Msolla, S.N.; Mduruma, Z.O. Estimate of Heritability for Maturity Characteristics of an Early x Late Common Bean (*Phaseolus Vulgaris* L.) Cross (TMO 216 x CIAT 16-1) and Relationships Among Maturity Traits with Yield and Components of Yield. *J. Agric. Sci.* **2007**, *8*, 11–18.
27. Moghaddam, S.F.; Mamidi, S.; Osorno, J.M.; Lee, R.; Brick, M.; Kelly, J.; Miklas, P.; Urrea, C.; Song, Q.; Cregan, P.; et al. Genome-Wide Association Study Identifies Candidate Loci Underlying Agronomic Traits in a Middle American Diversity Panel of Common Bean. *Plant Genome* **2016**, *9*, 1–21. [CrossRef] [PubMed]
28. Scully, B.T.; Wallace, D.H.; Viands, D.R. Heritability and correlation of biomass, growth rates, harvest index and phenology to the yield of common beans. *J. Am. Soc. Hortic. Sci.* **1991**, *116*, 127–130. [CrossRef]
29. Barroso, L.M.A.; Nascimento, M.; Nascimento, A.C.C.; Silva, F.F.; Serão, N.V.L.; Cruz, C.D.; Resende, M.D.V.; Silva, F.L.; Azevedo, C.F.; Lopes, P.S.; et al. Regularized quantile regression for SNP marker estimation of pig growth curves. *J. Anim. Sci. Biotechnol.* **2017**, *8*, 1–9. [CrossRef] [PubMed]
30. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
31. Infoteca-e: Repositório de Informação Tecnológica da Embrapa. Available online: https://www.infoteca.cnptia.embrapa.br/handle/doc/217045 (accessed on 30 May 2018).
32. Chiorato, A.F.; Carbonell, S.A.M.; Carvalho, C.R.L.; Barros, V.L.N.P.; Borges, W.L.B.; Ticelli, M.; Gallo, P.B.; Finoto, E.L.; Santos, N.C.B. 'IAC IMPERADOR': Early maturity "carioca" bean cultivar. *Crop Breed. Appl. Biotechnol.* **2012**, *12*, 297–300. [CrossRef]
33. IAPAR. Instituto Agronômico Do Paraná—IAPAR. Available online: http://www.iapar.br/modules/conteudo/conteudo.php?conteudo=1960 (accessed on 30 May 2018).
34. Carneiro, J.E.S.; Abreu, A.F.B.; Ramalho, M.A.P.; de Paula, T.J.; Del Peloso, M.J.; Melo, L.C.; Pereira, H.S.; Pereira Filho, I.A.; Martins, M.; Vieira, R.F.; et al. BRSMG Madrepérola: Common bean cultivar with late-darkening carioca grain. *Crop Breed. Appl. Biotechnol.* **2012**, *12*, 281–284. [CrossRef]
35. Wang, K.; Li, W.D.; Zhang, C.K.; Wang, Z.; Glessner, J.T.; Grant, S.F.A.; Zhao, H.; Hakonarson, H.; Price, R.A. A genome-wide association study on obesity and obesity-related traits. *PLoS ONE* **2011**, *7*, e18939. [CrossRef]
36. Cook, B.L.; Manning, W.G. Thinking beyond the mean: A practical guide for using quantile regression methods for health services research. *Shanghai Arch. Psychiatry* **2013**, *25*, 55–59.
37. Heckman, J.J. Sample selection bias as a specification error. *Econometrica* **1979**, *47*, 153–161. [CrossRef]