Extension Of The Bayesian Alphabet For Genomic Selection

D. Habier[*], R. L. Fernando[*], K. Kizilkaya[*†] and D. J. Garrick[*]

Introduction

Meuwissen et al. (2001) presented two hierarchical Bayesian models for genomic selection termed BayesA and BayesB that exploit linkage disequilibrium (LD) better than Least-Squares or Ridge-Regression (Habier et al., 2007, 2010; Meuwissen et al., 2001), and thus gave higher accuracies of genomic estimated breeding values (GEBVs). The marker effects have normal priors conditional on locus-specific variances, and these variances have a scaled inverse chi-square prior with known parameters. As pointed out by Gianola et al. (2009), a drawback of this model hierarchy is that the fully-conditional posteriors of the locus-specific variances have only one additional degree of freedom compared to their priors irrespective of the amount of data, and this conflicts with Bayesian learning. A special characteristic of BayesB is that it fits only a fraction of the available markers to meet the assumption that many of them have zero effect. However, the prior probability that a marker has zero effect, π, is assumed to be known in BayesB.

The objective of this study was to present two Bayesian model averaging approaches that address the drawbacks of BayesA and BayesB and also treat π as an unknown. GEBV accuracies were estimated using data from the North-American Holstein population to compare BayesA, BayesB and the proposed models.

Material and methods

Statistical models. The general statistical model can be written as

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{u} + \sum_{k=1}^{K} \boldsymbol{x}_k a_k + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector of trait phenotypes of N individuals in training, μ is the overall mean, $\boldsymbol{u}$ is a vector with polygenic effects of all individuals in the pedigree, K is the number of single nucleotide polymorphisms (SNPs), $\boldsymbol{x}_k$ is a $N \times 1$ vector of genotypes at SNP k, a_k is the additive effect of that SNP, and $\boldsymbol{e}$ is a vector of residual effects. SNP genotypes are coded as the number of copies of one of the SNP alleles, i.e., 0, 1 or 2.

[*]Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, IA 50011, USA

[†]Department of Animal Science, Adnan Menderes University, Aydin 09100 Turkey

The differences between the Bayesian models used here center around the prior specifications for a_k, which is zero with probability π and normal given the variance with probability (1-π). In BayesA and BayesB, each locus has its own variance with a known scaled inverse chi-square prior, which results in the drawbacks explained earlier. The models termed BayesC and BayesD are modifications of BayesB to overcome those drawbacks: In BayesC, a common effect variance replaces the locus-specific effect variances, whereas in BayesD the scale parameter of the scaled inverse chi-square priors of the locus-specific effect variances is treated as an unknown with Gamma(1,1) prior. As the unknown scale parameter is common to all loci, information from all the loci contribute to its posterior and through it to the posteriors of the locus-specific variances. In addition, π is treated as an unknown with uniform (0,1) prior in BayesC and BayesD, which is emphasized by using the terms BayesCπ and BayesDπ. In contrast, π = 0 in BayesA, while the value selected for BayesB was 0.99 in this study. Other prior specifications were as described in Meuwissen et al. (2001) for BayesA and BayesB.

Statistical analyses. Markov-Chain Monte Carlo (MCMC) was used to infer the model parameters, where μ, $\boldsymbol{u}$, a_k, σ_e^2, π (if unknown) and the common effect variance of BayesC were sampled by Gibbs steps. The locus-specific effect variances of BayesA, BayesB and BayesDπ, in contrast, were sampled by Metropolis-Hastings steps. The starting value for π was 0.5. The MCMC algorithms were run for 100,000 iterations with a burn in of 50,000 iterations.

Available data. The available data set consisted of 8,239 progeny tested North-American Holstein bulls that were genotyped with the Illumina Bovine50K array. The phenotypes were de-regressed breeding values for the traits milk yield, fat yield, protein yield and somatic cell score. Pedigree information was available to analyze the maximum additive-genetic relationships (a_{max}) of bulls in validation to the bulls in training as described by Habier et al. (2010), and to model the polygenic effects in (1). A total of 40,789 SNPs were selected for the analyses based on minor allele frequency (> 3%), fraction of missing genotypes (< 5%) , and rate of mismatches between genotype pairs of sires and sons (< 5%).

Training and validation data. Bulls born between 1995 and 2004 were used for training, whereas 115 bulls born between 1953 and 1975 were used for validation. This was done to keep the genetic relationships between bulls in training and validation as small as possible so that the accuracy of GEBVs was mainly due to LD information revealing the potential of genomic selection. As the size of the training data set affects the utilization of LD information, 1,000 and 4,000 training bulls were randomly selected from the bulls born between 1995 and 2004. GEBV accuracy was estimated as the correlation between GEBVs and de-regressed breeding values divided by the average accuracy of the de-regressed breeding values of the validation bulls. Standard errors were estimated with the formula $\sqrt{(1-\rho^2)/(n-2)}$, where ρ denotes GEBV accuracy and n=115 validation bulls.

Results and discussion

Additive-genetic relationships between training and validation bulls were small due to a gap of about three generations between the bulls of both data sets. The lower quartile, median, and upper quartile of the 115 a_{max} values of the validation bulls were 0.015, 0.05 and 0.08, respectively, and none of the validation bulls had a higher a_{max} value than 0.123.

Table 1 shows the accuracy of GEBVs obtained for the 115 bulls born between 1953 and 1975

depending on the Bayesian method used to estimate SNP effects and training data size for the traits milk yield, fat yield, protein yield and somatic cell score.

Table 1: GEBV accuracy[a] for 115 bulls born between 1953 and 1975 depending on the Bayesian method used to estimate SNP effects, the quantitative trait and the number of training bulls born between 1995 and 2004

Trait	Training data size	BayesA	BayesB[b]	BayesCπ	BayesDπ
Milk yield	1,000	0.42	0.29	0.37	0.43
	4,000	0.44	0.42	0.41	0.45
Fat yield	1,000	0.48	0.51	0.49	0.47
	4,000	0.52	0.49	0.55	0.53
Protein yield	1,000	0.15	0.10	0.15	0.14
	4,000	0.18	0.15	0.17	0.17
Somatic cell score	1,000	0.14	0.12	0.14	0.12
	4,000	0.28	0.18	0.24	0.23

[a] standard errors: 0.08-0.09
[b] $\pi = 0.99$

BayesA, which fitted all SNPs and had the statistical drawbacks described by Gianola et al. (2009), gave the highest accuracies for protein yield and somatic cell score, and similar accuracies compared to BayesCπ and BayesDπ for the other two traits. With the exception of fat yield, BayesA was also clearly better than BayesB, which fitted only about 400 SNPs in each round of the MCMC algorithm (Table 2).

Table 2: Posterior mean of (1-π) multiplied by the number of SNPs used in the analyses depending on the Bayesian method used to estimate SNP effects, the quantitative trait and the number of training bulls

Trait	Training data size	BayesB[a]	BayesCπ	BayesDπ
Milk yield	1,000	404	1,180	13,982
	4,000	436	2,162	13,329
Fat yield	1,000	402	487	13,533
	4,000	441	1468	13,513
Protein yield	1,000	403	13,942	14,430
	4,000	438	6,723	13,512
Somatic cell score	1,000	398	5,057	12,962
	4,000	428	3,261	13,941

[a] $\pi = 0.99$

BayesB was better than the other methods only for fat yield and when the training data set contained 1,000 bulls. However, as training data size increased this advantage vanished and BayesB had the lowest accuracy of all methods for fat yield. The explanation may be that loci with smaller effects contributed to the predictions by the other methods, but this did not happen in BayesB due to the high value of π. This is supported by the increasing number of SNPs fitted into the model with training data size, which was observed for BayesCπ for fat yield (Table 2). BayesCπ tended to give better accuracies than BayesDπ for all traits but

milk yield for which BayesDπ clearly outperformed BayesCπ. The number of SNPs fitted to the model was nearly constant in BayesDπ, whereas BayesCπ was sensitive to both trait and training data size. This demonstrates the different mechanisms of estimating SNP effects in the two methods. In BayesDπ, the locus-specific variances dominate over π resulting in poor mixing of π. Furthermore, the increase in the number of SNPs fitted with training data size in BayesCπ for milk yield and fat yield may be due to higher power to detect QTL, while the decay in protein yield and somatic cell score may indicate that fewer SNPs were falsely detected.

GEBV accuracies could not have been affected by additive-genetic relationships much, hence they were mainly due to LD information. Habier et al. (2010) estimated the accuracy of GEBVs due to LD using 3,868 German Holstein bulls and BayesB with $\pi = 0.99$. Most of these bulls were born between 1998 and 2004, and 60% descend from a North-American Holstein sire, which reveals the high genetic relationship between the German and the North-American Holstein population. GEBV accuracies obtained by BayesB compared to those in Habier et al. (2010) were similar for milk yield, comparable for fat yield when the training data size was greater than 1,000, but lower for protein yield and somatic cell score. The increase of accuracy with training data size, which was largest here for somatic cell score, but rather small for the yield traits, tended to be higher in Habier et al. (2010). The difficulties in comparing the accuracies found here to those in Habier et al. (2010) is that there might be genotype-environment interactions, because the environment in which the daughters of the bulls born before 1975 have been tested might be different from the environment of the last decade. Moreover, selection and genetic drift may have changed the LD structure in the population so that the accuracies of this study do not represent the GEBV accuracies due to LD in the current population.

Conclusion

The statistical drawbacks of BayesA and BayesB did not impair the accuracy of GEBVs. None of the Bayesian methods analyzed outperformed all other methods across all traits and training data sizes, and therefore the best method must be determined for each quantitative trait separately. BayesA appears to be a good choice, but this finding should be verified by predicting GEBVs across breeds to fully exclude the effect of additive-genetic relationships and the other concerns with this validation set. BayesB, BayesDπ and especially BayesCπ have the advantage of providing information about the genetic architecture of the quantitative trait and identifying QTL positions by model frequencies of SNPs.

References

Gianola, D., de los Campos, G., Hill, W. G., et al. (2009). *Genetics*, 183(1):347–363.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). *Genetics*, 177(4):2389–2397.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., et al. (2010). *Genet. Sel. Evol.*, 42:5.

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