Avaliação da capacidade preditiva de metodologias de seleção genômica ampla em diferentes cenários simulados de arquiteturas genômicas

Marcelo Silva de Freitas[1], Diogo Anastácio Garcia[1], Mônica Corrêa Ledur[2], Jane de Oliveira Peixoto[2], Rodrigo José de Almeida Torres[1]

[1]BRF/SA; Rodovia BR 277, nº 3001, CEP 82305-100, Curitiba-PR. e-mail: marcelo.freitas@brf-br.com.
[2]Embrapa Suínos e Aves; Concórdia – SC.

Resumo: Foram simulados 35000 SNPs e 19996 animais, de acordo com duas arquiteturas genômicas distintas, com o objetivo de avaliar a capacidade preditiva de diversas metodologias de seleção genômica ampla. Em um cenário de arquitetura genômica homogênea, as acurácias variaram de 0.57 a 0.66. Em geral, as metodologias exibiram resultados semelhantes, indicando que, para características poligênicas, é possível escolher metodologias que são mais fáceis de implantar e são mais eficientes. Para a característica sob uma arquitetura genômica heterogênea, as acurácias variaram de 0.67 a 0.91. As metodologias bayesianas apresentaram as maiores acurácias nesse cenário heterogêneo e são, provavelmente, a melhor opção para a avaliação genômica de características controladas por poucos genes. A decisão sobre que metodologia de seleção genômica ampla implantar, em um programa de melhoramento, deve ser baseada no conhecimento a priori sobre a arquitetura genômica da característica de interesse.

Palavras–chave: acurácia; polimorfismos de base única; valor genético genômico

Evaluation of the predicting ability of whole genomic selection methodologies under different simulated genomic architecture scenarios

Abstract: 35000 SNPs and 19996 animals were simulated, based on two different genomic architectures, in order to evaluate the predicting ability of several whole genomic selection methodologies. For the homogeneous genomic architecture scenario, accuracies ranged from 0.57 to 0.66. In general, methodologies showed similar accuracies, indicating that, for polygenic traits, it is possible to choose methodologies that are easier to implement and are more efficient. For the heterogeneous genomic architecture trait, accuracies ranged from 0.67 to 0.91. Bayesian methodologies presented the higher accuracies on this heterogeneous scenario and are probably the best option for the genomic evaluation of traits controlled by a few genes. The decision about which whole genomic selection methodology to implement in a breeding program should be based on a priori knowledge about the genomic architecture of the trait of interest.

Keywords: accuracy; genomic breeding value; single nucleotyde polymorphism

Introduction

Several whole genomic selection methodologies (WGS) have been proposed, with different approaches regarding *a priori* information about the genomic architecture of the trait. According to Resende et al. (2012), an optimum WGS methodology should accommodate a genomic architecture, perform shrinkage and variable selection. Due to the great variety of WGS methodologies available, it is necessary to identify which methodologies would produce higher accuracy genomic breeding values. The objective in this study was to evaluate several WGS methodologies predicting abilities under two different genomic architecture scenarios.

Material e Methods

Two genomic architectures were simulated, where one was composed by 35000 single nucleotide polymorphisms (SNPs) with small and very similar effects (homogeneous), and the other composed by 1750 SNPs with large, medium and small effects, and 33250 SNPs with null effects (heterogeneous). SNPs were independent, but were in different levels of linkage disequilibrium with the genes that influenced the simulated traits. Founders received random genotypes for each SNP, and the other animals received genotypes according to their parents' genotypes, based on the occurrence probabilities of each genotype, following a biallelic heritage model. Having the genotypes of each animal, and allelic substitution effects for each SNP, genotype effects were created per SNP, considering that all SNPs were additive. This sequence was performed for all 19996 animals and 35000 SNPs to create genomic breeding values (GEBVs) for each animal, which was equal to the sum of genotype effects of each

individual on each SNP. Residual effects for each animal were sampled from a Gaussian distribution with mean 0 and variance σ_e^2. Phenotypes were created as a function of the sum of the trait mean, GEBV and residual of each animal. Fixed effects were not simulated. To compose the training population, 1000 animals were randomly sampled from two generations. For the genomic evaluations and prediction of GEBVs, software AlphaBayes (Hickey e Tier, 2009), BGLR (Perez et al., 2010), GS3 (Legarra et al., 2013) and BLUPF90 (Misztal et al., 2012) and its respective methodologies were used, as described on Table 1.

Table 1. Whole genomic selection software and methodologies.

Methodologies	AlphaBayes	BGLR	GS3	BLUPF90
Bayes A	X[a]	X		
Bayes B	X	X		
Fast Bayes B	X			
Bayes Cπ	X	X	X	
GBLUP	X	X	X	
Bayesian LASSO		X		
Improved Bayesian LASSO			X	
BLUP				X
ssGBLUP				X
ssGBLUPw				X

[a]: X stands for methodology implemented in the software

To obtain *a posteriori* means for the hyperparameters and GEBVs using bayesian methodologies, 200000 gibbs sampling iterations were run, with a burn in period of 100000 and thinning interval of 10 iterations. A single trait model was fitted, for the Bayesian methodologies and GBLUP, which included the effects of the mean, as fixed, and additive SNP and residual, as random effects. For the weighted ssGBLUP, 20 iterations were performed, as proposed by Wang et al. (2012), but with standard parameters as proposed by Misztal et al. (2012). A single trait model was fitted, for the BLUP and ssGBLUP methodologies, which included the effects of the mean, as fixed, and additive animal effect and residual, as random effects. Accuracy, represented as the Pearson correlation between true and predicted genomic breeding values, of the 1000 animals from the training population, was the parameter to evaluate the predicting ability of each methodology, on each one of the simulated scenarios.

Results and Discussion

Whole genomic selection accuracies are exhibited on Table 2. For the homogeneous genomic architecture scenario, accuracies ranged from 0.57 to 0.66, where standard ssGBLUP and traditional BLUP showed slightly superior results, probably due to the use of phenotypes from genotyped and ungenotyped animals, and due to the assumption of an infinitesimal model. In general, methodologies showed similar accuracies, indicating that for polygenic traits, it is possible to choose methodologies that are easier to implement and are more efficient. In this context, standard ssGBLUP and GBLUP are suitable for application on breeding programs, where traits of economic interest are usually polygenic. For the heterogeneous genomic architecture trait, accuracies ranged from 0.67 to 0.91, where GBLUP showed the lowest accuracy, because it assumes *a priori* a homogeneous genomic architecture. Standard ssGBLUP and traditional BLUP also presented lower accuracies than most of the bayesian methodologies, but slightly superior to GBLUP, due to the use of all phenotypes. With weighted ssGBLUP, it was possible to obtain a maximum accuracy of 0.79, which was superior to some of the bayesian methods, due to its iterative process, which performs indirect variable selection by shrinkage. The methodologies dependent on the parameter π, which represents the proportion of SNPs in linkage disequilibrium with the genes that influenced the trait (0.05 in this scenario), showed higher accuracies (greater or equal to 0.88), due to the fact that π *a priori* (Bayes B and Fast Bayes B) or estimated (Bayes Cπ) were close to the simulated π. BGLR Bayes A was almost as efficient as AlphaBayes Bayes B and Fast Bayes B, even assuming a priori that 100% of the SNPs were in linkage disequilibrium with the genes that regulated the trait, due to the reduction of the effects of most of the minor SNPs to zero by shrinkage. Resende et al. (2012) also observed that Bayes A can be as effective as other Bayesian methodologies without necessarily doing direct variable selection. Bayesian Lasso and Improved Bayesian Lasso showed low accuracies, due to the nature of its *a priori* distribution for the SNP effects (Resende et al., 2012). Bayesian methodologies presented the higher accuracies on this heterogeneous scenario and are probably the best option for the genomic evaluation of traits controlled by few genes.

Table 2. Predicting ability as correlations between true and predicted breeding values, for each methodology, based on homogeneous and heterogeneous genomic architecture scenarios.

Software - Methodology	Homogeneous	Heterogeneous
Alpha Bayes - Bayes A	0.64	0.77
Alpha Bayes - Bayes B	0.62	0.91
Alpha Bayes - Bayes Cπ	0.64	0.91
Alpha Bayes - Fast Bayes B	0.62	0.90
Alpha Bayes - GBLUP	0.62	0.67
BGLR - Bayes A	0.64	0.88
BGLR - Bayes B	0.64	0.88
BGLR - Bayesian Lasso	0.64	0.69
BGLR - GBLUP	0.65	0.68
BLUPF90 - Standard ssGBLUP	0.66	0.71
BLUPF90 - Traditional BLUP	0.66	0.70
BLUPF90 - Weighted ssGBLUP	0.57	0.79
GS3 - Bayes Cπ	0.64	0.73
GS3 - GBLUP	0.64	0.69
GS3 - Improved Bayesian Lasso	0.64	0.69

Conclusions

The decision about which WGS methodology to implement in a breeding program should be based on the knowledge *a priori* about the genomic architecture of the trait of interest. Although it is expected that most production traits are polygenic, some traits could require a specific adjustment, to better accommodate for a more heterogeneous genomic architecture.

References

Hickey JM, Tier B. AlphaBayes: user manual. UNE, Australia; 2009.

Legarra, A., A. Ricardi, and O. Filangi. 2013. GS3: Genomic Selection—Gibbs Sampling—Gauss Seidel (and BayesCp). Accessed Mar. 31, 2013. http://snp.toulouse.inra.fr/~alegarra/manualgs3_last.pdf

Misztal, I., Aguilar, I., Tsuruta, S., Legarra Albizu, A. (2012). Adaptation of BLUPF90 package for genomic computations. In: Book of Abstracts of the 63rd Annual Meeting of the European Federation of Animal Science (p. 135-135). EAAP Book Abstracts, 18 (1ère Ed.). Presented at 63. Annual Meeting of the European Association of Animal Production (EAAP), Bratislava, Slovaquie (2012-08-27 - 2012-08-31). Wageningen, NLD : Wageningen Academic Publishers. DOI : 10.3920/978-90-8686-761-5 http://prodinra.inra.fr/record/209421

Pérez, P., G. de los Campos, J. Crossa, and D. Gianola, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression package in R. Plant Gen. 3: 106–116.

Resende, M.D.V., Silva, F.F., Lopes, P.S., Azevedo, C.F. 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 Estatítica. 2012. 291 p. http://www.det.ufv.br/ppestbio/corpo_docente.php

Wang,H., Misztal,I., Aguilar,I., Legarra,A., and Muir,W.M. (2012). Genome wide association mapping including phenotypes from relatives without genotypes. Genet.Res. 94, 73–83.doi:10.1017/S0016672312000274.