DOI: http://doi.org/10.1590/1678-992X-2020-0202

ISSN 1678-992X

SCIENTIA AGRICOLA

Research Article

Evaluation of Bayesian methods of genomic association via chromosomic regions using simulated data

Leísa Pires Lima[1]*, Camila Ferreira Azevedo[1], Marcos Deon Vilela de Resende[2], Moysés Nascimento[1], Fabyano Fonseca e Silva[3]

[1]Universidade Federal de Viçosa – Depto. de Estatística, Av. Peter Henry Rolfs, s/n – 36570–000 – Viçosa, MG – Brasil.
[2]Embrapa Café, Parque Estação Biológica (PqEB), Av. W3 Norte (Final) – 70770-901 – Brasília, DF – Brasil.
[3]Universidade Federal de Viçosa – Depto. de Zootecnia – Viçosa, MG – Brasil.
*Corresponding author <leisa.lima@ufv.br>

Edited by: Marcin Kozak

Received July 06, 2020
Accepted February 18, 2021

ABSTRACT: The development of efficient methods for genome–wide association studies (GWAS) between quantitative trait loci (QTL) and genetic values is extremely important to animal and plant breeding programs. Bayesian approaches that aim to select regions of single nucleotide polymorphisms (SNPs) proved to be efficient, indicating genes with important effects. Among the selection criteria for SNPs or regions, selection criterion by percentage of variance can be explained by genomic regions (%var), selection of tag SNPs, and selection based on the window posterior probability of association (WPPA). To also detect potentially associated regions, we proposed measuring posterior probability of the interval (PP_{int}), which aims to select regions based on the markers of greatest effects. Therefore, the objective of this work was to evaluate these approaches, in terms of efficiency in selecting and identifying markers or regions located within or close to genes associated with traits. This study also aimed to compare these methodologies with single–marker analyses. To accomplish this, simulated data were used in six scenarios, with SNPs allocated in non–overlapping genomic regions. Considering traits with oligogenic inheritance, WPPA criterion followed by %var and PP_{int} criteria were shown to be superior, presenting higher values of detection power, capturing higher percentages of genetic variance and larger areas. For traits with polygenic inheritance, PP_{int} and WPPA criteria were considered superior. Single–marker analyses identified SNPs associated only in oligogenic inheritance scenarios and was lower than the other criteria.

Keywords: genomic regions, molecular markers, genetic variance

Introduction

The development of new sequencing and genotyping technologies has promoted the growth of molecular genetics, enabling breeding programs to carry out genome–wide association studies (GWAS) between quantitative trait loci (QTL) and genetic values of individuals. The selection of marker groups has been identified as a major advantage for GWAS because these groups tend to capture a higher proportion of the genetic variance, identifying more complex relationships between markers (Moore et al., 2010). According to Fan et al. (2011) and Fernando et al. (2017), to select the associated regions' methods using the Bayesian approach is preferable as it offers significant advantages, such as the possibility of incorporating a prior knowledge and simultaneous estimation of marker effects.

Based on this, criteria using the Bayesian approach to select markers and associated regions that do not require major computational efforts are proposed. Among the existing criteria, selection by the percentage of variance explained by genomic regions (%var), selection criteria of tag single nucleotide polymorphisms (tag SNPs), and selection based on window posterior probability of association (WPPA) are the most notable. These approaches consider the genetic variance and differ in the criteria used for selecting the regions and thresholds determined for selection.

Further, to detect potentially associated regions and select SNPs with greater effects, we propose measuring the posterior probability of the interval (PP_{int}). The selection of these SNPs becomes viable, since, biologically, it is expected that S close to a QTL will have a greater effect because they are both close to the causal mutation (Habier et al., 2011; Meuwissen et al., 2016). Thus, PP_{int} is based on the number of iterations of the Markov Chain Monte Carlo (MCMC), in which the region has at least one SNP with an effect magnitude greater than the value of the third quartile, considering the entire distribution of the absolute effects in that iteration.

Given the above, this paper aimed to propose the measure of PP_{int} and compare it to the tag SNP, %var, and WPPA approaches to determine the efficiency in selecting and identifying markers or regions that are located within or near genes associated with traits of interest. This study also aimed to compare the results obtained by these methodologies with single–marker analyses. To achieve this objective, we used simulated data that considers six different scenarios, with SNPs allocated in non–overlapping genomic regions.

Materials and Methods

Simulated Data

The data set was simulated using the Genes software program (Cruz, 2013). The genome consisted of ten linking

groups, with 20 centimorgans (cM) and 200 markers in each group. The SNPs were considered to be distributed approximately equidistant in the genome. In the analysis of gene linkage, an F1 generation was simulated, in which one parent was a dominant homozygote and the other, a recessive homozygote. From the genotypes of the F1 population, a population of F2 mapped with 1,000 individuals was generated, considering 5,000 gametes, attributing the entire linkage disequilibrium (LD) to the linkage group. Quantitative traits were simulated considering a zero degree of dominance, mean equal to 100, and heritability levels which were selected to represent traits with high (h_a^2 = 0.50 and h_a^2 = 0.60), moderate (h_a^2 = 0.30 and h_a^2 = 0.40), and low (h_a^2 = 0.10 and h_a^2 = 0.20) heritabilities. Three genetic architectures were generated using three, ten, and one hundred loci controlling the trait, which explained equal parts of the genetic variance, these QTL being distributed in the regions covered by the markers. In the first architecture, a case was considered in which three QTL were randomly distributed among the ten chromosomes. In the second, ten controlling loci of the trait were assumed, in which one QTL was assigned to each of the ten chromosomes. In the third, traits controlled by many genes with small effects were considered, in which ten QTL were distributed in each of the ten chromosomes, totaling 100 QTL. Additionally, according to Goddard et al. (2011), the proportion of genetic variation associated with the QTL explained by the markers (r_{mq}^2) was ascertained by:

$$r_{mq}^2 = \frac{n}{n + n_{QTL}} \quad (1)$$

where n was the number of SNPs and n_{QTL}, the number of QTL.

In this way, six different scenarios were used in the analyses: three genetic architectures × two different levels of heritability in each architecture. The description of the scenarios is presented in Table 1. Each type of scenario was simulated ten times to assess the efficiency of the methods, according to Lima et al. (2019). Thus, the measures used were calculated in each repetition of the simulation and thereafter the mean and standard error of these values were obtained.

BayesDπ

The BayesDπ method allowed a (1 – π) percentage of marker effects to be equal to zero, leading to a lower number of marker effects to be estimated, raising the accuracy of the estimation process since many of the markers do not have genetic effects or are not in LD with QTL (Habier et al., 2011). Consider the following linear model proposed by Meuwissen et al. (2001):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Wm} + \boldsymbol{e} \quad (2)$$

where $\boldsymbol{y}$ was the vector of phenotypes ($N \times 1$, N, the number of individuals), μ was the general mean of the trait, $\mathbf{1}$, a vector of the same dimension of $\boldsymbol{y}$ with all elements equal to 1, $\boldsymbol{m}$, the vector of additive genetic effects of the markers ($n \times 1$, n, the number of markers), $\boldsymbol{W}$ ($N \times n$), the additive incidence matrix and $\boldsymbol{e}$ ($N \times 1$), the vector of errors of the model with $e \sim N(0, \sigma_e^2)$ with σ_e^2 being the error variance. The $\boldsymbol{W}$ matrix was coded according to Vitezica et al., 2013.

In this method it was considered that a fraction (1 – π) of the markers had no effect on the trait and the remaining fraction π had effects with a prior distribution given by a normal with a specific variance $\sigma_{m_j}^2$ for each marker being the variance of each marker from a scaled inverse chi-square distribution with υ degrees of freedom and scale parameter S_m^2. Thus, the equation used was:

$$m_j \left| \pi \sim \pi N\left(0, \sigma_{m_j}^2\right) + (1-\pi) N\left(0, \sigma_{m_j}^2 = 0\right) \right. \quad (3)$$

$$\sigma_{m_j}^2 \left| \pi \sim Scale - inv - \chi^2\left(\upsilon, S_m^2\right) \right. \quad (4)$$

where $\pi \left| m_j, \sigma_{m_j}^2, \sigma_e^2, S_m^2 \sim U[0,1] \right.$ and $S_m^2 \sim Gamma(\alpha, \beta)$ being α and β, respectively, as the hyperparameters of the a prior distribution. Note that this method allowed the specification of a prior distribution for the π probability and S_m^2 hyperparameter, considering them unknown parameters of the model for limiting the influence of subjectivity in the selection of markers and the shrinkage factor. The a posterior mean of the total genetic variance of the markers was given by $\hat{\sigma}_g^2 = \sum_j 2p_j q_j \hat{\sigma}_{m_j}^2$, where p_j and q_j were the allele frequencies associated with "A" and "a" the alleles, respectively, of the j-th marker.

BayesDπ can be advantageously used in GWAS in relation to other Bayesian approaches because it assumes a specific variance for each locus to obtain information about the genetic architecture of the trait and estimates the value of probability π, which is considered unknown

Table 1 – Description of scenarios with the proportion of quantitative trait loci (QTL) variation explained by the SNPs (r_{mq}^2), genetic architecture, number of QTL and narrow-sense heritability (h^2).

Scenarios	r_{mq}^2	Genetic architecture	Number of QTL	h^2
Scenario 1	0.99	3 QTL on 10 chromosomes	3	0.50
Scenario 2		3 QTL on 10 chromosomes	3	0.60
Scenario 3	0.99	1 QTL in each of the 10 chromosomes	10	0.30
Scenario 4		1 QTL in each of the 10 chromosomes	10	0.40
Scenario 5	0.95	10 QTL on each of the 10 chromosomes	100	0.10
Scenario 6		10 QTL on each of the 10 chromosomes	100	0.20

and considers that most of the markers have small effects (or zero effect), except for those closer to the causal mutation that would have more influential effects, and bring more biological meaning to the analyses (Habier et al., 2011).

For inference about the posterior distribution of the estimated effects of SNPs, 320,000 iterations were used for the MCMC algorithms, of which 20,000 were discarded (burn-in) to guarantee the heating of the chain and selection of one in ten iterations (thin). Convergence analysis was performed using the criterion proposed by Geweke (1992).

Formation of Regions

On each chromosome, SNPs were allocated to non-overlapping genomic regions of defined sizes, according to the mean LD between the markers and the QTL itself. Linkage disequilibrium values vary between zero and one, referring to the absence or complete LD, respectively. According to Zhao et al. (2007) and Viana et al. (2016), the effectiveness of locating QTL using LD between markers and QTL depends on the extent of LD and how it decreases according to the distance between markers and QTL in a population. In this study, two extensions of LD with values between 0.64 and 0.81 were considered thresholds when determining the size of the regions. These values were chosen because they represent a high correlation of 0.80 and 0.90, respectively, between the QTL and markers. The shortest distance that LD provided between the QTL and marker following these two established LD extensions was used to define the size of the region in the scenarios. Thus, for each scenario, two sizes of region were used according to the extent of LD considered. For the LD extension of 0.64, sizes of 4 cM, 5 cM, and 2.7 cM were established for scenarios 1 and 2, 3 and 4, and, 5 and 6, respectively. For LD of 0.81, sizes found for the respective scenarios were 2 cM, 1.5 cM, and 1 cM. The six scenarios were analyzed considering these region sizes and the five approaches to selecting SNPs/regions, single-marker analyses, %var, tag SNP, WPPA, and PP_{int}.

Comparison of methodologies

To verify the efficiency of the analyzed criteria, the following measures described below were calculated:

i) False positive (FP) consisted of declaring a marker/region as associated, when in fact this marker/region was not in LD with the QTL and was defined by the ratio between the number of SNPs/regions considered associated and that have no effect on the trait, and, equally, the number of SNPs/regions that have no effect on the trait.

ii) Detection power (PD) consisted of declaring a marker/region effect associated when this marker/region was actually in LD with the QTL, and was defined as the ratio between the number of SNPs/regions considered associated and that affect the trait and, equally, the number of SNPs/regions that affect the trait.

iii) Percentage of genetic variance recovered was based on the percentage of genetic variance captured by the SNPs/regions and was obtained by the ratio between the genetic variance of the SNPs/regions considered associated and a posterior mean of total genetic variance. According to Peters et al. (2012), genomic regions that contribute with greater genetic variances were considered those most associated with the trait of interest.

iv) Area under the curve obtained between false positive rates and detection power was calculated using the receiver operating characteristic curve (ROC) proposed by Metz (1978) to also compare the criteria. Previous studies (Gage et al., 2018; Liu et al., 2016) used ROC curves or similar visual aids to assess the effectiveness of different methods in GWAS. In an ROC curve, the detection power values are plotted against the false positive rate and, thus, the criterion that provides the highest area value below the curve is considered superior. The use of the area to compare the results in a single statistic allows for direct comparison of the results of the GWAS of traits with different simulation parameters (Gage et al., 2018).

Consequently, the methodology that presents lower rates of false positives, greater detection power, that captures a greater proportion of the genetic variance, and that has a larger area under the ROC curve was considered the most suitable for GWAS.

Selection Criteria for Regions or SNPs

Selection by proportion of genetic variance explained by genomic regions – %var

The selection of SNP groups using the proportion of genetic variance explained by genomic regions (%var), was initially proposed by Wang et al. (2014) as an efficiency measure for comparing methods of selecting regions in GWAS. For the effects of estimated SNPs, the genetic variance associated with the k-th region was calculated using:

$$\hat{\sigma}^2_{gk} = \sum_{j \in k} 2p_j q_j \hat{m}^2_j \quad (5)$$

where $\hat{m}_j$ was the allelic substitution effect of the j-th SNP belonging to the k-th region as estimated by BayesDπ. The explanation percentage of the genetic variance for each region was obtained as follows:

$$\%var = \frac{\hat{\sigma}^2_{gk}}{\hat{\sigma}^2_g} \times 100 \quad (6)$$

with $\hat{\sigma}^2_g$ being the posterior mean of the total genetic variance considering all markers.

The regions that presented proportion values of the genetic variance higher than the ratio between the posterior mean of the total genetic variance and the number of regions considered were selected as associated regions and subsequently used to explore and determine the possible QTL. A grid of values ranging from 100 % to 200 % of the ratio between the posterior mean of the total genetic variation and the number of regions was considered and the one that returned the least difference between the detection power and confidence level (an associated region was not declared when this region was not actually in LD with QTL) was considered in the results.

Selection of tag SNPs using Bayesian methods

This approach consisted of identifying associated regions in the genome and subsequently, selecting SNPs in those regions that supposedly had high LD with the QTL. In this method, the SNPs were allocated in genomic regions but not overlapping and thus, the regions that potentially contained QTL associated with the trait of interest were called top windows. Top windows were identified according to a threshold defined in terms of the contribution of the genetic variance of the markers that could be obtained through the estimated effects of all SNPs via BayesDπ. Sollero et al (2017) considered all regions that explained proportions of genetic variance greater than five times the threshold described by Schurink et al. (2012) as top windows, which is obtained through the quotient between the a posterior mean of the total genetic variance by the number of regions considered. In this criterion, the same grid of values of the percentage of the posterior mean of the total genetic variance used in the %var criterion was considered, and, thus, the regions that presented genetic variances above this threshold were considered top windows. Again, the percentage used and considered in the analyses were those that provided the least difference between the power to detect an associated region and the level of confidence.

The effects of SNPs considered associated within each top window were called tag SNPs and were selected using the inclusion frequencies (IF) of the SNPs in the model, which is, the ratio between the number of saved iterations of the MCMC that include the SNP in question in the model and the total number of iterations saved. Thus, for each top window, the IF values were calculated for all SNPs and the tag SNPs selected were those SNPs that had the highest IF value. To also select the tag SNPs within each top window, *t*-like statistics (TL) were used, considering an approach similar to IF to assess the consistency of the SNPs effects. This measure is given by the absolute value of the posterior mean effects of the markers (only for the chains that included the SNP in the model) divided by the respective standard deviations of these effects, an approximation of Student's t-statistic. Based on this statistic, SNPs considered significant ($p < 0.05$) within each top window were also identified as tag SNPs.

Selection by the window posterior probability of association – WPPA

The WPPA measure was implemented using the effects of SNPs estimated via BayesDπ and obtained from the proportion of the genetic variance explained by the markers of each genomic region. The genomic variance associated with the k-th region was estimated, in this context, by means of:

$$\hat{\sigma}^2_{g_k} = \sum_{j \in k} 2p_j q_j \hat{m}^2_j \quad (7)$$

where $\hat{m}_j$ was the allelic substitution effect of the j-th SNP belonging to the k-th region estimated by BayesDπ and p_j and q_j, the allele frequencies. From this, the proportion of the genetic variance explained by the markers in the k-th region, denoted by q_k, was defined as:

$$q_k = \frac{\hat{\sigma}^2_{g_k}}{E\left(\sigma^2_{g_k}\right)} \quad (8)$$

where $E(\sigma^2_{g_k}) = \sum_{j \in k} 2p_j q_j E(m^2_j)$ (in the absence of dominance), with

$$E(m^2_j) = \frac{(\sigma^2_g)}{n\bar{H}}$$

and σ^2_g, the genetic variance of the markers, n the number of SNPs, and $\bar{H}$ the mean of $2p_j q_j$. If $q_k > 1$, there was a causative mutation within the k-th region, since it had a greater than expected effect under the hypothesis of an equal distribution of genetic variance across the genome (Bennewitz et al., 2017; Peters et al., 2012). Thus, the WPPA measure was formulated from the ratio between the number of samples, where q_k was greater than one and the number of samples saved.

Regions that had a WPPA above a pre-established threshold were selected as associated regions. According to Fernando and Garrick (2013) and Fernando et al. (2017), if WPPA values greater than 0.95 are used to declare associated regions, this will result in a proportion of false positives below 0.05. Bennewitz et al. (2017) considered the levels 0.85, 0.95, and 0.99, and found that the power to detect an associated region decreased with the increase of these levels. In this study, several threshold levels ranging from 0.50 to 1.00, with an increment of 0.01, were tested in the analyses and the value that provided the least difference between the power to detect a region and the confidence level was considered.

Selection by a posterior probability of interval – *PP$_{int}$*

Biologically, it is expected that SNPs close to a QTL will have a greater effect being close to the causal mutation (Habier et al., 2011; Meuwissen et al., 2016) and for this reason it becomes viable to select these SNPs in the search for associations between markers and QTL through the PP_{int} measure that represents the probability SNPs with

great effects are included in the region. In addition, according to Resende et al. (2008), QTL can be located by adding the absolute effects of SNPs within each region and the regions with the largest sums of these absolute effects are likely to contain a QTL or be adjacent to a region containing a QTL and, thus, the position of the QTL can be found, and the discovery of QTL with great effect is facilitated. According to these authors, if there is no QTL in a given region, all estimates of the effects of SNPs within it will be small in magnitude. Thus, to also detect regions associated with the traits of interest, a new selection approach based on the effects of SNPs obtained in the MCMC samples was proposed and called a posterior probability of interval (PP_{int}). PP_{int} represented the probability of SNPs with large effects being included in the region and was calculated by the ratio between the number of iterations in which a given region had at least one SNP with an effect magnitude greater than the value of the third quartile, considering the entire distribution of the absolute effects in that iteration and the number of samples saved.

Regions with PP_{int} values greater than a pre-specified threshold were selected as associated. In this study, threshold values also varying from 0.50 to 1.00 with an increment of 0.01 were tested and, thus, the value that provided the least difference between the detection power and the confidence level was selected. This threshold was chosen by the researcher and directly reflects the posterior probability of a QTL being in the region.

Single-marker analyses

The mixed linear model of single-markers was used to estimate the effect of the j-th marker on the phenotype and was defined by:

$$\boldsymbol{y} = 1\mu + \boldsymbol{Z}_1\boldsymbol{g} + \boldsymbol{W}_j m_j + \boldsymbol{e} \quad (9)$$

where $\boldsymbol{y}$, μ, and $\boldsymbol{e}$ have been defined previously; $\boldsymbol{g}$ ($N \times 1$) was the vector of polygenic genetic effects with an incidence matrix $\mathbf{Z}_1$(N x N), being $\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$ where $\boldsymbol{G}$ was the additive genomic relationship matrix, σ_g^2, the polygenic variance, m_j the scalar, referring to the fixed effect of the j-th marker, and $\boldsymbol{W}_j$ the incidence vector of the j-th marker. The matrix of additive genomic relationship was given by (VanRaden, 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{W}\boldsymbol{W}'}{\sum_{j=1}^{n} 2p_j q_j} \quad (10)$$

To estimate the polygenic genetic effects and the effect of the j-th marker, the mixed model equations (Henderson, 1973) were given by:

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\boldsymbol{Z}_1 & \mathbf{1}'\boldsymbol{W}_j \\ \boldsymbol{Z}_1'\mathbf{1} & \boldsymbol{Z}_1'\boldsymbol{Z}_1 + \boldsymbol{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2} & \boldsymbol{Z}_1'\boldsymbol{W}_j \\ \boldsymbol{W}_j'\mathbf{1} & \boldsymbol{W}_j'\boldsymbol{Z}_1 & \boldsymbol{W}_j'\boldsymbol{W}_j \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{g}} \\ \hat{m}_j \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\boldsymbol{y} \\ \boldsymbol{Z}_1'\boldsymbol{y} \\ \boldsymbol{W}_j'\boldsymbol{y} \end{bmatrix} \quad (11)$$

where the components of variance, σ_g^2 and σ_e^2, were estimated via the restrict maximum likelihood (REML).

After estimating the effect of the j-th marker, the Wald test was performed to test the existence of an association between the marker and QTL. Thus, the null hypothesis (H_0) was defined as when "the j-th marker had no effect on the phenotype", and the alternative hypothesis (H_a) defined as "the j-th marker affected the phenotype", that is, the j-th marker and the QTL were found in LD. However, this statistical analysis suffers from the occurrence of a high rate of false positives due to the occurrence of multiple tests. An alternative for controlling this fact is to monitor the number of false positives in relation to the total number of positive results through the false discovery rate (FDR) as presented by Fernando et al. (2004). One way of considering FDR in the significance test is through correction in the p-value, called the q-value (Storey and Tibshirani, 2003).

The sizes of regions obtained based on the LD between the marker and QTL were used to define a region that determined the number of SNPs that really affected the trait. Thus, the SNPs that were distant to the QTL on the chromosome, lower than these thresholds, were considered associated SNPs and were used to calculate false positive rates, detection power, percentage of the variance, and area under the ROC curve.

Computational Resources

The entire implementation of the methods used was performed based on R software (R Development Core Team, 2019) through the GenomicLand visual interface (Azevedo et al., 2019). Convergence analysis of the effects of SNPs was estimated via the BayesDπ and was performed using the *coda* package (Plummer et al., 2006). In single-marker analysis, the *sommer* package was used (Covarrubias-Pazaran, 2016). The codes and data are available at https://www.licae.ufv.br/codes-for-association-analysis/.

Results and Discussion

The results, considering the LD extension of 0.64 for determining the sizes of the regions, are shown in Table 2. In this simulation study, to identify superior procedures for the selection of associated regions, we first selected the criteria that presented the highest value points for detection power. These criteria were considered preferable, since they disclosed the true proportion of regions that had been detected and were actually associated (Bennewitz et al., 2017). The results in Table 2 revealed that for scenarios with oligogenic genetic inheritance (traits controlled by a few genes with greater effects - scenarios 1, 2, 3, and 4), the WPPA criterion followed by the %var and PP_{int} criteria were higher than the tag SNP criterion, presenting higher and similar point values for the power to detect associated regions and, consequently, to capture higher

Table 2 – Size of the regions (distance) in centimorgans (cM) found through the linkage disequilibrium (LD) between the markers and quantitative trait loci (QTL) in each scenario using the LD extension of 0.64 and the means and standard errors of the estimated π probability via BayesDπ, false positive rates (FP), detection power (PD), percentage of genetic variance recovered (PE), area under the receiver operating characteristic curve (ROC) and threshold for selecting regions obtained by criteria %var, tag SNP, WPPA and PP_{int}.

Scenarios	π	Distance	Criterion	FP	Power	PE	Area	Threshold
1	0.32 ± 0.03	4	%var	0.02 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.06 ± 0.01	2.90 ± 0.20
			tag SNP	0.00 ± 0.00	0.02 ± 0.00	0.15 ± 0.03	0.01 ± 0.00	0.66 ± 0.09
			WPPA	0.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.90 ± 0.03
			PP_{int}	0.06 ± 0.06	0.77 ± 0.10	0.85 ± 0.07	0.85 ± 0.04	0.93 ± 0.03
2	0.16 ± 0.04		%var	0.01 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	0.02 ± 0.01	3.77 ± 0.33
			tag SNP	0.00 ± 0.00	0.02 ± 0.00	0.33 ± 0.04	0.00 ± 0.00	1.82 ± 0.29
			WPPA	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.60 ± 0.16	0.75 ± 0.07
			PP_{int}	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.60 ± 0.16	0.77 ± 0.07
3	0.45 ± 0.00	5	%var	0.16 ± 0.02	0.80 ± 0.03	0.90 ± 0.02	0.10 ± 0.01	1.12 ± 0.04
			tag SNP	0.00 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.58 ± 0.03
			WPPA	0.18 ± 0.02	0.79 ± 0.02	0.89 ± 0.01	0.89 ± 0.02	0.96 ± 0.00
			PP_{int}	0.40 ± 0.07	0.62 ± 0.06	0.68 ± 0.06	0.47 ± 0.04	1.00 ± 0.00
4	0.46 ± 0.00		%var	0.14 ± 0.02	0.82 ± 0.03	0.91 ± 0.01	0.09 ± 0.02	1.87 ± 0.12
			tag SNP	0.00 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.94 ± 0.04
			WPPA	0.17 ± 0.03	0.87 ± 0.02	0.94 ± 0.01	0.91 ± 0.02	0.96 ± 0.00
			PP_{int}	0.49 ± 0.09	0.78 ± 0.07	0.84 ± 0.05	0.44 ± 0.07	1.00 ± 0.00
5	0.46 ± 0.00	2.7	%var	0.00 ± 0.00	0.48 ± 0.01	0.63 ± 0.01	0.00 ± 0.00	0.73 ± 0.02
			tag SNP	0.00 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.76 ± 0.02
			WPPA	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.89 ± 0.01	0.85 ± 0.01
			PP_{int}	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.89 ± 0.04	0.84 ± 0.01
6	0.47 ± 0.00		%var	0.00 ± 0.00	0.49 ± 0.01	0.65 ± 0.01	0.00 ± 0.00	1.36 ± 0.02
			tag SNP	0.00 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	1.39 ± 0.02
			WPPA	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.01	0.84 ± 0.00
			PP_{int}	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.01	0.84 ± 0.00

Scenarios with oligogenic inheritance – 3 QTL: Scenario 1 (h^2 = 0.50) and Scenario 2 (h^2 = 0.60); 10 QTL: Scenario 3 (h^2 = 0.30) and Scenario 4 (h^2 = 0.40). Scenarios with polygenic inheritance – 100 QTL: Scenario 5 (h^2 = 0.10) and Scenario 6 (h^2 = 0.20); %var: proportion of genetic variance explained by genomic regions; WPPA: window posterior probability of association and PP_{int}: posterior probability of interval.

percentages of explanation of the genetic variance. For scenarios 1 and 2, the power to detect associated regions and the percentage of explanation of the variance were highest for the WPPA and %var criteria, elucidating the superiority of these methods in this genetic architecture.

Using a more general selection procedure, a copy of the one used by Gage et al. (2018), the WPPA criterion stood out in relation to the area under the ROC curve in scenarios with oligogenic inheritance, providing higher values compared to other methods. For false positive rates in scenarios 1 and 2, all criteria were similar, providing values of zero or close to zero. However, in scenarios 3 and 4, tag SNP was the method that showed superiority. Initially, to select tag SNPs, the frequency of inclusion of SNPs in the model (IF) and the *t*–like statistic (TL) were used. However, TL measures did not identify significant SNPs within the top windows and for this reason, only the results referring to the tag SNPs selected by the IF measure are shown in Table 2.

For the scenarios considering polygenic inheritance (traits controlled by many genes with small effects – scenarios 5 and 6), the PP_{int} and WPPA criterion were more efficient, presenting maximum values of power in detecting regions and percentage of variance explained. These criteria were also superior with respect to the area under the ROC curve, providing larger areas compared to the %var and tag SNP criteria for these scenarios. Note that the %var criterion was lower in these scenarios, proving efficient only for traits controlled by a few genes. With regard to the rate of false positives, in these scenarios, all methods analyzed obtained the lowest possible values (zero), indicating that no SNP/region was declared associated when it was not.

The WPPA and PP_{int} criteria showed lower power values in scenarios 3 and 4 than in other scenarios, which may have been influenced by the size of the region, which was the largest among all scenarios. Notably, in the smallest distance size considered (scenarios 5 and 6), these criteria stood out. Braz et al. (2019) and Bennewitz et al. (2017) reported on the influence of the size of the windows on detection power. In their studies considering sliding windows to select haplotypes associated with the bovine genome, Braz et al. (2019) used a mixed linear model, showing that smaller window sizes detect more associated regions and that larger window sizes may be more likely to introduce analytical problems, resulting in an excessive number of haplotypes, creating noise and computer memory problems. Furthermore, the results

obtained by Braz et al. (2019) corroborate those obtained above, in which power decreased with increasing window size. However, they contradict the results obtained by Bennewitz et al. (2017), where they verified an increase in power with an increase in the size of the windows.

For the PP_{int} criterion, detection power and the percentage of explanation of the variance increased with the increase in heritability for all scenarios with oligogenic inheritance (scenarios 1 for 2 and scenarios 3 for 4). However, the area under the ROC curve decreased with the increase in this heritability in these scenarios for this method. The same trend was found in the power and percentage verified for the WPPA criterion in scenarios 3 and 4 and for the tag SNP in scenarios 1 and 2. In the other scenarios, the values of power and percentage of variance were similar in relation to the increase in heritability for all criteria, including in scenarios with polygenic inheritance (scenario 5 to scenario 6). These results are in agreement with Shin and Lee (2015), in which they compared the statistical power according to the heritability for oligogenic and polygenic traits and found that the power difference between a heritability of 0.30 and 0.50 increased in the scenario containing 20 causal variants but decreased when there were 100. According to Shin and Lee (2015), the power estimated empirically from the simulation study would be applicable to GWAS for quantitative traits with known genetic parameters, predicting the degree of false negative associations.

The power values for %var criterion decreased according to the increase in the number of loci controlling the trait, indicating again that this method can be used advantageously for scenarios with oligogenic inheritance. However, it becomes inferior when considering traits with polygenic inheritance. The PP_{int} criterion stood out in scenarios controlled by many loci (polygenic inheritance), presenting lower rates of false positives, higher values for detection power, higher percentages of explanation of variance, and larger areas next to the WPPA criterion, which was also superior. Thus, the PP_{int} and WPPA criteria can be widely used in GWAS for inheritance, especially considering that the inheritance of most agronomically important traits are controlled by many genes, which individually have small or rare alleles (Yang et al., 2010).

Bennewitz et al. (2017) reported that the WPPA criterion seemed to be an inadequate approach to control the rate of false positives, since it was not built for this purpose and for this reason, it sometimes presents values at very high levels for this measure. Conversely, as observed in this study, this method captured greater proportions of the genetic variance and, in most cases, had greater power to detect associated regions. Fernando et al. (2017) also stated that the high threshold values for WPPA can compromise false positive rates, which corroborates the results found under certain scenarios in this study (scenarios 3 and 4). The same can be observed for the thresholds considered in the PP_{int} criterion under the same scenarios. Note that for the calculation of PP_{int}, we used regions that had SNPs with effects greater than the third quartile; however, in future studies, the possibility of determining an improvement in the results should be verified when considering other quartiles.

The WPPA measure, which was different from PP_{int} criterion, considers, for the calculation of the genetic variance of the regions, the allele frequency used to obtain the effects of the SNPs and the mean heterosis under the assumption of an equal distribution of the additive genetic variance and the dominance variance in all SNPs. Consideration of the allelic frequency in this criterion becomes feasible, since the detection power of SNPs is also determined by this measure (Shan and Purcell, 2014). According to these authors, a low allelic frequency influences a low detection power, unless there are relatively greater effects of SNPs. In addition, considering heterosis in GWAS can also satisfactorily affect the detection power of SNPs (Vidoti et al., 2019).

For all scenarios considered, the tag SNP criterion obtained false positive rates equal to zero. However, this criterion was the one that provided the least power to detect associated SNPs, compared to the other criteria. These results corroborate the information reported by Schmid and Bennewitz (2017), in which they stated that the decrease in the number of false positives in GWAS can compromise power. However, for scenarios 1, 2, 5, and 6, the criteria WPPA and PP_{int} also presented false positive rates equal to zero and additionally, they were efficient in terms of detection power. According to Li et al. (2014), the occurrence of false positives in GWAS can be controlled but this is only possible at the expense of reducing the power to detect true positives or statistical power. In other words, establishing a strict threshold for the association criterion is an effective way to control the rate of false positives. However, this also reduces the number of true positives detected. A desirable solution would be to reduce false positives, without compromising the detection power of the analysis, as performed by the WPPA and PP_{int} criteria in these studied scenarios.

The results also revealed that the π probability obtained by the BayesDπ method varied from 0.16 to 0.47 between the scenarios, indicating that the number of markers that are supposed to be in LD with the QTL varied from 320 to 940. According to Fernando and Garrick (2013), higher π values may be more discriminatory for the identification of QTL with the greatest effect, which is an important factor for the selection of SNPs. Additionally, Sollero et al. (2017), in studies to select tag SNPs related to tick resistance in cattle breeds, found that the decrease in the π probability value may cause an increase in the proportion of genetic variance explained by the SNPs, in accordance with the results obtained here for tag SNP criterion.

The results, considering 0.81 as a threshold for determining the regions in LD, are shown in Table 3. As for the detection power and percentage of explanation of the variance, the results were similar to those found using the LD of 0.64 (Table 2), in which the criteria

Table 3 – Size of the regions (distance) in centimorgans (cM) found through the linkage disequilibrium (LD) between the markers and quantitative trait loci (QTL) in each scenario using the LD extension of 0.81 and the means and standard errors of the estimated π probability via BayesDπ, false positive rates (FP), detection power (PD), percentage of genetic variance recovered (PE), area under the receiver operating characteristic curve (ROC) and threshold for selecting regions obtained by criteria %var, tag SNP, WPPA and PP_{int}.

Scenarios	π	Distance	Criterion	FP	Power	PE	Area	Threshold
1	0.32 ± 0.03	2	%var	0.03 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.08 ± 0.01	1.68 ± 0.09
			tag SNP	0.00 ± 0.00	0.05 ± 0.00	0.21 ± 0.03	0.00 ± 0.00	0.27 ± 0.05
			WPPA	0.01 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.01	0.89 ± 0.03
			PP_{int}	0.22 ± 0.11	1.00 ± 0.00	1.00 ± 0.00	0.89 ± 0.06	0.91 ± 0.03
2	0.16 ± 0.04		%var	0.02 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.04 ± 0.01	2.11 ± 0.16
			tag SNP	0.00 ± 0.00	0.04 ± 0.00	0.38 ± 0.03	0.00 ± 0.00	0.86 ± 0.15
			WPPA	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.60 ± 0.16	0.72 ± 0.06
			PP_{int}	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.59 ± 0.16	0.73 ± 0.06
3	0.45 ± 0.00	1.5	%var	0.18 ± 0.02	0.83 ± 0.02	0.88 ± 0.02	0.17 ± 0.01	0.39 ± 0.02
			tag SNP	0.01 ± 0.00	0.06 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.06 ± 0.00
			WPPA	0.19 ± 0.04	0.76 ± 0.06	0.82 ± 0.05	0.85 ± 0.02	0.96 ± 0.00
			PP_{int}	0.07 ± 0.01	0.53 ± 0.06	0.63 ± 0.06	0.75 ± 0.03	0.99 ± 0.00
4	0.46 ± 0.00		%var	0.15 ± 0.02	0.85 ± 0.02	0.91 ± 0.01	0.18 ± 0.01	0.67 ± 0.04
			tag SNP	0.01 ± 0.00	0.06 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.11 ± 0.01
			WPPA	0.12 ± 0.02	0.83 ± 0.04	0.89 ± 0.03	0.91 ± 0.01	0.96 ± 0.00
			PP_{int}	0.10 ± 0.01	0.74 ± 0.06	0.82 ± 0.04	0.83 ± 0.02	0.99 ± 0.00
5	0.46 ± 0.00	1	%var	0.33 ± 0.01	0.42 ± 0.02	0.52 ± 0.02	0.07 ± 0.00	0.32 ± 0.01
			tag SNP	0.02 ± 0.00	0.04 ± 0.00	0.05 ± 0.00	0.00 ± 0.00	0.30 ± 0.01
			WPPA	0.46 ± 0.04	0.56 ± 0.05	0.62 ± 0.04	0.59 ± 0.01	0.93 ± 0.00
			PP_{int}	0.54 ± 0.04	0.75 ± 0.05	0.79 ± 0.05	0.63 ± 0.01	0.95 ± 0.00
6	0.47 ± 0.00		%var	0.31 ± 0.01	0.43 ± 0.01	0.54 ± 0.01	0.07 ± 0.00	0.57 ± 0.01
			tag SNP	0.01 ± 0.00	0.04 ± 0.00	0.05 ± 0.00	0.00 ± 0.00	0.54 ± 0.01
			WPPA	0.42 ± 0.03	0.56 ± 0.04	0.65 ± 0.03	0.60 ± 0.01	0.92 ± 0.00
			PP_{int}	0.53 ± 0.05	0.76 ± 0.06	0.80 ± 0.05	0.64 ± 0.00	0.95 ± 0.00

Scenarios with oligogenic inheritance – 3 QTL: Scenario 1 (h^2 = 0.50) and Scenario 2 (h^2 = 0.60); 10 QTL: Scenario 3 (h^2 = 0.30) and Scenario 4 (h^2 = 0.40). Scenarios with polygenic inheritance – 100 QTL: Scenario 5 (h^2 = 0.10) and Scenario 6 (h^2 = 0.20); %var: proportion of genetic variance explained by genomic regions; WPPA: window posterior probability of association and PP_{int}: posterior probability of interval.

WPPA, %var, and PP_{int} were superior to the criteria of selection by tag SNPs. In scenarios 5 and 6, the PP_{int} criterion was the most efficient when compared to the others, once again showing its superiority in scenarios with polygenic inheritance. As regards the areas on the ROC curve, the rates of false positives, the detection power according to the increase in heritability, and the same results as those obtained previously were also observed. The resultant rate of false positives is in line with what was discussed by Moore et al. (2010), which highlighted the advantage of considering groups of markers together, since these sets tend to capture a greater proportion of the genetic variance.

Comparing Tables 2 and 3, in relation to the LD extensions considered, the results revealed that for the criteria %var, WPPA, and PP_{int}, the detection power and the percentage of variance explained increased with the decrease in LD extension from 0.81 to 0.64, for the two scenarios analyzed with polygenic inheritance. The sizes of regions obtained, considering the LD extension of 0.64, exceeded those found with an extension of 0.81, and thus, the results reported when we increased the sizes of the regions there seems to be an increase in power for these three criteria. These results corroborate those obtained by Bennewitz et al. (2017), in which there was also an increase in power with the increase in the size of the regions. However, the optimum size of the genomic regions may differ from one study to the next or different QTL in the same study, depending on the extent of LD between the markers and QTL, the effective size of the population, and the detection power of each approach (Braz et al., 2019; Guo et al., 2016).

For scenarios with oligogenic inheritance, only the tag SNP criterion presented different power values between the two extensions of the LD, verifying an increase in these values with an increase in threshold. The results also revealed that the threshold obtained in the PP_{int} and WPPA criteria decreased according to the increase in the sizes of the regions for scenarios with polygenic inheritance. However, Guo et al. (2016), also using a procedure to select regions with GWAS in pigs, found that for regions with sizes above 5 Megabases, there was no increase in the values for this threshold.

The results for the single-marker analysis are shown in Table 4 and reveal that this procedure identified SNPs associated only in scenarios with oligogenic inheritance in which three QTL were randomly distributed among the ten chromosomes (scenarios 1 and 2). The detection

Table 4 – Extensions of linkage disequilibrium (LD) used to determine the distances between SNPs, means and respective standard errors of false positive rates (FP), detection power (PD), percentage of genetic variance recovered (PE), the area under the receiver operating characteristic curve (ROC) estimated by the single–marker analysis and also the threshold used to select significant SNPs.

Scenarios	LD	Distance	FP	Power	PE	Area	Threshold
1	0.64	4	0.00 ± 0.00	0.07 ± 0.01	0.13 ± 0.02	0.01 ± 0.00	0.05 ± 0.00
	0.81	2	0.00 ± 0.00	0.15 ± 0.02	0.13 ± 0.02	0.00 ± 0.00	0.05 ± 0.00
2	0.64	4	0.00 ± 0.00	0.11 ± 0.01	0.20 ± 0.02	0.00 ± 0.00	0.05 ± 0.00
	0.81	2	0.00 ± 0.00	0.21 ± 0.02	0.20 ± 0.02	0.00 ± 0.00	0.05 ± 0.00

Scenarios with oligogenic inheritance – 3 QTL: Scenario 1 (h^2 = 0.50) and Scenario 2 (h^2 = 0.60).

power, considering an LD of 0.64, was always lower than that obtained using an LD of 0.81. As regards the false positive rate, this method presented values equal to zero, indicating that the method considering oligogenic effects was efficient in identifying SNPs only when they were truly associated with the traits of interest. However, the area obtained under the ROC curve was zero for both scenarios.

Compared to the criteria considered in this study, the single–marker analysis was lower than the %var, WPPA, and PP_{int} criteria, with less detection power, lower variance explained percentages, and smaller areas in scenarios 1 and 2 for both LD levels analyzed. However, this method showed higher values of power than the tag SNP criterion and similar percentages of explanation for the genetic variance. According to Resende et al. (2017), the single–marker method can capture a higher percentage of the genetic variance due to the fact that it generally overestimates the effects of the tags, since the estimation process is not done simultaneously.

Conclusions

Considering traits with oligogenic genetic inheritance, the WPPA criteria, followed by the %var and PP_{int} criteria, were shown to be superior to the *tag* SNP criterion presenting higher values of detection power, capturing higher percentages of genetic variance, and larger areas under the ROC curve. For traits with polygenic inheritance, the PP_{int} and WPPA criteria were considered superior to the others for the LD extension of 0.64 and the LD of 0.81, only PP_{int} stood out as being more efficient. The single–marker analysis method identified SNPs associated only in oligogenic inheritance scenarios and was lower than the %var, WPPA, and PP_{int} criteria. In general, the PP_{int} and WPPA criteria can be widely used in GWAS, especially considering that the inheritance of most agronomically important traits are controlled by many genes, which, individually, have small or rare alleles.

Acknowledgments

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

Authors' Contributions

Conceptualization: Lima, L.P.; Azevedo, C.F.; Resende, M.D.V.; Nascimento, M.; Silva, F.F. **Data acquisition:** Lima, L.P.; Azevedo, C.F.; Resende, M.D.V. **Data Analysis:** Lima, L.P.; Azevedo, C.F. **Design of methodology:** Lima, L.P., A;evedo, C.F.; Resende, M.D.V.; Nascimento, M.; Silva, F.F. **Software development:** Lima, L.P.; Azevedo, C.F. **Writing and Editing:** Lima, L.P.; Azevedo, C.F.; Resende, M.D.V.; Nascimento, M.

References

Azevedo, C.F.; Nascimento, M.; Fontes, V.C.; Resende, M.D.V.D.; Cruz, C.D. 2019. GenomicLand: software for genome-wide association studies and genomic prediction. Acta Scientiarum. Agronomy 41: e45361.

Bennewitz, J.; Edel, C.; Fries, R.; Meuwissen, T.H.; Wellmann, R. 2017. Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. Genetics Selection Evolution 49: 1-13.

Braz, C.U.; Taylor, J.F.; Bresolin, T.; Espigolan, R.; Feitosa, F.L.; Carvalheiro, R.; Oliveira, H.N. 2019. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. BMC Genetics 20: 1-12.

Covarrubias-Pazaran, G. 2016. Genome-assisted prediction of quantitative traits using the R package sommer. PloS One 11: e0156744.

Cruz, C.D. 2013. Genes: a software package for analysis in experimental statistics and quantitative genetics. Acta Scientiarum. Agronomy 35: 271-276.

Fan, B.; Onteru, S.K.; Du, Z.Q.; Garrick, D.J.; Stalder, K.J.; Rothschild, M.F. 2011. Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. PloS One 6: e14726.

Fernando, R.L.; Garrick, D. 2013. Bayesian methods applied to GWAS. p. 237-274. In: Gondro, C.; van der Werf, J.; Hayes, B. Genome-wide association studies and genomic prediction. Humana Press, Totowa, NJ, USA.

Fernando, R.; Toosi, A.; Wolc, A.; Garrick, D.; 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, R.L.; Nettleton, D.; Southey, B.R.; Dekkers, J.C.M.; Rothschild, M.F.; Soller, M. 2004. Controlling the proportion of false positives in multiple dependent tests. Genetics 166: 611-619.

Gage, J.L.; De Leon, N.; Clayton, M.K. 2018. Comparing genome-wide association study results from different measurements of an underlying phenotype. G3: Genes, Genomes, Genetics 8: 3715-3722.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. Bayesian Statistics 4: 641-649.

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

Guo, X.; Su, G.; Christensen, O.F.; Janss, L.; Lund, M.S. 2016. Genome-wide association analyses using a Bayesian approach for litter size and piglet mortality in Danish Landrace and Yorkshire pigs. BMC Genomics 17: 468.

Habier, D.; Fernando, R.L.; Kizilkaya, K.; Garrick, D.J. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Henderson, C.R. 1973. Sire evaluation and genetic trends. Journal of Animal Science 1973(Symposium): 10-41.

Li, M.; Liu, X.; Bradbury, P.; Yu, J.; Zhang, Y.M.; Todhunter, R.J.; Zhang, Z. 2014. Enrichment of statistical power for genome-wide association studies. BMC Biology 12: 73.

Lima, L.P.; Azevedo, C.F.; Resende, M.D.V.D.; Viana, J.M.S.; Oliveira, E.J.D. 2019. Triple categorical regression for genomic selection: application to cassava breeding. Scientia Agricola 76: 368-375.

Liu, X.; Huang, M.; Fan, B.; Buckler, E.S.; Zhang, Z. 2016. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genetics 12: e1005767.

Metz, C.E. 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8: 283-298.

Meuwissen, T.; Hayes, B.; Goddard, M. 2016. Genomic selection: a paradigm shift in animal breeding. Animal Frontiers 6: 6-14.

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

Moore, J.H.; Asselbergs, F.W.; Williams, S.M. 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445-455.

Peters, S.O.; Kizilkaya, K.; Garrick, D.J.; Fernando, R.L.; Reecy, J.M.; Weaber, R.L.; Silver, G.A.; Thomas, M.G. 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.

Plummer, M.; Best, N.; Cowles, K.; Vines, K. 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6: 7-11.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Resende, M.D.V.; Lopes, P.S.; Silva, R.L.; Pires, I.E. 2008. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético = Genome wide selection (GWS) and maximization of genetic improvement efficiency. Pesquisa Florestal Brasileira n. 56: 63-67 (in Portuguese, with abstract in English).

Resende, R.T.; Resende, M.D.V.; Silva, F.F.; Azevedo, C.F.; Takahashi, E.K.; Silva-Junior, O.B.; 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.

Schmid, M.; Bennewitz, J. 2017. Invited review: genome-wide association analysis for quantitative traits in livestock–a selective review of statistical models and experimental designs. Archiv fuer Tierzucht 60: 335-346.

Schurink, A.; Wolc, A.; Ducro, B.J.; Frankena, K.; Garrick, D.J.; Dekkers, J.C.; van Arendonk, J.A. 2012. Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. Genetics Selection Evolution 44: 31.

Shin, J.; 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.

Sollero, B.P.; Junqueira, V.S.; Gomes, C.C.; Caetano, A.R.; Cardoso, F.F. 2017. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. Genetics Selection Evolution 49: 1-15.

Storey, J.D.; Tibshirani, R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100: 9440-9445.

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

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

Vidotti, M. S. et al. 2019. Additive and heterozygous (dis) advantage GWAS models reveal candidate genes involved in the genotypic variation of maize hybrids to Azospirillum brasilense. PloS one, v. 14, n. 9, p. e0222788.

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

Wang, Q.; Tian, F.; Pan, Y.; Buckler, E.S.; Zhang, Z. 2014. A SUPER powerful method for genome wide association study. PloS One 9: e107684.

Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D. R; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; Goddard, M.E. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42: 565.

Zhao, H.; Nettleton, D.; Dekkers, J.C. 2007. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. Genetics Research 89: 1-6.