

Inclusion of covariables in genome-wide selection models for prediction accuracy






Abstract – The objective of this work was to evaluate models using the significant single nucleotide polymorphisms (SNPs) detected by marker-assisted selection and genome-wide association, as a fixed effect in the models commonly used in genome-wide selection for F_2 population, in comparison with models using all SNPs. For all models, the Bayesian ridge regression method was used. Comparisons between the models were carried out to evaluate the phenotypic and genotypic prediction ability, phenotypic accuracy, selection gain, coincidence index, and processing time. Both methods failed to accurately identify true quantitative trait loci (QTL). The selection based only in the QTL identified by the studied methods elected individuals of low genetic value. The use of a genome-wide selection model – with the significant SNPs found by the genome-wide association as a fixed effect, and the remaining SNPs as a random effect – was the suitable strategy to select superior individuals with high accuracy. The introduction of QTL already described for a given trait into the genome-wide selection model allows of the selection of superior individuals with greater precision.

Index terms: genomic prediction, genome-wide association, marker-assisted selection study, prediction accuracy.

Inclusão de covariáveis em modelos de seleção genômica ampla para acurácia de predição

Resumo – O objetivo deste trabalho foi avaliar modelos que utilizam os nucleotídeos de polimorfismo único significativos (SNPs), encontrados por seleção assistida por marcadores e associação genômica, como um efeito fixo em modelos comumente utilizados na seleção genômica ampla para a população F_2 , em comparação com o modelo que utiliza todos os SNPs. Utilizou-se para todos modelos o método bayesiano de regressão de crista. Para as comparações entre os modelos, avaliaram-se a capacidade de predição fenotípica e genotípica, a acurácia fenotípica, o ganho de seleção, o índice de coincidência e o tempo de processamento. Ambos os métodos não conseguiram identificar com precisão os verdadeiros *loci* de características quantitativas (QTL). A seleção baseada apenas nos QTL identificados pelos métodos avaliados elegeu indivíduos de baixo valor genético. O uso de um modelo de seleção genômica ampla – com os SNPs significativos encontrados pela associação genômica como um efeito fixo, e os SNPs restantes como um efeito aleatório – foi a estratégia adequada para selecionar indivíduos superiores com alta precisão. A introdução de QTL já descritos para uma dada característica no modelo de seleção genômica ampla permite a seleção de indivíduos superiores com maior precisão.

Termos para indexação: predição genômica, associação genômica ampla, seleção assistida por marcadores, acurácia de predição.

Leonardo de Azevedo Peixoto⁽¹⁾ ,
Paulo Eduardo Teodoro⁽²⁾ ,
Larissa Pereira Ribeiro Teodoro⁽²⁾ ,
Cosme Damião Cruz⁽³⁾ , and
Leonardo Lopes Bhering⁽³⁾ 

⁽¹⁾ Iowa State University, Research Scientists,
716 Farm House, LN, 50011-1051, Ames, IA,
USA. E-mail: leonardo@iastate.edu

⁽²⁾ Universidade Federal de Mato Grosso
do Sul, Campus Chapadão do Sul,
Departamento de Agronomia, S/N, Rodovia
MS-306, Km 105, Zona Rural, CEP 79560-
000 Chapadão do Sul, MS, Brazil. E-mail:
eduteodoro@hotmail.com,
larissa.uems@gmail.com

⁽³⁾ Universidade Federal de Viçosa,
Departamento de Biologia Geral,
Avenida Peter Henry Rolfs, S/N, Campus
Universitário, CEP 36570-900 Viçosa, MG,
Brazil. E-mail: cdcruz@ufv.br,
leonardo.bhering@ufv.br

✉ Corresponding author

Received
October 02, 2023

Accepted
September 11, 2024

How to cite
PEIXOTO, L. de A.; TEODORO, P.E.;
TEODORO, L.P.R.; CRUZ, C.D.; BHERING,
L.L. Inclusion of covariables in genome-wide
selection models for prediction accuracy.
Pesquisa Agropecuária Brasileira, v.59,
e03534, 2024. DOI: <https://doi.org/10.1590/S1678-3921.pab2024.v59.03534>.

Introduction

Since its inception, plant breeding has been based on the visual selection of individuals, that is, the selection is based only on phenotypic value. With advances in molecular genetics and genomics, other new strategies and, consequently, new criteria, have been developed to make the breeding cycle faster and, the selection of individuals, more efficient. The first marker-based method was molecular marker-assisted selection (MAS), which is based on the information of a probable quantitative trait loci (QTL), in which some are identified as responsible for the expression of a certain phenotypic characteristic. MAS has shown to be efficient for traits governed by few genes with large effect. For traits controlled by many small-effect genes, this method has proved to be inappropriate (Gregorio et al., 2013).

Genome-wide selection (GWS) described by Meuwissen et al. (2001) is an alternative to solve the limitations found by MAS for quantitative traits. The GWS models are based on the estimation of the genomic breeding value, using a large number of markers and the phenotypic value of the individuals (Meuwissen et al., 2001). This genomic breeding value is then used for selecting superior individuals. At first, the effects of markers are estimated using training population, in which individuals are genotyped and phenotyped. These marker effects are then used to estimate the genomically estimated breeding value (GEBV) in the validation populations, in which individuals are genotyped and phenotyped. Then, these marker effects can be used to estimate the GEBV in a test population (population whose individuals are just genotyped).

Another MAS limitation is the need for a linkage map that can only be created from structured populations. An alternative that becomes possible after the development of SNP markers is an association mapping. Association mapping has the potential to find and map QTL within the genome, besides identifying causal polymorphism within genes that may be responsible for the difference between two phenotypes (Palaisa et al., 2003). Thus, it is developed as a method capable of identifying significant QTL from a large number of markers covering the entire genome known as genome-wide association studies (GWAS) (Pritchard et al., 2000).

Although MAS, GWS, and GWAS are distinct methods, they have the same ultimate goal that is to

improve the selection accuracy and help breeders to select superior genotypes. An alternative to using these methods simultaneously is to consider the QTL identified by the MAS and GWAS as a fixed effect, and the other markers as a random effect for the original GWS model. However, studies that have shown the use of these methods simultaneously (Bernardo, 2014; Spindel et al., 2015; Arruda et al., 2016) did not take into account the effect of heritability on genomic prediction. In addition to these, it is important to evaluate two more complete models: one using all simulated QTL as a fixed effect for the GWS model; and other using only the two QTL with the greatest effect on the characteristic as fixed effect for the GWS. Then, it is necessary to compare them with the standard model used in genomic selection (Bayesian ridge regression – BRR).

Therefore, the objective of this work was to evaluate models, using the significant single nucleotide polymorphisms (SNPs) detected by marker-assisted selection and genome-wide association as a fixed effect in the models commonly used in genomewide selection for F_2 population, in comparison with the Bayesian ridge regression.

Materials and Methods

This study was developed in the laboratory of biometrics of the Universidade Federal de Viçosa, in the state of Minas Gerais, Brazil, in 2018. An F_2 population was simulated by using the GENE software module (Cruz, 2013), which allowed to generate information on the genome, genotypes of the genitors, controlled crossover populations, and quantitative traits data.

A genome consisting of 15 linkage groups was performed similarly to that of a diploid species $2n = 2x = 30$. Each linkage group was simulated with 150 cM, consisting of 300 codominant and biallelic markers, equally spaced (0.5 cM), totaling 4,500 marks.

Contrasting homozygous parents were simulated, for which parent 1 was coded as carrier of an A1 allele (received code 2), and parent 2 was coded as carrier of the alternative allele A2 (received code 0) for all existing markers.

The F_2 population was generated from the self-crossing of individuals from the F_1 population. For the formation of the first individual of the F_2

population, each individual of the F_1 population produced 5,000 gametes and, when 2 of these gametes were found at random, the first individual of the F_2 population was generated. This process was repeated until the formation of all individuals in the population.

The simulated F_2 population was coded with 0, 1, and 2, for which 0 corresponded to homozygous individuals (A_2A_2), 1 to heterozygous individuals (A_1A_2), and 2 to homozygous individuals (A_1A_1), for a given locus.

For the simulation of quantitative traits, a value corresponding to the probability generated by a binomial distribution each trait was used, and it was controlled by 100 QTL randomly distributed in the genome. The effect of each QTL was defined by $A_1A_1=\mu + a$; $A_1A_2=\mu + d$; $A_2A_2=\mu - a$, where: 'a' is the coded effect of the homozygote, and 'd' is the coded effect of the heterozygote.

The genotypic value (GV) of each individual was defined by the equation, in which PVG is the proportion of genetic variance explained by each QTL:

$$GV = \sum_{i=1}^n (PVG / QTL_i \times QTL_i \text{ effect}),$$

The environmental effect (EE) was assumed to be uncorrelated with the genotype value and was estimated following a distribution $N(0, \sigma^2)$. Heritability traits were simulated at 20%, 40%, 60%, and 80%. The $\hat{\sigma}_g^2$ was calculated as being the variance of the genotypic value of the individuals of F_2 population. Therefore, the phenotypic value was obtained by

$$PV = u + GV + EE,$$

where: u 100 is the mean; and PV is the phenotypic value.

After forming the population, the mapping process stages followed, starting with the analysis of segregation of individual loci. Chi-square (χ^2) tests were applied to verify if the markers segregated as expected in a F_2 population. It was also verified if all linkage groups were restored, with size, distance, and order of the markers, which should make it possible to conclude that it was an F_2 population with the desired simulation properties.

To verify the accuracy of the MAS and GWAS methods for finding significant SNPs, a comparison

was made between the simulated QTL and the QTL found by these methods.

The analyses for QTL detection considering MAS concepts involved the interval mapping (IM) method (Lander & Botstein, 1989). This analysis was performed using the package `qtl` in R. All SNPs whose logarithm of odds (LOD) was greater than 3 were selected to be used as fixed effect in the GWS model.

The GWS method used was the best linear unbiased prediction (BRR) from the Bayesian ridge regression, which aims to estimate the effect for each of the covariables (markers SNPs) included in the model. The BGLR package (Pérez & de los Campos, 2014) was used to process the BRR method. This method was chosen because in the BGLR package it is possible to model fixed effects and random effects, and the Bayesian method requires less computational time. Seven models of GWS were used, as in the following descriptions.

Model 1 (MAS): in this model, the significant SNPs found by molecular marker-assisted selection (MAS) for each trait under analysis were modeled as fixed. The MAS analyses were performed using the 'cim' function in the `qtl` package of the R program (R Core Team, 2015) with the following equation:

$$GEBV = \hat{\mu} + \hat{\beta}X,$$

where: μ is the genotype mean; β is the effect vector of each significant SNP found by the MAS analyses (fixed effect); and X is a marker matrix composed only of the significant SNPs.

Model 2 (GWAS): in this model, the significant SNPs found by the genome-wide association studies (GWAS) for each trait under analysis were modeled as fixed. The GWAS analyses were performed using the `gwas` function in the `rrBLUP` package in the R program with the following equation:

$$GEBV = \hat{\mu} + \hat{\beta}X,$$

where: μ is the genotype mean; β is the effect vector of each significant SNP found by the GWAS (fixed effect) analyses; and X is the marker array composed only of the significant SNPs.

Model 3 (GWS): in this model, all SNPs were modeled as random, using the following equation:

$$GEBV = \hat{\mu} + \hat{\alpha}W,$$

where: μ is the genotype mean; α is the additive effect vector of each SNP (random effect); and W is the marker array composed of all SNPs.

Model 4 (M_G): in this model, the significant SNPs found by MAS, for each trait under analysis, were modeled as fixed effect, and the remaining SNPs were modeled as random effect, using the following equation:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W,$$

where: μ is the genotype mean; β is the effect vector of each significant SNP found by the MAS analyses (fixed effect); X is the marker matrix composed only of the significant SNPs; α is the additive effect vector of each SNP (random effect); and W is the marker array composed of all nonsignificant SNPs in the MAS analyses.

Model 5 (G_G): in this model, the significant SNPs found by GWAS, for each trait under analysis, were modeled as fixed effect, and the remaining SNPs were modeled as random effect by the following equation:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W,$$

where: μ is the genotype mean; β is the effect vector of each significant SNP found by the GWAS (fixed effect) analyses; X is the marker matrix composed only of the significant SNPs; α is the additive effect vector of each SNP (random effect); and W is the marker array composed of all nonsignificant SNPs in the GWAS analyses.

Model 6 (H_G): in this model, the two QTL with the greatest effect simulated for each trait under analysis were modeled as fixed effect, and the remaining SNPs were modeled as random effect using the following equation:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W,$$

where: μ is the genotype mean; β is the effect vector of the two most significant QTL by the simulation process (fixed effect); X is the marker array composed only of the two QTL; α is the additive effect vector of each SNP (random effect); and W is the marker array composed of all SNPs, except for the two QTL with the greatest effect on the traits.

Model 7 (Q_G): in this model, all simulated QTL for each trait under analysis were modeled as fixed effect,

and the remaining SNPs were modeled as random effect with the following equation:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W,$$

where: μ is the genotype mean; β is the effect vector of all QTL (fixed effect); X is the marker array composed of all QTL; α is the additive effect vector of each SNP (random effect); and W is the marker array composed of all SNPs, except for the QTL.

To compare the models proposed in this work, some parameters were estimated as phenotypic predictive ability (PPA) and genotypic predictive ability (GPA). The PPA was achieved by the Pearson correlation between the genomically estimated breeding values (GEBV) by the models and the phenotypic value. The GPA was achieved by the Pearson correlation between the GEBV estimated by the models and the genotypic value.

Phenotypic accuracy (PA) was calculated as follows:

$$PA = PPA / (h^2)^{1/2},$$

where: h^2 is the heritability of the trait.

Selection gain (SG %) was estimated using the following equation:

$$SG = 100 (SD h^2) / m_0,$$

where: SD is the selection differential that was estimated as

$$SD = m_s - m_0,$$

where: m_s is the mean of the selected individuals; and m_0 is the mean of the initial population. A selection percentage of 20% was considered.

The coincidence index (CI) was calculated as follows:

$$CI = 100 (NIS/TN)$$

where: NIS is the number of individuals selected on the basis of the phenotypic value that were the same selected on the basis of the GEBV; and TN is the total number of selected individuals.

The maximum selection gain (SG_{\max}) was estimated as:

$$SG_{\max} = \bar{X}_S - \bar{X}_0$$

where: \bar{X}_s is the genotypic mean of the individuals selected on the basis of the simulated genetic values (true values); and \bar{X}_o is the genotypic mean of the original population.

Results and Discussion

The number of QTL found by GWAS was lower than those of the simulated total, since out of 100 QTL, 66 had a significant effect on traits, while the number of QTL found by MAS was higher than that of the simulated total (Table 1). Few QTL identified by the GWAS were actually QTL, which means that most of the SNPs identified as QTL were not true QTL. MAS was able to identify almost 50% of the simulated QTL, but showed a very high number of significant SNPs in the wrong position.

The models G_G, GWAS, Q_G, and H_G were superior to the other methods for estimating the phenotypic predictive ability for all evaluated heritabilities (Figure 1). For 60% and 80% heritabilities, the MAS and M_G models showed a phenotypic predictive ability close to zero, which is much lower than those of the other models that had values above 0.4.

The genotypic predictive ability estimated by the GWS, G_G, Q_G, and H_G models was higher than those estimated by the others (Figure 2) for all evaluated heritabilities. The GWAS model was lower than those of the MAS and M_G models for low heritability (20% and 40%), and higher for high heritability (60% and 80%). Genotypic predictive ability increased as heritability increased for all evaluated models, except for the MAS and M_G ones.

Table 1. Number of quantitative trait loci (QTL) detected (NQD) and number of QTL detected in the correct position (NQDCP) for different heritabilities, by the methods of molecular marker-assisted selection (MAS) and genome-wide association studies (GWAS).

Model and heritability	NQD	NQDCP
MAS – $h^2 = 20\%$	2.221	52
GWAS – $h^2 = 20\%$	7	1
MAS – $h^2 = 40\%$	1.807	47
GWAS – $h^2 = 40\%$	11	3
MAS – $h^2 = 60\%$	1.957	51
GWAS – $h^2 = 60\%$	24	5
MAS – $h^2 = 80\%$	2.096	52
GWAS – $h^2 = 80\%$	27	7

Following the same response observed for phenotypic predictive ability, the GWS, G_G, Q_G, and H_G models were superior for the phenotypic accuracy for all evaluated heritabilities (Figure 3). For 60% and 80% heritabilities, the MAS and M_G models displayed a phenotypic accuracy close to zero, which is much lower than the values of the other models which had values above 0.6. The Q_G model showed the highest values of phenotypic accuracy for heritabilities equal to, or greater than 40%.

In GWAS and G_G models, the accuracy and predictive capacity were equal to, or greater than the standard GWS model (Figures 1, 2, and 3). This happened because in these models the maximum number of markers used as fixed effect was 32. The same fact was observed in the model H_G, in which only the two QTL with the largest effect (Table 1) were used as fixed effect in the model.

The use of data obtained using the GWAS analysis within the GWS models can provide information on the genetic architecture of the studied trait and on the population structure being used in the breeding program, according to Spindel et al. (2015); these authors have shown that the use of significant markers found by GWAS, as a fixed effect in GWS models for grain yield, plant height, and flowering in rice, can show the presence of QTL with higher segregating effect in the breeding population, in which these QTL can be established as covariates in GWS models to improve accuracy. However, when the number of QTL is greater than 10, this effect can be contrary, that means, a decrease could happen in the value of accuracy (Bernardo, 2014). This fact explains why the predictive ability (Figure 1 and 2) and the accuracy (Figure 3) of the MAS and M_G models were lower than those of the others, since more than 10 marks were used as fixed effects in these models (Table 1).

Treating a QTL as a fixed effect can increase the prediction accuracy of the GWS models. However, if QTL is a false positive, it will actually decrease the prediction accuracy of the model. In the present study, several false positive numbers were observed for the MAS and M_G models, evidencing one more factors that made these models inferior to the others. Thus, it is preferable to treat false positives as random and, therefore, have their variance close to zero, instead of treating them as fixed, since they strongly influence the prediction of genetic values (Arruda et al., 2016).

The Q_G model clearly showed this fact, since, in this model, 100 QTL were used as a fixed effect, where all QTL had an effect on the traits. In other words, no

false positive was used as fixed effect, and values of predictive ability and accuracy were observed as equal to, or higher than the traditional GWS model, in which

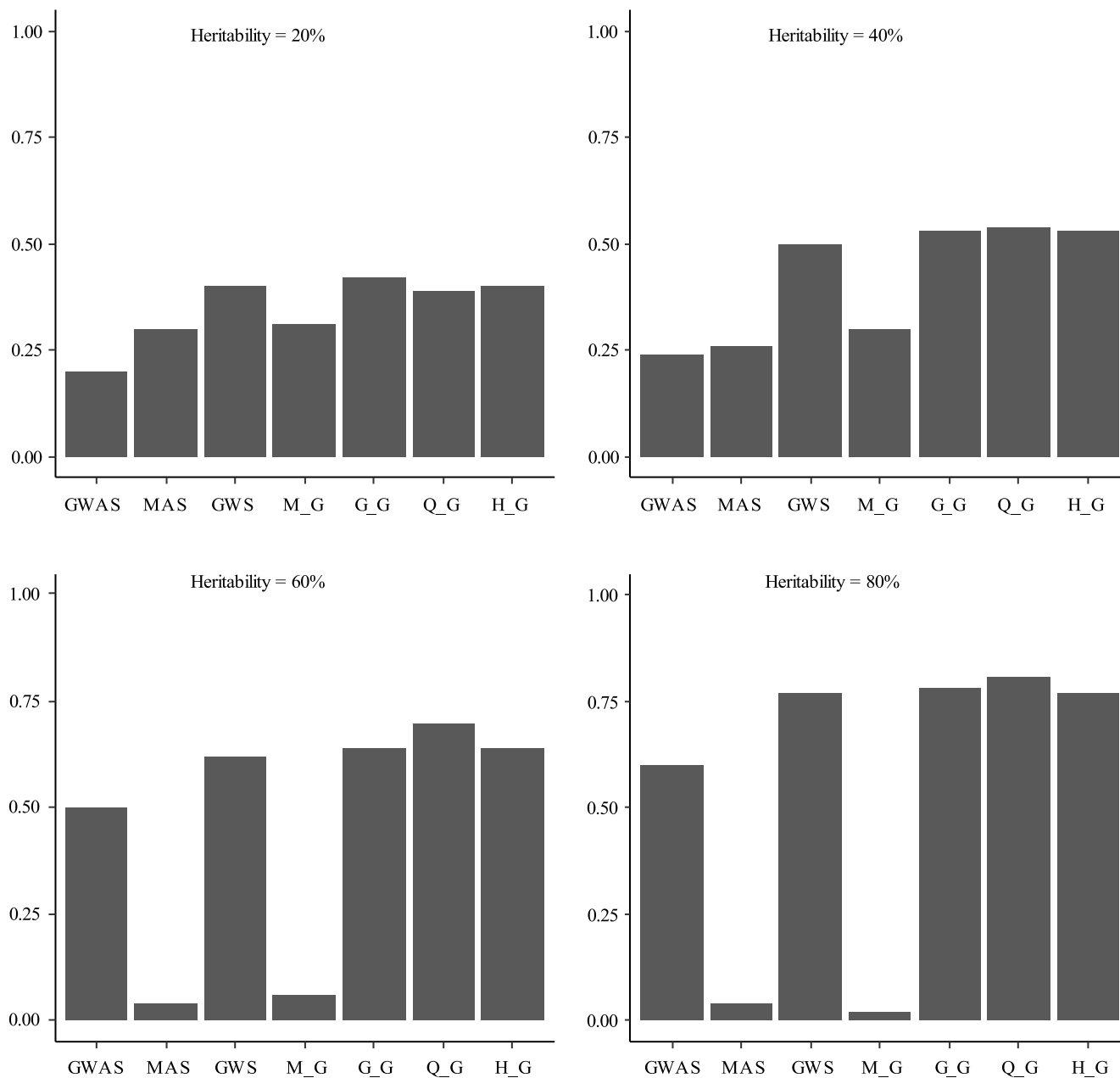


Figure 1. Comparison of the models using phenotypic predictive ability for characteristics with different heritabilities. Models: MAS, molecular marker-assisted selection; GWAS, genome-wide association studies; GWS, genome-wide selection; M_G, significant SNPs for each trait found by MAS were modeled as fixed effect, and the remaining SNPs were modeled as random effect; G_G, significant SNPs for each trait found by GWAS were modeled as fixed effect, and the remaining SNPs, as random effect; H_G, the two QTL with the greatest effect simulated for each trait were modeled as fixed effect, and the remaining SNPs, as random effect; Q_G, all simulated QTL for each trait were modeled as fixed effect, and the remaining SNPs, as random effect.

all SNPs were used as random effect (Figures 1, 2, and 3).

A single gene treated as a fixed effect in the GWS using RRBLUP will never be disadvantageous, except

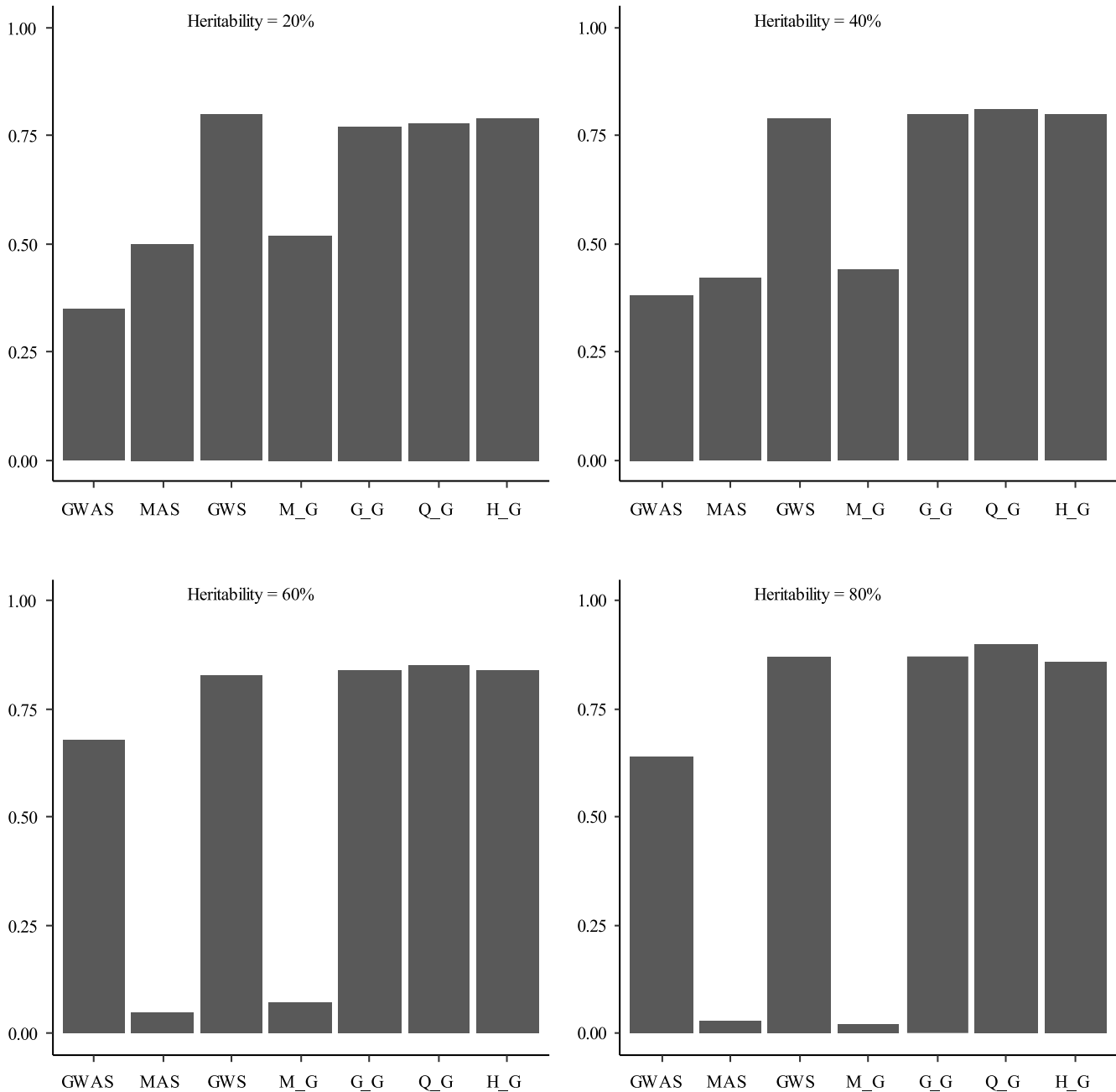


Figure 2. Comparison of the models using genotypic predictive ability for characteristics with different heritabilities. MAS, molecular marker-assisted selection; GWAS, genome-wide association studies; GWS, genome-wide selection; M_G, significant SNPs for each trait found by MAS were modeled as fixed effect, and the remaining SNPs were modeled as random effect; G_G, significant SNPs for each trait found by GWAS were modeled as fixed effect, and the remaining SNPs, as random effect; H_G, the two QTL with the greatest effect simulated for each trait were modeled as fixed effect, and the remaining SNPs, as random effect; Q_G, all simulated QTL for each trait were modeled as fixed effect, and the remaining SNPs, as random effect.

for some cases in which the variability explained by QTL is less than 10% (Bernardo, 2014). Thus, the use of significant QTL found by GWAS and MAS

methods is more influential in the prediction as greater is the effect of these QTL (Spindel et al., 2015). In the H_G model, the two QTL used as fixed effect showed

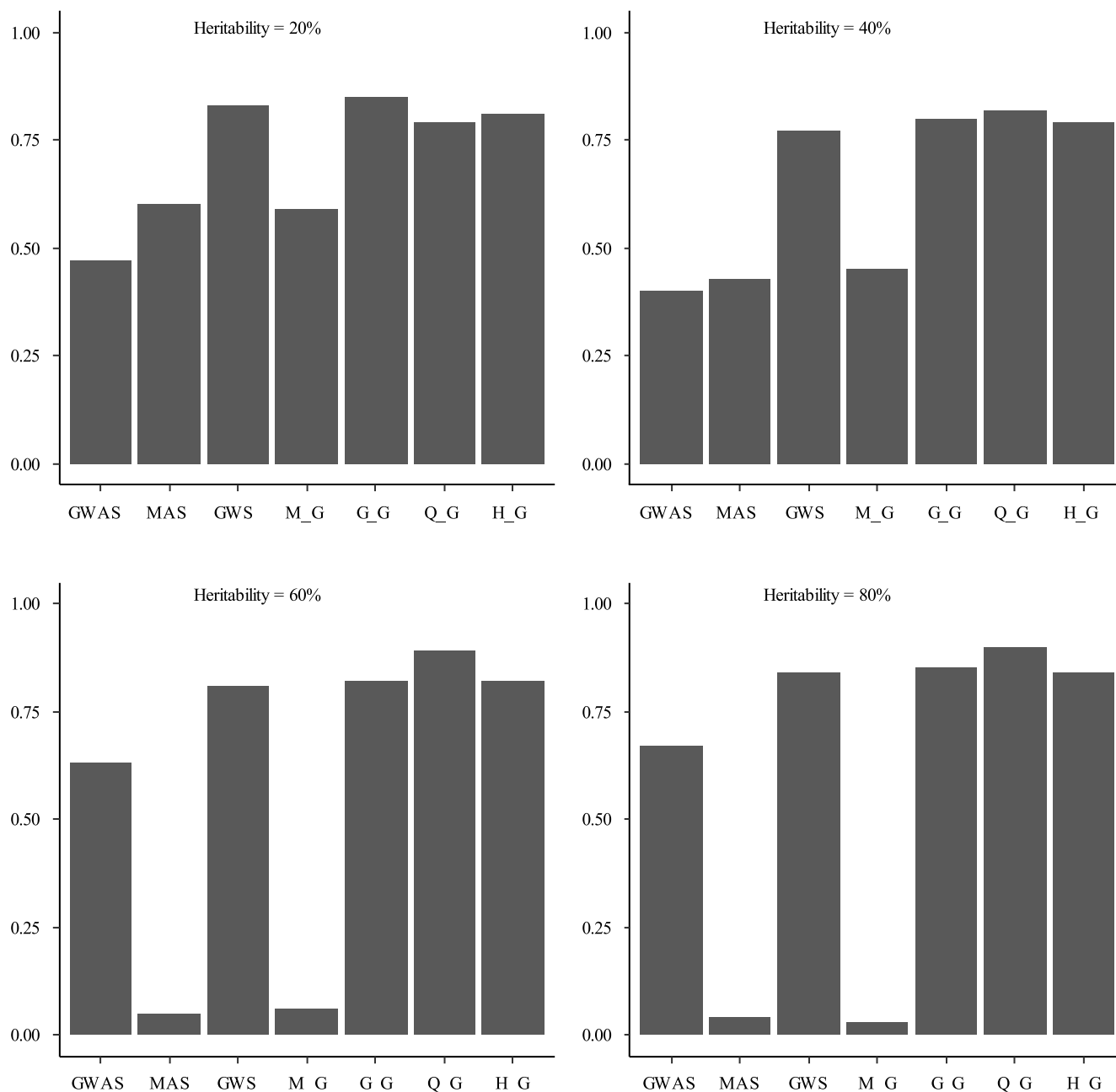


Figure 3. Comparison of the models via phenotypic accuracy for characteristics with different heritabilities. MAS, molecular marker-assisted selection; GWAS, genome-wide association studies; GWS, genome-wide selection; M_G, significant SNPs for each trait found by MAS were modeled as fixed effect, and the remaining SNPs were modeled as random effect; G_G, significant SNPs for each trait found by GWAS were modeled as fixed effect, and the remaining SNPs, as random effect; H_G, the two QTL with the greatest effect simulated for each trait were modeled as fixed effect, and the remaining SNPs, as random effect; Q_G, all simulated QTL for each trait were modeled as fixed effect, and the remaining SNPs, as random effect.

7.96% effect on the traits; this result show that the use of only two QTL with a greater effect on the trait can estimate values of prediction and accuracy similar to those of the models using a larger number of markers as fixed effect, as seen in the Q_G, G_G and M_G models (Figures 1, 2, and 3). This fact is important since, for many characteristics, some QTL of great effect are already known and, this way, these QTL can be introduced as a fixed effect in the GWS models. For

instance, in soybean, 18 QTL have been identified for tolerance to aluminum (Sharma et al., 2011), as well as one QTL for Asian soybean rust (Kim et al., 2012), and four QTL for drought tolerance (Carpentieri-Pipolo et al., 2012). Thus, all these previously identified QTL can be used as a fixed effect in the GWS model.

The model GWAS showed the greatest coincidence index for characteristics of low heritability (20%) (Figure 4). However, for characteristics with 60% and

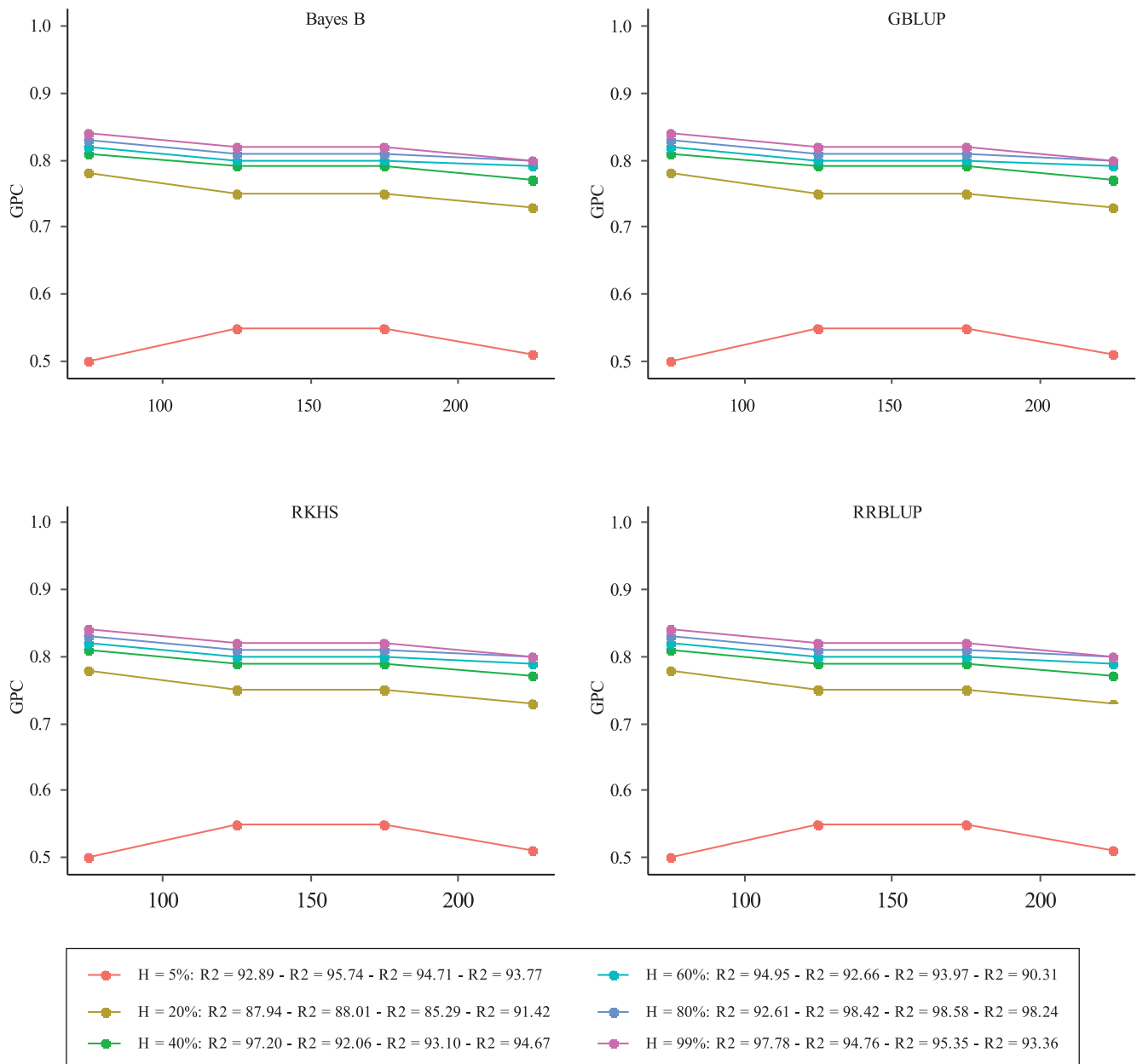


Figure 4. Comparison of the models using coincidence index for characteristics with different heritabilities (H).

80% heritabilities, the GWS, G_G, Q_G, and H_G models displayed higher coincidence indices. As the heritability of the characteristic increased, the coincidence index also increased, for all evaluated models, except for the MAS and M_G ones that remained practically constant (ranging from 20% to 40%).

Selection gain estimated by MAS and M_G models showed higher estimates for gain (Figure 5). However, the selection gain was overestimated by the MAS and M_G models for 60% and 80% heritabilities when compared to the maximum selection gain (Table 2), which was greater in the lower the heritability of the characteristic.

The models GWAS and MAS were observed as faster, considering processing time, than the others for all evaluated characteristics (Figure 6). However, as heritability increased, the processing time in the MAS model also increased, making it more time-consuming than the GWAS model. The other models had a very similar processing time, except for M_G that was the most time-consuming model.

The large number of markers considered as fixed effect in the model due to false positives also affected the selection gain (Figure 5 and Table 1), which was overestimated in the MAS and M_G models. However, GWAS and G_G models, in which the number of markers considered as fixed effect did not exceed 32, the genetic gain was not overestimated (Figure 5 and Table 2), as well as in the Q_G and H_G models. The coincidence index was another parameter affected by the number of false negative QTL as fixed effect in the model. For the MAS and M_G models, the coincidence index was lower than those of the other models and decreased as the heritability of the characteristic increased, since the number of false positives also increased (Figure 4). Therefore, knowing the genetic architecture of the trait under study can be important in the application of the correct GWS model and, consequently, to increase accuracy.

Considering the genetic architecture of the traits through the GWAS and MAS analyses can greatly improve the accuracy of the GWS models, since the higher effect of the qtl will be treated irrespective of the markers with smaller or no effect. Using simulated data, Bernardo (2014) verified that in traits with moderate to high heritability, QTL with effect greater than 30% as a fixed effect in the GWS model can increase the relative efficiency based on the selection

gain from 7% to 21%. However, when QTL with less than 5% effect was used in the model, the relative efficiency decreased, showing that markers that explain a small fraction of the genetic variance should be treated as random effect in the GWS model. As in present work, almost all genes that were considered as fixed effect explained a very small fraction of the genetic variance (less than 5%), no significant increase of predictive ability, accuracy, and gain with selection was observed and, in many cases, the value of these estimates decreased (Figure 1, 2, 3, and 5). Thus, the genetic architecture difference between the different species, as well as the genetic architecture difference between the characteristics of economic importance within the main species will influence the accuracy of the GWS models (Spindel et al., 2015). This fact is important for most major agricultural crops. In a study with GWAS analysis of maize, McMullen et al. (2009) found innumerable genes of lower effect controlling the main agricultural characteristics of this species. In rice, many large effect QTL have been found by GWAS and MAS (Chen et al., 2014).

For the characteristics governed by a smaller number of genes with higher effect, MAS and M_G models can be superior to the other models, since they can capture a large part of the total genetic variance present in only a few QTL (Arruda et al., 2016). Spindel et al. (2015) verified that MAS was superior to GWS for flowering time in rice. This trait is governed by few genes with large effect, whereas GWS was superior to MAS for grain yield, characteristic that is governed by a large number of genes of small effect. Arruda et al. (2016) compared the MAS, GWS, and M_G models in six characteristics associated with resistance to *fusarium* in wheat, and found that GWS had a predictive accuracy (0.4–0.9) higher than that of the MAS model (<0.3); however, when they used the QTL found in the MAS as a fixed effect in the GWS model (M_G model), the prediction accuracy was higher than that of the GWS model. Therefore, the performance of each method depends very much on the characteristic to be analyzed and its genetic structure, thus, a deep knowledge is necessary on the characteristic in study, to choose the most appropriate model, in order to increase its predictive ability and prediction accuracy.

An important feature of Bayesian ridge regression (BRR), the standard model used in the present study,

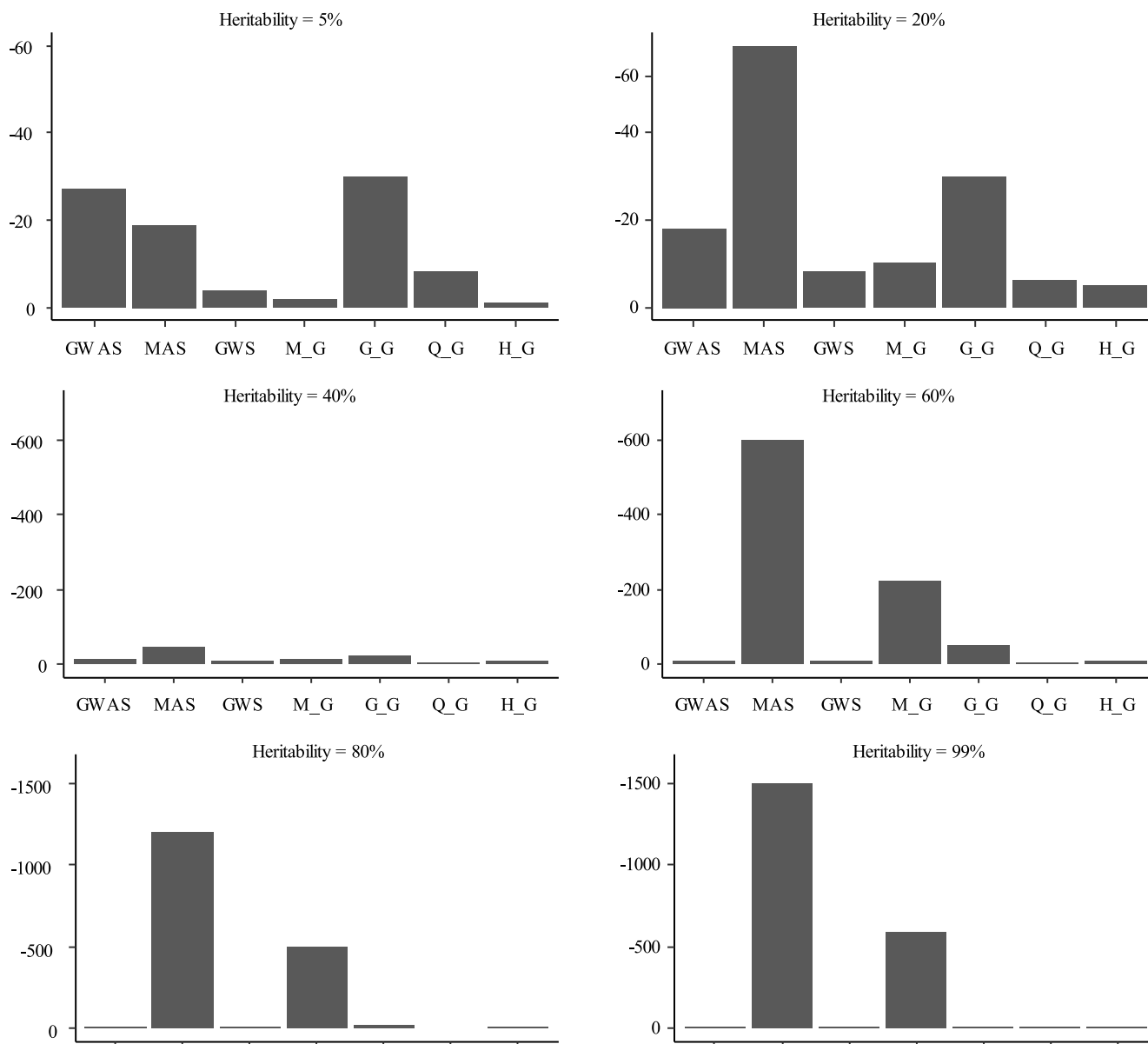


Figure 5. Comparison of the models using selection gain for characteristics with different heritabilities. MAS, molecular marker-assisted selection; GWAS, genome-wide association studies; GWS, genome-wide selection; M_G, significant SNPs for each trait found by MAS were modeled as fixed effect, and the remaining SNPs were modeled as random effect; G_G, significant SNPs for each trait found by GWAS were modeled as fixed effect, and the remaining SNPs, as random effect; H_G, the two QTL with the greatest effect simulated for each trait were modeled as fixed effect, and the remaining SNPs, as random effect; Q_G, all simulated QTL for each trait were modeled as fixed effect, and the remaining SNPs, as random effect.

Table 2. Maximum selection gain for the characteristics evaluated with different heritabilities.

Parameter	h ² = 20%	h ² = 40%	h ² = 60%	h ² = 80%
\bar{X}_o	151.27	147.71	151.05	153.38
X_s	77.88	101.43	115.42	124.41
GS _{max} (%)	48.52	31.33	23.59	18.89

Parameter: h², heritability; \bar{X}_o , genotypic mean of the population; X_s , genotypic mean of the selected individuals; GS_{max}, maximum selection gain.

is that all marks have the same genetic variance. However, this is practically impossible to happen in the characteristics of agronomic importance. Thus, markers that have an effect on the trait may be underestimated, and markers that have no effect may be being overestimated. Considering this context, placing the high-effect QTL indicated by MAS and

GWAS, as a fixed effect in the GWS model, ensures that these QTL are estimated more realistically, which consequently increases the predictive ability

and accuracy of the model (Spindel et al. 2015). This underestimation of the QTL with large effect may affect the selection response for several cycles in

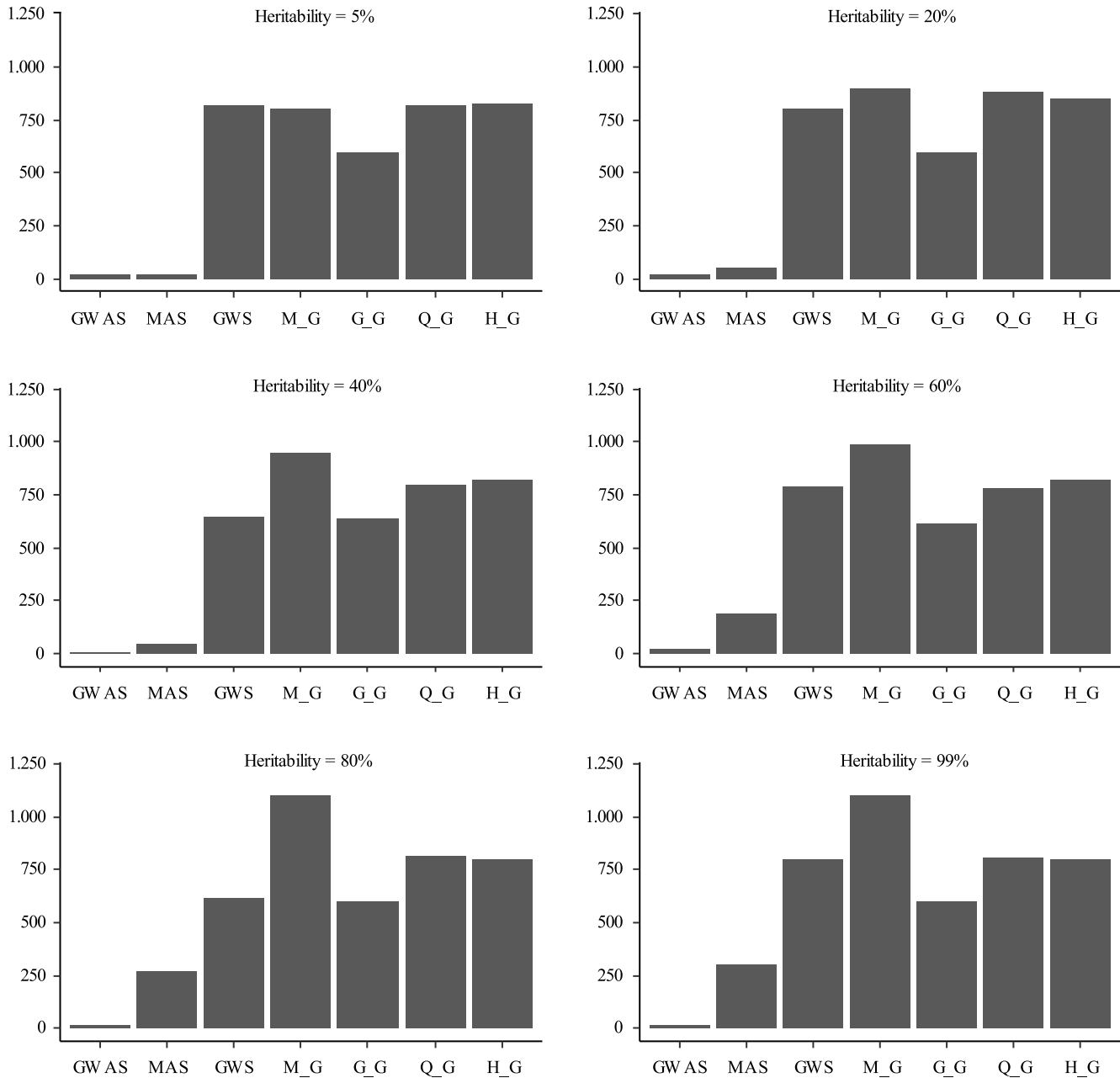


Figure 6. Comparison of models using processing time, in seconds, for characteristics with different heritabilities. MAS, molecular marker-assisted selection; GWAS, genome-wide association studies; GWS, genome-wide selection; M_G, significant SNPs for each trait found by MAS were modeled as fixed effect, and the remaining SNPs were modeled as random effect; G_G, significant SNPs for each trait found by GWAS were modeled as fixed effect, and the remaining SNPs, as random effect; H_G, the two QTL with the greatest effect simulated for each trait were modeled as fixed effect, and the remaining SNPs, as random effect; Q_G, all simulated QTL for each trait were modeled as fixed effect, and the remaining SNPs, as random effect.

the breeding program (Combs & Bernardo, 2013). The prediction accuracy increased, when QTL were treated as a fixed effect for rust resistance in wheat, according to Rutkoski et al. (2014). In a comparison between the GWS and MAS models, for 13 agronomic traits in wheat, Heffner et al. (2011) verified that the phenotypic and genotypic predictive abilities were 28% higher in the GWS. However, some studies show that, depending on the trait, an increase of accuracy can not occur when fixed effects are placed in the model, observed by Rutkoski et al. (2014) in a report on wheat resistance to *fusarium*. Besides, the authors verified that the accuracy in MAS was superior to that of GWS. Zhao et al. (2014) compared GS and MAS for plant height in wheat and verified that the predictive ability was the same.

The use of a genome-wide selection model with the significant markers found by the GWAS as a fixed effect, and the other markers as random effect, is a good strategy to select superior individuals with high accuracy in F_2 populations. Moreover, the introduction of QTL previously described as a fixed effect in the selection model, for the characteristic under study, allows of the selection of superior individuals more accurately. The results of the present study allow to choose the best GWS models to be applied more accurately and precisely in F_2 population in breeding.

Conclusions

1. The introduction of QTL already described for a given trait into the genome-wide selection model allows of the selection of superior individuals with greater precision.

2. The use of a genome-wide selection model with the significant markers found by the GWAS as a fixed effect, and the other markers, as random effect, is a good strategy to select superior individuals with high accuracy in the F_2 populations.

References

ARRUDA, M.P.; LIPKA, A.E.; BROWN, P.J.; KRILL, A.M.; THURBER, C.; BROWN-GUEDIRA, G.; DONG, Y.; FORESMAN, B.J.; KOLB, F.L. Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). **Molecular Breeding**, v.36, art.84, 2016. DOI: <https://doi.org/10.1007/s11032-016-0508-5>.

BERNARDO, R. Genomewide selection when major genes are known. **Crop Science**, v.54, p.68-75, 2014. DOI: <https://doi.org/10.2135/cropsci2013.05.0315>.

CARPENTIERI-PIPOLO, V.; PIPLOLO, A.E.; ABDEL-HALEEM, H.; BOERMA, H.R.; SINCLAIR, T.R. Identification of QTLs associated with limited leaf hydraulic conductance in soybean. **Euphytica**, v.186, p.679-686, 2012. DOI: <https://doi.org/10.1007/s10681-011-0535-6>.

CHEN, J.; LI, X.; CHENG, C.; WANG, Y.; QIN, M.; ZHU, H.; ZENG, R.; FU, X.; LIU, Z.; ZHANG, G. Characterization of epistatic interaction of QTLs LH8 and EH3 controlling heading date in rice. **Scientific Reports**, v.4, art.4263, 2014. DOI: <https://doi.org/10.1038/srep04263>.

COMBS, E.; BERNARDO, R. Genomewide selection to introgress semidwarf maize germplasm into US Corn Belt inbreds. **Crop Science**, v.53, p.1427-1436, 2013. DOI: <https://doi.org/10.2135/cropsci2012.11.0666>.

CRUZ, C.D. Genes: a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum. Agronomy**, v.35, p.271-276, 2013. DOI: <https://doi.org/10.4025/actasciagron.v35i3.21251>.

GREGORIO, G.B.; ISLAM, M.R.; VERGARA, G.V.; THIRUMENI, S. Recent advances in rice science to design salinity and other abiotic stress tolerant rice varieties. **SABRAO Journal of Breeding Genetics**, v.45, p.31-41, 2013.

HEFFNER, E.L.; JANNINK, J.-L.; IWATA, H.; SOUZA, E.; SORRELLS, M.E. Genomic selection accuracy for grain quality traits in biparental wheat populations. **Crop Science**, v.51, p.2597-2606, 2011. DOI: <https://doi.org/10.2135/cropsci2011.05.0253>.

KIM, K.-S.; UNFRIED, J.R.; HYTEN, D.L.; FREDERICK, R.D.; HARTMAN, G.L.; NELSON, R.L.; SONG, Q.; DIERS, B.W. Molecular mapping of soybean rust resistance in soybean accession PI 561356 and SNP haplotype analysis of the *Rpp1* region in diverse germplasm. **Theoretical and Applied Genetics**, v.125, p.1339-1352, 2012. DOI: <https://doi.org/10.1007/s00122-012-1932-5>.

LANDER, E.S.; BOTSTEIN, D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics**, v.121, p.185-199, 1989. DOI: <https://doi.org/10.1093/genetics/121.1.185>.

MCMULLEN, M.D.; KRESOVICH, S.; VILLEDA, H.S.; BRADBURY, P.; LI, H.; SUN, Q.; FLINT-GARCIA, S.; THORNSBERRY, J.; ACHARYA, C.; BOTTOMS, C.; BROWN, P.; BROWNE, C.; ELLER, M.; GUILL, K.; HARJES, C.; KROON, D.; LEPAK, N.; MITCHELL, S.E.; PETERSON, B.; PRESSOIR, G.; ROMERO, S.; OROPEZA ROSAS, M.; SALVO, S.; YATES, H.; HANSON, M.; JONES, E.; SMITH, S.; GLAUBITZ, J.C.; GOODMAN, M.; WARE, D.; HOLLAND, J.B.; BUCKLER, E.S. Genetic properties of the maize nested association mapping population. **Science**, v.325, p.737-740, 2009. DOI: <https://doi.org/10.1126/science.1174320>.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, p.1819-1829, 2001. DOI: <https://doi.org/10.1093/genetics/157.4.1819>.

- PALAISSA, K.A.; MORGANTE, M.; WILLIAMS, M.; RAFALSKI, A. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. **The Plant Cell**, v.15, p.1795-1806, 2003. DOI: <https://doi.org/10.1105/tpc.012526>.
- PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v.198, p.483-495, 2014.
- PRITCHARD, J.K.; STEPHENS, M.; ROSENBERG, N.A.; DONNELLY, P. Association mapping in structured populations. **American Journal of Human Genetics**, v.67, p.170-181, 2000. DOI: <https://doi.org/10.1086/302959>.
- R CORE TEAM. **R**: a language and environment for statistical computing. 2015. Available at: <<http://www.R-project.org/>>. Accessed on: Oct. 13 2024.
- RUTKOSKI, J.E.; POLAND, J.A.; SINGH, R.P.; HUERTA-ESPINO, J.; BHAVANI, S.; BARBIER, H.; HOUSE, M.N.; JANNINK, J.-L.; SORRELLS, M.E. Genomic selection for quantitative adult plant stem rust resistance in wheat. **Plant Genome**, v.7, art.plantgenome2014.02.0006, 2014. DOI: <https://doi.org/10.3835/plantgenome2014.02.0006>.
- SHARMA, A.D.; SHARMA, H.; LIGHTFOOT, D.A. The genetic control of tolerance to aluminum toxicity in the 'Essex' by 'Forrest' recombinant inbred line population. **Theoretical and Applied Genetics**, v.122, p.687-694, 2011. DOI: <https://doi.org/10.1007/s00122-010-1478-3>.
- SPINDEL, J.; BEGUM, H.; AKDEMIR, D.; VIRK, P.; COLLARD, B.; REDOÑA, E.; ATLIN, G.; JANNINK, J.-L.; MCCOUCH, S.R. Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **PLoS Genetics**, v.11, e1004982, 2015. DOI: <https://doi.org/10.1371/journal.pgen.1004982>.
- ZHAO, Y.; METTE, M.F.; GOWDA, M.; LONGIN, C.F.H.; REIF, J.C. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. **Heredity**, v.112, p.638-645, 2014. DOI: <https://doi.org/10.1038/hdy.2014.1>.
-