Bayesian Models for Predicting Genomic Breeding Values in a Swedish-Finnish Red Breed Cattle Population

E. Rius-Vilarrasa[*], R.F. Brøndum[§], I. Strandén[†], B. Guldbrandtsen[§], E. Strandberg, M.S. Lund[§] and W.F. Fikse

Introduction

Existence of population-wide linkage disequilibrium (LD) between molecular markers and quantitative trait loci (QTL) has enabled the successful prediction of an animal's genomic breeding value (GBV). Since the introduction of this concept by Meuwissen et al. (2001), several statistical methods have been suggested and cross-validation techniques have been used to test models' predictive ability of GBV. Based on a real multi-breed reference population, Hayes et al. (2009) reported higher predictive ability when Bayesian methods were used to calculate GBV compared to the best linear unbiased prediction method. Extensions to the Bayesian models have included a polygenic effect with little impact on the accuracy of GBV but increased persistency of its accuracy over generations (Solberg et al. 2009).

Despite the increasing interest for these methods, the influence of prior information on the model parameters may have a significant impact on the outcome (Gianola et al. 2009). The aim of this paper is to investigate on a real population of Swedish and Finnish Ayrshire Red dairy cattle the influence of model specification on the reliability of GBV as well as the impact of including a polygenic effect in the model.

Material and methods

Phenotypes and markers. Data available included 2,986 Swedish Red (SRB) and Finnish Ayrshire (FAY) progeny tested bulls born between 1986 and 2005. These bulls belonged to 224 half-sib families, of which 150 had five or more sons. Official Nordic breeding values (EBV) for production traits (milk, fat and protein index), udder health, fertility, calving direct, and body-conformation were used as response variables to estimate marker effects.

Marker data were available from genotyping of bulls using Illumina Bovine SNP50 BeadChip (Illumina, San Diego, CA), which included 54,001 single nucleotide

Department of Animal Breeding and Genetics, PO Box 7023, Swedish University of Agricultural Sciences, 750 07 Uppsala, Sweden

[§] Aarhus University, Faculty of Agricultural Sciences, Dept. of Genetics and Biotechnology, Blichers Allé 20, P.O. BOX 50, DK-8830 Tjele, Denmark

[†] MTT Agrifood Research Finland, Biotechnology and Food Research, Biometrical Genetics, FIN-31600 Jokioinen, Finland

polymorphisms (SNP) markers. Markers were excluded if their minor allele frequency was less than 5%, the proportion of animals called for a genotype at this locus was less than 95%, or the average call rate score was less than 0.65. Individuals were deleted from the analysis if their GeneCall score was less than 0.60. SNP on the X chromosome were excluded leaving a total of 38,315 SNP for analysis.

Statistical analyses. Bayesian inference using iBay software (Janss 2009) to perform the Gibbs sampling was used to estimate marker effects. Official Nordic EBVs were used as response variables and single marker SNP were used as predictors. The models used were:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{m} \mathbf{X}_{\mathrm{i}}\mathbf{q}_{i}v_{i} + \mathbf{e} \qquad [1] \qquad \mathbf{y} = \mathbf{1}\mu + \mathbf{a} + \sum_{i=1}^{m} \mathbf{X}_{\mathrm{i}}\mathbf{q}_{\mathrm{i}}v_{i} + \mathbf{e} \qquad [2]$$

Where ***y*** is a vector of EBVs, **1** is a vector of ones, ***μ*** is the intercept, ***m*** is the number of markers, $\boldsymbol{X_i}$ is the vector of genotypes of marker *i* for all individuals, $\boldsymbol{q_i}$ is the vector of scaled SNP effects of marker *i* modeled as random with $\boldsymbol{q_i} \sim N\ (\mathbf{0}, \mathbf{I})$, $\boldsymbol{v_i}$ is a scaling factor which models the variance explained by the *i*th marker and ***e*** is vector of residuals $\boldsymbol{e} \sim N\ (\mathbf{0}, \mathbf{W}\sigma_e^2)$ where **W** is a diagonal matrix with weights to account for EBVs accuracy. Model 2 was an extension of model 1 and included a polygenic effect (***a***); the relationship matrix was calculated based on 12 generations of pedigree. Prior distributions for the scaling factors ($\boldsymbol{v_i}$) included: 1) a common prior assumed to be a positive truncated normal distribution with null mean and variance σ_g^2, estimated from the data assuming a flat prior distribution and 2) a mixture of positive truncated normal distributions with either small ($\sigma_{g_0}^2$) or large ($\sigma_{g_1}^2$) variance. The proportion of markers sampled from the "small" distribution was set by specifying π_0 whereas π_1 ($\pi_1 = 1 - \pi_0$) refers to the proportion of markers sampled form the "large" distribution. A non-informative prior distribution was assumed for $\sigma_{g_1}^2$ while $\sigma_{g_0}^2$ was set to a fixed value. Starting values for $\sigma_{g_0}^2$ were obtained using the following rule of thumb (Janss 2009): 0.01×(variance of the trait)/(π_0×(number of markers)). Accordingly, four scenarios were investigated: 1) $\pi_0 = 60\%$ with $\sigma_{g_0}^2 = 4\times10^{-5}$ and 2) $\pi_0 = 80\%$, 3) $\pi_0 = 90\%$ and 4) $\pi_0 = 95\%$ all with $\sigma_{g_0}^2 = 3\times10^{-5}$, respectively. The same prior settings were used for the model including the polygenic effect. Marker effects were then estimated from all five different scenarios and the GBV computed as:

$$\mathbf{GBV}_k = \hat{\mu} + \sum_{i=1}^{m} \mathbf{X}_{\mathrm{i(k)}}\hat{\mathbf{q}}_i\hat{v}_i \qquad [1] \qquad \mathbf{GBV}_k = \hat{\mu} + \sum_{i=1}^{m} \mathbf{X}_{\mathrm{i(k)}}\hat{\mathbf{q}}_i\hat{v}_i + \mathbf{a} \qquad [2]$$

Predictive ability was evaluated by applying a two-fold cross validation leaving out data for two birth year cohorts at a time and using the remaining data for the estimation of marker effects. The youngest animals born between 1998 and 2005 were left out and used as test data. Within-year square correlation (R^2) between EBV and GBV for the bulls in the test data was used as a measure of GBV reliability.

Results and discussion

Reliabilities of GBVs ranged from 0.16 to 0.42 (Table 1). Except for body-conformation, GBV reliabilities in this study were lower than the ones reported by Lund and Su (2009) for the Danish-Swedish Holstein population. However the two studies agree that the model with a common prior distribution for scaling factors had a slightly higher predictive ability compared to the model with a mixture prior distribution. Only for some traits the mixture model led to an increase of GBV reliabilities over the common prior model (bold in Table 1).

Table 1. Within-year square correlations (R^2) between EBV and GBV for genomic models with different prior distribution settings of scaling factors

Traits	Mixture $\pi_0 = 60\%$	Mixture $\pi_0 = 80\%$	Mixture $\pi_0 = 90\%$	Mixture $\pi_0 = 95\%$	Common Prior
Milk	**0.29**	0.28	0.22	0.18	0.29
Fat	0.33	0.34	0.29	0.26	0.36
Protein	0.24	0.23	0.18	0.16	0.25
Fertility	0.28	0.30	0.27	0.24	0.32
Calving direct	**0.32**	**0.35**	0.33	0.30	0.32
Body-conf.	0.38	**0.42**	0.39	0.36	0.39
Udder-health	0.25	0.25	0.24	0.21	0.28
Mean	0.30	0.31	0.27	0.24	0.32

For the models using a mixture prior distribution, reliabilities of GBV tended to decrease with an increasing proportion (π_0) of markers with small effects (Table 1). The largest difference was for milk where R^2 decreased from 0.29 ($\pi_0 = 60\%$) to 0.18 ($\pi_0 = 95\%$). In general, a fairly large influence of prior distributions on GBV reliabilities is shown in Table 1. However with increasing size of the reference populations and the advent of a higher density SNP chip the use of a mixture prior distribution model might become more advantageous compared to other methods for predicting GBV (Meuwissen 2009). In this case, the prior influences on inferences and prediction could be mitigated by estimating the mixing probability assuming a beta prior, as reported in a recent study by Gianola et al.(2009).

Including a polygenic effect in the models with mixture prior distribution for scaled factors increased GBV reliability for most of the traits when $\pi_0 = 60\%$ but not when $\pi_0 = 80\%$ and 90% (Table 2). However the mixture model with a polygenic effect did not yield GBV reliabilities larger than for the common prior model, except for some traits (in bold in Table 2). Using EBVs as response variables might have influenced these results. A bull’s EBV depends to some extent on the parents EBV, especially for low heritable traits. Pedigree relationships are used to compute EBVs and therefore some of the genomic variance captured by the marker effects might capture some family relationships. Consequently, when a polygenic effect is included in the model the genetic variance which was previously explained entirely by the marker effects is now partitioned into marker effects and pedigree information without necessarily having to change the total variance explained by the model. Alternative response variables such as de-regressed EBVs or daughter yield deviations

(VanRaden and Wiggans 1991) might have showed a more clear benefit of including a polygenic effect in the model compared to using EBVs as response variables.

Table 2. Within year square correlations between EBV and GBV for genomic models with different prior distribution for scaling factors and including a polygenic effect

Traits	Mixture + Polygenic $\pi_0 = 60\%$	Mixture + Polygenic $\pi_0 = 80\%$	Mixture + Polygenic $\pi_0 = 90\%$	Mixture + Polygenic $\pi_0 = 95\%$	Common Prior
Milk	0.29	0.26	0.22	0.21	0.29
Fat	0.36	0.33	0.29	0.27	0.36
Protein	0.23	0.18	0.15	0.14	0.25
Fertility	0.30	0.27	0.24	0.24	0.32
Calving direct	**0.33**	**0.33**	0.31	0.29	0.32
Body-conf.	**0.40**	0.39	0.38	0.36	0.39
Udder-health	0.28	0.25	0.20	0.20	0.28
Mean	0.31	0.29	0.26	0.24	0.32

Although including a polygenic effect did not widely increase GBV reliabilities, false marker associations due to family relationships are likely to be reduced, increasing the probability for real marker associations to be estimated based purely on the LD between the markers and the QTL, therefore providing a higher persistence on the accuracy of GBV over generations (Solberg et al. 2009).

Conclusion

This study showed the influence of prior information on the GBV's reliabilities. No single model specification performed best for all traits. Fairly small changes on GBV reliabilities were observed when a polygenic effect was included in the models.

Acknowledgements

This study was financed by the Swedish Farmers' Foundation for Agricultural Research and the Finnish Ministry of Agriculture and Forestry (grant 2886/502/2008). The genotypes were funded by the Danish Directorate for Food, Fisheries and Agri Business, VikingGenetics, Nordic Genetic Evaluation, and Aarhus University (grant 3412-08-02253).

References

Gianola, D., de los Campos, G., Hill, W. G., et al. (2009). *Genetics,* 183: (1) 347-363
Hayes, B., Bowman, P., Chamberlain, A., et al. (2009). *Genet. Sel. Evol.,* 41: (1) 51
Janss, L. (2009). *iBay manual 1.46*
Lund, M. S. and Su, G. (2009). *Interbull Bul. 39, Uppsala, Sweden,*
Meuwissen, T. (2009). *Genet. Sel. Evol.,* 41: (1) 35
Meuwissen, T., Hayes, B. and Goddard, M. (2001). *Genetics,* 157: 1819 - 1829
Solberg, T., Sonesson, A., Woolliams, J., et al. (2009). *Genet. Sel. Evol.,* 41: (1) 53
VanRaden, P. M. and Wiggans, G. R. (1991). *J. Dairy Sci.,* 74: (8) 2737-2746