CAMILA FERREIRA AZEVEDO

RIDGE, LASSO AND BAYESIAN ADDITIVE-DOMINANCE GENOMIC MODELS AND NEW ESTIMATORS FOR THE EXPERIMENTAL ACCURACY OF GENOME SELECTION

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS – BRASIL
2015

CAMILA FERREIRA AZEVEDO

RIDGE, LASSO AND BAYESIAN ADDITIVE-DOMINANCE GENOMIC MODELS AND NEW ESTIMATORS FOR THE EXPERIMENTAL ACCURACY OF GENOME SELECTION

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 26 de outubro de 2015.

Júlio Sílvio de Sousa Bueno Filho

José Marcelo Soriano Viana

Fabyano Fonseca e Silva
(Coorientador)

Moysés Nascimento
(Coorientador)

Marcos Deon Vilela de Resende
(Orientador)

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

A994r
2015

Azevedo, Camila Ferreira, 1988-
Ridge, lasso and bayesian additive-dominance genomic models and new estimators for the experimental accuracy of genome selection / Camila Ferreira Azevedo. - Viçosa, MG, 2015.
xi, 108f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: Marcos Deon Vilela de Resende.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Genética molecular. 2. Genética - Métodos estatísticos. 3. Marcadores genéticos. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 576.50727

Ao meu pai, Sergio;
por ter dedicado a
vida a suas filhas.

AGRADECIMENTOS

A Deus por me amparar nos momentos difíceis, dar força interior para superar as dificuldades e indicar o caminho nas horas incertas.

A Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, por proporcionar a realização de um curso de excelência.

Aos meus pais, Marilia e Sergio, pelo amor incondicional, pelos ensinamentos, pela dedicação e confiança.

À minha irmã, Lorena, pelas palavras de carinho e de incentivo e por sempre estar ao meu lado.

Ao meu noivo Vitor, pela paciência, companheirismo, amor, carinho e incentivo em todos os momentos.

Às minhas queridas avós Arlene e Marilza, por todas as orações.

À minha querida família, por me apoiar em todos os momentos.

Aos meus amigos do PPESTBIO, em especial a Laís e a Gabi, pelos ótimos momentos, pelas trocas de experiência, pelas palavras de conforto e de motivação.

Aos meus amigos da Matemática, em especial Amanda, Izabela, Victor e Marcelo, pela valiosa amizade, por sempre se fazerem presentes e por torcerem sempre por mim.

Ao Doutor e orientador Marcos Deon Vilela de Resende, pelos ensinamentos, confiança, dedicação, por contribuir para o meu crescimento profissional e por ser também um exemplo de pesquisador admirável e generoso.

Aos Doutores e coorientadores Fabyano Fonseca e Silva e Moyses Nascimento, pelos saberes transmitido, pela confiança, disponibilidade, incentivo e generosidade.

Ao Doutor José Marcelo Soriano Viana pela disponibilidade e pela imprescindível ajuda na execução deste trabalho.

Agradeço novamente, ao Doutor e orientador Marcos Deon Vilela de Resende e ao Doutor e coorientador Fabyano Fonseca e Silva, por terem me proporcionado oportunidades profissionais que enriqueceram minha formação e por terem me orientado na realização de inúmeras conquistas. Agradecer nunca será o suficiente, mas fica aqui registrada minha eterna gratidão.

Aos membros da banca examinadora, Prof. Doutor Fabyano Fonseca e Silva, Prof. Doutor José Marcelo Soriano Viana, Prof. Doutor Júlio Sílvio de Sousa Bueno Filho, Prof. Doutor Marcos Deon Vilela de Resende, Prof. Doutor Moyses Nascimento, pela disponibilidade e pelas valiosas sugestões para o enriquecimento deste trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria e atuais colegas de trabalho, por contribuírem para minha formação acadêmica, por me auxiliarem e me conduzirem à prática docente, em especial aos Professores e amigos Ana Carolina Campana Nascimento, Moyses Nascimento, Paulo Roberto Cecon e Antonio Policarpo Souza Carneiro.

As secretárias e amigas do Departamento de Estatística, Anita e Carla, pela amizade, incentivo e carinho.

A FAPEMIG, pela concessão da bolsa de estudos.

Enfim, muito obrigada a todos aqueles que de certa forma contribuíram para o meu crescimento pessoal e profissional, para a realização de um sonho e a concretização deste trabalho.

BIOGRAFIA

CAMILA FERREIRA AZEVEDO, filha de Marilia Assis Ferreira Azevedo e de Sergio de Rezende Azevedo, nasceu em Bom Jesus do Itabapoana, Rio de Janeiro, em 15 de abril de 1988.

Em maio de 2006, ingressou no curso de Licenciatura em Matemática na Universidade Federal de Viçosa, Viçosa-MG, graduando-se em julho de 2010. Em agosto do mesmo ano, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 26 de julho de 2012.

Em abril de 2013, iniciou o curso de Doutorado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de tese em 26 de outubro de 2015.

SUMÁRIO

RESUMO

AZEVEDO, Camila Ferreira, D.Sc., Universidade Federal de Viçosa, outubro de 2015. **Modelos genômicos aditivo-dominante via abordagem Ridge, Lasso e Bayesiana e novos estimadores para a acurácia experimental da Seleção Genômica.** Orientador: Marcos Deon Vilela de Resende. Coorientadores: Fabyano Fonseca e Silva e Moyses Nascimento.

A principal contribuição da genética molecular no melhoramento é a utilização direta das informações de DNA no processo de identificação de indivíduos geneticamente superiores. Sob esse enfoque, idealizou-se a seleção genômica ampla (Genome Wide Selection – GWS), a qual consiste na análise de um grande número de marcadores SNPs (Single Nucleotide Polymorphisms) amplamente distribuídos no genoma. Este trabalho de simulação apresenta uma abordagem completa para a seleção genômica por meio de adequados modelos genéticos incluindo efeitos aditivos e devido à dominância, que são essenciais para a seleção de clones e de cruzamentos, bem como para melhorar a estimativa de efeitos aditivos para a seleção. Até o momento, as abordagens via Ridge Bayesiana e Lasso para modelos aditivo-dominante não foram avaliados e comparados na literatura. Neste trabalho, foram avaliados o desempenho de 10 modelos de predição aditivo-dominante (incluindo os modelos existentes e propostas de modificação). Um novo método Bayesiano/Lasso modificado (chamado BayesA* B* ou t-BLASSO) obteve melhor desempenho na estimação de valores genéticos genômicos dos indivíduos, em todos os quatro cenários (dois níveis de herdabilidades × duas arquiteturas genéticas). Os métodos do tipo BayesA*B* apresentaram melhor capacidade para recuperar a razão entre a variância de dominância e a variância aditiva. Além disso, o papel das três fontes de informação da genética quantitativa (chamadas de desequilíbrio de ligação, co-segregação e relações de parentesco) na seleção genômica foram elucidadas pela decomposição da herdabilidade e da acurácia nos três componentes, mostrando suas relações com a estrutura de populações e o melhoramento genético, a curto e longo prazo. Além disso, neste trabalho de simulação também foi desenvolvido dois novos estimadores para a acurácia preditiva da seleção genômica. O trabalho propõe e avalia o desempenho e a eficiência destes novos estimadores chamados estimador regularizado (RE) e estimador híbrido (HE). O estimador regularizado leva em consideração tanto a herdabilidade genômica quanto a herdabilidade da

característica, além da capacidade preditiva. Enquanto, o estimador híbrido (HE), combina as acurácias experimental e esperada. As comparações entre RE e HE com o estimador tradicional (TE) foram feitas sob quatro procedimentos de validação. Em geral, RE apresentou acurácias mais próximas aos valores paramétricos, principalmente quando há seleção de marcadores. RE também foi menos tendencioso e mais preciso, com desvios padrão menores do que o estimador tradicional. Diante dos resultados, o TE pode ser usado apenas com a validação independente, em que tende a ter um melhor desempenho do que RE, embora superestimando a acurácia. O estimador híbrido (HE) provou ser muito eficaz na ausência de validação. Enquanto, que a validação independente mostrou-se superior em relação aos procedimentos de Jacknife, perseguindo melhor a acurácia paramétrica com ou sem seleção de marcador. As seguintes inferências podem ser feitas de acordo com o estimador de acurácia e tipo de validação: (i) a acurácia mais provável: HE sem validação; (ii) a maior acurácia possível (acurácia superestimada): TE com validação independente; (iii) a menor acurácia possível (acurácia subestimada): RE com validação independente.

ABSTRACT

AZEVEDO, Camila Ferreira, D.Sc., Universidade Federal de Viçosa, October, 2015. **Ridge, Lasso and Bayesian additive-dominance genomic models and new estimators for the experimental accuracy of Genome Selection.** Adviser: Marcos Deon Vilela de Resende. Co-advisers: Fabyano Fonseca and Silva and Moyses Nascimento.

The main contribution of molecular genetics is the direct use of DNA information to identify genetically superior individuals. Under this approach, genome-wide selection (GWS) can be used with this purpose. GWS consists in analyzing of a large number of SNP markers widely distributed in the genome. This simulation work presents a complete approach for genomic selection by using adequate genetic models including dominance effects, which are essential for selecting crosses and clones as well as for improving the estimation of additive effects for parent selection. To date, the approaches via Ridge, Lasso and Bayesian additive-dominance models have not been evaluated and compared in the literature.The performance of 10 additive-dominance prediction models (including current ones and proposed modifications) were evaluated. A new modified Bayesian/Lasso method (called BayesA*B* or t-BLASSO) performed best in the prediction of genomic breeding value of individuals, in all the four scenarios (two heritabilities × two genetic architectures). The BayesA*B*-type methods showed better ability for recovering the dominance variance/additive variance ratio. Also, the role of the three quantitative genetics information sources (called linkage disequilibrium, co-segregation and pedigree relationships) in genomic selection were elucidated by decomposing the heritability and accuracy in the three components and showing their relations with the structure of populations and the genetic improvement in the short and long run. Moreover, this simulation work also, we developed the new estimators for the prediction accuracy of genomic selection. The work proposes and evaluates the performance and efficiency of these new estimators called regularized estimator (RE) and hybrid estimator (HE). The regularized estimator takes in consideration both the genomic and trait heritabilities, in addition to the predictive ability. The hybrid estimator (HE), combines both experimental and expected accuracies. The comparisons of the RE and HE with the traditional (TE) were done under four

validation procedures. In general, the new estimator presented accuracies closer to the parametric ones, mainly when selecting markers. It was also less biased and more precise, with smaller standard deviations than the traditional estimator. The TE can be used only with independent validation, where it tends to perform better than RE, although overestimating the accuracy. The hybrid estimator (HE) proved to be very effective in the absence of validation. The independent validation showed to be superior over the Jacknife procedures, chasing better the parametric accuracy with or without marker selection. The following inferences can be made according to the accuracy estimator and kind of validation: (i) most probable accuracy: HE without validation; (ii) highest possible accuracy: TE with independent validation; (iii) lowest possible accuracy: RE with independent validation.

GENERAL INTRODUCTION

Recently there was advancement in molecular genetics which promoted an rapid evolution of sequencing and genotyping technologies. The molecular genetics can benefit the plant and animal breeding in terms of the identification of genetically superior individuals by using directly the information from DNA. The use of molecular markers allows an increase in selection efficiency and speed in obtaining desirable genetic gains.

The single nucleotide polymorphisms (SNPs) genetic markers are the most used in genomic prediction, due to its low mutation rate, codominance and abundance. Due to the availability of high density of SNPs markers in the genome, Meuwissen et al. (2001) developed the genome-wide selection (GWS), as an approach to accelerating the breeding cycle. The GWS consists in an analysis of a large number of markers widely distributed in the genome, capturing the genes affecting quantitative traits of interest.

It is possible to assume that some markers are in linkage disequilibrium (LD) with quantitative trait loci (QTL), enabling, together with phenotypic data, its direct use in the estimation of the genetic value of individuals subject to selection, including individuals who have not yet been phenotyped. However, the number of markers is generally much larger than the number of genotyped and phenotyped individuals and that markers are highly correlated, which requires appropriate statistical methods that enable estimability and regularization properties (Gianola et al., 2003).

A complete approach for genomic wide selection also involves reliable statistical genetics models and methods and adequate size and structure of estimation

and validation populations. Reports on these topics are common for additive genetic models but not for additive-dominance models. However, it is known that dominance estimation is essential specially for vegetative propagated species (Denis and Bouvet, 2013) and crossed populations, where the mating allocation including both additive and dominance is an effective way of increasing genetic gain capitalizing on heterosis (Toro and Varona, 2010; Wellmann and Bennewitz, 2012). Additive-dominance models are able to capture both effects, allowing effective selection for parents, crosses and for clones. This allows taking full advantage of genomic selection in perennials and asexually propagated crops and also in crossed animals. In addition, the inclusion of dominance in the prediction model may improve the accuracy of genomic prediction when dominance effects are present (Wang et al., 2014; Hu et al., 2014; Su et al., 2012).

Wellmann and Bennewitz (2012) presented theoretical genetic models for Bayesian genomic selection with dominance and concluded that dominance enhances the analysis and several advantages are aggregated. However, Bayesian, Lasso and Ridge regression approaches have not been compared for additive-dominance models yet. Zeng et al. (2013), Muñoz et al. (2014), Su et al. (2012) and Denis and Bouvet (2013) and Wang and Da (2014) applied only the G-BLUP method, which is an equivalent model (Goddard et al., 2009), to ridge regression (RR-BLUP). On the other hand, Wellmann and Bennewitz (2012) applied only the Bayesian methods of Meuwissen et al. (2001) with modifications. Toro and Varona (2010) evaluating the introduction of the dominant effects in the model using the Bayes A. Lasso methods seem to be unused with dominance models for variance components in genomic selection.

In addition, the evaluation of the methods applied to genomic selection consist in one of the main lines of research in this area and the main measure for evaluating the efficiency of the prediction of genomic breeding values is the experimental accuracy (after obtaining data). However, traditional accuracy estimators (Legarra et al., 2008; Hayes et al., 2009a) have some practical and theoretical inconsistencies, which lead erroneous conclusions in some circumstances. Estaghvirou et al. (2013) carried out a comparative study among alternative accuracy estimators for GWS, but such estimators only differed from the estimator of Legarra et al. (2008) and Hayes et al. (2009a) in the different ways of estimating the trait heritability. In fact, to date, there are no proposed estimators for the experimental accuracy of genomic selection, but only for the expected (before obtaining data) accuracy (Daetwyler et al., 2008; 2010; Resende, 2008; Goddard, 2009; Goddard et al., 2011; Hayes et al. 2009b).

Given the above, a key to the success of practical application of genomic selection is the use of appropriate methodologies and evaluating measure and the use correct of information in the estimation population. Thus, this simulation work presents a complete approach for genomic selection by using adequate genetic models including dominance effects, evaluates 10 estimation methods (including Bayesian, Lasso and Ridge regression approaches) for fitting additive-dominance genomic models for GWS and decomposes genomic heritability and accuracy in terms of the three quantitative genetics information compounds linkage disequilibrium (LD), co-segregation (CS) and pedigree relationships (PR). In addition, this thesis also propose and evaluate the performance and efficiency of the two new estimators (called regularized and hybrid) for the accuracy of GWS.

CHAPTER 1
LITERATURE REVISION

1. Statistical Methods for Additive-Dominance Models

Statistical methods for genomic selection can be divided in three groups: explicit regression methods (Ridge, Bayesian and Lasso regression, etc), implicit regression methods (Kernel, Reproducing Kernel Hilbert Spaces - RKHS, Neural Networks, etc), dimensionality reduction methods (principal components regression, partial least square regression, independent components regression, etc). The explicit regression methods encompassing the penalization and Bayesian methods such as ridge regression (RR), Bayesian RR (BayesRR), Bayesian LASSO (BLASSO) and BayesA and B of Meuwissen et al. (2001). These are the main approaches being applied in practical genomic selection (De los Campos et al, 2012; Resende Jr., et al. 2012; Gianola, 2013; Lehermeier et al., 2013). The Bayesian methods are Bayesian linear regression that differs in the priors adopted while sharing the same model (Gianola, 2013). Daetwyler et al. (2012) recommended comparing accuracy and bias of new methods to results from genomic best linear prediction and a variable selection approach with specific variance components for each locus such as BayesB, because, together, these methods are appropriate for a range of genetic architectures.

1.1. Statistical concepts

An ideal bayesian prediction method of genomic breeding values (GBV) mainly depends on the prior distribution chosen for the coefficients. Thus, for a better understanding of Bayesian methods is necessary to address the following topics related to shrinkage and probability distributions.

1.1.1. Shrinkage

The shrinkage methods consist in constraints on the size of coefficient estimates, or equivalently, shrinks the coefficient estimates towards zero relative to the least squares estimates. These methods leads to an economy in the degrees of freedom (regularization property) and leads to stable estimates, allowing for the estimability of the parameters in the n >> N case, where n is the covariables numbers and N is the observations number, and when there is multicollinearity among the variables. Thus, this property is very important to high dimensionality case. The shrinkage effect considers the sample size and variations of random and residual effects.

Among the methodologies applied to genomic selection, the penalization and Bayesian methods are also called shrinkage methods. The shrinkage is implicit in Bayesian Inference and the degree of shrinkage is controlled by the assumed prior distribution to markers effects. The bayesian prediction of genomic breeding values (obtained by SNPs effects - m) is based in the bayesian estimation, which is equivalent the conditional mean of the genetic value given the individuals genotype at each QTL (Resende et al., 2012). Thus, based on markers and considering each QTL separately, the markers effects (m) are given by the conditional expectation $\hat{m} = E(m|\ y)$. The appropriate bayesian estimator is obtained by Bayes' theorem and is given by:

$$\hat{m} = E(m|\ y) = \frac{\int_{R_m} m f(y|m)\, f(m) dm}{\int_{R_m} f(y|m)\, f(m) dm},$$

where $f(y|m)$ is the data likelihood function and $f(m)$ is the prior distribution to the QTLs effects m. Therefore, the estimator above shows that the method is

dependent of assumed prior distribution to the markers effects (or putative QTLs) and consequently the degree of shrinkage ascribed.

The presence of QTL is analyzed in many positions, as there are thousands SNPs widely distributed in the genome (Goddard and Hayes, 2007), but it is known that not all markers are in linkage disequilibrium with the QTL. Thus, the ideal prior distribution $f(m)$ should have a high density in $f(0)$, produced by shrinkage of each prior distribution. Therefore, the shrinkage leads many effects equal to zero and consequently a more favorable statistical condition, i. e., more individuals (N) to estimate less markers effects (n) (the lower the ratio $\frac{n}{N}$, the better estimation process). In addition, learning about genetic architecture without contamination from effects of the prior does not take place whenever $N \ll n$ (Gianola, 2013).

1.1.2. Probability distributions

1.1.2.1. Normal distribution

Normal distributions are extremely important in statistics, since the most known statistical procedures assume that the continuous random variables follows this distribution. The normal probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < X < +\infty$$

where the parameter of location μ is the mean or expectation of the distribution, the parameter of scale σ is the standard deviation and σ^2 is the variance.

The Figure 1 shows three normal distributions. The red distribution has a mean of -2 and a standard deviation of 0.5, the distribution in blue has a mean of 0 and a standard deviation of 1 (standardized normal distribution), and the distribution

in yellow has a mean of 1 and a standard deviation of 2.5. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and tails less dense.

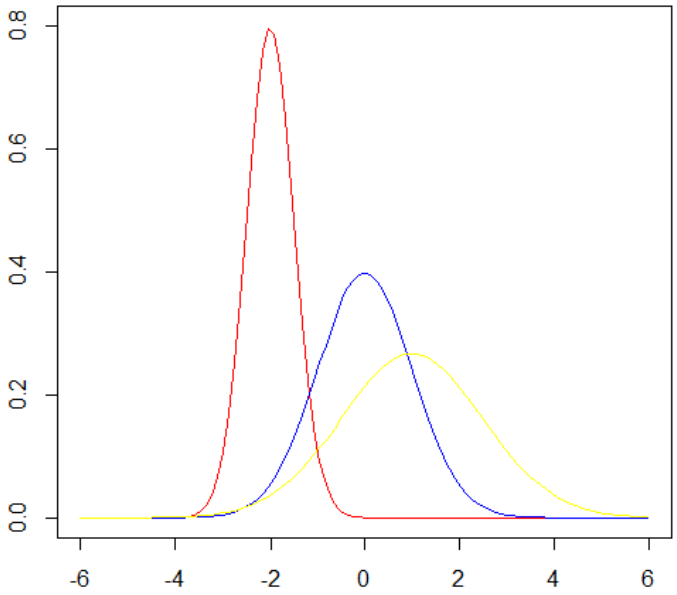

Figure 1. Normal distributions differing in the mean (-2, 0 and 1 – red, blue and yellow, respectively) and standard deviation (0.5, 1 and 1.5 – red, blue and yellow, respectively).

1.1.2.2. Student's t-distribution

The Student's t-distribution (or t-distribution) is also a member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. However, the larger the sample, the more the distribution resembles a normal distribution. Student's t-distribution has the probability density function given by:

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}, \quad -\infty < x < +\infty$$

where v is the number of degrees of freedom and Γ is the gamma function. The mean of t random variable is 0 for $v > 0$ and variance is $\frac{v}{v+2}$ for $v > 2$ or ∞ for $1 < v \leq 2$.

The Figure 2 shows t distributions with 4 (blue), 6 (red), and 12 (yellow) degrees of freedom and the standard normal distribution (green). It can be seen that as the number of degrees of freedom increases the tails of the distributions become heavier. Notice also, that the normal distribution has relatively more scores in the center of the distribution and the t distribution has relatively more in the tails. The t-distribution is therefore leptokurtic, meaning that it is more prone to producing values that fall far from its mean and more extreme values. That is, most effects are close to zero, but some extremely big. The t-distribution approaches the normal distribution as the degrees of freedom increases.

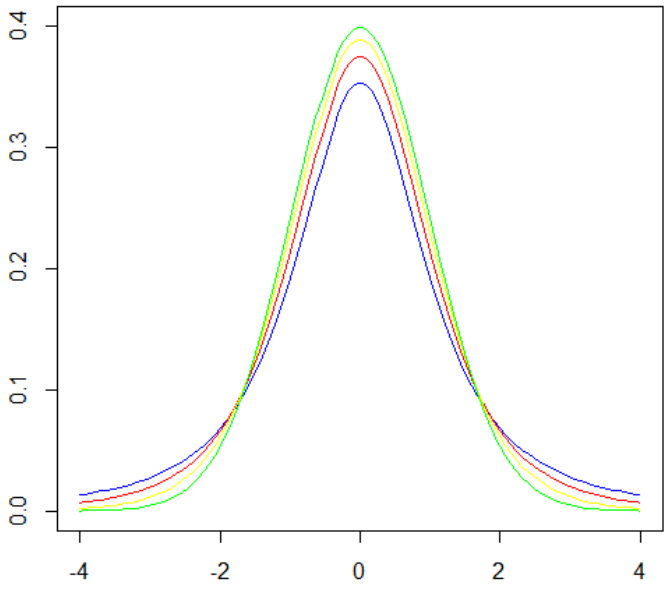

Figure 2. The t-distributions with 2, 4, and 10 (blue, red and yellow, respectively) degrees of freedom and the standard normal distribution (green).

1.1.2.3. Double Exponential distribution

The double exponential distribution is a continuous probability distribution, also sometimes called the Laplace distribution. The double exponential distribution can be thought of as two exponential distributions (with an additional location parameter) spliced together back-to-back. The general formula for the probability density function of the double exponential distribution is:

$$f(x) = \frac{1}{2\beta} e^{-\frac{|x-\mu|}{\beta}}, \quad -\infty < x < +\infty$$

where μ is a location parameter and $\beta > 0$ is a scale parameter. The mean of random variable is μ and variance is $2\beta^{\sim 2}$.

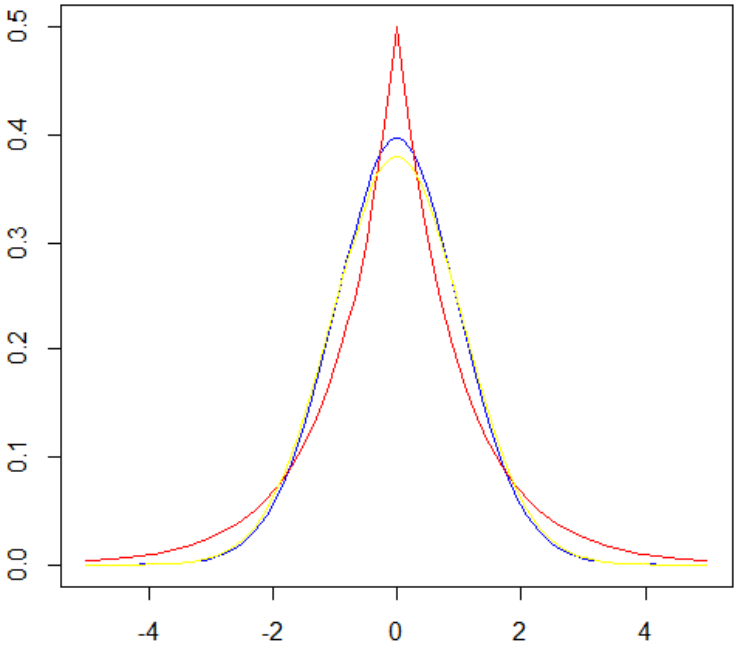

Figure 3. Densities of double exponential (red), normal (yellow) and t (blue) distributions, all with zero means and variances equal to unity.

The double exponential (red), normal (yellow) and t (blue) distributions can be compared in Figure 3. It is observed that the double exponential distribution has a

higher mass density at zero and thicker tails, meaning that it is more prone to producing values closer to zero, and the more extreme values than the normal distribution. The double exponential distribution has a higher density than the t distribution.

1.1.2.4. Scaled inverse chi-square distribution

The scaled inverse chi-squared distribution is the distribution for $x = \frac{1}{y^2}$, being y^2 is a sample mean of the squares of ν independent normal random variables that have mean 0 and inverse variance $\frac{1}{\sigma^2} = S^2$. The distribution is therefore parametrized by the two quantities *v* and S^2, referred to as the number of chi-squared degrees of freedom and the scaling parameter, respectively. The probability density function of the scaled inverse chi-squared distribution is given by:

$$f(x) = \frac{\left(S^2 \frac{v}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} \frac{e^{\left(-\frac{vS^2}{2x}\right)}}{x^{1+\frac{v}{2}}}, \; x > 0.$$

The mean of distribution is $\frac{vS^2}{v-2}$ for $v > 2$ and the variance of distribution is $\frac{2v^2S^4}{(v-2)(v-4)}$ for $v > 4$. The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom.

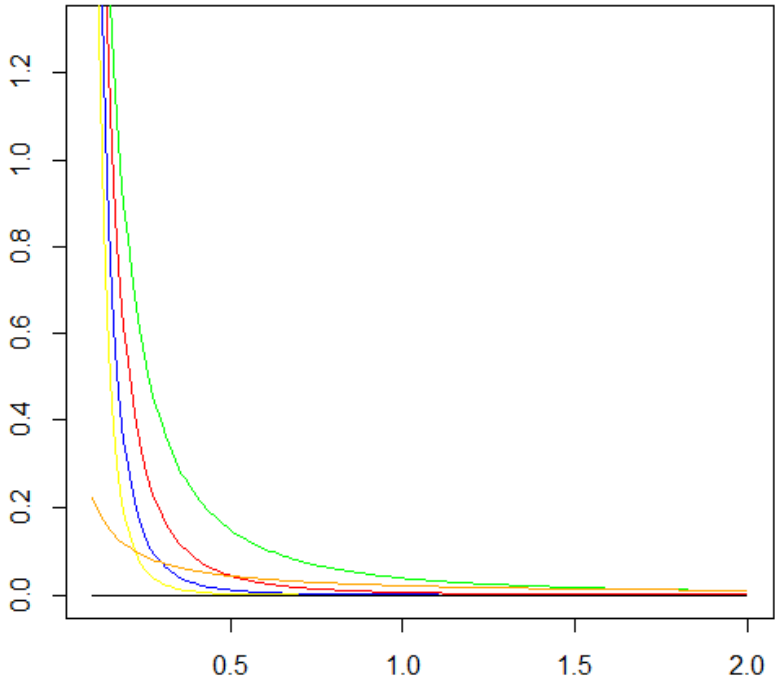

Figure 4. The scaled inverse chi-squared distributions with 0, 0.05, 2, 4, 6, and 8 degrees of freedom and all with scale parameter equal to 0.0429.

The Figure 4 shows scaled inverse chi-squared distributions with 0 (black), 0.05 (orange), 2 (green), 4 (red), 6 (blue), and 8 (yellow) degrees of freedom (ν) and all with scale parameter (S^2) equal to 0.0429. As the number of degrees of freedom approaches zero (orange and black curve) the scaled inverted chi-square distribution becomes approximately equivalent to uniform distribution. However, when the numbers of degrees of freedom deviates from 0, for example 2 (green curve), describe a heavy-tailed distribution. And as the number of degrees of freedom increases the distribution is replaced by a lower density tails.

1.1.3. Additive-Dominance Model for the REML/G-BLUP method

A mixed linear model for individual additive breeding values (u_a) and dominance deviations (u_d) is as follow:

$$y = Xb + Zu_a + Zu_d + e,$$

where y (N × 1, N is number of phenotype and genotype individuals) is a vector of phenotypic observations or corrected phenotypes vector; b (N × 1) is a vector of fixed effects (when using corrected phenotypes is reduced to a unit vector); e (N × 1) is the random error vector, $e \sim N(0, I\sigma_e^2)$ being σ_e^2 is the error variance ; X (N × p, p is number of fixed effects) and Z (N × N) are the incidence matrices for b, u_a and u_d, respectively. The variance structure given by $u_a \sim N(0, G_a\sigma_{u_a}^2)$, $u_d \sim N(0, G_d\sigma_{u_d}^2)$, ($\sigma_{u_a}^2$ is the additive variance, $\sigma_{u_d}^2$ is the dominance variance, G_a and G_d (N × N) are the genomic relationship matrices for additive and dominance effects.).

So mixed model equations to predict u_a and u_d through the G-BLUP method is equivalent to:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_a^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'X & Z'Z & Z'Z + G_d^{-1}\dfrac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u}_a \\ \hat{u}_d \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix} \quad (1)$$

Thus, the genomic breeding value (GBV) of the individual j (j = 1, … , N) is given by $\hat{GBV}_j = \hat{y}_i = \sum_{i=1}^{n} z_{ij}\hat{u}_{aj} + \sum_{i=1}^{n} z_{ij}\hat{u}_{dj}$.

An equivalent model (Goddard et al., 2010) at marker level is given by $y = Xb + ZWm_a + ZSm_d + e$ (2), where:

$u_a = Wm_a$;

$Var(Wm_a) = WI\sigma^2_{m_a}W' = WW'\sigma^2_{m_a}$;

$u_d = Sm_d$;

$Var(Sm_d) = SI\sigma^2_{m_d}S' = SS'\sigma^2_{m_d}$.

W (N × n, n is number of SNPs markers) and S (N × n) are, respectively, the incidence matrices for the vectors of additive (m_a) and dominance (m_d) marker genetic effects. The variance components associated to these effects are $\sigma^2_{m_a}$ and $\sigma^2_{m_d}$, respectively. The quantity m_a in one locus is the allele substitution effect given by $m_{ai} = \alpha_i = a_i + (q_i - p_i)d_i$ (i = 1 ,…, n), where p_i and q_i are allelic frequencies and a_i and d_i are the genotypic values of the homozygote and heterozygote, respectively, at the locus i. By its turn, the quantity m_d can be directly defined as $m_{di} = d_i$. The method admits the assumption that the effects of markers are considered random, normally distributed and homogeneous variance.

The matrices W and S are based on the values 0, 1 and 2 for the number of one the alleles at the i marker locus (putative QTL) in a diploid individual. Several parameterizations are available and the one that matches well with the classical quantitative genetics theory (Falconer and Mackay, 1996) is as follows (Van Raden, 2008; Da et al., 2014; Vitezica, 2013).

Fitting the individual genomic model is the same as fitting the traditional animal model but with the pedigree genetic relationship matrices A and D replaced by the genomic relationship matrices G_a and G_d for additive and dominance effects, respectively. The covariance matrix for the additive effects is given by

$G_a\sigma_a^2 = Var(Wm_a) = WW'\sigma_{ma}^2$, which leads to $G_a = WW'/(\sigma_a^2 / \sigma_{ma}^2)$ $= WW'/\sum_{i=1}^{n}[2p_i(1-p_i)]$, since $\sigma_a^2 = \sum_{i=1}^{n}[2p_i(1-p_i)]\sigma_{ma}^2$. The covariance matrix for the dominance effects is given by $G_d\sigma_d^2 = Var(Sm_d) = SS'\sigma_{md}^2$. So $G_d = SS'/(\sigma_d^2 / \sigma_{md}^2) = SS'/\sum_{i=1}^{n}[2p_i(1-p_i)]^2$, since $\sigma_d^2 = \sum_{i=1}^{n}[2p_i(1-p_i)]^2\sigma_{md}^2$. The correct parameterization in W and S is as follows, according to marker genotypes in a locus i.

$$W = \begin{cases} \text{If MM, then } 2-2p \rightarrow 2q \\ \text{If Mm, then } 1-2p \rightarrow q-p \\ \text{If mm, then } 0-2p \rightarrow -2p \end{cases} \qquad (2)$$

$$S = \begin{cases} \text{If MM, then } 0 \rightarrow -2q^2 \\ \text{If Mm, then } 1 \rightarrow 2pq \\ \text{If mm, then } 0 \rightarrow -2p^2 \end{cases} \qquad (3)$$

The G-BLUP method retains all the covariables (SNPs markers) leading to complex models. Thus, produces good results for the case in which it has many markers with small effect.

1.1.4. Bayesian Ridge Regression (BRR) Method

A Bayesian additive-dominance G-BLUP or Bayesian Ridge Regression (BRR) method assigning flat prior (the priori flat is the noninformative priori) distributions for variance components (i.e. with degrees of freedom equal to -2 which turns out the inverted chi-square into a uniform distribution). The RR are linear functions of the data and its normal prior distributions assumed leads to a homogeneous shrinkage through the markers, which results in many coefficients

(additive and dominance effects) close to zero but not zero, since the infinitesimal model assumes loci with many small effects. BRR is expected to produce similar results as the G-BLUP and RR-BLUP.

1.1.5. Ridge Regression with heterogeneity of variances (RR-HET)

A additive-dominance Ridge Regression (RR-BLUP) method can also be implemented considering the heterogeneity of variances between markers, called RR-HET. By this modification, the shrinkage is no longer homogeneous and becomes specific in accordance with the effect size and variance of the marker. The matrices with specific variances for each marker, $D_a = \mathrm{diag}(\tau^2_{a1}, \tau^2_{a2}, ..., \tau^2_{an})$ and $D_d = \mathrm{diag}(\tau^2_{d1}, \tau^2_{d2}, ..., \tau^2_{dn})$, the elements τ^2_{ai} and τ^2_{di} can be obtained through Bayesian methods. The additive and dominance genetic variance of each marker locus is, simply, given by $\sigma^2_{mai} = \tau^2_{ai}$ and $\sigma^2_{mdi} = \tau^2_{di}$ (with $i = 1, 2, ..., n$), respectively. Therefore, for additive and dominance genetic variance, using the relationships $\sigma^2_a = \sum_{i=1}^{n} 2p_i(1-p_i)\sigma^2_{mai}$ and $\sigma^2_d = \sum_{i=1}^{n} [2p_i(1-p_i)]^2\sigma^2_{mdi}$, we have $\sigma^2_a = \sum_{i=1}^{n} 2p_i(1-p_i)\tau^2_{ai}$ and $\sigma^2_d = \sum_{i=1}^{n} [2p_i(1-p_i)]^2\tau^2_{di}$.

1.1.6. BayesA and BayesB methods

The BayesA and BayesB methods (described by Meuwissen et al., 2001) are advantageous because they can potentially provide information on the genetic architecture of the quantitative trait. Additionally, the Bayesian methods theoretically provide higher accuracy because they force many effects of chromosomal segments

to values near 0 (BayesA) or 0 (BayesB) and the estimated effects are dictated by the prior distributions the QTLs effects.

1.1.7. BayesA method

Meuwissen et al. (2001) presented a methodology for estimating model parameters, in which different components of variance are assigned to each segment (SNP marker) considered in the analysis. The BayesA assume few genes of large effects and many genes of small effect, in other words, specific shrinkage for each marker. This is expected to produce similar results as the RR-HET.

The BayesA method assume the conditional distribution of each marker effect (given its variance) to follow a normal distribution, i.e, $m_{ai} \mid \sigma^2_{mai} \sim N(0, \sigma^2_{mai})$. The variances of the marker effects are assumed to be a scaled inverse chi-square distribution with v degrees of freedom and scale parameter $S^2_{m_a}$, i.e, $\sigma^2_{mai} \sim \chi^{-2}(v_{ma}, S^2_{ma})$. This implies that a larger number of markers presents small effects and a small number of markers presents big effects. Thus, the main advantage of using the scaled inverse chi-square distribution as prior for the variance components is, that when the data have normal distribution, a posterior distribution is also a scaled inverse chi-square distribution.

The use of a mixture of normal (for the effects) and scaled inverse chi-square (for variances) distributions leads to univariate t-distribution of the marker effects with mean zero (Sorensen and Gianola 2002), as follows:

$$p(m_{ai} \mid v_{ma}, S^2_{ma}) = \int_R N(0, \sigma^2_{mai}) \chi^{-2}(v_{ma}, S^2_{ma}) d\sigma^2_{mai} \propto \left(1 + \frac{m^2_{ai}}{v_{ma} S^2_{ma}}\right)^{-(v_{ma}+1)/2},$$

$$m_{ai} \mid v_{m_a}, S^2_{m_a} \sim t(0, v_{m_a}, S^2_{m_a}).$$

Gianola et al. (2009) proved that fitting a variance by locus in this way is equivalent to postulating t distribution for all loci. Thus, the identification of relevant marker effects is more likely in t-BayesA model than in normal-RR-BLUP model. The distributions used in the construction of the joint posterior density allow the use of the Gibbs sampler (Geman and Geman, 1984) to generate samples from the joint a posteriori density (and therefore, the marginal distribution of interest). At the end of the MCMC (Markov Chain Monte Carlo) process, estimates of the effects of each marker are obtained, and therefore the estimates of genomic values of each individual are $\hat{GBV}_j = \hat{y}_j = \sum_{i=1}^{n} w_{ij}\hat{m}_{ai} + \sum_{i=1}^{n} s_{ij}\hat{m}_{di}$ (j = 1 ,…N).

The $S^2_{m_a}$ parameter can be calculated from the additive variance according to Habier et al. (2011). Then, for the additive marker genetic effects we have $E(\sigma^2_{mai}) = \frac{s^2_{ma} v_{ma}}{v_{ma} - 2}$ and $S^2_{ma} = \frac{E(\sigma^2_{ma})(v_{ma} - 2)}{v_{ma}}$. The expectation $E(\sigma^2_{mai})$ is equivalent to $E(\sigma^2_{mai}) = \frac{\sigma^2_a}{\sum_{i=1}^{n} 2 p_i (1 - p_i)}$, where σ^2_a is the additive genetic variance of the trait and p_i is the frequency of marker allele i. As result we get $E(\sigma^2_{mai}) = \frac{\sigma^2_a}{\sum_{i=1}^{n} 2 p_i (1 - p_i)} \frac{(v_{ma} - 2)}{v_{ma}}$.

Alternative values of the parameters of the scaled inverse chi-square distribution are $v_{ma} = 4.012$ or 4.2 and $S^2_{ma} = 0.002$ or 0.0429 (Meuwissen et al., 2001). This describes a moderately leptokurtic distribution. Any value higher than 4 can be used for v_{ma} . This suffices for having an informative distribution. Values

equal or lower than 4 turn the a priori distribution into a flat and non-informative one.

For the residual effects we have $e|\sigma_e^2 \sim N(0,\sigma_e^2)$ and $\sigma_e^2 \sim v_e S_e^2 \chi_{v_e}^2$. Also $E(\sigma_e^2) = \frac{s_e^2 v_e}{v_e - 2}$ and $S_e^2 = \frac{E(\sigma_e^2)(v_e - 2)}{v_e}$. The expectation $E(\sigma_e^2)$ is equivalent to $E(\sigma_e^2) = \tilde{\sigma}_e^2$. So, $S_e^2 = \tilde{\sigma}_e^2 \frac{(v_e - 2)}{v_e} = \tilde{\sigma}_e^2 \frac{(4.2-2)}{4.2}$, where $\tilde{\sigma}_e^2$ is a prioi value of σ_e^2. Zeng et al. (2013) used v equal 4 for both genetic and environmental effects and the true simulated values for $E(\sigma_e^2)$ and $E(\sigma_{ma}^2)$.

For dominance effects at the intra-population level, the distributions are similar as described for additive effects. Then:

$m_{di}|\sigma_{mdi}^2 \sim N(0,\sigma_{mdi}^2)$ for the marker dominance effects;

$\sigma_{mdi}^2 \sim \chi^{-2}(v_{md}, S_{md}^2)$ for the marker dominance variance;

being the marginal the prior distribution for marker dominance effects given by $m_{di}|v_{m_d}, S_{m_d}^2 \sim t(0, v_{m_d}, S_{m_d}^2)$.

Additive and dominance variances are given by $\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i)m_{ai}^2$ and $\sigma_d^2 = \sum_{i=1}^{n} [2p_i(1-p_i)]^2 m_{di}^2$, respectively, according to parameterizations in W and S. The full conditional distributions for the parameters of the BayesA model were presented in details by Zeng et al. (2013).

1.1.8. BayesB method

The problem in the BayesA method is the fact that the variances distribution of haploid effects does not exhibit a mass density at 0. This feature would be of interest to this distribution, since most of the segments have no genetic variance (do not present segregation). This assumption leads to a more favorable statistical condition $\frac{n}{N}$. BayesB has the same assumptions of BayesA for a π fraction of SNPs and assumes that $(1-\pi)$ of the SNPs have no effect, where π is adopted subjectively (or by BayesCpi). It uses a priori density of QTLs effects with mass density in $\sigma^2_{ma_i} = 0$ (and $\sigma^2_{md_i} = 0$) with probability π and $\sigma^2_{mai} \sim \chi^{-2}(v_{ma}, S^2_{ma})$ (and $\sigma^2_{mdi} \sim \chi^{-2}(v_{md}, S^2_{md})$) with probability $(1-\pi)$.

In this case, the prior distributions for the additive and dominant marker effects, respectively, are given by:

$$m_{ai} \mid \pi, \sigma^2_{mai} \sim (1\text{-}I_{ai}) N(0, \sigma^2_{mai} = 0) + I_{ai} N(0, \sigma^2_{mai})$$

$$m_{di} \mid \pi, \sigma^2_{mdi} \sim (1\text{-}I_{di}) N(0, \sigma^2_{mdi} = 0) + I_{di} N(0, \sigma^2_{mdi})$$

where the indicator variable I ($I_{ai} = (0,1)$ and $I_{di} = (0,1)$), so the distributions of $I_a = (I_{a1} ... I_{an})$ and $I_d = (I_{d1} ... I_{dn})$ are binomial with a probability π. If $\pi = 1$, we have BayesA.

In principle it is possible to construct a Gibbs sampler for this approach, but in doing so the Markov chain does not visit all the necessary sample space, since the marginal posterior distribution of $\sigma^2_{ma_i}$ (and $\sigma^2_{md_i}$) do not have the form of a known probability distribution. Thus, it is necessary to use the Metropolis-Hastings algorithm (Gelman et al., 2004), which can generate sequential samples as a mean of

bringing a distribution from which there is no direct sampling. The full conditional distributions for the parameters of the BayesB model were presented in details by Zeng et al. (2013).

1.1.9. LASSO method

The Bayesian regression can be used in situations where there are more markers (covariate) than observations, once the a priori distributions impose the regularization in the model fitting in a way of shortening the regression coefficients (shrinkage). An interesting way to impose this shortening is by LASSO regression (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1996), which combines variable selection (model interpretability) and regularization via shrikaging of the regression coefficients. In general, this consists in obtaining estimates of regression coefficients (2) by solving the following optimization problem:

$$\min\{[y-(1\mu+\sum_{i=1}^{n}w_i m_{ai}+\sum_{i=1}^{n}s_i m_{di})]'[y-(1\mu+\sum_{i=1}^{n}w_i m_{ai}+\sum_{i=1}^{n}s_i m_{di})]+$$
$$\lambda_a\sum_{i=1}^{n}|m_{ai}|+\lambda_d\sum_{i=1}^{n}|m_{di}|\}$$

where $\sum_{i=1}^{n}|m_{ai}|,\sum_{i=1}^{n}|m_{di}|$ are the sum of the absolute values of the regression coefficients, λ_a and λ_d are parameters that controls the strength of the smoothing, so that when $\lambda_a=0$ (or $\lambda_d=0$) there is no regularization. The LASSO solution allows to $N-1$ non-zero regression coefficients, where N is the number of individuals.

According Resende et al. (2012) the choice of lambda (λ) is very important because it influences the markers group size with different effect from 0 (selected

markers). Thus, as λ approaches zero the solution converges to the method of least squares (no selection and unstable predictive covariates). While high values of λ greatly reduce the values of the regression coefficients. Thus, for the optimally calculation of λ, Usai et al. (2009) proposed an algorithm based in cross-validation, which is called minimum angle regression (LARS).

1.1.10. BLASSO method

The Bayesian version of the LASSO regression (BLASSO - Park and Casella, 2008) for genomic selection was designed by De Los Campos et al. (2009). The BLASSO includes a common variance term to model both terms, the residuals and genetic effects of the markers. Therefore, using the same model equation used previously for the estimation of the markers effects, we have:

$$e \mid \sigma^2 \sim MVN(0, I\sigma^2)$$

$$m_{ai} \mid \lambda_a, \sigma^2 \sim \prod_i \left(\frac{\lambda_a}{2\sigma} \right) e^{\left[\frac{-\lambda_a |m_{ai}|}{\sigma} \right]}$$

$$m_{di} \mid \lambda_d, \sigma^2 \sim \prod_i \left(\frac{\lambda_d}{2\sigma} \right) e^{\left[\frac{-\lambda_d |m_{di}|}{\sigma} \right]},$$

where MNV represents a multivariate normal distribution, λ is the "sharpness" parameter and the σ^2 prior distribution consists of a scaled inverse chi-square. The parameter λ_a (or λ_d) can be estimated from the data by MCMC methods (using a non-informative prior) and MCEM (Monte Carlo EM - does not require a priori information).

The prior distribution used in Bayesian LASSO shows greater density of mass at zero and more robust tails, putting further shrinkage over the regression

coefficients next to 0 and less shrinkage on regression coefficients distant from zero. Thus, posterior averages are estimated, producing very small values, but not zero as the original LASSO. Using a formulation in terms of an augmented hierarchical model, we have:

$$p(m_{ai} \mid \tau_a) \sim N(0, D_a \sigma^2), \; p(m_{di} \mid \tau_d) \sim N(0, D_d \sigma^2)$$

$$p(\tau_a^2 \mid \lambda_a^2) = \prod_i \left(\frac{\lambda_a^2}{2} \right) e^{\left[\frac{-\lambda_a^2 \tau_{ai}^2}{2} \right]} \text{ and } p(\tau_d^2 \mid \lambda_d^2) = \prod_i \left(\frac{\lambda_d^2}{2} \right) e^{\left[\frac{-\lambda_d^2 \tau_{di}^2}{2} \right]}$$

where $D_a = \text{diag}(\tau_{a1}^2, \tau_{a2}^2, ..., \tau_{an}^2)$ and $D_d = \text{diag}(\tau_{d1}^2, \tau_{d2}^2, ..., \tau_{dn}^2)$. This leads to double exponential distribution of the marker effects (Park and Casella, 2008), as follows:

$$p(m_{ai} \mid \lambda_a^2) = \int_R N(0, \sigma^2 \tau_{ai}^2) Exp\left(\frac{\lambda_a^2}{2} \right) d\tau_{ai}^2 \propto \frac{1}{2\left(\frac{\sigma}{\lambda_a} \right)} e^{\left(-\frac{m_{ai}}{\sigma / \lambda_a} \right)},$$

$$p(m_{di} \mid \lambda_d^2) = \int_R N(0, \sigma^2 \tau_{di}^2) Exp\left(\frac{\lambda_d^2}{2} \right) d\tau_{ai}^2 \propto \frac{1}{2\left(\frac{\sigma}{\lambda_d} \right)} e^{\left(-\frac{m_{di}}{\sigma / \lambda_d} \right)}$$

$$m_{ai} \mid \lambda_a^2 \sim \text{DoubleExp}\left(0, \frac{\sigma}{\lambda_a} \right) \text{ and } m_{di} \mid \lambda_d^2 \sim \text{DoubleExp}\left(0, \frac{\sigma}{\lambda_d} \right).$$

Bayesian Lasso are advantageous compared to Bayesian methods of Meuwissen et al. (2001) because it is asymptotically free of priori information, i.e., provides better learning from data (Gianola, 2013; Gianola et al., 2009). This occurs because in the hierarchical models, such as BLASSO a priori information is assigned to the hyperparameters so that the influence of this information disappears asymptotically (Resende et al., 2012).

The additive and dominance genetic variance of each marker locus is given by $\sigma_{mai}^2 = \tau_{ai}^2 \sigma^2$ and $\sigma_{mdi}^2 = \tau_{di}^2 \sigma^2$ (with $i = 1, 2, ..., n$), respectively. Therefore, additive

and dominance genetic variance, using the relationships $\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i)\sigma_{ai}^2$ and $\sigma_d^2 = \sum_{i=1}^{n} [2p_i(1-p_i)]^2\sigma_{di}^2$, we have:

$$\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i)\tau_{ai}^2\sigma^2 \text{ and } \sigma_d^2 = \sum_{i=1}^{n} [2p_i(1-p_i)]^2\tau_{di}^2\sigma^2 .$$

The full conditional distributions for the parameters of the BLASSO model were presented in details by De Los Campos et al. (2009).

1.1.11. IBLASSO method

Legarra et al. (2011) proposed the improved BLASSO (IBLASSO) method, which uses two terms of variance, one for shaping the residuals and the other for shaping the genetic effects (additive and dominant) of markers. The parameters of IBLASSO is equivalent to the original LASSO of Tibshirani (1996), but the implementation is bayesian and does not assumes that the incidence matrices markers (W and S) are standardized.

Using a formulation in terms of an augmented hierarchical model including the extra variance components τ_{ai}^2 and τ_{di}^2 associated with each marker locus, we have:

$$e \mid \sigma^2 \sim MVN(0, I\sigma^2)$$

$$p(m_{ai} \mid \tau_a) \sim N(0, D_a), \; p(m_{di} \mid \tau_d) \sim N(0, D_d)$$

$$p(\tau_a^2 \mid \lambda_a^2) = \prod_i \left(\frac{\lambda_a^2}{2}\right) e^{\left[\frac{-\lambda_a^2\tau_{ai}^2}{2}\right]} \text{ and } p(\tau_d^2 \mid \lambda_d^2) = \prod_i \left(\frac{\lambda_d^2}{2}\right) e^{\left[\frac{-\lambda_d^2\tau_{di}^2}{2}\right]}$$

where $D_a = \mathrm{diag}(\tau_{a1}^2, \tau_{a2}^2, ..., \tau_{an}^2)$, $D_d = \mathrm{diag}(\tau_{d1}^2, \tau_{d2}^2, ..., \tau_{dn}^2)$ and the σ^2 prior distribution consists of a scaled inverse chi-square. This leads to double exponential distribution of the marker effects (Legarra et al., 2011), as follows:

$$p(m_{ai} \mid \lambda_a^2) = \int_R N(0, \tau_{ai}^2) \mathrm{Exp}\left(\frac{\lambda_a^2}{2}\right) d\tau_{ai}^2 \propto \frac{\lambda_a}{2} e^{\left(-|m_{ai}|\frac{\lambda_a}{2}\right)},$$

$$p(m_{di} \mid \lambda_d^2) = \int_R N(0, \tau_{di}^2) \mathrm{Exp}\left(\frac{\lambda_d^2}{2}\right) d\tau_{di}^2 \propto \frac{\lambda_d}{2} e^{\left(-|m_{di}|\frac{\lambda_d}{2}\right)},$$

$$m_{ai} \mid \lambda_a^2 \sim \mathrm{DoubleExp}\left(0, \frac{1}{\lambda_a}\right) \text{ and } m_{di} \mid \lambda_d^2 \sim \mathrm{DoubleExp}\left(0, \frac{1}{\lambda_d}\right).$$

The additive and dominance genetic variance of each marker locus is, simply, given by $\sigma_{mai}^2 = \tau_{ai}^2$ and $\sigma_{mdi}^2 = \tau_{di}^2$ (with $i = 1, 2, ..., n$), respectively. Therefore, additive and dominance genetic variance, using the relationships $\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i)\sigma_{mai}^2$ and $\sigma_d^2 = \sum_{i=1}^{n} [2p_i(1-p_i)]^2 \sigma_{mdi}^2$, we have:

$$\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i)\tau_{ai}^2 \text{ and } \sigma_d^2 = \sum_{i=1}^{n} [2p_i(1-p_i)]^2 \tau_{di}^2 .$$

The full conditional distributions for the parameters of the IBLASSO model were presented in details by Legarra et al. (2011).

Table 1. Distributions of genetic effects on RR-BLUP, Bayes and Lasso methods. (Resende et al., 2012).

Method	Prior distribution of effects	Prior distribution of variances	Posterior distribution of variances
Bayesian RR	Normal with common variance	non informative scaled inverse chi-squared	scaled inverse chi-squared
Bayesian RR-HET	Normal with heterogeneous variances	non informative scaled inverse chi-squared	scaled inverse chi-squared
Bayes A	Normal with heterogeneous variances (t given prior scaled inverse chi-squared for the variances)	scaled inverse chi-squared (equivalent to Bayes B when $\pi = 1$)	scaled inverse chi-squared
Bayes B	Normal with heterogeneous variances (t given prior scaled inverse chi-squared for the variances)	Mixture of distributions 0 with probability $(1-\pi)$ and scaled inverse chi-squared with probability π	scaled inverse chi-squared
Lassos	Double exponential	Double exponential	Inverse Gamma

2. References

Da Y., Wang C., Wang S., Hu G. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. **PLoS One**, 30,9(1):e87666.

Daetwyler H. D., Kemper K. E., Van Der Werf J. H. J., Hayes B. J. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. **Journal of Animal Science**, 90:3375–3384.

Daetwyler, H. D., B. Villanueva, J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. **PLoS One**, 3:e3395.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, et al. 2010. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, 185:1021–1031.

De Los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D., Callus M. P. L. 2012. Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, 193:327-345.

De Los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K., Cotes J. M. 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. **Genetics**, 182(1):375-385.

Denis M., Bouvet J. M. 2013. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. **Tree Genetics and Genomics**, 9:37-51.

Estaghvirou, S. B. O., J. O. Ogutu, T. Schulz-Streeck, C. Knaak, M. Ouzunova, A. Gordillo, H.P. Piepho. 2013. Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. **BMC Genomics**, 14, 860.

Falconer D. S., Mackay T. F. C. 1996. **Introduction to Quantitative Genetics**, Ed 4. Longmans Green, Harlow, Essex, UK.

Gelman A., Carlin J. B., Stern H. S., Rubin D. B. 2004. **Bayesian Data Analysis**. Chapman & Hall, London.

Geman S., Geman D. 1984. Stochastic relaxation, Gibbs distribution and the bayesian restoration of imagens. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 6:721-741.

Gianola D. 2013. Priors in whole-genome regression: the bayesian alphabet returns. **Genetics**, 194(3):573-96.

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

Gianola, D., Perez-Enciso M., Toro M. A. 2003. On marker-assisted prediction of genetic value: beyond the ridge. **Genetics**, 163:347-365.

Goddard M. E., Hayes B. J. 2007. Genomic selection. **Journal of Animal Breeding and Genetics**, 124:323-330.

Goddard M. E., Hayes B. J., Meuwissen T. H. E. 2010. Genomic selection in livestock populations. **Genetics Research**, 92:413–421.

Goddard M. E., Wray N. R., Verbyla K., Visscher P. M. 2009. Estimating effects and making predictions from genome-wide marker data. **Statistical Science**, 24:517-529.

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

Hayes, B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. 2009a. Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, 92: 433-443.

Hayes, B. J., P. M. Visscher, M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research, 91:47-60.

Hu G., Wang C., Da Y. 2014. Genomic heritability estimation for the early life-history transition related to propensity to migrate in wild rainbow and steelhead trout populations. **Ecology and Evolution**, 4(8): 1381–1388.

Legarra A., Robert-Granié C., Croiseau P., Guillaume F., Fritz S. 2011. Improved Lasso for genomic selection. **Genetics Research**, 93(1):77-87.

Legarra A., Robert-Granie C., Manfredi E., Elsen J. M. 2008. Performance of genomic selection in mice. **Genetics**, 180:611-618.

Lehermeier C., Wimmer V., Albrecht T., Auinger H. J., Gianola D., Schimid V. J. Schön C. C. 2013. Sensitivity to prior specification in Bayesian genome-based prediction models. **Statistical Applications in Genetics and Molecular Biology**, 12(3):375-391.

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.

Muñoz P. R., Resende Jr M. F. R., Gezan S. A., Resende M. D. V., De los Campos G, Kirst M. Huber D., Peter G. F. 2014. Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. **Genetics**, 198:1759-1768.

Park T., Casella G. 2008. The Bayesian LASSO. **Journal of the American Statistical Association**, 103(482):681-686.

Resende Jr M. F. R., Valle P. R. M., Resende M. D. V., Garrick D. J., Fernando R. L., Davis J. M., Jokela E. J., Martin T. A., Peter G. F., Kirst M. 2012. Accuracy of genomic selection methods in a standard dataset of loblolly pine. **Genetics**, 190:1503 - 1510.

Resende M. D. V., Silva F. F., Lopes P. S., Azevedo C. F. 2012. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística, 291 p. Disponível em: <http://www.det.ufv.br/ppestbio/corpo_docente.php>.

Resende, M. D. V. de, P. S. Lopes, R. L. Silva, I. E. Pires. 2008. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, 56:63-78.

Sorensen D., Gianola D. 2002. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. New York: Springer Verlag 740 p.

Su G., Christensen O. F., Ostersen T., Henryon M., Lund M. S. 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. **PLoS One**, 7(9):e45293.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistics Society Series B**, 58:267-288.

Toro M. A., Varona L. 2010. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, 42:33.

Usai M. G., Goddard M. E., Hayes B. J. 2009. LASSO with cross-validation for genomic selection. **Genetics Research**, 91(6): 427-36.

Van Raden P. M. 2008. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, 91(11):4414-4423.

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(4):1223-30.

Wang C., Da Y. 2014. Quantitative Genetics Model as the Unifying Model for Defining Genomic Relationship and Inbreeding Coefficient. **PLoS ONE**, 9(12):e114484.

Wang C., Prakapenga D., Wang S., Puligurta S., Runesha H. B., Da Y. 2014. GVCBLUP: a computer package for genomic prediction and variance

component estimation of additive and dominance effects. **BMC Bioinformatics**, 15:270.

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

Zeng J., Toosi A., Fernando R. L., Dekkers J. C. M., Garrick D. J. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genetics Selection Evolution**, 45:11.

CHAPTER 2

Published as original paper to BMC Genetics

Reference: AZEVEDO, C. F.; Resende, M.D.V.; SILVA, F. F.; VIANA, J. M. S.; VALENTE, M. S.; RESENDE, M. F. R.; MUNOZ, P. Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genetics (Online), v. 16, p. 105, 2015.

RIDGE, LASSO AND BAYESIAN ADDITIVE-DOMINANCE GENOMIC MODELS

Abstract

Background

A complete approach for genome-wide selection (GWS) involves reliable statistical genetics models and methods. Reports on this topic are common for additive genetics models but not for additive-dominance models. This paper aimed at: (i) comparing the performance of 10 additive-dominance prediction models (including current ones and proposed modifications), fitted through Bayesian, Lasso and Ridge regression approaches; (ii) decomposing genomic heritability and accuracy in terms of the three quantitative genetics information compounds named linkage disequilibrium (LD), co-segregation (CS) and pedigree relationships or family structure (PR). The simulation study considered two broad sense heritability levels (0.30 and 0.50, associated to narrow sense heritabilities of 0.20 and 0.35, respectively) and two genetic architecture (the first one consisting of small gene effects and the other one with mixed inheritance model with five major genes) for the traits.

Results

G-REML/G-BLUP and a modified Bayesian/Lasso (called BayesA*B* or t-BLASSO) methods performed best in the prediction of genomic breeding and total genotypic values of individuals, in all the four scenarios (two heritabilities × two

genetic architectures). The BayesA*B*-type methods showed better ability for recovering the dominance variance/additive variance ratio. The decomposition of genomic heritability and accuracy revealed the following descending importance order of information: LD, CS and PR not captured by markers, being these last two very close.

Conclusions

Amongst the 10 models/methods evaluated, the G-BLUP, BAYESA*B* (-2,8) and BAYESA*B* (4,6) methods presented the best results and were found to be adequate for accurately predicting genomic breeding and total genotypic values as well as for estimating additive and dominance in additive-dominance genomic models.

Key words: Dominance Genomic Models; Bayesian Methods; Lasso Methods; Selection Accuracy.

1. Background

Genome wide selection (GWS) concerns to phenotype prediction and relies on simultaneous prediction of a great amount of molecular markers effects, thus characterizing as a new paradigm in quantitative genetics (Meuwissen et al., 2001; Gianola et al., 2009) and plant and animal breeding (Goddard and Hayes, 2007; Meuwissen, 2007; Van Raden, 2008; Resende et al., 2008; Grattapaglia and Resende, 2011; Endelman and Jannink, 2012; Resende Jr. et al., 2012a and b).

Also in genome-wide association studies (GWAS) the simultaneous prediction of marker effects are now of common use (Visscher at al., 2010; Yang et al, 2010; Goddard et al., 2009) and so GWS methods are also being applied to human genetics, gene discovery and association genetics.

Recent methodologies for GWS and GWAS has been evaluated through simulation studies (Piccoli et al., 2014; Talluri et al., 2014). Simulation and practical results with additive models in GWS with several organisms are common (Van Raden et al. , 2009; Hayes et al., 2009; Grattapaglia and Resende, 2011; Silva et al., 2011; Resende et al., 2012; Resende Jr. et al., 2012a and b; Oliveira et al., 2012; Zapata-Valenzuela et al., 2013; Muñoz et al., 2014). But additive-dominance models are much less common (Zeng et al., 2013; Muñoz et al., 2014; Su et al., 2012; Denis and Bouvet, 2013).

Hill et al. (2008), Bennewitz and Meuwissen (2010) and Wellman and Bennewitz (2012) discussed the relevance of dominance models for the Quantitative Genomics and Genetics. Wellmann and Bennewitz (2012) presented theoretical genetic models for Bayesian genomic selection with dominance and concluded that dominance enhances the analysis and several advantages are aggregated.

Wang and Da (2014) established the correct definitions of genomic relationships and inbreeding, which came to unify the prediction models for additive-dominance genomic selection. Da et al. (2014) and Wang et al. (2014) presented software for additive-dominance models in the framework of the G-BLUP method. Zapata-Valenzuela et al. (2013) also tried to fit additive-dominance models but convergence failed for the main traits.

Dominance estimation is essential specially for vegetative propagated species (Denis and Bouvet, 2013) and crossed populations, where the mating allocation including both additive and dominance is an effective way of increasing genetic gain capitalizing on heterosis (Toro and Varona, 2010; Wellmann and Bennewitz, 2012). Additive-dominance models are able to capture both effects, allowing effective selection of parents, crosses and clones. This allows taking full advantage of

genomic selection in perennials and asexually propagated crops and also in crossed animals.

Bayesian, Lasso and Ridge regression approaches have not been compared for additive-dominance models yet. Zeng et al. (2013), Muñoz et al. (2014), Su et al. (2012) and Denis and Bouvet (2013) and Wang and Da (2014) applied only the G-BLUP method, which is an equivalent model (Goddard et al., 2009), to ridge regression (RR-BLUP). On the other hand, Wellmann and Bennewitz (2012) applied only the Bayesian methods of Meuwissen et al. (2001) with modifications (mixture of two t distributions, one of them with a small variance). Toro and Varona (2010) evaluating the introduction of the dominant effects in the model using the Bayes A. Lasso methods seem to be unused with dominance models for variance components in genomic selection. The study of partition of accuracy and heritability due to the three quantitative genetics information, linkage disequilibrium (LD), co-segregation (CS) and pedigree relationships (PR) has been explored only by Habier et al. (2013).

Given the scarcity of papers on dominance genomic models in the literature and, aiming at the increasing the knowledge and the enrichment of the discussions of such an essential topic in this field, this paper has a two-fold objective: (i) to evaluate 10 estimation methods (including Bayesian, Lasso and Ridge regression approaches) for fitting additive-dominance genomic models for GWS; (ii) to decompose genomic heritability and accuracy in terms of the three quantitative genetics information compounds LD, CS and PR.

2. Methods

2.1. Simulated datasets

Two random mating populations in linkage equilibrium were crossed generating a population (of size 5,000, coming from 100 families) with linkage disequilibrium (LD), which was subjected to five generations of random mating without mutation, selection or migration. The resultant population is an advanced generation composite, which presents Hardy-Weinberg equilibrium and LD. According to Viana (2004), the LD value (Δ) in a composite population is $\Delta_{ab} = \left(\frac{1-2\theta_{ab}}{4}\right)\left(p_a^1 - p_a^2\right)\left(p_b^1 - p_b^2\right)$, where a and b are two SNPs, two QTLs, or one SNP and one QTL, θ is the frequency of recombinant gametes, and p^1 and p^2 are the allele frequencies in the parental populations (1 and 2). Notice also that the LD value depends on the allele frequencies in the parental populations. Thus, regardless of the distance between the SNPs and/or QTLs, if the allele frequencies are equal in the parental population, $\Delta = 0$. The LD is maximized ($|\Delta| = 0.25$) when $\theta = 0$ and $|p^1 - p^2| = 1$. In this case, the LD value is positive with coupling and negative with repulsion (Viana, 2004).

From the advanced generation of the composite, one thousand individuals were generated with diploid genomes having a length of 200 centimorgans (cM) (L = 2 Morgans) and assuming ten equally sized chromosomes, each one with two haplotypes. We simulated a marker density by assigning 2,000 equidistant SNP markers that were separated by 0.1 cM across the ten chromosomes. One hundred of the 2,000 markers were actually genes (QTL). A total of 1,000 individuals that came from the same generation and from 20 full-sib families (each one with 50 individuals) were genotyped and phenotyped. This simulation provides a typical

small effective population size (Ne = 39.22) and a large LD in the breeding populations. Ne of approximately 40 and the use of 50 individuals per family are typical values in elite breeding populations of plant species (Resende, 2002).

The QTLs were distributed in the regions covered by the SNPs. For each trait, we informed the degree of dominance (d/a being a and d are the genotypic values for one homozygote and heterozygote, respectively) and the direction of dominance (positive and/or negative). The obtained genotypic values for homozygotes were within the limits of Gmax = 100(m + a) and Gmin = 100(m - a) where m is the mean of genotypic values, which are the maximum and minimum values, respectively.

Goddard et al. (2011) presented the realized proportion (r^2_{mq}) of genetic variation explained by the markers as $r^2_{mq} = \frac{n}{n + n_{QTL}}$, where n_{QTL} is the number of QTL and n is the number of markers. With n = 2,000 markers and n_{QTL} = 100, we have $r^2_{mq} = 0.95$. An alternative Hayes et al. (2009) takes $n_{QTL} = 2NeL = 2(39.22)2 = 156.88$, where L is the total length of the genome (in Morgans), producing $r^2_{mq} = 0.93$. Another approach Sved (1971) provides r^2_{mq} as $r^2_{mq} = \frac{1}{1 + 4NeS} = \frac{1}{1 + 4(39.22)0.001} = 0.86,$ where S is the distance between markers (in Morgans). These values reveal that the genome was sufficiently saturated by markers.

Traits with two genetic architectures were simulated, one following the infinitesimal model and the other with five major effects genes accounting for 50% of the genetic variability. For the former, to each of 100 QTL one additive effect of small magnitude on the phenotype was assigned (under the Normal Distribution

setting). For the latter, small additive effects were assigned to the remaining 95 loci. The effects were normally distributed with zero mean and genetic variance (size of genetic effects) allowing the desired heritability level. The phenotypic value was obtained by adding to the genotypic value a random deviate from a normal distribution N($0, \sigma_e^2$), where the variance σ_e^2 was defined according to two levels of broad-sense heritability, 0.30 and 0.50, associated with narrow-sense heritabilities of approximately 0.20 and 0.35, respectively. Heritability levels were chosen to represent one trait with low heritability and another with moderate heritability, which addressed the cases where genomic selection is expected to be superior to phenotypic selection (Meuwissen et al., 2001). The magnitudes of the narrow-sense and broad-sense heritabilities are associated with an average degree of dominance level (d/a) of approximately 1 (complete dominance) in a population with intermediate allele frequencies. Simulations assumed independence of additive and dominance effects, with dominance effects having the same distribution as the additive effects (both were normally distributed with zero mean). In the simulation, it was also observed that marker alleles had minor allele frequency (MAF) greater than 5%.

The data were simulated using the RealBreeding software (Viana, 2011).

2.2. Scenarios

For the populations of full-sib families, four scenarios were studied: two broad-sense heritability levels (approximately 0.30 and 0.50) × two genetic architectures. The scenarios were analyzed using 10 statistical methods (Table 1).

Table 1. Softwares.

Method	Full Name of the Method	Class of Methods	DF1	DF2	Software
BRR (-2,-2)	Bayesian Ridge Regression	Bayesian	-2	-2	GS3
IBLASSO (4,-2)	Improved Bayesian Lasso	Bayesian Lasso	4	-2	GS3
IBLASSO (4,2)	Improved Bayesian Lasso	Bayesian Lasso	4	2	GS3
BAYESA*B* (-2,6)	IBLASSO with t distribution	Bayesian Lasso	-2	6	GS3
BAYESA*B* (4,6)	IBLASSO with t distribution	Bayesian Lasso	4	6	GS3
BAYESA*B* (-2,8)	IBLASSO with t distribution	Bayesian Lasso	-2	8	GS3
RR-HET (-2, -2)	RR-BLUP with heterogeneous variance	Ridge Regression	-2	-2	GS3
BLASSO (4,2)	Bayesian Lasso	Bayesian Lasso	4	2	BLR-R
G-BLUP	Genomic BLUP	Random Regression	-	-	GVC
Pedigree-BLUP	Pedigree-BLUP	Random Regression	-	-	Pedigreemm-R

Description of the fitted models and softwares used.

DF1: Degrees of Freedom of the chi-square prior distribution for the residual variance;

DF2: Degrees of Freedom of the chi-square prior distribution for genetic variance or shrinkage parameter.

2.3. Statistical Methods for Additive-Dominance Models

2.3.1. Additive-Dominance Model for the REML/G-BLUP method

A mixed linear model for individual additive breeding values (u_a) and dominance deviations (u_d) is as follow $y = Xb + Zu_a + Zu_d + e$, with variance structure given by $u_a \sim N(0, G_a\sigma^2_{u_a})$; $u_d \sim N(0, G_d\sigma^2_{u_d})$; $e \sim N(0, I\sigma^2_e)$. An equivalent model (Goddard et al., 2010) at marker level is given by $y = Xb + ZWm_a + ZSm_d + e$, where:

$u_a = Wm_a$;

$Var(Wm_a) = WI\sigma^2_{m_a} W' = WW'\sigma^2_{m_a}$;

$u_d = Sm_d$;

$Var(Sm_d) = SI\sigma^2_{m_d} S' = SS'\sigma^2_{m_d}$.

W and S are, respectively, the incidence matrices for the vectors of additive (m_a) and dominance (m_d) marker genetic effects. The variance components associated to these effects are $\sigma^2_{m_a}$ and $\sigma^2_{m_d}$, respectively. G_a and G_d are the genomic relationship matrices for additive and dominance effects. The quantity m_a in one locus is the allele substitution effect given by $m_{ai} = \alpha_i = a_i + (q_i - p_i)d_i$, where p_i and q_i are allelic frequencies and a_i and d_i are the genotypic values of one of the homozygotes and heterozygote, respectively, at the locus i. By its turn, the quantity m_d can be directly defined as $m_{di} = d_i$.

The matrices W and S, which will be defined later, are based on the values 0, 1 and 2 for the number of one of the alleles at the i marker locus (putative QTL) in a diploid individual. Several parameterizations are available and the one that matches well with the classical quantitative genetics theory (Falconer and Mackay, 1996) is as follows (Van Raden, 2008;Vitezica, 2013; Wang and Da, 201; Da et al., 2014).

Fitting the individual genomic model is the same as fitting the traditional animal model but with the pedigree genetic relationship matrices A and D replaced by the genomic relationship matrices G_a and G_d for additive and dominance effects, respectively. The covariance matrix for the additive effects is given by $G_a\sigma^2_a = Var(Wm_a) = WW'\sigma^2_{ma}$, which leads to $G_a = WW'/(\sigma^2_a / \sigma^2_{ma})$ $= WW'/\sum_{i=1}^{n}[2p_i(1-p_i)]$, since $\sigma^2_a = \sum_{i=1}^{n}[2p_i(1-p_i)]\sigma^2_{ma}$. The covariance matrix for

the dominance effects is given by $G_d\sigma_d^2 = Var(Sm_d) = SS'\sigma_{md}^2$. So $G_d = SS'/(\sigma_d^2 / \sigma_{md}^2) = SS'/\sum_{i=1}^{n}[2p_i(1-p_i)]^2$, since $\sigma_d^2 = \sum_{i=1}^{n}[2p_i(1-p_i)]^2\sigma_{md}^2$. The correct parameterization in W and S is as follows, according to marker genotypes in a locus i.

$$W = \begin{cases} \text{If MM, then } 2-2p \rightarrow 2q \\ \text{If Mm, then } 1-2p \rightarrow q-p \\ \text{If mm, then } 0-2p \rightarrow -2p \end{cases} \quad (1)$$

$$S = \begin{cases} \text{If MM, then } 0 \rightarrow -2q^2 \\ \text{If Mm, then } 1 \rightarrow 2pq \\ \text{If mm, then } 0 \rightarrow -2p^2 \end{cases} \quad (2)$$

Additive-dominance G-BLUP method was fitted using the GVC-BLUP software (Wang et al., 2014) via REML through mixed model equations.

2.3.2. Bayesian Ridge Regression (BRR) Method

A Bayesian additive-dominance G-BLUP or Bayesian Ridge Regression (BRR) method was fitted using the GS3 software (Legarra et al., 2013) via MCMC-REML/BLUP assigning flat (i.e. with degrees of freedom equal to -2 which turns out the inverted chi-square into a uniform distribution) prior distributions for variance components (the priori flat is the noninformative priori).

2.3.3. BayesA and BayesB methods

The BayesA and BayesB methods (described by Meuwissen et al., 2001) are advantageous because they can potentially provide information on the genetic architecture of the quantitative trait.

In these methods, specific variances are allowed to each locus. Additionally, BayesB performs variable selection, since the majority of the markers are not in LD with the genes. Thus, a set of markers associated to a trait must be identified. BayesB method subjectively determines π, the proportion of markers having effects. Using the indicator variable I, in the BayesA and BayesB model the additive genetic effect of individual j is defined as $a_j = \sum_{i=1}^{n} m_{ai} w_{ij} I_{ai}$, where $I_{ai} = (0,1)$. The distribution of $I_a = (I_{a1} ... I_{an})$ is binomial with a probability π, which is 1 for BayesA and subjectively determined for BayesB. The quantities for w_{ij} are elements of the marker genotype matrix W. Dominance effects are coded in a similar way $d_j = \sum_{i=1}^{n} m_{di} s_{ij} I_{di}$.

These Bayesian methods assume the conditional distribution of each marker effect (given its variance) to follow a normal distribution, i.e, $m_{ai} | \sigma^2_{mai} \sim N(0, \sigma^2_{mai})$. The variances of the marker effects are assumed to be a scaled inverse chi-square distribution with v degrees of freedom and scale parameter $S^2_{m_a}$, i.e, $\sigma^2_{mai} \sim \chi^{-2}(v_{ma}, S^2_{ma})$. This implies that a larger number of markers presents small effects and a small number of markers presents big effects. This leads to univariate t-distribution of the marker effects with mean zero (Sorensen and Gianola, 2002). Gianola et al. (2009) proved that fitting a variance by locus in this way is equivalent to postulating t distribution for all loci. Thus, the identification of relevant marker effects is more likely in t-BayesA model than in normal-RR-BLUP model.

For the Bayes methods, the marginal prior distribution for additive marker effects is $m_{ai} | v_{m_a}, S^2_{m_a} \sim t(0, v_{m_a}, S^2_{m_a})$. The combination of normal (for marker effects)

and inverse chi-square distributions (for variances) leads to a t distribution for m_{ai}, and so with a longer tail than that for normal distribution. In this paper, the values 6 and 8 were assigned for v as to provide sufficiently thick tails associated to t distributions (Gianola, 2013) and $S^2_{m_a}$ was calculated from the additive variance according to Habier et al. (2011).

The value of $S^2_{m_a}$ can be derived according to the expected value of a random variable distributed as scaled inverse chi-square. This mathematical expectation for a generic variance component σ^2 is given by $E(\sigma^2) = \frac{S^2 v}{v-2}$, which is the variance of the prior. So the scale parameter is given by $S^2 = \frac{E(\sigma^2)(v-2)}{v}$. Then, for the additive marker genetic effects we have $E(\sigma^2_{mai}) = \frac{s^2_{ma} v_{ma}}{v_{ma}-2}$ and $S^2_{ma} = \frac{E(\sigma^2_{ma})(v_{ma}-2)}{v_{ma}}$. The expectation $E(\sigma^2_{mai})$ is equivalent to $E(\sigma^2_{mai}) = \frac{\sigma^2_a}{\sum_{i=1}^{n} 2 p_i (1-p_i)}$, where σ^2_a is the additive genetic variance of the trait and p_i is the frequency of marker allele i. As result, $E(\sigma^2_{mai}) = \frac{\sigma^2_a}{\sum_{i=1}^{n} 2 p_i (1-p_i)} \frac{(v_{ma}-2)}{v_{ma}}$.

Alternative values of the parameters of the scaled inverse chi-square distribution are $v_{ma} = 4.012$ or 4.2 and $S^2_{ma} = 0.002$ or 0.0429 (Meuwissen et al., 2001). This describes a moderately leptokurtic distribution. Any value higher than 4 can be used for v_{ma}. This suffices for having an informative distribution. Values

equal or lower than 4 turn the a priori distribution into a flat and non-informative one.

For the residual effects we have $e|\sigma_e^2 \sim N(0,\sigma_e^2)$ and $\sigma_e^2 \sim v_e S_e^2 \chi_{v_e}^2$. Also $E(\sigma_e^2) = \frac{s_e^2 v_e}{v_e - 2}$ and $S_e^2 = \frac{E(\sigma_e^2)(v_e - 2)}{v_e}$. The expectation $E(\sigma_e^2)$ is equivalent to $E(\sigma_e^2) = \tilde{\sigma}_e^2$. So, $S_e^2 = \tilde{\sigma}_e^2 \frac{(v_e - 2)}{v_e} = \tilde{\sigma}_e^2 \frac{(4.2-2)}{4.2}$, where $\tilde{\sigma}_e^2$ is a priori value of σ_e^2. Zeng et al. (2013) used v equal 4 for both genetic and environmental effects and the true simulated values for $E(\sigma_e^2)$ and $E(\sigma_{ma}^2)$.

For dominance effects at the intra-population level, the distributions are similar as described for additive effects. Then:

$m_{di} | \sigma_{mdi}^2 \sim N(0,\sigma_{mdi}^2)$ for the marker dominance effects;

$\sigma_{mdi}^2 \sim \chi^{-2}(v_{md}, S_{md}^2)$ for the marker dominance variance;

being the marginal the prior distribution for marker dominance effects given by $m_{di} | v_{m_d}, S_{m_d}^2 \sim t(0, v_{m_d}, S_{m_d}^2)$.

Additive and dominance variances are given by $\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i) m_{ai}^2$ and $\sigma_d^2 = \sum_{i=1}^{n} [2p_i(1-p_i)]^2 m_{di}^2$, respectively, according to parameterizations in W and S. The full conditional distributions for the parameters of the BayesA and BayesB models were presented in details by Zeng et al. (2013).

2.3.4. BayesA*B* or IBLASSO$_t$ method

According to Lehermeier et al. (2013), a strong influence of prior parameters on the predictive ability was seen in BayesA and BayesB models. The variation of

the scale parameters S^2_{ma} and S^2_{md} in these methods had a strong impact on prediction. Choosing a too large scale (S^2_{ma} or S^2_{md}) for the prior distribution of variance led to an overfitting of the data and a too small scale parameter led to underfitting, due to too much shrinkage of the effects. In both cases the predictive ability is considerably reduced. To obtain good predictive abilities an adequate choice of hyperparameters is necessary to prevent both over- and underfitting.

The differences between the explicit regression GWS methods are mainly due to: type and extent of the shrinkage imposed by the method; ability of learning from data; influence of prior distributions. In the N <<< n (n is number of markers and N is the individuals observation) case, learning from data is hardy to verify as the data (likelihood) do not dominate the posterior distribution. Thus given the same sampling model postulated by the methods, likelihood shrinkage properties are not too different and differences in posterior inferences between these methods must be due to the fact that priors are influential and very different (Gianola, 2013). From this, it can be asserted that different methods can be fitted with the same machinery only altering somehow drastically the prior distribution.

The Bayesian LASSO provides better learning from data than BayesA and BayesB (Gianola, 2013; Gianola et al., 2009). The difference between the Bayesian LASSO and the Bayesian approaches (BayesA and BayesB) developed by Meuwissen et al. (2001) comes from the different specification of the prior variance of the marker-specific regression coefficient and the type and extent of shrinkage effected.

Basing on that, we chose to implement BayesA through BLASSO framework by specifying the prior distribution through appropriate degrees of freedom (6 and 8)

for the scaled inverse chi-square distribution associated to marker genetic variance (and then with the penalization parameter λ). This produces a t-like distribution, which is intermediate between the normal (of the RR-BLUP) and double exponential (of the LASSO) distributions and provides the desirable shrinkage estimates of QTL effects as does BayesA.

By fitting in this way (via BLASSO), BayesA can have better learning properties. This kind of improved BayesA can be called BayesA*. It can turn out to be Bayes B* if the BLASSO machinery effectively leads a great number of markers to zero effects. Then the method will be called Bayes A*-B* (or t-Bayesian Lasso) because it conjugates the priors of BayesA and the type and extent of shrinkage (covariable selection) of the BLASSO. Meuwissen et al. (2009) in their fast BayesB method changed the prior distribution of marker effects from Student-t to double exponential of Laplace, which has improved it, perhaps turning it closer to a BLASSO method. Kärkkäinen and Sillanpää (2012) discuss the interchange of Student-t and Laplace (DE) as prior distributions of marker effects. Another possible name for Bayes A*-B* is t-BLASSO, meaning Bayesian LASSO (Legarra et al., 2011) with t distribution as prior for marker effects.

Bayes A*-B* methods were fitted using the GS3 software (Legarra et al., 2013) via MCMC (Markov Chain Monte Carlo) assigning 6 and 8 degrees of freedom for the inverted chi-square distribution for genetic variance (and then with the penalization parameter λ), which turns out the prior for marker effects into a t distribution. This is expected to produce similar results as the Bayes methods of Meuwissen et al. (2001) but with the learning ability of the BLASSO. Additionally, the BLASSO is asymptotically free of prior information and more consistent than BayesB and is tuning free.

2.3.5. BLASSO and IBLASSO Methods

In the Bayesian LASSO (Park and Casella, 2008) the prior assigned to marker effects is a Laplace (double exponential, DE) distribution. All marker effects are assumed to be independently and identically distributed as DE. This prior assign the same variance or prior uncertainty to all marker effects, but it possesses thicker tails than the normal or Gaussian prior. Comparative discussions on the DE prior are in De los Campos et al. (2012) and Fang et al. (2012).

With two variance components (σ_e^2 and σ_{ma}^2) the model is called improved Bayesian LASSO (IBLASSO) (Legarra et al., 2011). The practical implementation of this model via Gibbs sampling, including the full posterior conditionals distributions was described by Legarra et al. (2011). The a priori distribution of σ_e^2 consists of an inverted chi-square with 4 degrees of freedom. The prior distribution for λ was deliberately vague, being a uniform between 0 and 1,000,000. For dominance effects similar distributions hold as described for additive effects.

Concerning the IBLASSO of Legarra et al. (2011), Gianola (2013) criticizes the choice of a uniform flat prior on the regularization parameter λ. Due to this, our paper used two alternative priors: such flat prior and also a prior with 4 degrees of freedom on the parameter λ as used with the BLASSO. Computations were performed in the GS3 Software.

From the relation $\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i)\sigma_{m_a}^2$ (Gianola et al., 2009), with the IBLASSO we have $\sigma_a^2 = \sum_{i=1}^{n} 2p_i(1-p_i)\,2/\lambda^2$. Since σ_a^2 is generally known, a priori information on λ can be given by $\lambda^2 = \sum_{i=1}^{n} 2p_i(1-p_i)\,2/\sigma_a^2$.

The form of the distribution of marker effects is accomplished by the parameter λ, which is related to marker genetic variation by the expression $\mathrm{Var}(m_a) = 2/\lambda^2$. This relation denotes that λ^2 plays a similar role as the inverse of the variance in the Gaussian models. With IBLASSO through GS3 software we define priors for variance components and the value of λ is set to $\lambda^2 = 2/\sigma_a^2$.

2.3.6. Ridge Regression with heterogeneity of variances (RR-HET)

A additive-dominance Ridge Regression (RR-BLUP) method can also be implemented considering the heterogeneity of variances between markers, called RR-HET. In the our paper, the matrices with specific variances for each marker, $D_a = \mathrm{diag}(\tau_{a1}^2, \tau_{a2}^2, ..., \tau_{an}^2)$ and $D_d = \mathrm{diag}(\tau_{d1}^2, \tau_{d2}^2, ..., \tau_{dn}^2)$, were obtained by the method BLASSO (-2, -2), using the GS3 software.

2.4. Fitting Models

Each type of population was simulated 10 times under the same parameter settings, which preserved the same features and provided samples that were effectively of the same conceptual population. Nine replicates were used as training populations, and one replicate was used as a validation population. The estimations based on each of the nine replicates were validated by obtaining estimates of the

parameter accuracy and bias. Validation and reference individuals belonged to the same population but to different families.

In each replicate, marker effects were estimated and used to estimate the genetic values of individuals in the tenth population. These estimated genetic values were correlated with the parametric genetic values of individuals of the tenth population, providing the accuracy values. The results from the nine analyses were averaged to obtain the final results. Accuracies were calculated as the correlation between parametric and estimated individual genetic effects computed for each replicate of the simulation and averaged across replicates. Heritabilities were also estimated nine times in each scenario and averaged.

For Bayesian methods, we used 120,000 iterations for the MCMC algorithms of the different models, with the first 20,000 iterations discarded as burn in. After every set of 10 iterations (thin) were performed, a sample was retained to calculate a posteriori statistics. Hence, 10,000 MCMC samples were used to construct the posterior densities. The convergence of the Markov chains was checked with a Geweke (1992) diagnostic and also by visualizing the trace plot and running repeated progressive analyses until convergence was met. Posterior distributions were plotted (Figure 1) to view the Bayesian learning of the methods. A summary of the fitted models is presented in Table 1.

The Bayesian methods fitted through GS3 software (Table 1) provide estimates of a_i and d_i. Then, following the parameterizations in the incidences matrices W and S shown before, additive and dominance variances are given by:

$$\sigma^2_{\mu_a} = \sum_{i=1}^{n} [2p_i(1-p_i)]\sigma^2_{ai} + \sum_{i=1}^{n} [2p_i(1-p_i)](q_i - p_i)^2\sigma^2_{di}$$

$$\sigma^2_{\mu_d} = \sum_{i=1}^{n} [2p_i(1-p_i)d_i]^2 = \sum_{i=1}^{n} [2p_i(1-p_i)]^2\sigma^2_{di},$$

respectively.

2.5. Methods for Computing Parametric Accuracies

Methods for computing parametric accuracies under the additive-dominance models were derived following Resende (2008), Grattapaglia and Resende (2011) and Resende et al. (2013). The following formulas were obtained.

Additive accuracy: $r_{a\hat{a}} = \sqrt{\frac{r^2_{mq}(Nr^2_{mq}h^2_a / n_{QTL})}{1 + Nr^2_{mq}h^2_g / n_{QTL}}}$

Dominance accuracy: $r_{d\hat{d}} = \sqrt{\frac{r^2_{mq}(Nr^2_{mq}h^2_d / n_{QTL})}{1 + Nr^2_{mq}h^2_g / n_{QTL}}}$

Genotypic accuracy: $r_{g\hat{g}} = \sqrt{r^2_{a\hat{a}} + r^2_{d\hat{d}}}$

where r^2_{mq} is the realized proportion of genetic variation explained by the markers, n_{QTL} is the number of QTL, N is the number of individuals in the estimation dataset, and h^2_a, h^2_d and h^2_g are additive, dominance and total heritability, respectively.

2.6. Decomposing the Quantitative Genetic Information

The three types of quantitative-genetic information can be defined as in Habier et al., 2013:

Linkage disequilibrium: refers to founder alleles from different loci in the same gamete, and the loci are in LD (not sampled independently, i.e., in population level disequilibrium) and describe genetic relationships between founders.

Co-segregation: refers to non founder alleles (not in LD and not identical by descent from the base population, not present in the base population) from different loci in the same gamete, and the loci are linked (not transmitted independently, i.e., in population level equilibrium but in within-family level disequilibrium).

Genetic relationships: statistical dependency between alleles from the same locus in different gametes. This kind of information is of three types: When associated with markers, it refers to parentage only on the marker loci and does not involving a linkage between markers and QTL; when associated with the pedigree of individuals in a model with both markers and pedigree, it refers to residual polygenic effects; when associated with the pedigree of individuals only, it refers to total polygenic effects.

G-BLUP makes use of the following: (i) co-segregation of QTL and markers due to linkage; (ii) pedigree genetic relationships between markers not linked to QTL; and (iii) LD between markers and genes to capture relationships at QTL (Habier et al., 2013). The genomic relationship matrix is called the realized relationship, as it describes IBD at SNP, assuming an ancient founder population. However, only genetic relationships at QTL matter.

The genomic relationship matrix includes LD, co-segregation and pedigree genetic relationships between markers not linked to QTL (for example, in structured populations). Habier et al. (2013) derived formulas for proving that all three sources of information are used by G-BLUP.

The data sets analyzed were as follows: overall (raw or without any correction of the phenotypes); within-family deviations across families (with correction of the phenotypes for family effects and analyzing families altogether); and within each family with posterior averaging (with correction of the phenotypes for family effects and analyzing one family at a time). The accuracy of genomic selection in the analysis using the within-each-family with posterior averaging dataset is due to LD and co-segregation. In the analysis using the dataset from within-family deviations across families, the accuracy is due only to LD, while the accuracy of the analysis with the overall dataset is due to family IBD relationships, LD and co-segregation.

3. Results

3.1. Comparison of Methods

In the evaluation of the methods the following quantities were subjected to comparisons: heritability and dominance/additive variation ratio (the best are the closest to parametric ones); accuracies (the best are the highest ones); bias (the best are the closest to 1). Thus, besides the accuracy and bias, the molecular heritability also can be seen as a measure of evaluation and comparison of the methods applied to GWS.

Results concerning the trait controlled by small gene effects with heritability 0.30 are presented in Table 2. It is shown the results for overall analysis applied in the raw dataset. It can be seen from the Table 2 that, out of the 10 methods, the BAYESA*B* (-2,8) method (or t-BLASSO) had seven best (b) criteria amongst the 7 classification criteria. It was followed by BAYESA*B* (4,6) software, which had six best criteria. The G-BLUP method fitted through GVC-REML was intermediate

and seems to overestimate a bit the dominance/additive variation ratio. Other intermediate methods were BRR (-2,-2) and BAYESA*B* (-2,6) .

Additive accuracies for alternative methods were 0.68, 0.63 and 0.53 for parametric GWS; GWS by the best methods and pedigree, respectively. Expected additive accuracy estimate of the parametric GWS obtained by using deterministic formula is 0.68 in this case. BayesA*B* methods and the G-BLUP method fitted through GVC-REML software were the best and gave accuracy of 0.63, which is close to the parametric one. These results reinforce the worth of GWS, which performed better than pedigree phenotypic selection (Table 2).

Figure 1 also corroborates the power of GWS in catching up the parametric individual genetic values (in dark). The methods which fitted and matched best the parametric values were the BayesA*B*-type methods and Bayesian Regression (in dark blue, brown, gray, red and green) as can be seen for additive effects in Figure 1. For dominance effects the best methods were the BayesA*B*-type methods (in dark blue, brown, gray). The Bayesian Regression (in red and green) did not follow these methods for the dominance effects.

The statistical distance between distributions can be measured by the metric of Kullback-Leibler (KL). But the extent of Bayesian learning can also be quantified by the Hellinger distance between the prior and posterior densities of marker effects. This metric is easier to interpret than KL. The Hellinger (H) distance or HD (according to Roos and Held, 2011) between priors is given by Lehermeier et al. (2013). In their study, LASSO models produced H values considerably higher than for BayesA and BayesB, indicating that there was less Bayesian learning in BayesA and BayesB, as there is more similarity between the marginal prior and posterior densities of the marker effects in those methods. However, a small distance between

prior and posterior density can also emerge if a perfect prior density is assigned, although this happens with probability close to zero if prior knowledge is scarce. For this reason, we did not compute this H distance but we plotted the posterior distributions featured in Figure 1, which showed that the intended a priori distributions were kept in the posteriori. So the desirable shrinkage properties of the methods have been met.

According to Gianola (2013), the effective Bayesian learning is even more difficult in BayesB than in BayesA, as it takes more information from the data to "neutralize" the prior of BayesB than does BayesA. Following him, although linear Bayesian regression models address the “small N, large n” problem, they are wrong mechanistically. Then, the claim that such methods help to understand genetic architecture is arguably inaccurate. BLASSO is a better learner and perhaps due to this, the BayesA*B* of the present paper fitted very well, recovering well the parametric values.

Lehermeier et al. (2013) found that when the prior parameter setting was appropriate, predictive abilities and accuracies of BayesA and BayesB were high and equal to those of Bayesian Ridge and Bayesian LASSO with random λ. As finding optimal parameters is not straightforward, they recommend finding hyperparameters iteratively via cross-validation (as we did in the present work), which have high computational costs. Then they advocated Bayesian models that are less sensitive with respect to the choice of hyperparameters such as BLASSO and Bayesian Ridge, models with a strong Bayesian learning ability.

Lehermeier et al. (2013) reported that little is known about the sensitivity of the Bayesian models with respect to prior and hyperparameter specification, because comparisons of prediction performance are mainly based on a single set of

hyperparameters. Our paper has varied these hyperparameters and showed that there are measurable differences coming from the different specifications. This is in accordance with literature. BayesA and BayesB hyperparameter settings had a stronger effect on prediction performance than with the BLASSO and Bayesian regression (Lehermeier et al., 2013). On the other hand, BayesCπ and IBLASSO were similar in accuracy for most of the traits (Colombani et al., 2013). According to Legarra et al. (2011), IBLASSO is more similar to BayesA and BayesCπ than BLASSO is.

In general, comparing with the parametric values, the methods for additive-dominance models underestimate a bit the narrow sense heritability. The G-BLUP fitted via GVC software overestimated a bit the dominance heritability. The best methods were able to capture well the dominance heritability but were not completely good for capturing the additive heritability, perhaps due to limited number of markers and/or imperfect LD. Dominance heritability was overestimated by G-BLUP and BLASSO and perfectly estimated by BayesA*-B*.

Wellmann and Bennewitz (2012) introduced new Bayesian linear regression models for genomic evaluation of quantitative traits that account for dominance effects of QTLs. These models are generalizations of Bayesian SSVS (Verbyla et al., 2009, also called BayesC), which includes only additive effects. The models of Wellmann and Bennewitz (2012) are called BayesD (D from dominance) and are modifications of BayesC (that is a mixture of two t distributions, one of them with a small variance). BayesB is the limiting case of BayesC when the small variance goes to zero. Bayesian models associated to mixture of distributions were criticized by Gianola (2013), who pointed out that MCMC methods with such models can never

converge. So we did not study these Bayesian D models. Nevertheless, comparisons including such methods are welcome.

Methods based on mixture of distributions such as Bayes B, Bayes C, Bayes R (Erbe et al., 2012) and Bayes S (Brondum et al., 2012) are richly parameterized and devised to inform about genetic architecture of the trait. However, Celeux et al. (2012) and Gianola (2013) reported that mixture models have failed to converge when using MCMC samplers. Estimation of the vector of parameters (θ) of a mixture basing on data is only valid when θ is likelihood identifiable (McLachlan and Peel, 2000), which means the data (likelihood function) dominating over the priori in the posterior distribution. Models richly parameterized induces strong identification deficit (Gianola, 2013). In genomic selection models (featured as a n>>N case, where n is the number of markers and N is the number of individuals), likelihood is hard to be dominating. Duchemin et al. (2012) reported over-parameterization problems of Bayes Cπ method.

Results concerning the trait controlled by mixed (major and small gene effects) inheritance model with heritability 0.30 are presented in Table 3. It can be seen from the Table 3 that the best methods were similar as in the small gene size effects case (Table 2), except that the G-BLUP method fitted through GVC-REML software outperformed the three BayesA*B* methods. G-BLUP was better for estimation of dominance effects and BayesA*B* methods were better for estimation of the dominance/additive variation ratio. Such methods proved to be robust to genetic architecture of the trait.

Results concerning the trait controlled by small gene effects with heritability 0.50 are presented in Table 4. It can be seen from the Table 4 that the best methods were the same as in Tables 2 and 3, i.e, the three BayesA*B* methods and the G-

BLUP method fitted through GVC-REML software. The methods were good for estimation of both, additive and dominance effects as well as the dominance/additive variation ratio. As expected accuracies for $h^2 = 0.5$ were higher than that for $h^2 = 0.3$ (Table 2).

Expected additive accuracy estimate of GWS obtained by deterministic formula is 0.73 in this case. BayesA*B* methods and the G-BLUP method fitted through GVC-REML software were the best with accuracy of 0.70.

Results in Table 5 were similar to those in Table 3, with G-BLUP outperforming the three BayesA*B* methods, except in recovering the dominance/additive variation ratio. G-BLUP proved to be specially better for estimation of dominance in a mixed inheritance model scenario.

3.2. Partition of accuracy due to the three quantitative genetics information

Aiming at partition of the quantitative genetic information, analyses were performed in raw (raw: uses family IBD relationships at markers not linked to QTL, LD and co-segregation), deregressed mendelian segregation or phenotype corrected for family structure (DMS: uses LD only), averaged within family analyses in the raw data (AWF: uses LD and co-segregation) and raw using pedigree instead of markers (uses co-segregation and family IBD relationships at QTL). Results for $h^2 = 0.5$ and mixed inheritance model are presented in Table 6 (method BayesA*B* (-2,8)).

From the genomic heritability (0.26) it can be seen that the main source of information is LD (0.16), followed by co-segregation (0.06) and family IBD relationships not linked to QTL (0.04). In the simulation, the proportion (r^2_{mq}) of genetic variation explained by markers exclusively in LD was high, around 90%. In

such a case, genetic variation is mainly due to LD in detriment of co-segregation and residual polygenic effects, corroborating the results.

From the pedigree heritability (0.20) it can be seen that the main source of information is individual IBD relationships (0.14), which is a fraction (0.875 = 0.14/0.16) of IBS-LD captured by markers, followed by co-segregation (0.06). These partitions are in accordance with results reported by Habier et al. (2013). The value 0.14 is not all necessarily originated from the 0.16 as the pedigree can be capturing some loci which markers are not. Accuracy estimates follows almost the same tendency.

Additive accuracy of related individuals ($r_{g\hat{g}r}$, using the raw dataset) was 0.69. It can also be given as function of accuracy due to pedigree ($r_{g\hat{g}ped}$) and the accuracy of unrelated individuals ($r_{g\hat{g}u}$), by: $r_{g\hat{g}r} = r_{g\hat{g}ped} + (1\text{-}r_{g\hat{g}ped})r_{g\hat{g}u}$ $= 0.45 + (1-0.45)\,0.52 = 0.73$, which is close to 0.69. It can be seen that the use of related individuals increases the accuracy.

Habier et al. (2013) proposed to apply pedigree analyses concepts which define founders in a recent past generation. According to those, genetic covariances between individuals at QTL result only from LD between markers and QTL measured in founders or from the fact that the training individuals are related by pedigree to selection candidates. This last aspect characterizes the co-segregation of alleles at markers and QTL which are linked, but not in disequilibrium. This is called linkage information and results in genetic relationships at QTL captured by markers.

Then, out of the three quantities captured by the G matrix, two, co-segregation and LD between markers and QTL, contribute to genetic gain in the short run. But only one, LD between markers and QTL, can lead to genetic gain in

the long run, i.e., several generations ahead. Co-segregation can not contribute too much for persistency of the accuracy of the GWS over generations without re-training, due to recombination leading to a loss of linkage information. The pedigree genetic relationships between markers not linked to QTL do not contribute to accuracy of the GWS and can lead to declining validated accuracy with increasing (in terms of marker number) training data size (Habier et al., 2013). This relationship always occurs when markers on one chromosome are found explaining variation at a QTL on another chromosome. This situation can never be explained by LD and co-segregation.

Co-segregation can be estimated by averaged within family analyses for each family. As there are pedigree genetic relationships in within family raw data in each family, the accuracy within each family in the case of linkage equilibrium between markers and QTL is exclusively due to co-segregation. With linkage disequilibrium (across families) between markers and QTL, LD is also captured. For the within family deviations across families, the accuracy is only due to LD between markers and QTL and does not include co-segregation.

Accuracy of GWS based on additive-genetic relationships can decline with increasing training data size and modeling polygenic effects via pedigree relationships jointly with genomic breeding values using Bayesian methods may prevent that decline (Habier et al., 2013). The decline can be explained by the fact that increasing the number of markers the shrinkage of marker effects in G-BLUP turns out to be stronger.

The number of markers to be used is called effective number of markers (Habier et al., 2013) and is a compromise between increasing information about the marker and decreasing the number of information per marker (Resende et al., 2010).

With increasing marker density more markers support the same QTL and so compensate stronger shrinkage. With long-range LD, QTL effects are captured by more markers than with short-range LD (Habier et al., 2013).

As G-BLUP cannot capture short-range LD information well, Habier et al. (2013) recommended Bayesian methods with t-distributed priors that are expected to capture LD better than G-BLUP (Fernando et al., 2007). Our results support this by showing that BayesA*-B*, which use t-distributed priors, was the best for recovering the dominance variance / additive variance ratio (Tables 2, 3 and 4).

4. Discussion

The so called BayesA*B* methods fitted by the GS3 software produced the best results, together with G-BLUP. The degrees of freedom associated with prior error variance were found to have little impact in the three BayesA*B* methods, and the greater impact comes from using adequate (6 or 8 instead of -2, 2 or 4) degrees of freedom for the marker variance associated with the shrinkage parameter. Using 6 or 8 degrees of freedom produced only small differences, the BayesA*B* (-2, 8) being slightly better. G-BLUP was as good as these BayesA*B* methods. Figure 2 and the associated table summarize the results and show the following final classification of methods: (i) best: G-BLUP; BAYESA*B* (-2,8); BAYESA*B* (4,6); (ii) intermediate: BRR (-2,-2); BAYESA*B* (-2,6); IBLASSO (4,2); and (iii) worst: IBLASSO (4,-2); RR-HET (-2 -2); BLASSO (4,2); Pedigree.

Matrix (G) calculated using molecular markers can be more efficient than the pedigree-based relationship matrix (A) as it can account for Mendelian sampling and segregation distortion (Nejati-Javaremi et al., 1997; Fernando, 1998). Carré et al.

(2013) proposed a method named mendelian segregation model in which base animals are predicted via pedigree and the descendants are predicted by G-BLUP.

In general, the Bayesian Ridge Regression (BRR) method provided good results. This is in accordance with Lehermeier et al. (2013), who reported that the Bayesian Ridge model with marker-homogeneous shrinkage was in all datasets among the models with the highest predictive ability. Also, they found that independently of the number of markers and observations, marker-specific shrinkage did not outperform marker-homogeneous shrinkage. And considering also the higher computing efforts of models with marker-specific shrinkage, they recommended Bayesian Ridge as a robust model for genome-based prediction. In line with this, most studies report that Bayesian shrinkage models perform similar to or slightly better than G-BLUP model (equivalent to the ridge regression model).

In BayesA and BayesB the degrees of freedom of the fully conditional posterior distribution of σ^2_{mai} are df + 1 (where df is the prior degrees of freedom), and thus only one degree of freedom higher than the prior degrees of freedom, independently of the number of observations (N) or markers (n) in the model (Gianola et al., 2009; Lehermeier et al., 2013). Differently, in Bayesian Ridge Regression, the degrees of freedom increase with the number of markers in the model. In genomic datasets, learning in the Bayesian methods is limited due to the n >> N situation. With next generation sequencing data, n will be even larger, and is expected to increase much more than N. Thus, models with a strong Bayesian learning ability such as the Bayesian Ridge and Bayesian Lasso will be useful (Lehermeier et al., 2013).

The accuracies were very close across methods for all effects (additive, dominance, although dominance effects were poorly estimated). This is in

accordance with literature results (De los Campos et al., 2012; Gianola, 2013), which point to similarity of several methods for prediction purposes, in terms of accuracy. Thus, the main criteria contributing for the differences among methods are bias (related to architecture learning), heritability estimation and dominance/additive variation capture.

The IBLASSO (4,-2) method, criticized by Gianola (2013) in terms of the chi-square number (-2) of degrees of freedom for markers variance, also had a poor performance in the present work, as well as the RR-BLUP-HET method using variance components results from the same IBLASSO (-2,-2) method. Trying to improve the results, in the sense of the BLASSO of De los Campos et al (2009), the degrees of freedom of the chi-square prior distributions for genetic variances were changed from -2 to 2 producing the IBLASSO (4,2) method, which was better than IBLASSO (4,-2) but poorer than BLASSO (4,2) fitted in the BLR software.

Concerning the estimation methods, 7 evaluation criteria were used. Accuracy did not differ so much, even with contrasting methods, corroborating the majority of reports in the literature (Hayes et al., 2009; Van Raden et al., 2009; De los Campos et al., 2012; Gianola, 2013). Unbiasedness or learning of the genetic architecture favored the methods fitted through Bayesian LASSO.

Across the 7 criteria, the additive-dominance BayesA*-B*-type or t-BLASSO methods (with 6 or 8 degrees of freedom on chi-square distribution for genetic variance and then for the penalization parameter) and G-BLUP performed best in over 5 criteria.

With increasing degrees of freedom of the chi-square distribution for variance components, the DE distribution for marker effects goes to the normal distribution, with the t distribution in between them. The Student t-distribution approximates the

normal distribution when the degree of freedom v increases, then G-BLUP can be considered as a limiting case of BayesA. The fitting of the BLASSO with new double exponential and t distributions has been considered recently (Fang et al., 2012). They proposed three new methods (improved double-exponential prior, improved Student's t prior and extended Bayesian LASSO) that outperform the traditional Bayesian LASSO.

The models Bayes/BLASSO we fitted differed in the prior specification for the marker effects with hyperparameters controlling the amount of shrinkage of the effects. As the degree of freedom v controls the thickness of the tails of a t-distribution, the choice of v had an large effect on the results. In our paper, the suitable parameters were inferred and chosen by the result of the cross validations.

According to Gianola (2013), even though BLASSO bears a parallel with the LASSO, it does not "kill" or remove markers from the model, contrary to what happens in variable selection approaches. BLASSO poses a leptokurtic prior, so it is expected to shrink small effects more strongly towards zero than the Gaussian prior, as opposed to inducing sparsity in the strict sense of the LASSO. In BLASSO, markers having tiny effects are "effectively", but not physically, wiped out of the model. Also, markers with strong effects receive a heavier weight in this overall measure of complexity (Gianola, 2013).

On the other hand, Fang et al. (2012) eported that Bayesian LASSO usually cannot effectively shrink the zero-effects QTL very close to zero. They concluded that the improved Student's t prior for the LASSO is able to effectively shrink toward zero the zero-effects QTL and the signals of QTL were very clear. The results reported by Fang et al. (2012) corroborate our choice of changing the DE to t distribution in BLASSO.

In our paper, the additive-dominance BayesA*-B*-type methods which use t-distributed priors, were best for recovering the dominance variance / additive variance ratio (Tables 2, 3 and 4). This property is of great relevance in keeping the real proportionality between dominance and additive effects in the estimates. The BRR method was best in this criterion in one situation (Table 5).

The ability to recover well the heritabilities can be more sensitive to discriminate methods. This is because heritabilities are more complex parameters than are the simple correlation coefficients (accuracies) (De los Campos et al., 2012). According to Makowsky et al. (2011), the heritability can be regarded as a measure of goodness of fit in the current dataset (projected to the base population) and predictive accuracy refers to prediction in future samples. Both are interdependent and the predictive accuracy (estimated by using a validation population) is able to capture over-fitting. The heritability estimates the proportion of phenotypic variance accounted for by true genetic values in the base population comprised by unrelated individuals. By its turn, the squared predictive accuracy estimates the proportion of phenotypic variance accounted by predicted genetic values in the sample, not in the base population. Then, it ignores inbreeding, relationships between individuals and estimation errors, not producing a consistent information about the magnitude of the heritability (Makowsky et al. ,2011).

The most probable true symmetrical distributions of genetic effects (genetic architecture) are normal (Gaussian), t (Studentian) and double exponential (Laplacean). So it is imperative to test these three distributions by assuming them as priors in the methods of analyses. This will reveal which assumed prior distribution is more adequate and/or robust. As we did in this paper, cross-validation is crucial in these settings, providing evidences of the adequacy and error magnitude of each

assumed prior distribution. According to Gianola (2013) each prediction generates an error, and this error will have a cross-validation distribution.

Wang and Da (2014) presented the traditional quantitative genetics model as the unifying model for definitions of the genomic relationship and inbreeding coefficients. According to them, theoretical differences between the existing and new definitions of genomic additive and dominance relationships were in the assumptions of equal SNP effects (equivalent to across-SNP standardization), equal SNP variances (equivalent to within-SNP standardization), and expected or sample SNP additive and dominance variances. These conclusions came to facilitate the understanding and comparison of alternative prediction and estimation methods.

As advocated by Wang and Da (2014), after their results, the need for methods comparisons is less evident. Our results showing the equivalence between several predictive methods corroborate their findings.

5. Conclusions

Amongst the 10 models/methods evaluated, G-BLUP, BAYESA*B* (-2,8) and BAYESA*B* (4,6) methods presented the best results and showed to be adequate for accurate prediction of genomic breeding and total genotypic values in additive-dominance genomic models.

6. References

Bennewitz J., Meuwissen T. H. E. 2010. The distribution of QTL additive and dominance effects in porcine F2 crosses. **Journal of Animal Breeding and Genetics**; 127(3):171-9.

Brondum R. F., Guosheng S., Lund M. S., Bowman P. J., Goddard M. E., Hayes B. J. 2012. Genome position specific priors for genomic prediction. **BMC Genomics**, 10;13:543.

Carré C., Gamboa F., Cros D., Hickey J. M., Gorjanc G., Manfredi E. 2013. Genetic prediction of complex traits: integrating infinitesimal and marked genetic effects. **Genetica**, 141(4): 239-246.

Celeux G., El Anbari M., Marin J. M., Robert C. P. 2012. Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. **Bayesian Analysis**, 7(2): 477-502.

Colombani C., Legarra A., Fritz S., Guillaume F., Croiseau P., Ducrocq V., Robert-Granjé C. 2013 Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbeliarde breeds. **Journal of Dairy Science**, 96(1):575-591.

Da Y., Wang C., Wang S., Hu G. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. . **PLoS ONE**, 9(1):e87666.

De los Campos G., Gianola D., Rosa G. J. M. 2009. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science**, 87:1883-1887.

De los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D., Callus M. P. L. 2012. Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, 193:327-345.

Denis M., Bouvet J. M. 2013. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. **Tree Genetics and Genomics**, 9: 37–51.

Duchemin S. I., Colombani C., Legarra A., Baloche G., Larroque H., Astruc J. M., Barillet F., Robert-Granié C., Manfredi E. 2012. Genomic selection in the French Lacaune dairy sheep breed. **Journal of Dairy Science**, 95: 2723-2733.

Endelman J. B., Jannink J. L. 2012. Shrinkage estimation of the realized relationship matrix. **Genes, Genomes, Genetics**, 2:1405-1413.

Erbe M., Hayes B. J., Matukumali L. K., Goswami S., Bowman P. J., Reich C. M., Mason B. A., Goddard M. E. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. **Journal of Dairy Science**, 95: 4114-4129.

Falconer D. S., Mackay T. F. C. 1996. **Introduction to Quantitative Genetics**, Ed 4. Longmans Green, Harlow, Essex, UK.

Fang M., Jiang D., Li D., Yang R., Fu W., Pu L., Gao H., Wang G., Yu L. 2012. Improved LASSO priors for shrinkage quantitative trait loci mapping. **Theoretical and Applied Genetics**, 124:1315-1324.

Fernando R. L. 1998. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. **Proceedings of the 6th World Congress on Genetics Applied to Livestock Production**, Armidale, NSW, Australia, 26:329-336.

Fernando R. L., Habier D., Stricker C., Dekkers J. C. M., Totir L. R. 2007. Genomic selection. **Acta Agriculturae Scandinavica**, 57(4):192-195.

Geweke J. 1992. **Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments**. In Bayesian Statistics 4 (ed JM Bernardo, JO Berger, AP Dawid, and AFM Smith). Clarendon Press, Oxford, UK.

Gianola D. 2013. Priors in whole-genome regression: the bayesian alphabet returns. **Genetics**, 194(3):573-96.

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

Goddard M. E., Hayes B. J. 2007. Genomic selection. **Journal of Animal Breeding and Genetics**, 124:323-330.

Goddard M. E., Hayes B. J., Meuwissen T. H. E. 2010. Genomic selection in livestock populations. **Genetics Research**, 92:413-421.

Goddard M. E., Wray N. R., Verbyla K., Visscher P. M. 2009. Estimating effects and making predictions from genome-wide marker data. **Statistical Science**, 24:517-529.

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

Grattapaglia D., Resende M. D. V. 2011. Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, 7:241-255.

Habier D., Fernando R. L., Garrick D. J. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. **Genetics**, 194(3):597-607.

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

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. 2009. Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, 92: 433-443.

Hill W. G., Goddard M. E., Visscher P. M. 2008. Data and theory point to mainly additive genetic variance for complex traits. **PLoS Genetics**, 29:4(2):e1000008.

Kärkkäinen H. P., Sillanpää M. K. 2012. Back to basis for Bayesian model building in genomic selection. **Genetics**, 191: 969-987.

Legarra A., Ricard A., Filangi O. 2013. **GS3 Genomic Selection – Gibbs Sampling – Gauss Seidel (and BayesCπ)**. Avaliable in: <http://genoweb.toulouse.inra.fr/~alegarra/manualg3_last.pdf>. Accessed Jun 2013.

Legarra A., Robert-Granié C., Croiseau P., Guillaume F., Fritz S. 2011. Improved Lasso for genomic selection. **Genetics Research**, 93(1):77-87.

Lehermeier C., Wimmer V., Albrecht T., Auinger H. J., Gianola D., Schmid V. J., Schön C. C. 2013. Sensitivity to prior specification in Bayesian genome-based prediction models. **Statistical Applications in Genetics and Molecular Biology**, 12(3):375-391.

Makowsky R., Pajewski N. M., Klimentidis Y. C., Vazquez A. I., Duarte C. W., Allison D. B., De los Campos G. 2011. Beyond missing heritability: prediction of complex traits. **PLoS Genetics**, 7(4):e1002051.

McLachlan G., Peel D. 2000. **Finite mixture models**. John Wiley & Sons, New York.

Meuwissen T. H. E. 2007. Genomic selection: marker assisted selection on genome-wide scale. **Journal of Animal Breeding and Genetics**, 124:321-322.

Meuwissen T. H. E., Goddard M. E. 2010. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. **Genetics**, 185: 623-631

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.

Meuwissen T. H. E., Solberg T. R., Shepherd R., Woolliams J. A. 2009. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. **Genetics Selection Evolution**, 41(1):2.

Muñoz P. R., Resende Jr M. F. R., Gezan S. A., Resende M. D. V., De los Campos G, Kirst M. Huber D., Peter G. F. 2014. Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. **Genetics**, 198:1759-1768.

Nejati-Javaremi A., Smith C., Gibson, J. P. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, 75:1738 - 1745.

Oliveira E. J., Resende M. D. V., Santos V. S., Ferreira C. F., Fachardo G. A., Silva M. S., Oliveira L. A., Vildoso C. I. A. 2012. Genome-wide selection in cassava. **Euphytica**, 187:263-276.

Park T., Casella G. 2008. The Bayesian LASSO. **Journal of the American Statistical Association**, 103(482):681-686.

Piccoli M. L., Braccini J., Cardoso F. F., Sargolzaei M., Larmer S. G., Schenkel F. S. 2014. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. **BMC Genetics**, 15:157.

Resende Jr M. F. R., Valle P. R. M., Acosta J. J., Peter G. F., Davis J. M., Grattapaglia D., Resende M. V. D., Kirst M. 2012b. Accelerating the

domestication of trees using genomic selection: accuracy of prediction models across ages and environments. **New Phytologist**, 193:617-624.

Resende Jr MFR, Valle PRM, Resende MDV, Garrick DJ, Fernando RL, Davis J. M., Jokela E. J., Martin T. A., Peter G. F., Kirst M. 2012a. Accuracy of genomic selection methods in a standard dataset of loblolly pine. **Genetics**, 190:1503-1510.

Resende M. D. V. 2008. **Genômica Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais**. Colombo: Embrapa Florestas, 330 p.

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. **Pesquisa Florestal Brasileira**, 56:63-78.

Resende M. D. V., Resende Jr M. F. R., Aguiar A. M., Abad J. I. M., Missiaggia A. A., Sansaloni C., Petroli C. Grattapaglia D. 2010. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, 79 p.

Resende M. D. V., Silva F. F., Lopes P. S., Azevedo C. F. 2013. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística 291 p. Avaliable in: <http://www.det.ufv.br/ppestbio /corpo_ docente.php>.

Resende M.D.V. 2002. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica, 975p.

Resende MDV, Resende Jr MFR, Sansaloni C, Petroli C, Missiaggia AA, Aguiar A. M., Abad J. M., Takahashi E. K., Rosado A. M., Faria D. A., Pappas G. J. Jr, Kilian A., Grattapaglia D. 2012. Genomic Selection for growth and wood

quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, 194:116-128.

Roos M., Held L. 2011. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. **Bayesian Analysis**, 6(2): 259-278.

Silva F. F., Rosa G. J. M., Guimarães S. E. F., Lopes P. S., De los Campos G. 2011. Three-step Bayesian factor analysis applied to QTL detection in crosses between outbred pig populations. **Livestock Science**, 142(1):210-215.

Sorensen D., Gianola D. 2002. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. New York: Springer Verlag, 740 p..

Su G., Christensen O. F., Ostersen T., Henryon M., Lund M. S. 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. **PLoS ONE**, 7(9): e45293.

Sved J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. **Theoretical Population Biology**, 2:125-141.

Talluri R., Wang J., Shete S. 2014. Calculation of exact p-values when SNPs are tested using multiple genetic models. **BMC Genetics**, 15:75.

Toro M. A., Varona L. 2010. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, 42:33.

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

Van Raden P. M., Van Tassell C. P., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor J. F., Schenkel F. S. 2009. Invited Review: Reliability of genomic predictions for North American dairy bulls. **Journal of Dairy Science**, 92(1):16-24.

Verbyla K. L., Hayes B. J., Bowman P. J., Goddard M. E. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. **Genetics Research**, 91:307-311.

Viana, J. M. S. 2004. Quantitative genetics theory for non-inbred populations in linkage disequilibrium**. Genetics and Molecular Biology**, 27(4):594-601.

Viana, J. M. S. 2011. **Programa para análises de dados moleculares e quantitativos RealBreeding**. Viçosa: UFV.

Visschier P. M., Yang J., Goddard M. E. 2010. A commentary on "Common SNPs explain a large proportion of the heritability for human height" by Yang et al. (2010). **Twin Research and Human Genetics**, 13(6):517–524.

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(4):1223-30.

Wang C., Da Y. 2014. Quantitative Genetics Model as the Unifying Model for Defining Genomic Relationship and Inbreeding Coefficient. **PLoS ONE**, 9(12):e114484.

Wang C., Prakapenga D., Wang S., Puligurta S., Runesha H. B., Da Y. 2014. GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. **BMC Bioinformatics**, 15:270.

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

Yang J., Benyamin B., Mcevoy B. P., Gordon S., Henders A. K. 2010. Common SNPS explain a large proportion of the heritability for human height. **Nature Genetics**, 42(7):565-569.

Zapata-Valenzuela J., Whetten R. W., Neale D., Mckeand S., Isik F. 2013. Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine. **G3**, 3(5): 909-916.

Zeng J., Toosi A., Fernando R. L., Dekkers J. C. M., Garrick D. J. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genetics Selection Evolution**, 45:11.

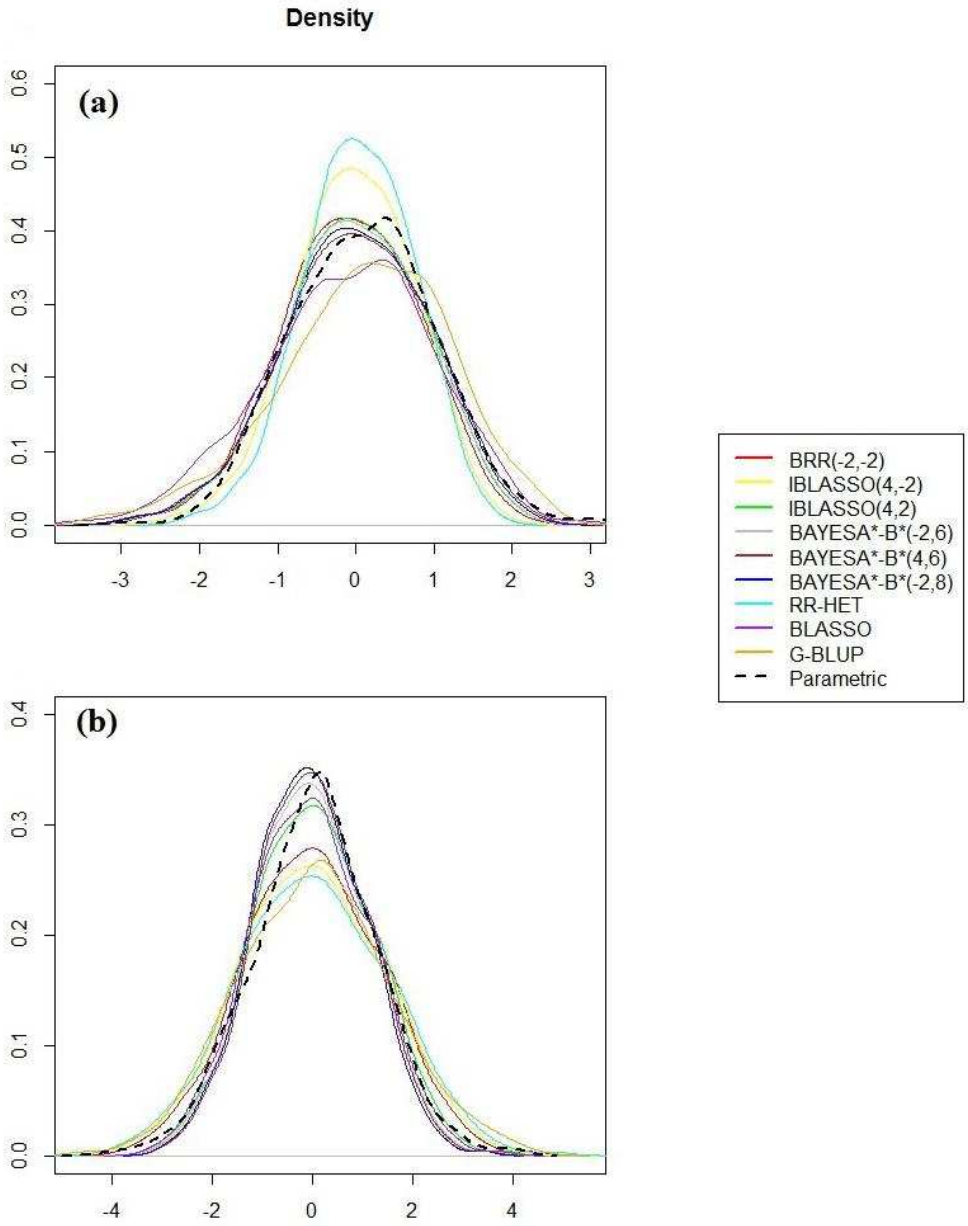

Figure 1. Posterior distributions. Parametric and predicted additive (a) and dominance (b) individual values (h2 = 0.30; small gene effects model).

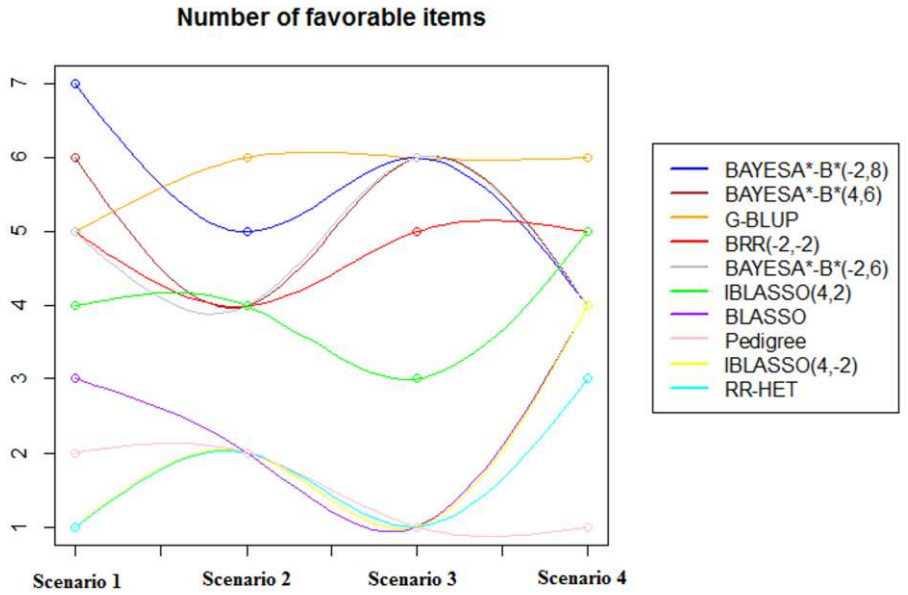

Figure 2. Favorable items. Number of favorable items in the four scenarios.

	Method	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Mean	Rank
Number of favorable items	G-BLUP	5	6	6	6	5.75	1
	BAYESA*-B*(-2,8)	7	5	6	4	5.50	2
	BAYESA*-B*(4,6)	6	4	6	4	5.00	3
	BRR(-2,-2)	5	4	5	5	4.75	4
	BAYESA*-B*(-2,6)	5	4	6	4	4.75	4
	IBLASSO(4,2)	4	4	3	5	4.00	5
	BLASSO(4,2)	3	2	1	4	2.50	6
	IBLASSO(4,-2)	1	2	1	4	2.00	7
	RR-HET(-2,-2)	1	2	1	3	1.75	8
	Pedigree	2	2	1	1	1.50	9

Table 2. Scenario 1: Results for the trait controlled by small gene effects with heritability 0.30.

Method	h2a	h2d	cor_a	byg_a	cor_d	byg_d	Vd/Va	Number of criteria scored as best
Parametric	0.21±0.01	0.10±0.01	0.68	-	0.48	-	0.48	-
BRR (-2,-2)	0.15[b]±0.05	0.12[b]±0.05	0.63[b]±0.03	1.40±0.33	0.31[b]±0.07	0.57[b]±0.23	0.77	5 [b]
IBLASSO (4,-2)	0.12±0.06	0.14±0.05	0.62[b]±0.03	2.41±1.82	0.28±0.06	0.46±0.24	1.19	1
IBLASSO (4,2)	0.14±0.06	0.10[b]±0.06	0.63[b]±0.03	1.86±1.14	0.29[b]±0.06	0.63[b]±0.42	0.81	4
BAYESA*B* (-2,6)	0.15[b]±0.06	0.10[b]±0.05	0.63[b]±0.03	1.51±0.57	0.29[b]±0.06	0.69[b]±0.42	0.67	5 [b]
BAYESA*B* (4,6)	0.15[b]±0.06	0.10[b]±0.05	0.63[b]±0.03	1.49[b]±0.56	0.29[b]±0.06	0.71[b]±0.43	0.65	6 [b]
BAYESA*B* (-2,8)	0.15[b]±0.05	0.09[b]±0.05	0.63[b]±0.03	1.44[b]±0.47	0.29[b]±0.06	0.72[b]±0.42	0.61[b]	7 [b]
RR-HET (-2, -2)	0.11±0.06	0.14±0.05	0.62[b]±0.03	2.43±1.74	0.28±0.05	0.44±0.23	1.24	1
BLASSO (4,2)	0.17[b]±0.09	0.13±0.02	0.63[b]±0.03	1.44±0.65	0.29[b]±0.05	3.20±5.34	0.74	3
G-BLUP	0.15[b]±0.05	0.13±0.06	0.63[b]±0.03	1.25[b]±0.35	0.31[b]±0.04	0.70[b]±0.30	0.83	5 [b]
Pedigree	0.16[b]±0.03	0.07±0.01	0.53±0.03	0.96[b]±0.19	0.05±0.02	0.20±0.11	-	2

b: best = highest +- 0.02 for h2a, h2d, cor a, cor d and Vd/Va; 0.5 to 1.5 for bya and byd; highest minus 2 for best criteria in the last column.

Table 3. Scenario 2: Results for the trait controlled by mixed (major and small gene effects) inheritance model with heritability 0.30.

Method	h2a	h2d	cor_a	byg_a	cor_d	byg_d	Vd/Va	Number of criteria best
Parametric	0.20±0.01	0.13±0.01	0.65	-	0.53	-	0.64	-
BRR (-2,-2)	0.13[b]±0.03	0.12[b]±0.06	0.63[b]±0.03	1.53[b]±0.29	0.33±0.04	0.65±0.22	0.94	4 [b]
IBLASSO (4,-2)	0.10±0.04	0.14[b]±0.05	0.64[b]±0.03	3.49±4.49	0.31±0.04	0.55±0.24	1.44	2
IBLASSO (4,2)	0.12[b]±0.04	0.11[b]±0.05	0.63[b]±0.03	2.26±2.22	0.32±0.05	0.71[b]±0.33	0.93	4 [b]
BAYESA*B* (-2,6)	0.13[b]±0.04	0.10±0.04	0.63[b]±0.03	1.53[b]±0.53	0.33±0.04	0.80[b]±0.32	0.73	4 [b]
BAYESA*B* (4,6)	0.13[b]±0.04	0.10±0.04	0.63[b]±0.03	1.54[b]±0.53	0.33±0.04	0.79[b]±0.32	0.74	4 [b]
BAYESA*B* (-2,8)	0.14[b]±0.04	0.09±0.04	0.63[b]±0.03	1.47[b]±0.48	0.33±0.04	0.83[b]±0.33	0.68[b]	5 [b]
RR-HET (-2, -2)	0.10±0.04	0.14[b]±0.05	0.64[b]±0.03	3.43±4.38	0.31±0.04	0.55±0.24	1.43	2
BLASSO (4,2)	0.10±0.03	0.16±0.07	0.63[b]±0.04	1.91±0.82	0.32±0.05	0.76[b]±0.61	1.63	2
G-BLUP	0.14[b]±0.03	0.13[b]±0.03	0.64[b]±0.04	1.26[b]±0.21	0.38[b]±0.04	0.84[b]±0.20	0.92	6 [b]
Pedigree	0.13[b]±0.02	0.09±0.01	0.46±0.04	0.89[b]±0.11	0.06±0.03	0.22±0.10	-	2

b: best = highest +- 0.02 for h2a, h2d, cor a, cor d and Vd/Va; 0.5 to 1.5 for bya and byd; highest minus 2 for best criteria in the last column.

Table 4. Scenario 3: Results for the trait controlled by equal gene effects with heritability 0.50.

Method	h2a	h2d	cor_a	byg_a	cor_d	byg_d	Vd/Va	Number of criteria best
Parametric	0.35±0.01	0.17±0.01	0.73	-	0.51	-	0.48	-
BRR (-2,-2)	0.25[b]±0.04	0.20[b]±0.03	0.69[b]±0.03	1.42[b]±0.23	0.36±0.04	0.54[b]±0.11	0.81	5 [b]
IBLASSO (4,-2)	0.22±0.06	0.22±0.04	0.69[b]±0.03	1.74±0.82	0.35±0.04	0.48±0.11	1.01	1
IBLASSO (4,2)	0.24±0.06	0.20[b]±0.04	0.69[b]±0.03	1.60±0.71	0.36±0.04	0.54[b]±0.14	0.82	3
BAYESA*B* (-2,6)	0.25[b]±0.06	0.18[b]±0.04	0.70[b]±0.03	1.53[b]±0.66	0.36±0.04	0.57[b]±0.15	0.73[b]	6 [b]
BAYESA*B* (4,6)	0.25[b]±0.06	0.18[b]±0.04	0.70[b]±0.03	1.52[b]±0.66	0.36±0.04	0.58[b]±0.15	0.72[b]	6 [b]
BAYESA*B* (-2,8)	0.26[b]±0.06	0.18[b]±0.04	0.70[b]±0.03	1.51[b]±0.64	0.36±0.04	0.59[b]±0.15	0.69[b]	6 [b]
RR-HET (-2 -2)	0.22±0.06	0.22±0.04	0.69[b]±0.03	1.76±0.83	0.35±0.04	0.48±0.11	1.02	1
BLASSO (4,2)	0.18±0.05	0.29±0.03	0.69[b]±0.03	1.69±0.45	0.35±0.03	0.46±0.08	1.59	1
G-BLUP	0.27[b]±0.03	0.20[b]±0.03	0.70[b]±0.02	1.17[b]±0.13	0.40[b]±0.04	0.74[b]±0.22	0.77	6 [b]
Pedigree	0.24±0.02	0.11±0.01	0.53±0.02	0.87[b]±0.09	0.04±0.02	0.12±0.06	-	1

b: best = highest +- 0.02 for h2a, h2d, cor a, cor d and Vd/Va; 0.5 to 1.5 for bya and byd; highest minus 2 for best criteria in the last column.

Table 5. Scenario 4: Results for the trait controlled by mixed (major and small gene effects) inheritance model with heritability 0.50.

Method	h2a	h2d	cor_a	byg_a	cor_d	byg_d	Vd/Va	Number of criteria best
Parametric	0.33±0.01	0.21±0.01	0.69	-	0.55	-	0.64	-
BRR (-2,-2)	0.25[b]±0.06	0.17±0.04	0.69[b]±0.02	1.36[b]±0.24	0.42±0.03	0.83[b]±0.18	0.67[b]	5
IBLASSO (4,-2)	0.24[b]±0.07	0.18±0.04	0.69[b]±0.02	1.44[b]±0.30	0.41±0.04	0.79[b]±0.20	0.74	4
IBLASSO (4,2)	0.25[b]±0.07	0.15±0.04	0.69[b]±0.03	1.35[b]±0.27	0.42±0.04	0.90[b]±0.26	0.61[b]	5
BAYESA*B* (-2,6)	0.26[b]±0.07	0.14±0.03	0.69[b]±0.03	1.31[b]±0.26	0.42±0.04	0.97[b]±0.03	0.55	4
BAYESA*B* (4,6)	0.26[b]±0.07	0.14±0.04	0.69[b]±0.03	1.31[b]±0.26	0.42±0.04	0.96[b]±0.28	0.55	4
BAYESA*B* (-2,8)	0.26[b]±0.07	0.14±0.04	0.69[b]±0.03	1.29[b]±0.25	0.42±0.04	0.99[b]±0.30	0.53	4
RR-HET (-2 -2)	0.23±0.07	0.17±0.04	0.69[b]±0.02	1.44[b]±0.30	0.41±0.04	0.80[b]±0.20	0.74	3
BLASSO (4,2)	0.23±0.08	0.21±0.06	0.68[b]±0.03	1.37[b]±0.35	0.41±0.03	0.86[b]±0.26	0.88	4
G-BLUP	0.25[b]±0.06	0.19±0.04	0.70[b]±0.02	1.25[b]±0.03	0.46[b]±0.02	0.94[b]±0.20	0.76	6
Pedigree	0.20±0.02	0.13±0.01	0.45±0.03	0.84[b]±0.11	0.08±0.03	0.24±0.10	-	1

b: best = highest +- 0.02 for h2a, h2d, cor a, cor d and Vd/Va; 0.5 to 1.5 for bya and byd; highest minus 2 for best criteria in the last column.

Table 6. Partition of accuracy due to the three quantitative genetics information for a trait controlled by mixed (major and small gene effects) inheritance model with heritability 0.50 (method BayesA*B* (-2,8)).

Information	Additive h^2	Composition of Information	Additive Accuracy	Composition of Accuracy
1: Raw	0.26	COSEG+ IBD-LD + F-IBD-R	0.69	Calculated from data
2: AWF	0.22	COSEG+LD	0.53	Calculated from data
3: DMS	0.16	LD	0.52	Calculated from data
4: (2) minus (3)	0.06	COSEG	0.10	$Sqr(0.53^2 - 0.52^2)$
5: (1) minus (2)	0.04	F-IBD-R	-	-
6: Pedigree-Raw	0.20	COSEG + I-IBD-R	0.45	Calculated from data
7: (6) minus (4)	0.14	I-IBD-R	0.43	$Sqr(0.45^2 - 0.10^2)$
9: Parametric	0.33	ALL	-	-

I-IBD-R: individual IBD relationships; **F-IBD-R:** family IBD relationships; Sqr: square root.

CHAPTER 3

REGULARIZED AND HYBRID ESTIMATORS FOR THE EXPERIMENTAL ACCURACY OF GENOME SELECTION

Abstract

Accuracy is the main measure for evaluating the efficiency of the prediction of genomic breeding values. The traditional estimator (TE) of the experimental accuracy weights the phenotypic predictive ability by the inverse of the square root of estimated trait heritability. This estimator has some inconsistencies such as, the higher the trait heritability the lower the prediction accuracy. Furthermore, it can lead to estimates of accuracy higher than one, which is outside of the parameter space. Several alternative estimators have been proposed based on the expected accuracy conditional on assumptions of factors affecting this accuracy. However, to date, there are no proposed alternative estimators for the experimental (after obtaining the data) accuracy of genomic selection. This paper proposes and evaluates (through simulations) the performance and efficiency of a regularized estimator (RE) for the accuracy of genome wide selection (GWS). The proposed method takes into account both the genomic and trait heritabilities, in addition to the predictive ability. Also a hybrid estimator (HE), combining both experimental and expected accuracies was proposed and evaluated. The simulation study considered two narrow sense heritabilities levels (around of 0.20 and 0.35) and two genetic architectures for traits (the first consisting of small gene effects and the second consisting of a mixed inheritance model with five major genes). The comparisons of TE, RE and HE were done under four validation procedures: independent validation (IV), ten-fold validation through Jacknife allowing different markers to be selected in each cycle (TFD), ten-fold validation through Jacknife with the same markers selected in each cycle

(TFS) and without validation (WV). For IV, at higher heritability (around 0.35), RE and TE performed similarly in terms of the distance between the estimated accuracy and the parametric accuracy, but the RE underestimated and TE overestimated the accuracy. For IV at lower heritability (around 0.20), TE showed a smaller distance from the parametric accuracy but it still overestimated the accuracy, while RE underestimated it. HE was poorer for IV, notable underestimating the accuracy. For TFD and TFS validations when the accuracy asymptote (with increasing markers number) is reached, RE and HE were approximately coincident and superior to TE in terms of both distance (smaller bias) and direction of the estimation (underestimation) in all the four scenarios. For WV the HE matched the parametric accuracy curve while RE achieved similar results when all the markers were used and TE overestimated it. In general, the regularized estimator presented accuracies closer to the parametric ones, mainly when selecting markers. It was also less biased and more precise, with smaller standard deviations than the traditional estimator. The TE can be used only with IV and even in this case it overestimates the accuracy. The hybrid estimator (HE) proved to be very effective in the absence of validation and in the Jacknife procedures but is not recommended for IV. The regularized estimator revealed that not only the predictive ability of GWS methods matters, but also their capacity of precisely estimating the genomic heritability. So not only the predictive ability and bias should be used in comparing methods but also the genomic heritability and accuracy by the new method should be used altogether. The independent validation showed to be superior over the Jacknife procedures, chasing better the parametric accuracy with or without marker selection. With the RE the TFS showed to be better than the TFD validation scheme. The following inferences can be made according to the accuracy estimator and kind of validation: (i) most probable accuracy: HE without validation; (ii) highest possible accuracy: TE with independent validation; (iii) lowest possible accuracy: RE with independent validation.

1. Introduction

Genome Wide Selection (GWS - Meuwissen et al., 2001) is a technique of exacerbated importance in plant and animal breeding, allowing efficiency in genetic evaluation and anticipation of genetic gains. It is based on genomic prediction performed through phenotypes and a large number n of molecular markers widely distributed in the genome. The genomic breeding values (GBV) of N individuals are predicted by appropriate functional models, which estimate the effect of each marker on phenotypes, allowing early identification of the genetically superior individuals. However, genomic prediction poses statistical challenges such as estimability, due to the high dimensionality problem (N<<n case), and multicollinearity between the covariates, since the molecular markers are highly correlated. Those challenges require the use of statistical methods to consider the regularization of the estimation process and/or the selection of covariates (Gianola et al., 2003).

To address these drawbacks, many statistical methods were proposed for genomic selection, such as penalized estimation methods (Meuwissen et al., 2001; Van Raden, 2008), Bayesian estimation (Meuwissen et al., 2001; De Los Campos et al., 2009; Habier et al., 2011; Legarra et al., 2011), implicit regression methods (Gianola et al., 2006; Gianola et al., 2009; De Los Campos et al., 2009) and methods with dimensional reduction (Moser et al., 2009; Solberg et al., 2009; Azevedo et al., 2014), among others (Hayes, 2013). These methods were typically evaluated based on their prediction accuracy.

Accuracy is the main measure for evaluating the efficiency of the prediction of GBV. An traditional estimator for the experimental accuracy of GWS was introduced by Legarra et al. (2008) and Hayes et al. (2009a). Such estimator is given by the ratio between the predictive ability and the square root of the trait heritability. However, in some circumstances, this estimator has inconsistencies such as the fact that the higher the trait heritability the lower

the accuracy and it can lead to estimates outside the parameter space (higher than 1). Estaghvirou et al. (2013) carried out a comparative study among alternative accuracy estimators for GWS, but such estimators only differed from the estimator of Legarra et al. (2008) and Hayes et al. (2009a) in the different ways of estimating the trait heritability. In fact, to date, there are no proposed estimators for the experimental (after obtaining data) accuracy of genomic selection, but only for the expected (before obtaining data) accuracy (Daetwyler et al., 2008; 2010; Resende, 2008; Goddard, 2009; Goddard et al., 2011; Hayes et al. 2009b).

Thus, we propose an estimator for the experimental accuracy called regularized estimator, which is given by the multiplication of the traditional estimator by the square root of the molecular heritability. This correction leads the accuracy estimator to within the parameter space (from 0 to 1) and produces less biased and more precise estimates. Furthermore, the predictive ability combined with the proportion of genetic variance explained by markers yields a hybrid estimator, which uses both the experimental information and the theoretical expectation.

In addition, the estimation of accuracy is linked to the validation form and the marker selection, since when thousands effects are estimated, there is a risk of over-parameterization, i.e., experimental errors in the data explaining markers effects (Meuwissen, 2007). Thus, it is necessary to study and assess the behavior of the accuracy estimators in different forms of validation and under markers selection.

Given the above, this paper had as objective to propose and evaluate the performance and efficiency of the two new estimators (called regularized and hybrid) for the accuracy of GWS, which includes the predictive ability and both the genomic heritability and pedigree-based heritability, considering also different validation forms and the selection of markers.

2. Methods

2.1. Simulated datasets

Two random mating populations in linkage equilibrium were crossed generating a population (of size 5,000, coming from 100 families) with linkage disequilibrium (LD), which was subjected to five generations of random mating without mutation, selection or migration. The resultant population is an advanced generation composite, which presents Hardy-Weinberg equilibrium and LD. According to Viana (2004), the LD value (Δ) in a composite population is $\Delta_{ab} = \left(\frac{1-2\theta_{ab}}{4}\right)\left(p_a^1 - p_a^2\right)\left(p_b^1 - p_b^2\right)$, where a and b are two SNPs, two QTLs, or one SNP and one QTL, θ is the frequency of recombinant gametes, and p^1 and p^2 are the allele frequencies in the parental populations (1 and 2). Notice also that the LD value depends on the allele frequencies in the parental populations. Thus, regardless of the distance between the SNPs and/or QTLs, if the allele frequencies are equal in the parental population, $\Delta = 0$. The LD is maximized ($|\Delta| = 0.25$) when $\theta = 0$ and $|p^1 - p^2| = 1$. In this case, the LD value is positive with coupling and negative with repulsion (Viana, 2004; Azevedo et al., 2015).

From the advanced generation of the composite, one thousand individuals were generated with diploid genomes having a length of 200 centimorgans (cM) (L = 2 Morgans) and assuming ten equally sized chromosomes, each one with two haplotypes. We simulated a marker density by assigning 2,000 equidistant SNP markers that were separated by 0.1 cM across the ten chromosomes. One hundred of the 2,000 markers were actually genes (QTL). A total of 1,000 individuals that came from the same generation and from 20 full-sib families (each one with 50 individuals) were genotyped and phenotyped. This simulation provides a typical small effective population size (Ne = 39.22) and a large LD in the breeding populations. Ne of approximately 40 and the use of 50 individuals per family are typical values in elite breeding populations of plant species (Resende, 2002).

The QTLs were distributed in the regions covered by the SNPs. For each trait, we informed the degree of dominance (d/a being a and d are the genotypic values for one homozygote and heterozygote, respectively) and the direction of dominance (positive and/or negative). The obtained genotypic values for homozygotes were within the limits of Gmax = 100(m + a) and Gmin = 100(m - a) where m is the mean of genotypic values, which are the maximum and minimum values, respectively.

Goddard et al. (2011) presented the realized proportion (r^2_{mq}) of genetic variation explained by the markers as $r^2_{mq} = \frac{n}{n + n_{QTL}}$, where n_{QTL} is the number of QTL and n is the number of markers. With n = 2,000 markers and $n_{QTL} = 100$, we have $r^2_{mq} = 0.95$. An alternative Hayes et al. (2009b) takes $n_{QTL} = 2NeL = 2(39.22)2 = 156.88$, where L is the total length of the genome (in Morgans), producing $r^2_{mq} = 0.93$. Another approach Sved (1971) provides r^2_{mq} as $r^2_{mq} = \frac{1}{1 + 4NeS} = \frac{1}{1 + 4(39.22)0.001} = 0.86,$ where S is the distance between markers (in Morgans). These values reveal that the genome was sufficiently saturated by markers.

Traits with two genetic architectures were simulated, one following the infinitesimal model and the other with five major effects genes accounting for 50% of the genetic variability. For the former, to each of 100 QTL one additive effect of small magnitude on the phenotype was assigned (under the Normal Distribution setting). For the latter, small additive effects were assigned to the remaining 95 loci. The effects were normally distributed with zero mean and genetic variance (size of genetic effects) allowing the desired heritability level. The phenotypic value was obtained by adding to the genotypic value a random deviate from a normal distribution $N(0, \sigma^2_e)$, where the variance σ^2_e was defined according to two levels of narrow-sense heritabilities around of 0.20 and 0.35, respectively. Heritability levels were

chosen to represent one trait with low heritability and another with moderate heritability, which addressed the cases where genomic selection is expected to be superior to phenotypic selection (Azevedo et al., 2015). The magnitudes of the narrow-sense and broad-sense heritabilities are associated with an average degree of dominance level (d/a) of approximately 1 (complete dominance) in a population with intermediate allele frequencies. Simulations assumed independence of additive and dominance effects, with dominance effects having the same distribution as the additive effects (both were normally distributed with zero mean). In the simulation, it was also observed that marker alleles had minor allele frequency (MAF) greater than 5%.

2.2. Scenarios

To evaluate the proposed accuracy estimators, we studied 4 different scenarios: two heritability levels (around 0.20 and 0.35) × two genetic architectures. These 4 scenarios were analyzed considering four forms of validation and three accuracy estimators. The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32. Each kind of population (or scenario) was simulated 10 times. Accuracies and genomic heritabilities were calculated for each replicate of the simulation and averaged across replicates. For validation forms based on folds the accuracies and genomic heritabilities were also averaged across validation cycles in each replicate of the population.

2.3. Traditional accuracy estimator (TE)

Considering the phenotypic model given by $y = \mu + g + e$, where μ is the overall mean, g is a genetic effect and e is an residual effect, an accuracy estimator for genomic selection ($r_{\hat{y}g}$) was proposed by Legarra et al. (2008) and Hayes et al. (2009a) and is given by $r_{\hat{y}g} = (r_{\hat{y}y}/h)$ or $r_{\hat{y}g} = (r_{\hat{y}y}/r_{\hat{a}a})$, respectively, where $r_{\hat{y}y}$ is the predictive ability of GWS represented by correlation between the phenotype y and predicted genomic breeding values $\hat{y}$, $r_{\hat{a}a}$ is the accuracy of the pedigree-based evaluation (a and $\hat{a}$ are the true additive genetic value and estimated additive genetic value, respectively) and h^2 is the heritability of the trait given by $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ being σ_g^2 the genetic variance and σ_e^2 the residual variance. It can be seen that *h* is itself the accuracy ($r_{\hat{a}a}$) of the individual phenotypic selection.

Consider the quantities $\hat{g}_M$ and g stand for the predicted genomic breeding value and true breeding value (with variance σ_g^2), respectively. The algebraic proof of this estimator can be made considering the corrected phenotypic values ($y = g + e$) and predicted genomic breeding values ($\hat{y}$) in the validation population and under the assumption that $\hat{y}$ only captures genetic effects ($\hat{y} = \hat{g}_M$), that is, does not explain any environmental effect (e). This assumption holds only when the validation is perfect (totally independent) and simultaneously LD (linkage disequilibrium) between marker and QTL is complete.

In this way, the covariance of these two variables is as follow:

$$Cov(\hat{y}, y) = Cov(\hat{g}_M, g + e) = Cov(\hat{g}_M, g) = \sigma_g^2 .$$

This leads to $Cov(\hat{g}_M, g) = \sigma_g^2$, meaning that $\hat{g}_M$ captures the whole of g.

And the variances of $\hat{y}$ and *y* are $Var(\hat{y}) = \sigma_{\hat{y}}^2 = Var(\hat{g}_M) = \sigma_{g_M}^2$ and $Var(y) = \sigma_y^2 = \sigma_g^2 + \sigma_e^2 = \sigma_g^2/h^2$, respectively.

Thus, the predictive ability equals the correlation value between $\hat{y}$ and *y* and is given by:

$$r_{\hat{y}y} = Cor(\hat{y}, y) = Cov(\hat{y}, y)/(\sigma_{\hat{y}}\,\sigma_y) = \sigma_g^2/(\sigma_{\hat{y}}\,\sigma_y) = \sigma_g^2/(\sigma_{\hat{y}}\,(\sigma_g/h)) = r_{\hat{y}g}h.$$

Therefore, the accuracy estimator is equal to $r_{\hat{y}g} = r_{\hat{y}y}/h$.

However, this estimator has some inconsistencies. In $r_{\hat{y}g} = r_{\hat{y}y}/h$ the larger *h* the lesser $r_{\hat{y}g}$, a fact that is not consistent with reality, because, theoretically, the higher the heritability of the trait the greater the accuracy of selection. Also, this estimator produces $r_{\hat{y}g}$ outside (greater than 1) of the parameter space from 0 to 1. Conceptually, it does not make sense that $r_{\hat{y}g}$ tend to infinity when *h* approaches zero. This can lead in practice to the situation where a population without genetic variability (heritability of zero) is able to show accuracy of reasonable magnitude with GWS, which is inconsistent with the genomic selection theory.

It is important to emphasize that the genomic or molecular heritability (h_M^2) must be equal or less than the heritability of the trait (h^2) as h_M^2 is a fraction of h^2 that is captured by markers, and the maximum limit of the squared accuracy of GWS is h_M^2 (De los Campos et al. 2013; 2014). The molecular heritability is given by $h_M^2 = \sigma_{g_M}^2/(\sigma_{g_M}^2 + \sigma_e^2)$, where $\sigma_{g_M}^2 = \sum_{i=1}^{n} 2p_iq_i\sigma_{m_i}^2$ is the additive genomic variance, $\sigma_{m_i}^2$, p_i and q_i are the variance and allele frequency of marker *i*, respectively (Gianola et al., 2009).

There is a need for correction in the traditional expression of the genomic accuracy. The literature (Estaghvirou et al., 2013) reports efforts to improve the estimate of *h* used in $r_{\hat{y}g} = r_{\hat{y}y}/h$, but not yet addressed the need to modify this expression. This paper addresses this issue.

2.4. Regularized estimator (RE)

The proposed regularized estimator is given by $r_{g_M g} = r_{\hat{y}y}\frac{h_M}{h}$ or $r_{g_M g} = r_{\hat{y}g}h_M$ where h_M is the square root of the genomic or molecular heritability. This estimator multiplies the proximity (distance) between *y* and $\hat{y}$ given by $r_{\hat{y}y}$, by the proximity (distance) between h_M and *h* (given by $\frac{h_M}{h}$), producing what is needed, i.e., the proximity (distance) between g_M and *g*. In this case, to multiply $r_{\hat{y}y}$ by $\frac{h_M}{h}$ decreases the accuracy, i. e., penalizes $r_{\hat{y}y}$, as it should be.

Under the assumption that $\hat{y}$ captures both genetic and environmental effects, i.e. $\hat{y} = \hat{g}_M + \hat{e}$, as the LD is imperfect ($Cov(\hat{g}_M, g) = \sigma^2_{g_M}$) and / or validation is not performed or is inaccurate which leads to covariance of errors ($Cov(\hat{e}, e) = \sigma^2_e$).

In this case, we have:

$$Cov(\hat{y}, y) = Cov(\hat{g}_M + \hat{e}, g + e) = Cov(\hat{g}_M, g) + Cov(\hat{e}, e) = Var(\hat{g}_M) + Var(e)$$
$$= \sigma^2_{g_M} + \sigma^2_e = Var(y_M)$$

The variances of the corrected phenotypic values *y* and predicted values $\hat{y}$ are equal to $Var(y) = \sigma^2_y = \sigma^2_g + \sigma^2_e = \sigma^2_g / h^2$ and $Var(\hat{y}) = \sigma^2_{y_M} = \sigma^2_{g_M} / h^2_M$, respectively.

Thus, the predictive ability equals the correlation between y and $\hat{y}$ given by:

$$r_{\hat{y}y} = Cor(\hat{y}, y) = Cov(\hat{y}, y)/(\sigma_{\hat{y}}\,\sigma_y) = \sigma^2_{y_M}/(\sigma_{y_M}\,\sigma_y) = \sigma^2_{y_M}\Big/\left(\frac{\sigma_{g_M}}{h_M}\frac{\sigma_g}{h}\right) =$$
$$= \frac{\sigma^2_{y_M} h_M h}{\sigma_{g_M}\sigma_g} = \frac{\sigma_{y_M} h \sigma_{g_M}}{\sigma_{g_M}\sigma_g}\frac{\sigma_{g_M}}{\sigma_{g_M}} = \frac{\sigma_{y_M} h \sigma^2_{g_M}}{\sigma_{g_M}\sigma_g\sigma_{g_M}} = \frac{\sigma_{y_M} h r_{g_M g}}{\sigma_{g_M}} = \frac{h r_{g_M g}}{h_M}$$

and, therefore, the accuracy estimator is equal to $r_{g_M g} = r_{\hat{y}y}\frac{h_M}{h}$.

The regularized estimator $r_{g_M g} = r_{\hat{y}y} \frac{h_M}{h}$ has good statistical property, producing $r_{g_M g}$ in the parameter space as product of two fraction, since $\frac{h_M}{h}$ is always smaller or equal to 1. Also, the lower $\frac{h_M}{h}$, the lower $r_{g_M g}$, fact which is consistent with reality. Thus, the inconsistencies in the traditional formula are corrected for. This estimator is conservative (produces smaller accuracies than the traditional formula) and is a function of three parameters, not two, also using genomic heritability. The ratio of the heritabilities takes into account the efficiency of markers in capturing QTL, i.e., it considers the degree of imperfection in LD.

2.5. Hybrid estimator (HE)

The hybrid estimator (HE) combines the experimental predictive ability and the theoretical expectation of $r_{mq}^2 = \sigma_{g_M}^2 / \sigma_g^2$, which also is a regularized estimator. Assuming $\sigma_{y_M}^2 \approx \sigma_y^2$, we have $h_M^2 / h^2 \approx \sigma_{g_M}^2 / \sigma_g^2$. Being $r_{mq}^2 = \sigma_{g_M}^2 / \sigma_g^2$ the proportion of g explained by markers, we have $h_M / h = r_{mq}$. Thus, using the regularized estimator (RE), the accuracy is given by $r_{g_M g} = r_{\hat{y}y} \frac{h_M}{h} = r_{\hat{y}y} r_{mq} = r^*_{g_M g}$, where $r_{mq}^2 = \frac{n}{n + M_e}$, n is the number of markers and $M_e = 2N_e L$, being N_e the effective population size and L the genome size in Morgans (Goddard et al., 2011; Meuwissen et al., 2011). It is noteworthy that r_{mq}^2 changes with the number of selected markers. The HE provides an accuracy estimate without requiring the estimation of h_M^2 and h^2. And, interestingly, it penalizes the selection of a very small number of markers.

A deterministic formula for the predictive ability is derived and presented in the Appendix 1. It states the relation between it and the genomic heritability.

2.6. Parametric Accuracy

The computing of parametric accuracies under the additive model was done by the formula (Resende et al.,2008; Grattapaglia and Resende, 2011; Goddard et al., 2011; Resende et al., 2014):

$r_{g\hat{g}} = \sqrt{\frac{r_{mq}^2(Nr_{mq}^2 h^2 / n_{QTL})}{1+ Nr_{mq}^2 h^2 / n_{QTL}}}$, where n_{QTL} is the number of QTL, N is number of genotyped and phenotyped individuals, h^2 is the trait heritability and r_{mq}^2 proportion of genetic variance explained by markers.

2.7. Supervised RR-BLUP

In the context of the genomic selection, according to Meuwissen et al. (2001), the basic linear model is as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathrm{X}\mathbf{m} + \mathbf{e}$$

where **y** is the phenotypes vector ($N\times 1$, being N the number of genotyped and phenotyped individuals), **1** is a vector with all entries equal to 1 ($N\times 1$), μ is the average of the trait, **m** is the vector of additive genetic markers effects ($n\times 1$, being n is the number of markers) with incidence matrix X ($N\times n$) and $\mathrm{m} \sim \mathrm{N}(0, \mathrm{I}\sigma_m^2)$ being σ_m^2 the variance of markers, **e** is the model of the residual vector with $\mathrm{e} \sim \mathrm{N}(0, \mathrm{I}\sigma_e^2)$ and σ_e^2 is residual variance. In this study, we evaluated the accuracies under RR-BLUP of Meuwissen et al. (2001), and RR-BLUP_B (Resende Jr et al., 2012) methods.

The RR-BLUP method was applied to each data set, considering the total number of markers. After this preliminary analysis, markers groups of size 50, 100, 150 and so on up to 2,000 were selected based on the highest magnitudes of their effects.

2.8. Estimation and validation populations

The estimation population is used to estimate the markers effects while the validation population is used to analyze the efficiency of the estimated effects in recovering individuals genomic values in an independent sample of the population.

An estimation-validation approach was cross-validation by the Jacknife procedure. According to this method, the original data set with 1,000 individuals was divided into 10 training data sets of size equal to 100 individuals (ten-folds procedure or TF) (Efron, 1982).

The genomic heritability was estimated as the average of heritabilities obtained in each of the cycles. The estimation of accuracy was done in two ways, first as the average of the predictive abilities obtained in each of the cycles allowing for different markers to be selected (TFD scheme) in each cycle, and second, restricting the selected markers to be the same (TFS scheme) in each cycle, after identifying them through the average effects across ten cycles. Considering the supervised RR-BLUP method the selection of markers was made through the greatest magnitudes of effects. However, the chosen sets of markers were alternatively variable in each cycle (form I of the Jacknife validation or TFD) or unique for all validation cycles (form II of the Jacknife validation or TFS).

In this study the following forms of validation were considered: (i) physically distinct (2 different subpopulations), called independent validation (IV); (ii) Jacknife procedure considering the forms I (TFD) and II (TFS); (iii) without validation (WV) with the population used for estimation and validation at the same time.

The data were simulated using the RealBreeding software (Viana, 2011). All computational routines of analyses were implemented in R software (R Development Core Team, 2010) using the rrBLUP package and mixed.solve function.

3. Results and Discussion

Results concerning the average accuracy using the regularized estimator (RE), the traditional estimator (TE) and the hybrid estimator (HE) for independent validation (900 and 100 individuals in the estimation and validation populations, respectively) are presented in Figure 1. For IV, at high heritability (0.37 and 0.32), RE and TE performed similarly in terms of the distance between the estimated accuracy and the parametric accuracy, but the RE underestimated and TE overestimated the accuracy, then favoring RE. For IV at low heritability (0.22 and 0.20), TE showed a smaller distance from the parametric accuracy but it still overestimated accuracy, while RE underestimated it. HE was poorer for IV, underestimating notable the accuracy.

The behavior of the parametric accuracy curve along with the number of the biggest effects markers in the analysis shows that an asymptote is reached around 500 markers and, after that, a constant value is kept up to the total number of 2,000 markers. For the IV, estimators of the accuracies approximately reach the maximum value also at the 500 selected markers. This result indicates that in average, 5 markers are enough to capture one QTL. After that point, the accuracy tends to stabilize. This occurred for all estimators, all scenarios in IV and also for the without validation case. These results were likely to the parametric case.

Results of the average accuracy by the regularized estimator (RE), traditional estimator (TE) and hybrid estimator (HE) by the Jacknife procedure (with k=100) considering the I and II validation forms (TFD and TFS, respectively) are presented in Figure 2. For all estimators the results showed notable greater distances from the parametric values than when

considering the independent validation (Figure 1). Then the independent validation showed to be superior over the Jacknife procedures. This is in accordance with the reports of Wray et al. (2013a, b). The independent validation showed to be superior over the Jacknife procedures, approaching better the parametric accuracy with or without marker selection. With the RE, the TFS showed to be better than the TFD validation scheme.

For the key value of 500 markers, RE produces accuracy values very close to parametric ones, in the scenarios 2 and 4 (Figure 1) and all scenarios in Figure 2. In these situations, the TE estimator overestimated the accuracies. In addition, the TE exceeded the parameter space (values greater than 1) in some scenarios (Figure 2). This is inadmissible for a good estimator. For the RE and the HE it never occurred. In the comparative study of Estaghvirou et al. (2013) selection of markers was not considered and even so out of space parameter estimates were obtained. This fact has also been observed in genomic selection analysis using real data (Resende et al., 2012).

For TFD and TFS validations when (number of markers equal to 500) the accuracy asymptote (with increasing markers number) is reached, RE and HE were approximately coincident and superior to TE in terms of both distance (smaller bias) and direction of the estimation (underestimation) in all the four scenarios. In this case (number of markers equal to 500), the parametric values were 0.64, 0.68, 0.61 e 0.65; the estimates for RE were 0.51, 0.55, 0.48 and 0.55; and the estimates for TE were 1.13, 0.99, 1.09 and 1.02, for the four scenarios, respectively. This shows that TE should not be used with cross-validation.

The behaviors of the genomic heritability (h^2_M) estimates through cross-validation were similar to the parametric accuracy curve by IV, reaching a maximum and then keeping constant (Figure 3). The predictive ability with cross-validation showed a different behavior, decreasing with increase on the numbers of markers (Figure 3).

For the TFD and TFS, estimators of the accuracies approximately reach the maximum value also at the 500 selected markers. After that they tend to decrease (Figure 2). The decay in accuracy with increasing number of markers is in accordance with Fernando et al. (2007) also working with RR and two thousand markers. As in Figure 3, Wray et al. (2013a, b) also states that the predictive ability decrease with increasing number n of independent markers according to the ratio $\frac{n}{N}$, decreasing with increase on n in relation to N (number of individuals). This result is due to the fact that increasing n causes greater variance of the estimated genetic relationship coefficients (Wray et al., 2013a, b). This decrease in predictive ability was also observed in practice associated to cross-validation for a few traits (Resende et al., 2012).

The results of the average accuracy by the regularized estimator (RE), the traditional estimator (TE) and approximate estimator (HE) without validation (WV) are presented in Figure 4. For WV the HE matched very closely the parametric accuracy curve while RE and TE overestimated it. However, when all markers were used, RE also achieved accuracies close to the parametric. In general, compared to TE, the regularized estimator presented accuracies closer to the parametric ones, mainly when selecting markers. It was also less biased and more precise, with smaller standard deviations than the traditional estimator. In all scenarios, the RE and the HE proves to be conservative with respect to estimation of accuracy. In theory, the RE assumes that validation was poorly made and the HE does not need validation. Thus, without validation, the best results were for HE (coincident with parametric value) and RE, in this order. Without marker selection the RE without validation matched perfectly with the parametric accuracy. By its turn, TE was the worst and exceeded the parameter space in the entire curve, proving to be completely inadequate when there is no

validation or when it is poorly made. Even in such cases, the use of the adequate accuracy estimator (other than TE) can guarantee a valid estimate.

The standard deviations of the accuracy estimates with the RE, TE and HE are presented in Figures 5 and 6 for independent validation and forms I (TFD) and II (TFS) of the Jacknife procedure, respectively. For independent validation the best (the ones with smaller standard deviation) estimators for the accuracy were HE, RE and TE, in this order, for all scenarios. For the Jacknife procedures the same tendencies were observed. Then RE and HE methods were by far more precise than TE.

The TE can be used only with IV and even in this case it overestimates accuracy and is less precise. The hybrid estimator (HE) proved to be very effective in the absence of validation and in the Jacknife procedures but is not recommended for IV. The regularized estimator revealed that not only the predictive ability of GWS methods matters, but also their capacity of precisely estimating the genomic heritability (Azevedo et al., 2015). So not only the predictive ability and bias should be used in comparing methods but also the genomic heritability and accuracy by the new method should be used altogether. The following inferences can be made according to the accuracy estimator and kind of validation: (i) most probable accuracy: HE without validation; (ii) highest possible accuracy: TE with independent validation; (iii) lowest possible accuracy: RE with independent validation.

4. References

Azevedo, C. F., F. F. Silva, M.D.V. Resende, M. S. Lopes, N. Duijvesteijn, S. E. F. Guimarães, P. S. Lopes, M. J. Kelly, J. M. S. Viana, E. F. Knol. 2014. Supervised independent component analysis as an alternative method for genomic selection in pigs. **Journal Animal Breeding and Genetics**, 131(6):452–461.

Azevedo, C. F., M. D .V. Resende, F. F. Silva, J. M. S. Viana, M. S. Valente, M. F. R. Resende Jr, P. Munoz. 2015. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics (Online)**, v. 16, p. 105.

Daetwyler, H. D., B. Villanueva, J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. **PLoS One**, 3:e3395.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, et al. 2010. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, 185:1021–1031.

De los Campos, G., D. A. Sorensen, D. Gianola. 2014. Genomic Heritability: What is it? In: **Proceedings of the 10th WCGALP**, Vancouver, Canada, August 17-22.

De los Campos, G., D. A. Sorensen. 2013. A Commentary on Pitfalls of Predicting Complex Traits from SNPs. **Nature Reviews Genetics**, 14(12):894-894.

De los Campos, G., D. Gianola, G. J. M. Rosa. 2009. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science**, 87:1883-1887.

Efron, B. 1982. **The jacknife, the bootstrap and other resampling plans**. Philadelphia: Society for Industrial and Applied Mathematics, 385p.

Estaghvirou, S. B. O., J. O. Ogutu, T. Schulz-Streeck, C. Knaak, M. Ouzunova, A. Gordillo, H.P. Piepho. 2013. Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. **BMC Genomics**, 14, 860.

Fernando, R. L., D. Habier, C. Stricker, J. C. M. Dekkers, L. R. Totir. 2007. Genomic selection. **Acta Agriculturae Scandinavica**, 57(4):192-195.

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

Gianola, D., M. Perez-Enciso, M. A. Toro. 2003. On marker-assisted prediction of genetic value: beyond the ridge. Genetics, 163:347-365.

Gianola, D., R. L. Fernando, A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, 173:1761–1776.

Goddard, M. E. 2009. Genomic selection: prediction of accuracy and maximization of long term response. **Genetica**, 136(2):245-257.

Goddard, M. E., B. J. Hayes, T. H. E Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal of Animal Breeding and Genetics**, 128(6):409-421.

Grattapaglia, D., M. D. V. Resende. 2011. Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, 7:241-255.

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

Hayes, B. J. 2013. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In: Gondro, C., J. Van Der Werf, B. Hayes (Eds.). **Genome-Wide Association Studies and Genomic Prediction**. Armidale.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, M. E. Goddard. 2009a. Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, 92: 433-443.

Hayes, B. J., P. M. Visscher, M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. **Genetics Research**, 91:47-60.

Legarra, A., C. Robert-Granie, E. Manfredi, J. M. Elsen. 2008. Performance of genomic selection in mice. **Genetics**, 180:611-618.

Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, S. Fritz. 2011. Improved Lasso for genomic selection. **Genetics Research**, 93(1):77-87.

Meuwissen, T. H. E, T. Luan, J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. **Journal of Animal Breeding and Genetics**, 128(6):429-39.

Meuwissen, T. H. E. 2007. Genomic selection: marker assisted selection on genome-wide scale. **Journal of Animal Breeding and Genetics**, 124:321-322.

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

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics Selection Evolution**, 41:41-53.

R Development Core Team. 2010. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, Available: <http://www.R-project.org>.

Resende Jr, M. F. R., P. R. M. Valle, M. D. V. Resende, D. J. Garrick, R. L. Fernando, J. M. Davis, E. J. Jokela, T. A. Martin, G. F. Peter, M. Kirst. 2012. Accuracy of genomic selection methods in a standard dataset of loblolly pine. **Genetics**, 190:1503-1510.

Resende M.D.V. 2002. **Genética biométrica e estatística no melhoramento de plantas perenes**. Brasília: Embrapa Informação Tecnológica, 975p.

Resende, M. D. V. de, P. S. Lopes, R. L. Silva, I. E. Pires. 2008. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, 56:63-78.

Resende, M. D. V., M. F. R. Resende Jr., C. Sansaloni, C. Petroli, A. A. Missiaggia, A. M. Aguiar, J. I. M. Abad, E. Takahashi, A. M. Rosado, D. Faria, G. Pappas, A. Kilian, D. Grattapaglia. 2012. Genomic Selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, 194:116-128.

Resende, R. M. S., M. Casler, M. D. V. Resende. 2014. Genomic Selection in Forage Breeding: Accuracy and Methods. **Crop Science**, 54:143.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, 41:299.

Sved, JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. **Theoretical Population Biology**, 2:125-141.

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

Viana, J. M. S. 2004. Quantitative genetics theory for non-inbred populations in linkage disequilibrium. **Genetics and Molecular Biology**, 27(4):594-601.

Viana, J. M. S. 2011. **Programa para análises de dados moleculares e quantitativos RealBreeding**. Viçosa: UFV.

Wray, N. R., J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, P. M. Visscher. 2013a. Author reply to A commentary on Pitfalls of predicting complex traits from SNPs. **Nature Reviews Genetics**, 14(12): 894-894.

Wray, N. R., J. Yang,B. J. Hayes,A. L. Price, M. E. Goddard, P. M. Visscher. 2013b. Pitfalls of predicting complex traits from SNPs. **Nature Reviews Genetics**, 14:507–515.

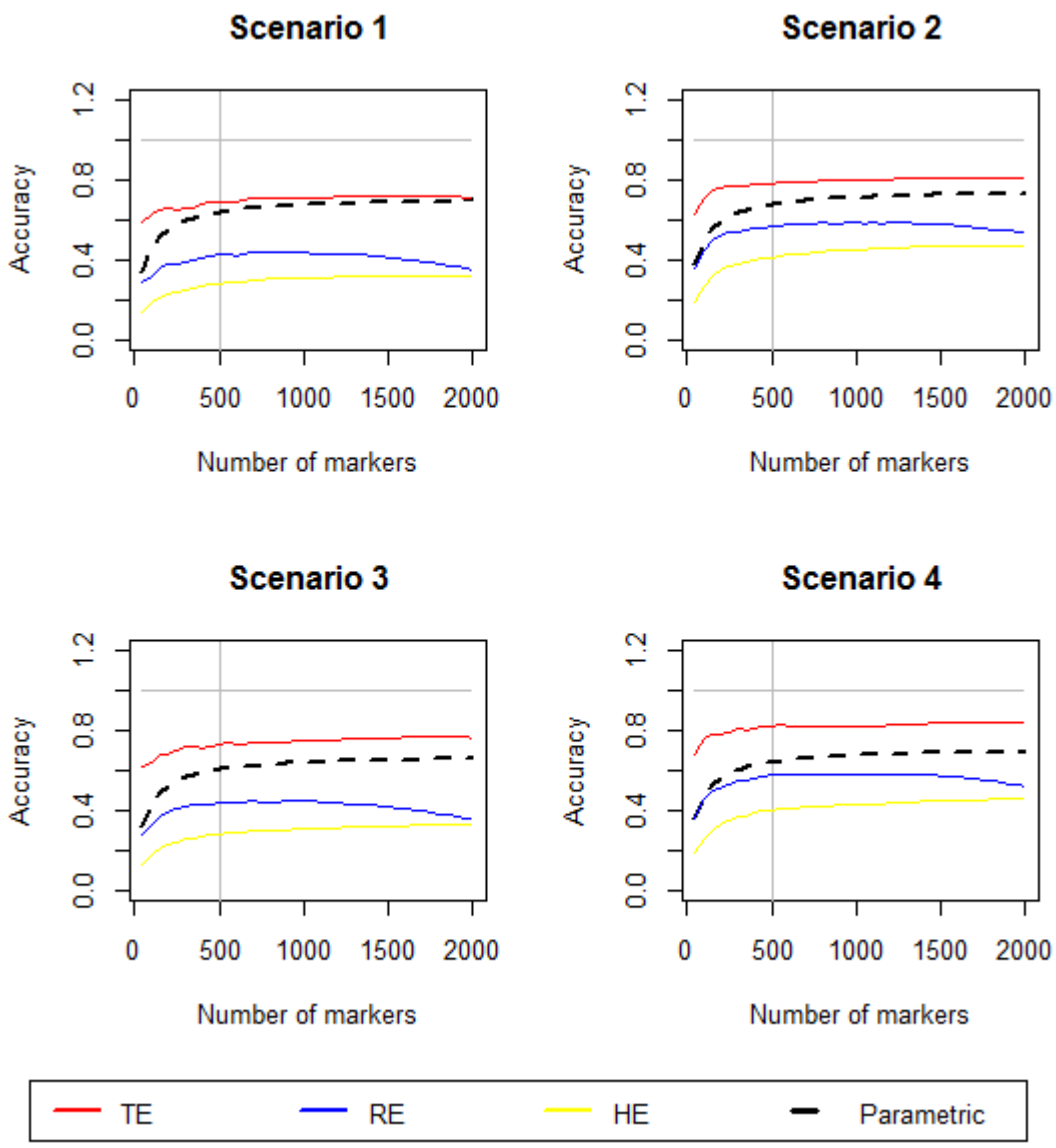

Figure 1. Average prediction accuracy by the regularized estimator (RE - Blue), the traditional estimator (TE - Red) and the hybrid estimator (HE - Yellow) for independent validation. The gray line color stands for maximum accuracy, which is one; and the ordinate in 500 refers to the number of markers wich maximizes the estimated accuracy. The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

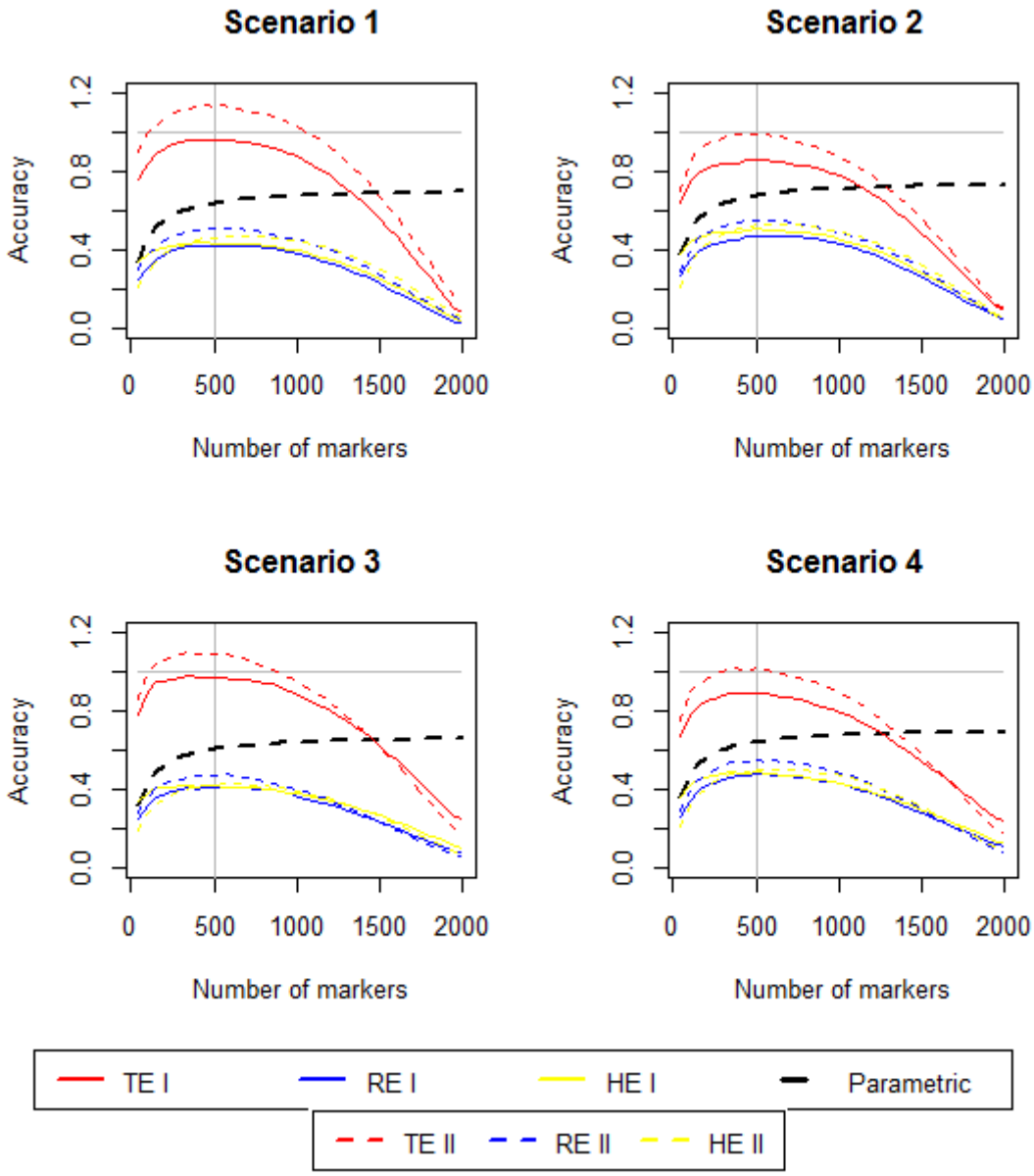

Figure 2 . Average prediction accuracy by the regularized estimator (RE – Blue), the traditional estimator (TE – Red) and the hybrid estimator (HE - Yellow) for Jacknife procedure considering forms I and II (TFD and TFS, respectively). The gray line color stands for maximum accuracy, which is one; and the ordinate in 500 refers to the number of markers wich maximizes the estimated accuracy. The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

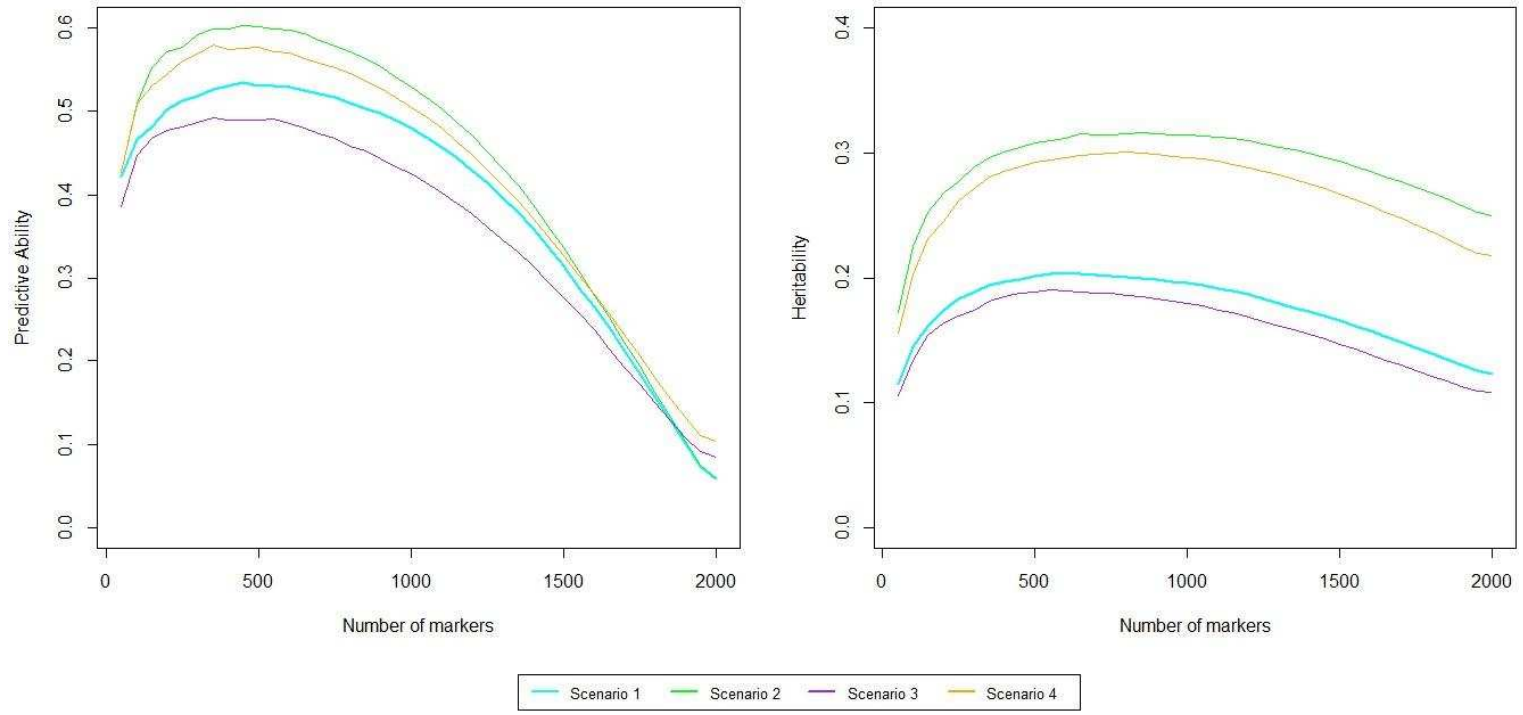

Figure 3. Behavior of the predictive ability and genomic heritability across selected groups of SNPs with cross-validation. The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

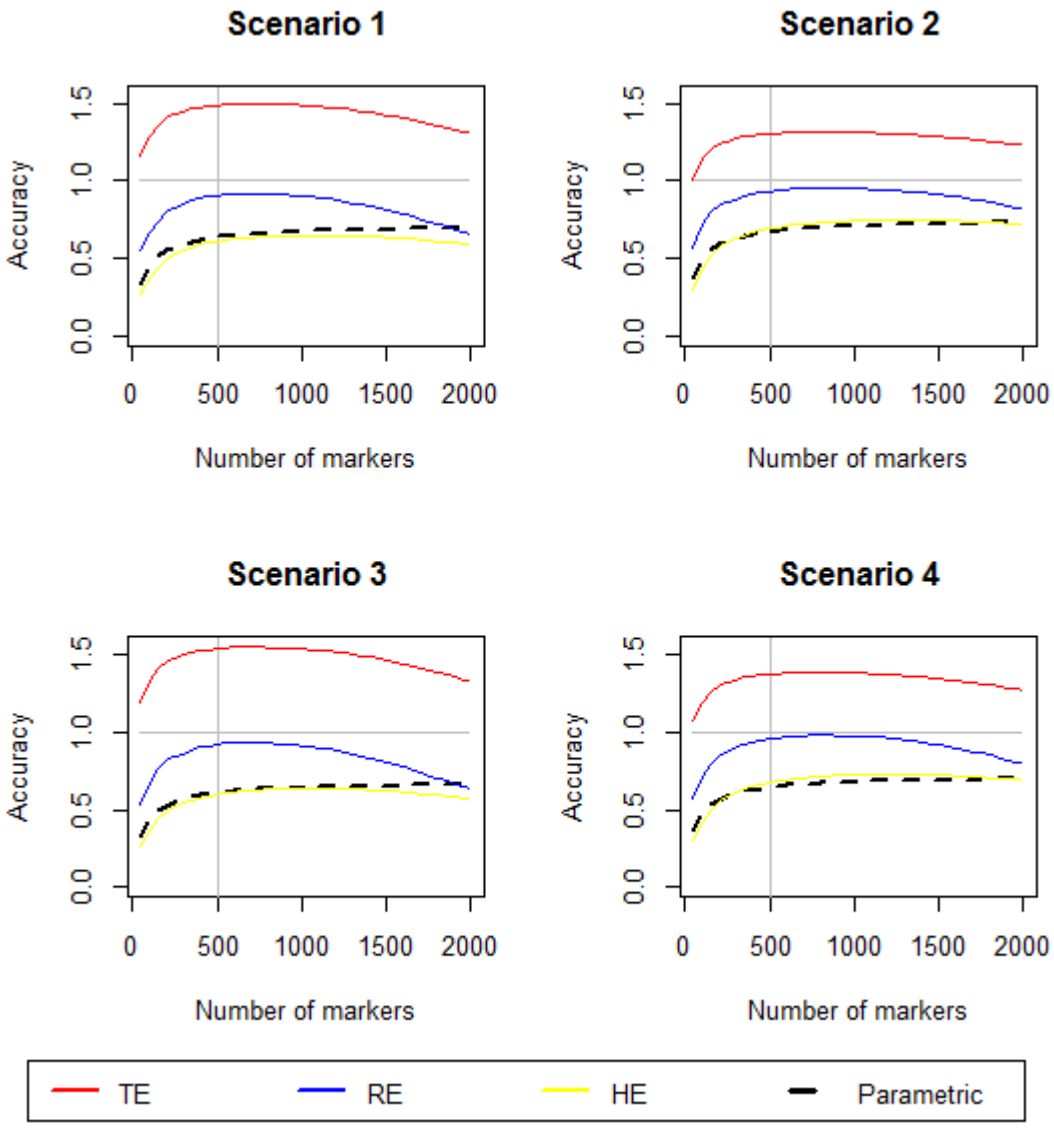

Figure 4. Average prediction accuracy by the regularized estimator (RE - Blue), the traditional estimator (TE - Red) and the hybrid estimator (HE - Yellow) without validation. The gray line color stands for maximum accuracy, which is one; and the ordinate in 500 refers to the number of markers wich maximizes the estimated accuracy.The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

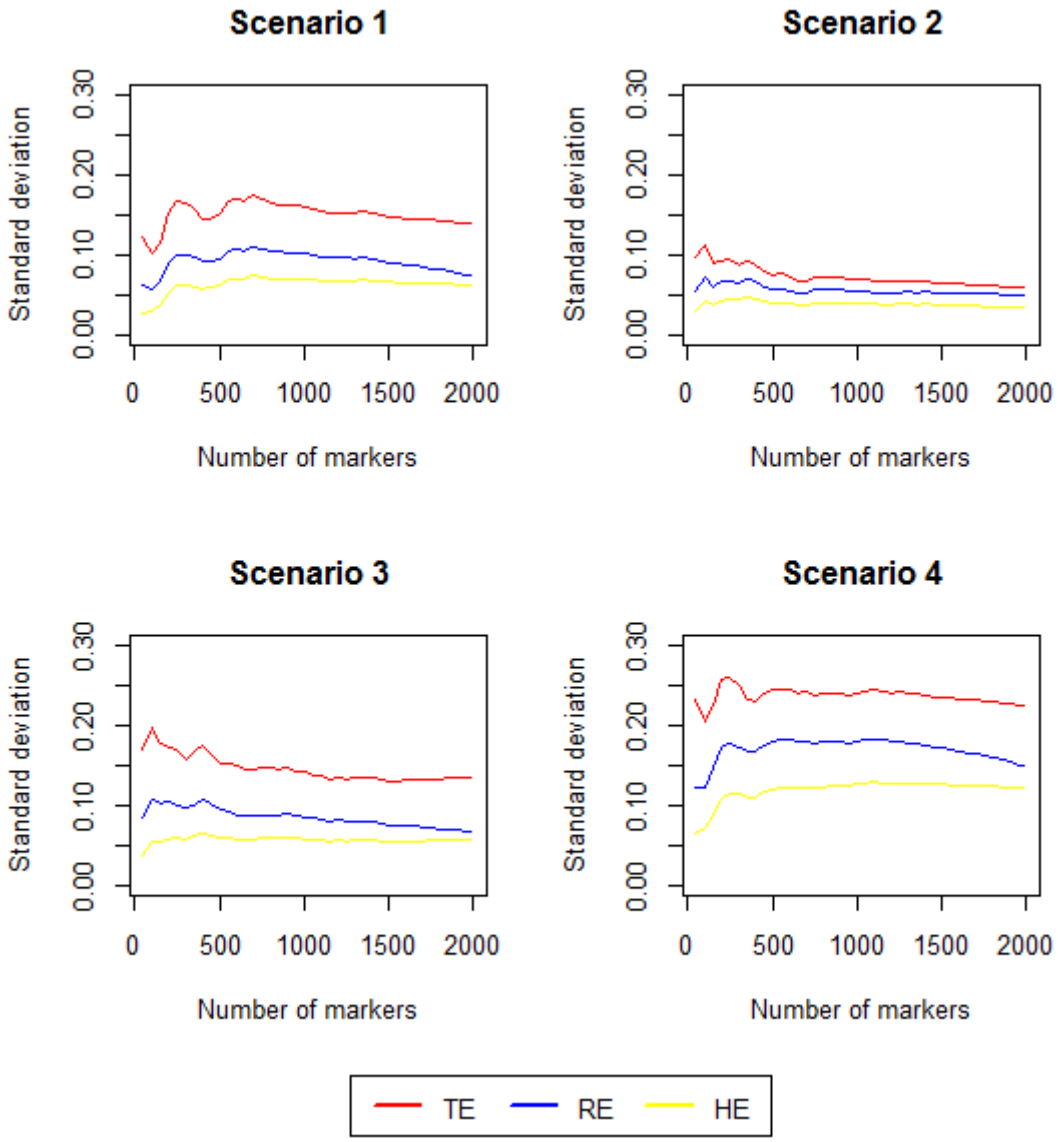

Figure 5. Standard deviation of the accuracy by the regularized estimator (RE - Blue) and the traditional estimator (TE - Red) and the hybrid estimator (HE - Yellow) for independent validation. The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

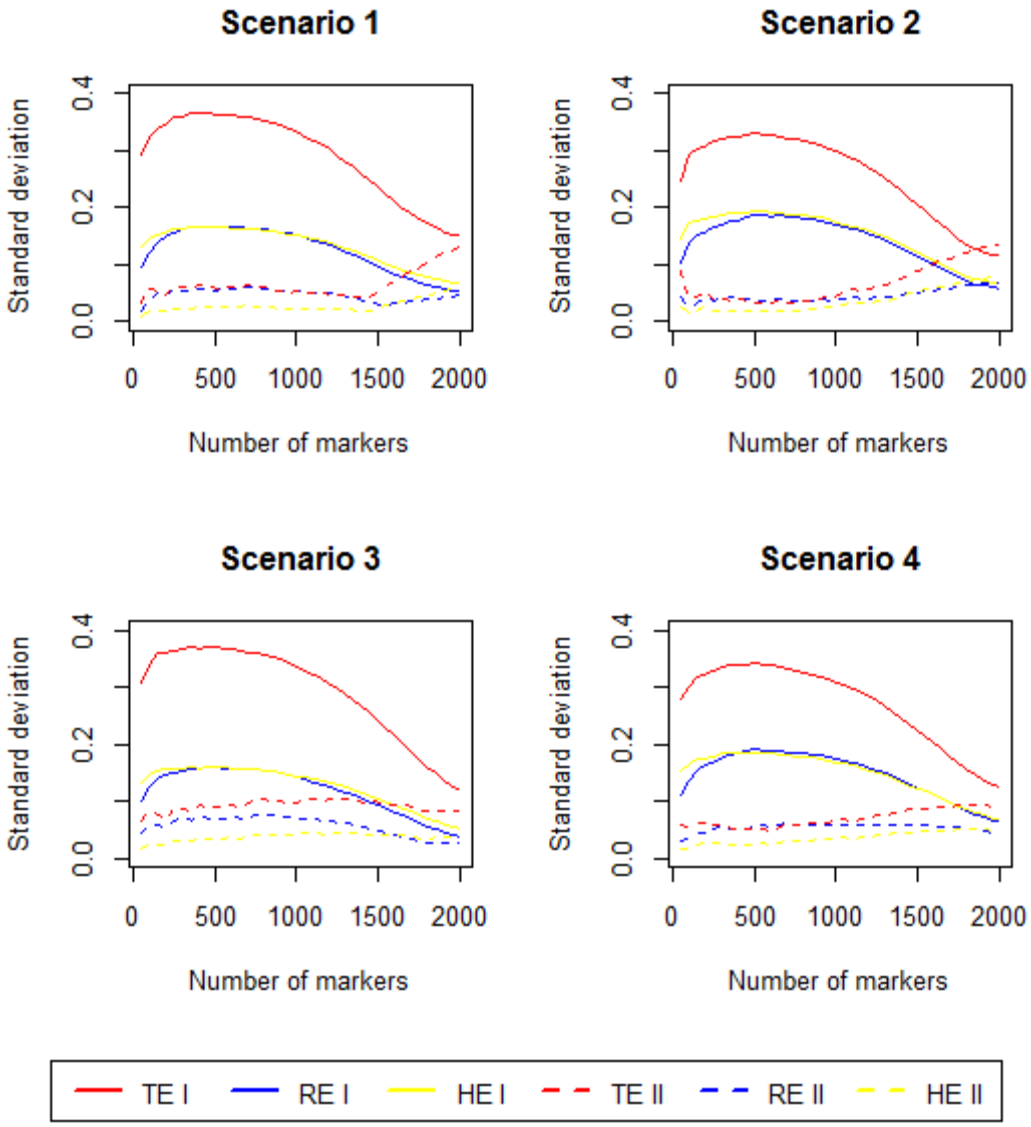

Figure 6. The standard deviation of by the regularized estimator (RE – Blue), the traditional estimator (TE – Red) and the hybrid estimator (HE - Yellow) for Jacknife procedure considering forms I and II (TFD and TFS, respectively). The scenarios were defined as: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effects genes and heritability 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability 0.32.

Appendix 1

Deterministic formula for the predictive ability $r_{\hat{y}y}$, molecular heritability h_M^2 and accuracy $r_{\hat{g}g}$ of GWS.

From the expression $r_{mq}^2 = \frac{h_M^2}{h^2}$ we get $h_M^2 = r_{mq}^2 h^2$ (1).

From the expression for the expected squared accuracy

$$r_{g\hat{g}}^2 = \frac{r_{mq}^2 (N r_{mq}^2 h^2 / n_{QTL})}{[1 + N r_{mq}^2 h^2 / n_{QTL}]} \quad (2)$$

substituting n_{QTL} for M_e we have $r_{g\hat{g}}^2 = \frac{r_{mq}^2 (N r_{mq}^2 h^2 / M_e)}{[1 + N r_{mq}^2 h^2 / M_e]}$ (3).

As $h_M^2 = r_{mq}^2 h^2$ (1) it follows that the squared experimental accuracy is $r_{g\hat{g}}^2 = \frac{r_{mq}^2 (N h_M^2 / M_e)}{[1 + N h_M^2 / M_e]}$ (4).

From (4) the squared accuracy of the RE we get the squared predictive ability as:

$$r_{\hat{y}y}^2 = r_{g\hat{g}}^2 \frac{h^2}{h_M^2} = \frac{r_{mq}^2 (N h_M^2 / M_e)}{[1 + N h_M^2 / M_e]} \frac{h^2}{h_M^2} = \frac{r_{mq}^2 (N h^2 / M_e)}{[1 + N h_M^2 / M_e]} = \frac{(N h_M^2 / M_e)}{[1 + N h_M^2 / M_e]} = \frac{N h_M^2}{N h_M^2 + M_e} \quad (5)$$

which simplifies to $r_{y\hat{y}}^2 = \frac{N h_M^2}{N h_M^2 + M_e} = \frac{h_M^2}{h_M^2 + \frac{M_e}{N}} = \frac{1}{1 + \frac{M_e}{N h_M^2}}$ (6), showing that it depends mainly

on h_M^2 and N. From (6) we get back to the squared experimental accuracy as

$$r_{g\hat{g}}^2 = r_{mq}^2 r_{\hat{y}y}^2 = \frac{h_M^2}{h^2} r_{\hat{y}y}^2 = \frac{h_M^2}{h^2 + \frac{M_e h^2}{N h_M^2}} \quad (7).$$

From the formula for the squared predictive ability via the RE, the genomic heritability is

given by $h_M^2 = \frac{r_{\hat{y}y}^2 M_e}{(1 - r_{\hat{y}y}^2) N}$ (8).