ARTICLE

Bayesian methods for genomic association of chromosomic regions considering the additive–dominance model

Crop Breeding and Applied Biotechnology
22(3): e418422310, 2022
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332022v22n3a33

Camila Ferreira Azevedo[1*], Leísa Pires Lima[1], Moyses Nascimento[1] and Ana Carolina Campana Nascimento[1]

Abstract: *Bayesian approaches applied in association studies select regions of single-nucleotide polymorphisms, indicating genes with important effects. The Bayesian methods differ in terms of the distribution assumed for the marker effects. Here, we used the window posterior probability of association to detect potential regions. The present study evaluated the efficiency of these methods in identifying regions located close to genes. Data were simulated in six scenarios. Considering the lack of dominance, BayesA was more efficient in the scenario with three QTLs. For scenarios with 10 or 100 QTLs, BayesCπ and BayesDπ were more efficient according to the false positive rate and detection power. Considering the presence of dominance, all methods were similar in the scenario with three QTLs, except in terms of accuracy. BayesDπ was superior in the scenario with 10 QTLs, while BRR was more efficient in the scenario with 100 QTLs.*

Keywords: *Genomic regions, molecular markers, association study, linkage disequilibrium*

INTRODUCTION

Genome-wide association studies (GWASs) allow for the detection of possible associations between quantitative trait loci (QTLs) and traits of interest using molecular markers with linkage disequilibrium (LD) (Resende et al. 2014). These analyses have prompted important advances in breeding programs by enabling the molecular control and biological processes of complex traits (Azevedo et al. 2019). However, in many studies, these analyses were performed using a single marker. Such analyses with single markers capture a smaller proportion of genetic variance and cannot identify complex associations among various markers; therefore, analyses of marker groups or select genomic regions are recommended (Moore et al. 2010).

The Bayesian methods were used by Fernando et al. (2017) to select associated regions. This approach allows for incorporating prior knowledge while simultaneously estimating marker effects, which increases the proportion of explained genetic variance. For instance, regions can be selected based on the Bayes factor, percentage of the recover variance by genomic regions, or window posterior probability of association (WPPA). Several criteria of region selection using Bayesian approaches were evaluated by Lima et al. (2022) and the authors recommended WPPA as the best index. However, in addition to the selection criteria, Bayesian models must be compared.

***Corresponding author:**
E-mail: camila.azevedo@ufv.br
ORCID: 0000-0003-0438-5123

Received: 05 March 2022
Accepted: 06 September 2022
Published: 20 September 2022

[1] Universidade Federal de Viçosa, Avenida PH Rolfs, s/n, Campus Universitário, 36570-000, Viçosa, MG, Brazil

Bayesian approach is used for several studies in genetic breeding (Azevedo et al. 2015, Evangelista et al. 2021). Specifically, Bayesian methods applied genomic prediction differ regarding the distribution assumed for marker effects, and the assumptions encompass the genetic architecture of the desirable trait (Gianola 2013). Specifically, Bayesian ridge regression (BRR) assumes a normal distribution with a single variance term for the marker effect. Meanwhile, Bayes A assumes a *t* distribution and different variance terms for each marker (Meuwissen et al. 2001). The Bayesian least absolute shrinkage and selection operator (BLASSO) assumes a double exponential distribution and different variance terms for each marker (de los Campos et al. 2009). BayesCp assumes a normal distribution with a single variance for the p fraction of the markers and no effect for other 1 − p markers (Gianola et al. 2009). BayesDp assumes a normal distribution with specific variance for the p fraction of the markers and no effect of the 1 − p fraction of the markers (Habier et al. 2011). BayesA*B*, proposed by Azevedo et al. (2015) for genomic prediction, involves BLASSO with a *t*-distribution; however, it can also be used in the context of the association.

To this end, the present study compared different Bayesian methods regarding their efficiency in selecting and detecting regions within or close to genes associated with the desirable traits. Specifically, we used simulated data for six different scenarios (three genetic architectures and two heritability levels, one with and the other without dominance) with single-nucleotide polymorphisms (SNPs) groups in non-overlapping regions.

MATERIAL AND METHODS

Simulated data

The simulated data scenarios were performed using the R package AlphaSimR (Gaynor et al. 2021), with 10 replicates each and 10 burn-in cycles. The simulated genome **consists of** 12 chromosome pairs for a hypothetical plant species and an F_2 population with 1,000 individuals. The diploid genome assumes that all chromosomes were of equal size with a genome length of 12 Morgans. Marker density was simulated by assigning 250 SNPs to each chromosome. Quantitative traits were simulated considering the absence or presence of dominance, mean of 0, and narrow-sense heritability, representing traits with high ($h^2_a = 0.50$), moderate ($h^2_a = 0.30$), and low ($h^2_a = 0.10$) heritability. Three genetic architectures were generated using 3, 10, and 100 loci controlling the trait, and these QTLs were randomly distributed in the first 10 chromosomes. The QTL effect was simulated using a normal distribution. The proportion of genetic variation associated with the QTL explained by the marker (r^2_{mq}) was calculated using the expression proposed by Goddard et al. (2011). Overall, six different scenarios were considered in the analyses: three genetic architectures × two heritability levels, one with and the other without dominance (Table 1).

Table 1. Description of scenarios (Sc.) with the proportion of variation in quantitative trait loci (QTL) explained by the SNPs (r^2_{mq}), genetic architecture, number of QTLs, narrow-sense heritability (h^2_a), dominance heritability (h^2_d) and list of statistics related to genotypic data: Mean (standard error) of the distance between markers (Dist. in Morgans), maximum of linkage disequilibrium (LD) per chromosome (maxLD), distance at which half of the maximum LD value is reached (d), number of regions (NREGIONS), and marker number (N) per region and the respective standard errors

Sc.	r^2_{mq}	Genetic architecture	Number of QTL	h^2_a	h^2_d	Dist.	maxLD	d	NREGIONS	N
1	0.9998	3 QTL on 12 chromosomes*	3	0.50	-	0.003(0.001)	0.41(0.01)	0.04(0.01)	308.10(1.15)	9.74(1.09)
2	0.9992	1 QTL in each of the 10 chromosomes*	10	0.30	-	0.004(0.001)	0.41(0.01)	0.04(0.01)	304.50(1.49)	9.85(1.10)
3	0.9917	10 QTL in each of the 10 chromosomes*	100	0.10	-	0.002(0.001)	0.40(0.01)	0.04(0.01)	302.80(0.47)	9.91(1.14)
4	0.9998	3 QTL on 12 chromosomes*	3	0.50	0.20	0.003(0.001)	0.50(0.01)	0.03(0.01)	165.70(6.54)	17.87(2.10)
5	0.9992	1 QTL in each of the 10 chromosomes*	10	0.30	0.10	0.005(0.001)	0.50(0.01)	0.03(0.01)	164.60(7.57)	18.23(2.00)
6	0.9917	10 QTL in each of the 10 chromosomes*	100	0.10	0.05	0.004(0.001)	0.50(0.01)	0.03(0.01)	164.11(7.54)	18.15(2.34)

* The last two chromosomes do not have QTL.

Statistical methods

The additive-dominant model is given by (Meuwissen et al. 2001):

$$y = 1\mu + Wm_a + Sm_d + e,$$

where y is the vector of phenotypes ($N \times 1$, N = number of individuals); μ is the mean; 1 is the vector of the same dimension as y, with all elements equal to unity; m_a is the vector of the additive effect of markers ($n \times 1$, n = number of markers); W ($N \times n$) is the incidence matrix of the additive effect of the markers; m_d is the vector of the dominant effect of the markers with $n \times 1$ dimension; S ($N \times n$) is the incidence matrix of the dominant effect of the markers; e ($N \times 1$) is the error vector ($e \sim N(0, I\sigma^2_e)$ and σ^2_e is the error variance); and I is the identity matrix. The marker incidence matrices were encoded as shown below (Vitezica et al. 2013).

The data distribution is defined by $y|\mu, m_a, m_d, \sigma^2_e \sim N(1\mu + Wm_a + Sm_d, I\sigma^2_e)$. However, in the Bayesian inference, the vectors of unknown parameters of the model are random quantities. Then, the probability distributions are assumed for them, which, in this context, are called prior distributions, such as $\mu \sim N(0, 10^8)$; $\sigma^2_e \sim v_e S^2_e \chi^{-2}$, where $v_e S^2_e \chi^{-2}$is the scaled inverse chi-squared distribution with the hyperparameters v_e and S^2_e. In the present study, $v_e = 5$ and $S^2_e = 0.5\sigma^2_y (v_e + 2)$, where σ^2_y is phenotypic variance (Perez and de los Campos 2014). The parameter vectors m_a and m_d of the model, which represent the marker effect vectors, also assumed prior distributions. The different prior distributions assumed for the markers effects are suitable for the different Bayesian methods proposed, and the selection of prior distributions plays a key role in the type of shrinkage of the estimates of marker effects (Azevedo et al. 2015).

BRR assumes a normal distribution of the marker effects (additive and dominant), with a mean of 0 and the same genetic variance as markers $\sigma^2_m(m|\sigma^2_m \sim N(0, I\sigma^2_m))$, which induces a homogeneous shrinkage of estimates. The common variance σ^2_m assumes a scaled inverted chi-square distribution with the hyperparameters v_m and S^2_m ($\sigma^2_m \sim v_m S^2_m \chi^{-2}$). In the present study, $v_m = 5$ and $S^2_m = \frac{0.5\sigma^2_y(v_m + 2)}{MS_W}$, where MS_W is the sum of sample variances of the columns of W (Perez and de los Campos 2014). BayesA assumes a normal distribution of the marker effects (additive and dominant) with a mean of 0 and specific genetic variance to the markers $\sigma^2_{mj}(m|\sigma^2_{mj} \sim N(0, I\sigma^2_{mj}))$, which induces a different shrinkage for each estimate. The specific variance σ^2_{mj} assumes a scaled inverted chi-square distribution with the hyperparameters v_m and S^2_m. In the present study, $v_m = 5$ and $S^2_m \sim Gamma(r,s)$, where $s = 1.1$ and $r = \frac{(s-1)MS_W}{0.5\sigma^2_y(v_m + 2)}$ (Perez and de los Campos 2014). The marginal prior distribution of the marker effects was t-student. BLASSO assumes a normal distribution of the marker effects, with a mean of 0 and a specific genetic variance to the markers $\tau^2_{mj}\sigma^2_e$ $(m|\tau^2_{mj}\sigma^2_e \sim N(0, I\tau^2_{mj}\sigma^2_e))$. The parameter τ^2_{mj} is assigned an exponential density with the rate parameters $\frac{\lambda^2}{2}$ and $\lambda^2 \sim Gamma(r,s)$. In the present study, $s = 1.1$ and $r = \frac{(s-1)}{2MS_W}$ (Perez and de los Campos 2014). The marginal prior distribution of the marker effects was double exponential. BayesA*B* involves BLASSO with a t-distribution (Azevedo et al. 2015). BayesCπ assumes that the 1 − π fraction of the markers has no genetic effect, while the π fraction has genetic effects assuming a mixture normal with the common variance σ^2_m, $m|\sigma^2_m, \pi \sim (1-\pi)\, N(0, \sigma^2_m = 0) + \pi N(0, \sigma^2_m)$, where $\sigma^2_m \sim v_m S^2_m \chi^{-2}$. For the parameter π is assigned the beta prior distribution $\pi \sim Beta(p_0,\pi_0)$. In the present study, $v_m = 5$, $S^2_m = \frac{0.5\pi\sigma^2_y(v_m + 2)}{MS_W}$, $p_0 = 10$, and $\pi_0 = 0.5$ (Perez and de los Campos 2014). BayesDπ assumes that the 1 − π fraction of the markers has no genetic, while the π fraction has genetics effects assuming a mixture of normal distribution with the specific variance σ^2_{mj}, $m|\sigma^2_{mj}, \pi \sim (1-\pi)\, N(0, \sigma^2_{mj} = 0) + \pi N(0, \sigma^2_{mj})$, where $\sigma^2_{mj} \sim v_m S^2_m \chi^{-2}$ and n $\pi \sim Beta(p_0,\pi_0)$. In the present study, $p_0 = 10$, $\pi_0 = 0.5$, $v_m = 5$, and $S^2_m \sim Gamma(r,s)$, where $s = 1.1$ and $r = \frac{(s-1)MS_W}{0.5\pi\sigma^2_y(v_m + 2)}$ (Perez and de los Campos 2014).

To the evaluation of the Bayesian methods, 320,000 iterations were performed with the Markov chain Monte Carlo (MCMC) algorithm, a burn-in period equal to 20,000 interactions, and a selection of 1 in 10 iterations (thinning). For convergence analysis, the criteria proposed by Geweke (1992) was used. The computational codes of the methods were based on R through GenomicLand visual interface (Azevedo et al. 2019).

Formation of regions

According to Viana et al. (2016), the size of genomic regions can be based on the average LD among markers. Pairwise estimates of LD (r^2, as measured using the squared correlation of incidence markers) were obtained to determine the region size. An LD value is pre-established by the researcher, and the shortest distance that provides this value is used to define the size of the region within of chromosome. Otyama et al. (2019) presented that an LD value < 0.20 tend toward equilibrium because LD is completely eroded. Resende et al. (2017) used this value to determine region length and claimed that this criterion is commonly used in the literature, together with the half-decay distance, for region formation (Rafalski 2002, Vos et al. 2017). The LD values were obtained using the *LD.decay* function of the R package sommer.

Selection based on WPPA

The window posterior probability of association of regions is based on the proportion of genetic variance explained by all markers of each genomic region. The portion of genomic estimated breeding values of individuals in the kth region was estimated as $\hat{g}_k = W_k\hat{m}_{a_k} + S_k\hat{m}_{d_k}$, where W_k and $\hat{m}_{a_k}$ are the additive marker incidence and estimate of the additive effect, respectively, and S_k and $\hat{m}_{d_k}$ are the dominant marker incidence and estimate of the dominant effect associated with the kth region, respectively. And the genomic variance of the kth region was estimated as $\hat{\sigma}^2_{g_k}$ = Var(g_k), and the proportion of genetic variance explained by the kth region (q_k) was defined as $q_k = \frac{\hat{\sigma}^2_{gk}}{E(\sigma^2_{gk})}$, where $E(\sigma^2_{g_k}) = \frac{\sigma^2_g}{n_r}$, where σ^2_g is the genetic variance of the trait and n_r is the number of regions. If $q_k > 1$, the kth region harbors a causative mutation (Peters et al. 2012, Bennewitz et al. 2017). WPPA is the ratio of the number of samples with $q_k > 1$ to the total number of samples. A grid with threshold values ranging from zero to one was tested.

Measures of comparison

The measures used to compare the Bayesian methods were: (i) accuracy ($r_{g,\hat{g}}$): correlation between the simulated and estimated genetic values; (ii) false positive rate (FP): portion of the regions detected by the methods but are not in LD with the QTL; (iii) detection power (PD): portion of the regions detected by the methods and are in LD with the QTL; (iv) percentage of genetic variance recovered by the regions; (v) area under the receiver operating characteristic (ROC) curve, i.e., the curve between the false positive rate and detection power (the method that provides the highest AUC value is considered more efficient for GWAS) (Lima et al. 2022); and (vi) number of regions detected as associated with chromosomes 11 and 12 (without QTLs). An optimal threshold should lead to the smallest distance between the point on the ROC curve and the point ($FP = 0$, $PD = 1$). The method that presents a lower false positive rate and higher detection power, recovers a greater portion of genetic variance, shows a larger AUC, and reports fewer associated regions on chromosomes 11 and 12 was considered the most suitable for GWAS. The measure denoted by rank summarizes the evaluation criteria applied to facilitate comparison among methods. It represents the sum of ranks of each method for each evaluation criterion. Therefore, the lower the sum, the more efficient the method.

RESULTS AND DISCUSSION

Definition of regions

The numbers of regions and markers in each region based on LD decay analysis for all scenarios are presented in Table 1. The distance that reached half of the maximum LD value between the markers was used to determine the region size ($r^2 = 0.20$). This measure has been used by several authors (e.g., Kim et al. 2007, Lam et al. 2010, Branca et al. 2011, Vos et al. 2017). The values obtained considering this threshold were similar across all scenarios. According to Vos et al. (2017), this threshold is promising for comparing LD decay among different studies, as it remains constant across simulated datasets. In scenarios without dominance (1, 2, and 3), the maximum LD values were slightly lower than those in scenarios with dominance. Consequently, shorter distances were obtained, providing more regions with fewer markers in each. According to Ge et al. (2018), different LD levels are relevant when dominance is considered. In addition, LD levels considered to determine the distances were close to 0.20, which is the commonly used threshold in the literature, because with this value, LD is expected to be completely eroded (Delourme et al. 2013, Resende et al. 2017).

Accuracy

Accuracy is not a measure typically presented in genomic association studies. However, since previous studies initially used Bayesian methods for genomic prediction, accuracy has become a crucial measure that must be reported. Contrary to the works of de los Campos et al. (2012) and Gianola (2013), we detected differences in the accuracy of different Bayesian methods, particularly in scenarios with fewer QTLs, in the present study. BRR presented higher accuracy in all scenarios, consistent with the report of Azevedo et al. (2015) that G-BLUP is the best prediction method, independent of the genetic architecture of the trait. Notably, both G-BLUP and BRR are based on the same infinitesimal model assumption.

Association Analysis

The results for scenarios without dominance are presented in Table 2. For scenario 1 (with three QTLs), BayesA gained prominence. This method simultaneously presented a lower false positive rate, higher detection power, greater proportion of variance explained, larger AUC, and fewer associated regions detected on chromosomes 11 and 12, being the best method in this scenario. The greater efficiency of BayesA than that of BLASSO, primarily for oligogenic traits, has been previously reported by Maturana et al. (2014) across different genetic architectures with simulated data, and it was re-confirmed in the present study. The *t* distribution of BayesA provides sufficiently thick tails that allow for capturing genes more easily, with a greater effect than normal and double exponential distributions (Azevedo et al. 2015).

In scenarios 2 and 3, BayesCπ and BayesDπ stood out for their efficiency. These methods allow direct selection of markers affecting the trait and detect potential QTL positions according to the frequency of non-null SNPs. According to Resende et al. (2014), methods that estimate the value of π to select a group of markers *a priori* are essential in association studies, since most markers are not in LD with QTLs. The advantages of BayesDπ in association studies have also been reported by Lima et al. (2022). In these scenarios, another method that stood out was BayesA, presenting relevant results, which, together with BayesCπ and BayesDπ, can be applied in GWAS. Our results corroborate the findings of Chen et al. (2017), who detected genomic regions considering simulated data and reported the superiority of BayesA and methods that select variables in detecting regions; according to the authors, these methods typically lead to much less shrinkage of large effects but a still greater shrinkage of small effects. Likewise, Chen et al. (2017) and Fernando et al. (2017) have stated that the use of variable selection procedures facilitates the determination of WPPA, which effectively maximizes the sensitivity and specificity in GWAS compared with inferences based on classic statistics.

In scenarios with 3 and 10 QTLs, BRR showed the poorest performance. In scenario with 100 QTLs, BLASSO and BayesA*B* were the less efficient methods. The threshold levels set for WPPA were the highest in some scenarios in which these methods performed poorly, corroborating the findings of Fernando et al. (2017), who stated that the false positive rate could be compromised by high threshold values for WPPA. The inferiority of BRR to the other Bayesian methods in GWAS has already been reported by Fernando and Garrick (2013), and it was re-confirmed in the present study. BayesA*B* may have been affected by the selection of degrees of freedom for marker variance, and it is associated

Table 2. Accuracy ($r_{g,\hat{g}}$), false positive rate (FP), detection power (Power), percentage of genetic variance recovered (PE), threshold for selecting regions obtained by window posterior probability of association (WPPA), area under the receiver operating characteristic (ROC) curve, number (N) of associated regions on chromosomes 11 and 12 (without QTLs), sum of the ranks of each method for each evaluation criterion (Rank), and the π_a fraction of the markers with the additive effect of BayesCπ and BayesDπ in scenarios without dominance

Scenarios	Method	π_a	$r_{g,\hat{g}}$	FP	Power	PE	WPPA	ROC	N	Rank
1	BRR		0.83(0.01)^a	0.18(0.06)^c	0.75(0.07)^a	96.07(1.19)^b	0.63(0.08)	0.79(0.02)^a	8.00(2.85)^b	10^w
	BayesA		0.75(0.02)^b	0.11(0.04)^a	0.75(0.07)^a	99.77(0.06)^a	0.39(0.08)	0.82(0.03)^a	3.40(1.71)^a	7^b
	BLASSO		0.82(0.02)^a	0.21(0.06)^d	0.83(0.07)^a	99.44(0.23)^a	0.51(0.09)	0.81(0.02)^a	9.80(3.10)^b	10
	BayesA*B*		0.76(0.02)^b	0.17(0.05)^b	0.80(0.07)^a	99.78(0.07)^a	0.36(0.06)	0.82(0.02)^a	6.70(2.67)^b	9
	BayesCπ	0.03(0.01)	0.72(0.02)^c	0.12(0.05)^a	0.68(0.08)^b	99.87(0.04)^a	0.51(0.13)	0.73(0.04)^b	4.50(2.30)^a	10
	BayesDπ	0.12(0.02)	0.73(0.02)^c	0.20(0.06)^c	0.78(0.07)^a	99.89(0.04)^a	0.42(0.12)	0.78(0.03)^a	7.50(2.42)^b	11
2	BRR		0.72(0.02)^a	0.62(0.08)^b	0.90(0.04)^a	93.52(2.97)^b	0.28(0.05)	0.57(0.04)^c	31.80(4.56)^b	11^w
	BayesA		0.69(0.02)^b	0.32(0.07)^a	0.68(0.08)^b	95.37(1.54)^b	0.29(0.04)	0.67(0.03)^b	14.70(4.01)^a	10
	BLASSO		0.72(0.02)^a	0.37(0.09)^a	0.69(0.09)^b	83.91(5.89)^c	0.42(0.07)	0.64(0.03)^b	17.90(4.95)^a	10
	BayesA*B*		0.71(0.02)^a	0.38(0.07)^a	0.69(0.08)^b	89.41(3.83)^c	0.35(0.05)	0.63(0.03)^b	18.60(4.35)^a	10
	BayesCπ	0.04(0.01)	0.67(0.03)^b	0.12(0.02)^a	0.67(0.04)^b	98.44(0.45)^a	0.23(0.03)	0.77(0.03)^a	4.00(1.26)^a	8^b
	BayesDπ	0.08(0.02)	0.67(0.03)^b	0.16(0.03)^a	0.72(0.05)^b	98.85(0.38)^a	0.22(0.03)	0.77(0.03)^a	6.60(2.39)^a	8^b
3	BRR		0.61(0.02)^a	0.48(0.04)^a	0.67(0.04)^c	78.78(2.95)^c	0.34(0.02)	0.59(0.01)^a	20.90(1.16)^a	10
	BayesA		0.56(0.03)^b	0.52(0.05)^b	0.71(0.05)^a	81.53(4.06)^b	0.29(0.02)	0.58(0.01)^a	23.70(1.90)^c	11
	BLASSO		0.55(0.04)^b	0.50(0.05)^b	0.67(0.05)^c	79.58(4.29)^b	0.30(0.03)	0.58(0.01)^a	22.20(2.06)^b	12^w
	BayesA*B*		0.56(0.03)^b	0.50(0.03)^b	0.68(0.04)^b	79.98(3.06)^b	0.32(0.02)	0.58(0.01)^a	23.00(1.41)^c	12^w
	BayesCπ	0.24(0.01)	0.59(0.03)^a	0.49(0.05)^a	0.67(0.05)^c	80.07(3.79)^b	0.30(0.02)	0.59(0.01)^a	21.60(1.89)^a	9^b
	BayesDπ	0.44(0.01)	0.59(0.03)^a	0.52(0.04)^b	0.70(0.04)^b	82.64(2.86)^a	0.27(0.01)	0.59(0.01)^a	22.00(1.26)^b	9^b

$^{w:}$ worst and $^{b:}$ best method

with the shrinkage parameter. According to Azevedo et al. (2015), proper selection of degrees of freedom is essential for the efficiency of the method; thus, new values should be analyzed in future studies.

The false positive rate and number of associated regions on chromosomes 11 and 12 (without QTL) increased with increase in the number of QTLs controlling the trait of interest, except when BayesCπ and BayesDπ were used in scenarios 1 and 2, in which equal or lower values were noted. Regarding the detection power of regions, percentage of genetic variance explained, and AUC, there were small yet similar decreases in the values with increase in the number of QTLs and consequent decrease in heritability. These results are consistent with the reports of Shin and Lee (2014); the authors used simulated data and found that statistical power estimates were affected by heritability, the proportion of the genetic variance and polygenic effects, and the number of causal variants.

The results for scenarios considering dominance are presented in Table 3. BLASSO was superior in scenarios controlled by three QTLs, because it jointly presented better results of the tested efficiency measures. In other scenarios, however, the values of all methods were similar for most measures, with the superiority of BLASSO being less evident. BLASSO, BayesA*B*, BayesCπ, and BayesDπ presented the maximum values of detection power for associated regions and percentage of genetic variance explained. In the scenario controlled by 10 QTLs, only BayesDπ stood out, because it presented a lower false positive rate, higher detection power value, greater percentage of variance explained, larger AUC, and fewer associated regions detected on chromosomes 11 and 12. BayesA*B* showed the poorest performance, with a lower detection power, smaller percentage of genetic variance recovered, and lower accuracy. Furthermore, in the scenario with 100 QTLs, BRR was more efficient. Contrary to that in the other scenarios, BayesDπ was the least efficient in this architecture.

For the scenario with three QTLs, the detection power for associated regions of all Bayesian methods analyzed increased in the presence of dominance, indicating the importance of considering these effects in association studies (Tables 2 and 3). Although less evident, the detection power also increased in the scenario with polygenic inheritance (100 QTLs). Wellmann and Bennewitz (2011) have elucidated the relevance of dominance effects, demonstrating that the dominance and additives effects are inter-dependent. Furthermore, Bennewitz et al. (2017) stated that the additive and dominance effects can often collectively increase the detection power of genomic regions. Even if only additive effects

Table 3. Accuracy ($r_{g,\hat{g}}$), false positive rate (FP), detection power (Power), percentage of genetic variance recovered (PE), threshold for selecting regions obtained by window posterior probability of association (WPPA), area under the receiver operating characteristic (ROC) curve, number (N) of associated regions on chromosomes 11 and 12 (without QTLs), sum of the ranks of each method for each evaluation criterion (Rank), and the π_a and π_d fractions of the markers with the additive and dominant effects of BayesCπ and BayesDπ in scenarios with dominance

Scenarios	**Method**	**π_a**	**π_d**	**$r_{g,\hat{g}}$**	**FP**	**Power**	**PE**	**WPPA**	**ROC**	**N**	**Rank**
4	BRR			0.85(0.02)^a	0.55(0.06)^a	0.95(0.05)^b	97.44(2.56)^b	0.30(0.05)	0.47(0.04)^b	54.20(4.77)^a	9
	BayesA			0.76(0.02)^c	0.56(0.06)^a	0.95(0.05)^b	97.36(2.64)^b	0.17(0.03)	0.46(0.04)^b	55.70(4.52)^a	11^w
	BLASSO			0.80(0.01)^b	0.57(0.06)^a	1.00(0.00)^a	100.00(0.00)^a	0.26(0.04)	0.49(0.05)^a	56.30(4.93)^a	7^b
	BayesA*B*			0.74(0.05)^c	0.60(0.07)^b	1.00(0.00)^a	100.00(0.00)^a	0.22(0.04)	0.48(0.04)^a	58.00(5.25)^b	10
	BayesCπ	0.46(0.01)	0.42(0.01)	0.74(0.03)^c	0.60(0.07)^b	1.00(0.00)^a	100.00(0.00)^a	0.26(0.04)	0.47(0.04)^b	58.10(5.08)^b	11^w
	BayesDπ	0.33(0.02)	0.45(0.01)	0.80(0.04)^b	0.61(0.07)^b	1.00(0.00)^a	100.00(0.00)^a	0.14(0.02)	0.47(0.04)^b	58.90(5.20)^b	10
5	BRR			0.78(0.02)^a	0.37(0.09)^c	0.62(0.07)^a	76.02(5.07)^a	0.50(0.08)	0.58(0.02)^b	10.60(2.86)^c	11
	BayesA			0.66(0.03)^c	0.28(0.06)^b	0.54(0.06)^b	70.81(6.66)^b	0.48(0.05)	0.59(0.02)^b	7.50(1.40)^b	13
	BLASSO			0.72(0.02)^b	0.37(0.08)^c	0.62(0.06)^a	77.36(4.58)^a	0.45(0.06)	0.57(0.02)^b	10.80(2.57)^c	13
	BayesA*B*			0.68(0.02)^c	0.26(0.07)^b	0.52(0.06)^c	69.40(5.63)^c	0.55(0.05)	0.58(0.02)^b	6.90(1.66)^a	14^w
	BayesCπ	0.48(0.01)	0.42(0.01)	0.69(0.02)^b	0.26(0.07)^b	0.52(0.06)^c	68.16(5.31)^c	0.56(0.05)	0.58(0.02)^b	6.70(1.61)^a	13
	BayesDπ	0.42(0.01)	0.45(0.01)	0.64(0.02)^d	0.25(0.07)^a	0.54(0.06)^b	76.47(6.36)^a	0.45(0.04)	0.60(0.02)^a	6.30(1.76)^a	10^b
6	BRR			0.71(0.01)^a	0.46(0.07)^a	0.71(0.05)^a	83.46(3.78)^a	0.35(0.05)	0.62(0.04)^a	37.50(7.14)^a	6^b
	BayesA			0.65(0.02)^b	0.47(0.06)^b	0.73(0.04)^b	85.06(3.06)^b	0.30(0.03)	0.62(0.04)^a	38.80(7.61)^a	10
	BLASSO			0.62(0.04)^c	0.43(0.05)^a	0.70(0.05)^a	83.53(3.54)^a	0.32(0.02)	0.62(0.03)^a	43.20(10.97)^b	9
	BayesA*B*			0.63(0.01)^c	0.47(0.06)^b	0.73(0.04)^b	85.06(3.06)^b	0.32(0.03)	0.62(0.04)^a	38.60(7.26)^a	11
	BayesCπ	0.45(0.01)	0.43(0.01)	0.68(0.01)^b	0.50(0.06)^c	0.75(0.04)^b	86.26(3.30)^b	0.30(0.03)	0.62(0.04)^a	43.90(9.14)^b	12
	BayesDπ	0.46(0.01)	0.45(0.01)	0.68(0.03)^b	0.52(0.07)^c	0.77(0.04)^c	87.84(2.99)^b	0.24(0.02)	0.62(0.04)^a	43.10(8.14)^b	13^w

$^{w:}$ worst and $^{b:}$ best method.

are of interest, dominance effects should not be neglected, because these effects are not completely independent. In addition, dominance is an important factor contributing to heterosis (Wellmann and Bennewitz 2012), which is of great significance to GWAS and can markedly affect the identification power of SNPs (Vidotti et al. 2019). For the scenario with 10 QTLs, detection power and percentage of genetic variance explained decreased when dominance was considered. Importantly, the number of associated regions on chromosomes 11 and 12 increased when dominance was considered. This was observed for all methods and in all scenarios, particularly in the scenario with three QTLs, in which this increase was more evident. However, according to Wang et al. (2004), if dominance effects are not correctly considered, the genetic progress of a breeding program may be even slower than when these values are neglected.

The probability π obtained by BayesCπ and BayesDπ for the scenarios without dominance ranged from 0.03 to 0.44, indicating that the number of markers that were supposedly in LD with the QTL ranged from 90 to 1,320. The π values were lower for BayesCπ criterion in all scenarios without dominance (scenarios 1, 2, and 3). However, among scenarios considering dominance, in the scenarios with 3 and 10 QTLs, the π values for additive effects were smaller for BayesDπ, while in the scenario with 100 QTLs, the values were similar for both methods. Thus, even in the scenarios with dominance, BayesCπ presented lower values. Notably, the π values were always higher in scenarios in which dominance was considered. According to Fernando and Garrick (2013), higher π values can be more discriminatory in identifying QTLs with major effects, which is an important element for SNP selection. In addition, no difference was observed in probability values for different mean degrees of dominance.

In the absence of dominance, BayesA, BayesDπ, and Bayes Cπ were more efficient in terms of the false positive rate and detection power in all scenarios. However, in the presence of dominance, no method stood out in more than one scenario. Thus, BayesDπ is more suitable for genomic association studies, particularly when the genetic architecture of the trait is unknown. In addition, BayesA*B* did not stand out in any simulated scenario in the present study and was, therefore, considered the least suitable in all scenarios.

ACKNOWLEDGMENTS

We thank the Brazilian National Council for Scientific and Technological Development (CNPq) and Coordination for the Improvement of Higher-Level Personnel (CAPES) for funding.

REFERENCES

Azevedo CF, Nascimento M, Fontes VCF, Silva FF, Resende MDV and Cruz CD (2019) GenomicLand: Software for genome-wide association studies and genomic prediction. **Acta Scientiarum. Agronomy 41**: e45361.

Azevedo CF, Resende MDV, Silva FF, Viana JMS, Valente MSF, Resende Jr MFR and Muñoz P (2015) Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics 16**: 1-13.

Bennewitz J, Edel C, Fries R, Meuwissen THE and Wellmann R (2017) Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. **Genetics Selection Evolution 49**: 7.

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, Ben C, Denny R, Sadowsky MJ, Ronfort J, Bataillon T, Young ND and Tiffin P (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. **Proceedings of the National Academy of Sciences 108**: E864-70.

Chen C, Steibel JP and Tempelman RJ (2017) Genome-wide association analyses based on broadly different specifications for prior distributions, genomic windows, and estimation methods. **Genetics 206**: 1791-1806.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD and Callus MPL (2012) 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, Weigel K and Cotes JM (2009) Predicting quantitative traits with regression Models for Dense Molecular Markers and Pedigree. **Genetics 182**: 375-385.

Delourme R, Falentin C, Fomeju BF, Boillot M, Lassalle G, André I, Duarte J, Gauthier V, Lucante N, Marty A, Pauchon M, Pichon JP, Ribière N, Trotoux G, Blanchard P, RiVière N, Martinant JP and Pauquet J (2013) High-density SNP-based genetic map development and linkage disequilibrium assessment in Brassica napus L. **BMC Genomics 14**: 1-18.

Evangelista JSPC, Peixoto MA, Coelho IF, Alves RA, Silva FF, Resende MDV, Silva FL and Bhering LL (2021) Environmental stratification and genotype recommendation toward the soybean ideotype: a Bayesian approach. **Crop Breeding and Applied Biotechnology 21**: e359721111.

Fernando R, Toosi A, Wolc A, Garrick D and Dekkers J (2017) Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. **Journal of Agricultural, Biological and Environmental Statistics 22**: 172-193.

Fernando RL and Garrick D (2013) Bayesian methods applied to GWAS. In Gondro C, van der Werf J and Hayes B (eds) **Genome-wide association studies and genomic prediction**. Humana Press, Totowa, p. 237-274.

Gaynor RC, Gorjanc G and Hickey JM (2021) AlphaSimR: an R package for breeding program simulations. **G3 Genes, Genomes, Genetics 11**: jkaa017.

Ge T, Chen CY, Neale BM, Sabuncu MR and Smoller JW (2018) Phenome-wide heritability analysis of the UK Biobank. **PLoS Genetics 14**: e1007228.

Geweke J (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernado JM, Berger JO, Dawid AP and Smith AFM (eds) **Bayesian statistics**. Clarendon Press, Oxford, p. 641-649.

Gianola D (2013) Priors in whole-genome regression: The Bayesian alphabet returns. **Genetics 194**: 573-596.

Gianola D, de los Campos G, Hill W, Manfredi E and Fernando R (2009) Additive genetic variability and Bayesian alphabet. **Genetics 183**: 347-363.

Goddard ME, Hayes BJ and Meuwissen TH (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal of Animal Breeding and Genetics 128**: 409-421.

Habier D, Fernando RL, Kizilkaya K and Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics 12**: 186.

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D and Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. **Nature Genetics 39**: 1151-1155.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, Weiming H, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SSM and Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. **Nature Genetics 42**: 1053-1059.

Lima LP, Azevedo CF, Resende MDV, Nascimento M and Silva FF (2022) Evaluation of Bayesian methods of genomic association via chromosomic regions using simulated data. **Scientia Agricola 79**: 3.

Maturana EL, Ibáñez-Escriche N, González-Recio O, Marenne G, Mehrban H, Chanock SJ, Goddard ME and Malats N (2014) Next generation modeling in GWAS: comparing different genetic architectures. **Human genetics 133**: 1235-1253.

Meuwissen TH, Hayes BJ and Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics 157**: 1819-1829.

Moore JH, Asselbergs FW and Williams SM (2010) Bioinformatics challenges for genome-wide association studies. **Bioinformatics 26**: 445-455.

Otyama PI, Wilkey A, Kulkarni R, Assefa T, Chu Y, Clevenger J, O'Connor DJ, Wright GC, Dezern SW, MacDonald GE, Anglin NL, Cannon EKS, Ozias-Akins P and Cannon SB (2019) Evaluation of linkage disequilibrium, population structure, and genetic diversity in the U.S. peanut mini core collection. **BMC Genomics 20**: 481.

Perez P and de los Campos G (2014) Genome-Wide regression and prediction with the BGLR statistical package. **Genetics 198**: 483-495.

Peters SO, Kizilkaya K, Garrick DJ, Fernando RL, Reecy JM, Weaber RL, Silver GA and Thomas MG (2012) Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers**. Journal of Animal Science 90**: 3398-3409

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. **Current Opinion in Plant Biology 5**: 94-100.

Resende MDV, Silva FF and Azevedo CF (2014) **Estatística matemática, biométrica e computacional**. Editora Suprema, Visconde do Rio Branco, 881p.

Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB and Grattapaglia D (2017) Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. **New Phytologist 213**: 1287-1300.

Shin J and Lee C (2015) Statistical power for identifying nucleotide markers associated with quantitative traits in genome-wide association analysis using a mixed model. **Genomics 105**: 1-4.

Viana JMS, Piepho HP and Silva FF (2016) Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations. **Scientia Agricola 73**: 243-251.

Vidotti MS, Lyra DH, Morosini JS, Granato ISC, Quecine MC, Azevedo JL and Fritsche-Neto R (2019) Additive and heterozygous (dis) advantage GWAS models reveal candidate genes involved in the genotypic variation of maize hybrids to *Azospirillum brasilense*. **PLoS One 14**: e0222788.

Vitezica ZG, Varona L and Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics 195**: 1223-1230.

Vos PG, Paulo MJ, Voorrips RE, Visser RG, van Eck HJ and van Eeuwijk FA (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. **Theoretical and Applied Genetics 130**: 123-135.

Wang J, van Ginkel M, Trethowan R, Ye G, DeLacy IH, Podlich DW and Cooper M (2004) Simulating the effects of dominance and epistasis on selection response in the CIMMYT Wheat Breeding Program using QuCim. **Crop Science 44**: 2006-2018.

Wellmann R and Bennewitz J (2011) The contribution of dominance to the understanding of quantitative genetic variation. **Genetics Research 93**: 139-154.

Wellmann R and Bennewitz J (2012) Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genetics Research 94**: 21-37.