Estimating Reliabilities of Genomic Breeding Values

M.P.L. Calus[a], H.A. Mulder[a], K. Verbyla[a,b] and R.F. Veerkamp[a]

[a] *Animal Breeding and Genomics Centre, Animal Sciences Group, Wageningen UR, The Netherlands*

[b] *Melbourne School of Land and Environment, The University of Melbourne, Parkville 3010, Australia*

Keywords: genomic breeding values, reliability

Introduction

The availability of genomic estimated breeding values (GEBV) allows the selection of young bulls with relatively high reliability before phenotypic performance of daughters is recorded. Reported absolute gains in reliability for those young bulls, compared to using pedigree indexes only, are up to 20% (Hayes *et al.,* 2009). In the countries where GEBV are currently used in the national evaluation, GEBV are calculated using the usual procedures, but replacing a pedigree based relationship matrix by a genomic relationship matrix (GRM) (Berry *et al.,* 2009, VanRaden, 2008). Subsequently, reliabilities of GEBV are obtained by inverting the left-hand sides of the mixed model equations (Berry *et al.,* 2009, VanRaden, 2008). An alternative method to obtain reliabilities is cross-validation (e.g. De Roos *et al.,* 2009). Since the availability of correct reliabilities of GEBV is important both for national and international genetic evaluations, the objective of this paper was to provide a method to validate reliabilities of GEBV.

Material and Methods

Reliabilities were estimated based on 1) prediction error variances obtained from the left-hand side of the mixed model equations (REL_LHS) and 2) from Monte Carlo simulation (REL_MC). To assess the bias in individual REL_LHS, a procedure is developed to obtain REL_MC, i.e. a reliability for each individual using Monte Carlo sampling. This procedure is based on the idea that for a given dataset, a large number of independent samples for both the breeding values and the phenotypes can be generated using Monte Carlo simulation (Fouilloux and Laloe, 2001).

In applications for ‘classical’ breeding values, x sets of true breeding values are simulated through the existing pedigree (Fouilloux and Laloe, 2001, Hickey *et al.,* 2009). Phenotypic records are consequently drawn from $N(\mathbf{0}, \mathbf{ZGZ'} + \mathbf{R})$. Similarly, we developed an approach for SNP data, that used the same data structure to simulate x sets of true breeding values and phenotypic records. The steps are (per replicate) as follows:

1. Draw some SNP loci from existing SNP data to be QTL, and assign simulated QTL effects to those loci. QTL effects are calculating using equation 1 (see simulated scenarios).

Repeat the following steps x times:

2. Sample the sign (- or +) of each QTL-locus with equal probability,
3. Calculate the true breeding value for each animal as the sum of all simulated QTL-effects, using the appropriate signs from step 2,
4. Predict GEBV for each animal with REML,
5. Calculate for each animal the prediction error variance (PEV) based on the inverse of the LHS and based on this a reliability (REL_LHS) as $REL\_LHS = 1 - (\ PEV\ /\ \hat{\sigma}_a^2\ )$, where $\hat{\sigma}_a^2$ is the estimated genetic variance.

Across the x generated data sets:

6. Calculate for each animal REL_MC as the squared correlation between its TBV and EBV across the x generated data sets,
7. Calculate $\overline{REL\_LHS}$ for each animal across the x generated data sets .

When the number of generated datasets is large enough, the distributions of $GEBV_i$ and $G\hat{E}BV_i$ (the simulated and estimated GEBV of animal i) converge to their true distributions (Hickey *et al.,* 2009). Therefore, when the number of samples is large enough, REL_MC approaches the 'true' reliability of an animal's GEBV. This allowed to evaluate REL_LHS by comparing $\overline{REL\_LHS}$ to REL_MC.

Simulated scenarios

The SNP data contained 576 cows with 43,080 SNPs after editing. In each run, the 400 oldest animals were used as reference population (i.e. training data), while the phenotypes of the youngest 176 animals were supposed to be unknown. Overall, 10 replicates were performed. Within each of the 10 replicates, the above mentioned steps were used to generate 5000 datasets, for four scenarios: 10, 100, 1000 or 43,080 QTL. In the first 3 scenarios, the SNP that were drawn to be QTL, were excluded from the SNP data to mimic real life where QTL are supposed to be between markers. In the fourth scenario, all SNP were used as QTL, and all of them were also used to calculate the GRM matrix. This mimics the situation of an infinitesimal model and avoids the loss of accuracy of GEBV due to that markers explain only part of the genetic variance. In all scenarios, the contribution of each QTL to the genetic variance was considered to be equal. Therefore, for QTL locus *i*, the simulated allele substitution effect was calculated as:

$$a_i = \sqrt{\frac{\sigma_a^2}{\#QTL \times 2p(1-p)}}$$

where σ_a^2 is the total additive genetic variance, *#QTL* is the number of simulated QTL and *p* is the frequency of one of both alleles at locus *i*. Two traits were simulated with heritabilities of 0.6 and 0.9.

A maximum of 5000 generated data sets was considered, where $\overline{REL\_LHS}$ was evaluated every 5 generated data sets. In each data set, the following model was used to estimate the GEBV in ASReml (Gilmour *et al.,* 2006):

$$y_i = \mu + GEBV_i + e_i$$

where GEBV and its variance were simultaneously estimated using REML. The GEBV were distributed as $N(\mathbf{0}, \mathbf{G}\hat{\sigma}_a^2)$. were **G** is a GRM calculated as $G = \frac{ZZ'}{2\sum p_i(1-p_i)}$, where **Z** contains the marker genotypes for all animals at all loci corrected for the allele frequencies per locus, and p_i is the frequency of one of both alleles at locus *i* (VanRaden, 2008).

Results

The results for the animals in the validation data are presented. Average reliabilities for those animals ranged were around 0.22 and 0.30, for heritabilities of 0.6 and 0.9, respectively. In all scenarios after 1,000 generated data sets both REL_MC and $\overline{REL\_LHS}$ hardly changed (results not shown). The correlation between REL_MC and REL_LHS came close to 1.0 within 5,000 generated data sets when either all or 1,000 SNPs were included as QTL (Figure 1). This correlation was much lower for 100 and 10 QTL, although still slightly increasing at 5,000 generated data sets for both scenarios. Interestingly, the scenarios with 100 or more QTL all yielded REL_LHS that were overestimated by 0.02 to 0.04 (Figure 2). This overestimation was lower for the scenario with 10 QTL, but in this scenario the correlation between REL_MC and REL_LHS was still quite low (Figure 1).

Discussion

The advantage of REL_LHS is that each animal gets an individual reliability, that may depend on the relationship to the reference population (Berry *et al.,* 2009), while reliabilities obtained using regular cross-validation assumes that GEBV of all animals without phenotypic information have equal reliability. One of the disadvantages of REL_LHS is that it assumes that all genetic variation is explained by the markers, while it is likely that an important part of the genetic variance is explained by loci in between the markers (VanRaden, 2008). This may lead to overestimation of REL_LHS. The cross-validation method makes no such assumptions. Our results indicated that for scenarios with 100 or more QTL with equal effect, 1,000 generated data sets is sufficient to assess the difference between REL_MC and REL_LHS. In these scenarios, the REL_LHS was indeed overestimated by 0.02 to 0.04, even when all genetic variance was explained by the markers. This study provided a method to obtain unbiased individual reliabilities for GEBV. An important unanswered question is how these REL_MC can be calculated from real data where the true breeding values and QTL are not known. For the model that was used in our study, with a GRM assuming equal contribution to the genetic variance, this can be done as follows. Replace steps 1, 2 and 3 by drawing the simulated breeding values from $N(\mathbf{0}, \boldsymbol{G}\hat{\sigma}_a^2)$. All other steps remain unchanged. A similar approach could be taken for models that allow unequal contributions of SNPs to the genetic variance, by adjusting *G* for SNP specific contributions to the genetic variance.

Although the presented approach to obtain REL_MC is computationally demanding, it allows to validate calculated reliabilities for GEBV, obtained using other methods.

Conclusions

The presented method provides a procedure to validate GEBV reliabilities. Applying this method showed that reliabilities based on the inverse of the left-hand side of the mixed model equations tend to be overestimated.

Acknowledgments

NWO – Casimir, CRV, and Hendrix Genetics are acknowledged for financial support. Sander de Roos, Chris Schrooten, Abe Huisman, Addie Vereijken and John Bastiaansen are thanked for useful comments on the set up of this study. The RobustMilk project (financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708) provided the genotypes of the animals.

References

Berry, D.P., Kearney, F. & Harris, B.L. 2009. Genomic Selection in Ireland. *Interbull Bulletin 39,* 29-33. Proc. of the Interbull International Workshop - Genomic Information in Genetic Evaluations, Uppsala, Sweden.

De Roos, A.P.W., Schrooten, C., Mullaart, E., Van der Beek, S., De Jong, G. & Voskamp, W. 2009. Genomic Selection at CRV. *Interbull Bulletin 39,* 47-50. Proc. of the Interbull International Workshop - Genomic Information in Genetic Evaluations, Uppsala, Sweden.

Fouilloux, M.N. & Laloe, D. 2001. A sampling method for estimating the accuracy of predicted breeding values in genetic evaluation. *Genet. Sel. Evol. 33(5),* 473-486.

Gilmour, A.R., Gogel, B.J., Cullis, B.R. & Thompson, R. 2006. *ASReml User Guide Release 2.0.* VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. 2009. Invited review: *Genomic selection in dairy cattle: Progress and challenges. 92(2)*, 433-443.

Hickey, J., Veerkamp, R., Calus, M., Mulder, H. & Thompson, R. 2009. Estimation of prediction error variances via Monte Carlo sampling methods using different formulations of the prediction error variance. *41(1),* 23.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91(11),* 4414-4423.

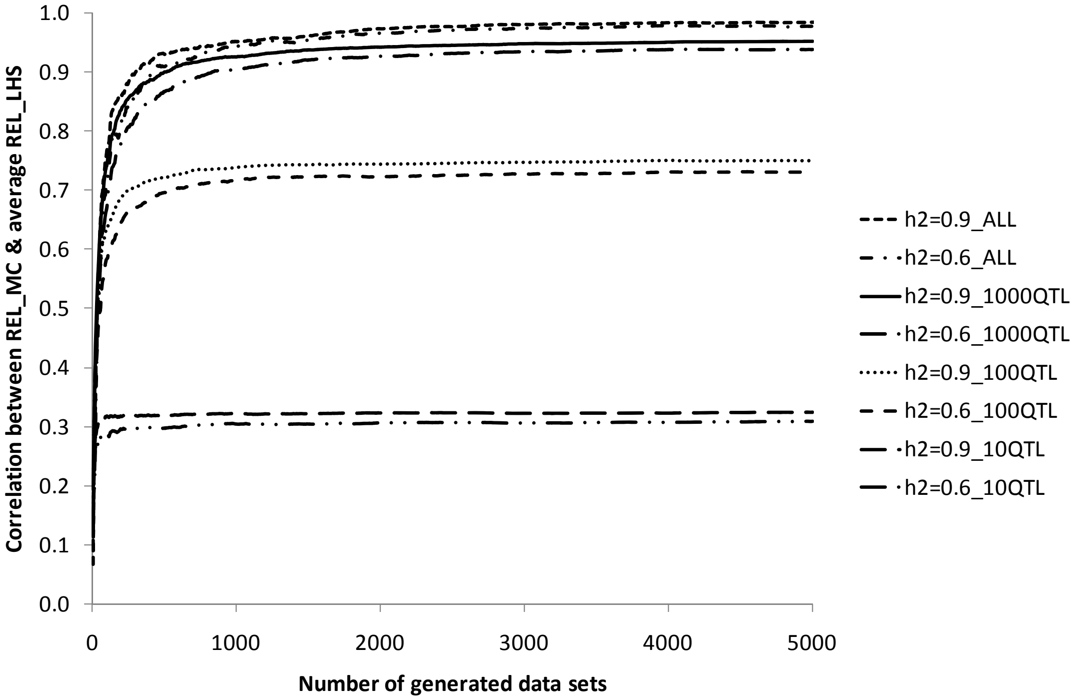

Figure 1. Correlation between REL_MC and average REL_LHS for the validation animals across numbers of generated data sets.

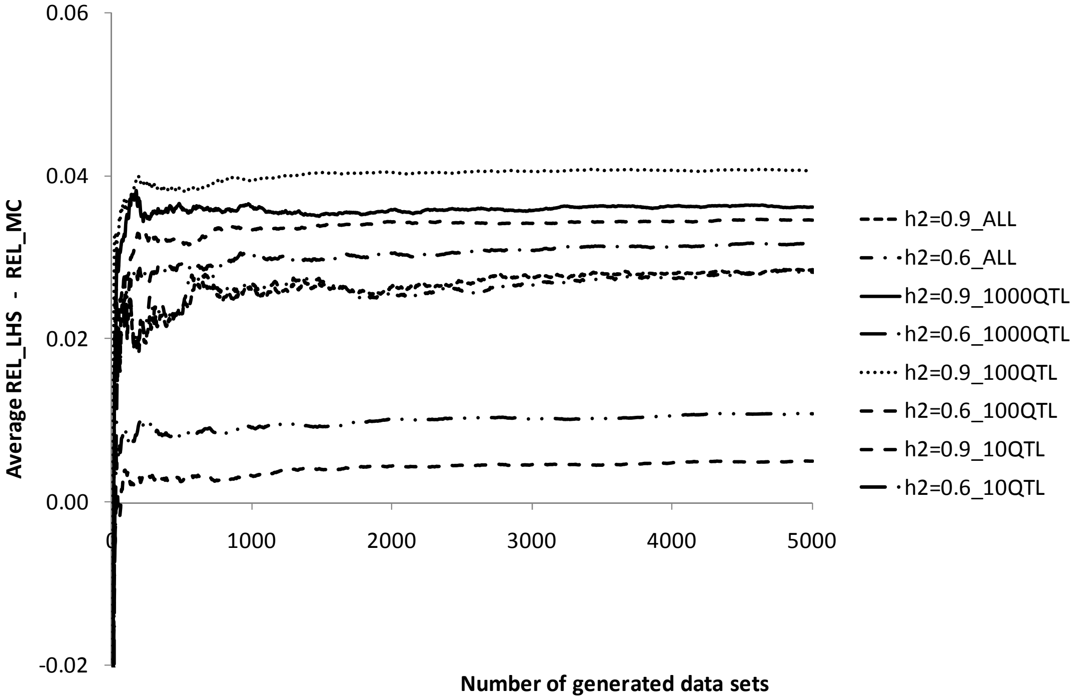

Figure 2. REL_MC minus REL_LHS for the validation animals across numbers of generated data sets.