Goodness of Fit Comparisons among Five Bayesian Models in Genome-Wide Association of Tick Resistance in Brazilian Hereford and Braford Beef Cattle

B.P Sollero [*,‡], **C. C. G. Gomes** [*], **V. M. Roso** [§], **R. H. Higa** [#], **M. J. Yokoo** [*], **L.L. Cardoso** [*,‡], **A. R. Caetano** [†,|], **F.F. Cardoso** [*,†]

[*]Embrapa Southern Region Animal Husbandry, Bagé, [‡]Coordination for the Improvement of Higher Level Personnel (CAPES/PNPD), Brasilia, [†]National Counsel of Technological and Scientific Development (CNPq), Brasilia, [§]Gensys Associated Consultants, Porto Alegre, [#]Embrapa Agriculture Informatics, Campinas, [|]Embrapa Genetic Resources & Biotechnology, Brasilia, Brazil

ABSTRACT: This study aimed to compare five models fitness and top effect SNPs obtained with three different Bayesian GWAS methods applied to cattle tick resistance in Braford and Hereford. After SNPs and sample's quality control analyses, 78% of the SNPs (41,045) were selected to be used simultaneously in the GWAS analysis of 3,455 animals. Among the tested models, Bayes C (BC) was the method showing the best goodness of fit, according to the posterior mean of the log-likelihood and the Deviance Information Criterion parameters, while the worst results were obtained with Bayes B (BB3). The standardized estimated squared-marker effects and the top ten SNPs ranked across tested models also favored Bayes C method, and highlighted SNPs in BTA5, BTA11 and BTA15, especially. Results indicate that further analyses to identify specific genes or genomic regions related to cattle tick resistance should be concentrated in these chromosomes.
Keywords: genomics; marker effects; standardization

Introduction

Brazil has the world's largest commercial herd of cattle and historically is one of the leading beef producer and exporter. The tick *Rhipicephalus (Boophilus) microplus* is among the main causes for losses in cattle production in Brazil, causing decreased performance of their hosts both directly by blood sucking and indirectly as vector of viral, bacterial and protozoal diseases (Machado et al. (2010)). Recent technological advances in molecular biology and quantitative genetics have enabled the advancement of knowledge on the genetic mechanisms of tick resistance. In this context, genome wide association studies (GWAS) are ideal methods to discover major genes responsible for controlling complex traits (Zhang et al. (2012)) and more recently, in order to identify more complex relationships, a shift to more sophisticated multi-SNP (Single Nucleotide Polymorphism) approaches has taken place (Moore, 2010), either applying frequentist statistics or through Bayesian inference methods. The latter, even though it requires increased computational power provides a flexible approach to solving high dimensional problems due to its ability to incorporate prior knowledge and its unified probabilistic approach of data analysis. More specifically, Bayesian methods like Bayes B (BB) (Meuwissen et al. (2001)), Bayes C (BC) (Habier et al. (2011)) and Bayes Lasso (BL) (Park and Casella (2008)) provide flexible penalizing strategies to estimate marker effects compared to the typical normal distribution assumption imposed by some other methods (e.g RR-BLUP, Hoerl and Kennard (1979)). Therefore, comparison of results obtained by different Bayesian GWAS methods, in terms of number of associated SNPs and ranking of their effects may contribute to understand tick resistance in Hereford and Braford beef cattle. Bayesian methods were originally adopted and are largely used for genomic prediction of breeding values (Meuwissen et al. (2001), Cleveland et al. (2010)), however, in recent years they have been applied for GWAS as well (Zare et al. (2014)). In this paper, we aimed to compare model fitness and top effect SNPs obtained by different Bayesian GWAS methods applied to tick-count data from Braford and Hereford cattle raised in Brazil.

Materials and Methods

Animal Sample and Data. The animals sampled from Hereford and Braford cattle breeds (N=3,545) were derived from eight different herds belonging to the Delta G Connection breeding program located in the Rio Grande do Sul state of Brazil. Tick counts over one side of the body of the animals were registered twice or three times consecutively during the post-weaning period of the animals, which were born between 2008 and 2011. In the total, 10,673 tick counts were considered for further analyses. Blood, hair or semen samples were used for DNA extraction and genotyping with the Illumina BovineSNP50 SNP Chip (54,609 SNPs).

Quality control analysis. To proceed the Quality Control (QC) of the samples and markers, scripts on the R software version 3.0.2 (R Core Team, 2013) were developed using the snpStat package (Clayton (2012)). The criteria and limits of SNP exclusion were minor allele frequency (MAF<0.03), deviation from Hardy-Weinberg equilibrium ($P<10^{-7}$) and SNPs in the same position or highly correlated (r>0.98). For the QC of the samples genotyped, those which presented heterozigozity higher than 3 standard deviation were excluded, as well, those which fall in the call rate criteria (<90%, considered as DNA samples with low quality), those presenting more than 99.5% of identical geno-

types and those due to errors in sex recorded. Sex chromosome "x" was considered only to analyze this last criterion. The log-transformed tick counts were analyzed to estimate variance components and breeding values using the BLUPf90 family of programs (Misztal et al., (2002)). Deregressed estimated breeding values (dEBV) and weighting information were calculated according to Garrick (2009) and used to estimate the marker effects in the GWAS analysis.

Statistical analyses. Comparisons of model fitness and complexity of three Bayesian methods (BB, BC and BL) for estimating SNP effects on tick resistance trait were applied to the samples using all markers simultaneously through R/BGLR package (de los Campos and Rodriguez (2013)). According to the prior densities available in BGLR to determine the type of shrinkage applied to markers effect estimates, the double exponential (DE) density, which has higher mass at zero and ticker tails than the normal density, was used in the BL model. On the other hand, a mixture point of mass at zero and a scale-t slab is used in the BB model and finally a mixture point of mass at zero and a Gaussian slab was used in the BC model (de los Campos and Rodriguez (2013)). In the Markov Chain Monte Carlo (MCMC) implementation, inferences were based on 40,000 samplers obtained from the posterior distribution after discarding 10,000 samplers as burn in, with no thinning. In BC, the value of π (proportion of SNPs with null effect) was deemed unknown and jointly estimated with other parameters of this model. For BB, the number of *prior counts* used to control how informative is the prior on π was set to 10^9, effectively fixing π at its prior value, which were alternatively specified as the BC posterior mean (BB1), or as 0.05 (BB2) or 0.95 (BB3), as proposed by Saatchi et al. (2013). The posterior mean of the log-likelihood, the estimated number of effective parameters (pD) and the Deviance Information Criterion (DIC, Spiegelhalter et al. (2002)) of each model were compared. Additionally, Manhattan plots of the standardized estimated squared-marker effects were presented, and the rank of the highest ten SNP's effect were compared among the models by Pearson's correlation.

Results and Discussion

Quality control analysis. After SNPs quality control criteria, 78% of SNPs (41,045) were selected to be used simultaneously in the GWAS analysis. In the total, 2,803 Braford and 652 Hereford (n=3,455) were selected to proceed further analyses, which corresponded to 98% out of 3,545 samples.

Bayesian Methods and Models Comparison. The lowest DIC model choice criterion value was observed for BC model (Table 1), followed by BB1 and BB2 models (tested with similar π values; 0.035 and 0.05), then the BL model and finally BB3 model (π=0.95), which showed the highest value. The posterior means of the log-likelihood (pMLogLik) followed a similar pattern to DIC, despite the higher number of effective parameters (pD) estimated for the best fit models (Table 1). Therefore, BC model resulted in a better adjusted model, even though more parameterized than the others. These two parameters (pD and pMLogLik) help to identify the most promising model, but once DIC balances goodness of fit and complexity, model BC was favored. BB1 and BB2 models, which consider higher proportion of SNPs with non-null effect (or lower π value), fitted the data similarly, and were considerably better than BB3 with π=0.95. According to Onteru et al. (2013), Bayes B method was used for GWAS in residual feed intake in pigs rather than a Bayes C approach because of its better performance for QTL mapping with 1 Mb genomic windows and concluded that genomic selection Bayesian methods are more powerful than frequentist methods to detect association by GWAS. On the contrary, the current study corroborates to Porto Neto et al. (2011), reflecting that BC model may be proper to study cattle tick resistance in a genomics perspective, where the majority of markers seems to explain a small portion of the phenotypic variation of the trait.

Table 1. Proportion of SNPs with no effect (π), Deviance Information Criterion (DIC), estimated effective number of parameters (pD) and posterior mean of the log-likelihood (pMLogLik) for each Bayesian model proposed: Bayes C (BC), Bayes B (BB) and Bayes Lasso (BL).

Model	pi (π)	DIC	pD	pMLogLik
BC	0.035[&]	-410.6	474.4	442.5
BB1	0.035[¥]	-406.0	445.7	425.9
BB2	0.05	-405.5	449.0	427.3
BL		-405.3	446.7	426.0
BB3	0.95	-307.8	105.4	206.6

[&] π value estimated after running Bayes C (BC).
[¥] π value suggested by BC used in BB1 model.

Table 2. The mean of estimated squared-marker effects (bHat^2), the mean (SbHat) and maximum ($SbHat_m$) of standardized estimated squared-marker effects obtained by each model.

Model	bHat^2	SbHat	$SbHat_m$
BC	7.036E-09	1.120E-02	2.789E-01
BB1	6.209E-09	1.004E-02	1.562E-01
BB2	6.291E-09	9.882E-03	1.526E-01
BL	6.222E-09	1.061E-02	2.401E-01
BB3	3.796E-10	2.413E-04	2.242E-02

Marker effects. Comparisons of the standardized estimated squared-marker effects (SbHat) reveled that BC model presented the highest average, as well as the highest maximum ($SbHat_m$) value (SNP of higher effect) relative to other models (Table 2). BL model presented the second highest SbHat and $SbHat_m$, followed by BB1 and BB2 models. Finally, BB3 model presented the lowest values for both estimates. Figure 1 presents the two Manhattan plots resulting from the two most extreme models (BC and BB3).

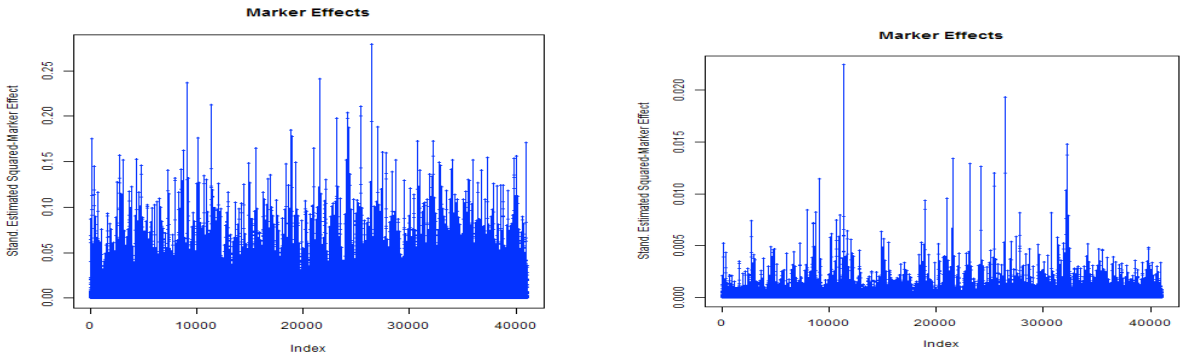

Figure 1. Manhattan plots representing standardized estimated squared-marker effects of the BC and BB3 models, respectively. X axes represents the 41,050 markers evaluated across the genome in order of chromosomes (BTA). w

According to the results, we compared the ten SNPs with highest SbHat in each of the five analyses. After ranking them, BC and BL models presented the same nine SNPs (among the top ten) in a very similar order (r = 0.87). More specifically, it was observed that the top three SNPs (in BTA15, BTA11 and BTA5, respectively) were ranked in the same way in both models. The others seven SNPs between BC and BL models were located in BTA6, BTA13, BTA14 and again, in BTA15. Even presenting four to seven SNPs, among the top ten, the others models (BB1, BB2 and BB3) ranked them in a very different order and with lower average estimates relatively to BC and BL. Machado et al. (2010), through a whole genome scan using microsatellite markers, identified spec
ific QTLs in BTA5 and BTA11 associated with tick resistance mechanisms of cattle. The ability to find SNPs or genomic regions associated with a trait is the main purpose of GWAS. The agreement between top SNPs in terms of the size of their effects in two out of the three Bayesian methods tested, provides evidence of association of those markers with tick resistance in cattle, and suggests that Bayes B method may not be the best choice to select SNPs with high effects associated with cattle tick resistance. Peters et al. (2012) also inferred that Bayes C offers advantages, as the results tend to be more biologically realistic than estimation of locus-specific variances influenced by SNP frequencies. It also worth to emphasize that the values of the parameters considered in the present work for the Bayesian methods (e.g π) crucially impact GWAS results and Genomic Selection (GS) as well. SNPs of most interest in a GWAS are those showing the strongest evidence of association, so we focused on the question of choosing SNPs to follow up, suggesting deeper investigations in those chromosomes reported, especially BTA15. In order to assess the functional relevance, The cattle QTLdb database (Hu et al. (2013)) needs to be examined to find out if any top SNP identified here overlap with a previously described bovine quantitative trait locus (QTL) for tick resistance.

Conclusion

Results confirm that Bayesian methods are highly useful for analyzing large SNP datasets in GWAS studies aiming at identifying more informative molecular markers to assist breeding. In the present work, Bayes C was the method of choice in terms of goodness of fit and identification of SNPs with high effects related to cattle tick resistance.

Literature Cited

Clayton, D. (2013). snpStats: R package version 1.12.0.

Cleveland, M. A., Hickey, J. M. and Forni, S. (2012). G3. 2:429-435.

De los Campos, G. and Rodriguez, P. P. (2013). BGLR: R package version 1.0.2.
http://CRAN.Rproject.org/package=BGLR.

Garrick, J. D., Taylor, J. F. and Fernando, R. L. (2009). Genet. Sel. Evol. 41:55.

Habier, D., Fernando, R. L., Kizilkaya et al. (2011). BMC Bioinf. 12:186.

Hoerl, A. E. and Kennard, R. W. (1970). Technometrics. 12:55–67.

Hu, Z. L., Park, C. A., Wu, X. L. et al. (2013). Nucl. Acid Res. 41:871–879.

Machado, M. A., Azevedo, A. L. S., Teodoro, L. R. et al. (2010). BMC Genomics. 11:280.

Meuwissen, T. H. E, Hayes, B. and Goddard, M .E. (2001). Genetics, 157:1819-1829.

Misztal, I., Tsuruta, S., Strabel, T., et al. (2002). in Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. Comm. 28–07.

Moore, J. H., Asselbergs, F. W. and. Williams, S. M. (2010). Bioinformatics. 26: 445–455.

Onteru, S. K., Gorbach, D. M., Young, J. M. et al. (2013). PLoS ONE 8: e61756.

Park, T and Casella, G. (2008). J. Am. Stat. Assoc. 103:681–686.

Peters, S. O., Kizilkaya, K., Garrick, D. J. et al. (2012). J Anim Sci. 90:3398-3409.

Porto Neto, L. R., Jonsson, N. N., D'Occio, M. J. et al (2011). Vet. Parasit.180:165-172.

Saatchi, M., Ward, J. and Garrick, D. J. (2013) J. Anim. Sci. 91:1538-1551.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. et al. (2002). J. of the Roy. Stat. Soc. B, 64:583-639.

Zare, Y., Shook, G. E., Collins, M. T. et al. (2014). PLoS ONE 9: e88380.

Zhang, H., Wang, Z., Wang, S. et al. (2012). J. of Anim. Sci. and Biot. 3:26.