Comparison of Reliabilities of Direct Genomic Values

M.P.L. Calus, H.A. Mulder and R.F. Veerkamp
Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, The Netherlands

Keywords: direct genomic values, reliability

Introduction

In some countries where direct genomic values (DGV) are currently used in the national evaluation, DGV are calculated by replacing a pedigree based relationship matrix by a genomic relationship matrix (G) (e.g. Berry *et al.*, 2009; VanRaden, 2008). In other countries, methods are used where SNP effects are individually explicitly estimated and summed across the genome for each animal (e.g. De Roos *et al.*, 2009).

There are two common methods to obtain reliabilities of DGV: 1) by inverting the left-hand side of the mixed model equations and 2) using cross-validation. The first method assumes that genetic variance is completely explained by all SNPs, neglecting that SNPs may capture only part of the QTL-effects. Cross-validation allows estimating reliabilities for groups of animals, without making such assumptions. The disadvantage of cross-validation is that no individual reliabilities are obtained. The disadvantage of inverting the left-hand side of the mixed model equations is that reliabilities tend to be overestimated (Calus *et al.*, 2009) next to the computational burden to invert large and dense matrices.

The objective was to compare both methods to calculate reliabilities, to gain more insight in which factors cause differences between them.

Simulated Data

Data was simulated to allow comparing both types of reliabilities to the true reliability. An effective population size of 500 animals was simulated, of which half of the animals were female and the other half male. This structure was kept constant for 1000 generations. Mating was performed by drawing the parents of an animal randomly from the animals of the previous generation. Therefore, each animal had on average two offspring in the next generation. The simulated genome spanned 5 M. Ten thousand bi-allelic loci were simulated, evenly spaced, across 5 chromosomes. In the first generation, animals received at random alleles 1 or 2 with equal chance. In the 1000 generations thereafter, each locus had a mutation rate of 2.5×10^{-5}. A mutation caused an allele 1 to become 2, and vice versa. In total, on average across the 5 replicates, 5,575 loci were segregating in the last 4 generations. These four generations will be from here on referred to as generations 1 to 4.

Two hundred loci that were segregating in generations 1 to 4, scattered evenly across the genome, were drawn to be QTL loci. These QTL were used to simulate a trait with a heritability of 0.9, reflecting average offspring performance such as daughter yield deviations (VanRaden and Wiggans, 1991) or de-regressed proofs (Sigurdsson and Banos, 1995). For example, this value is equivalent to a trait in dairy cattle with a heritability at the phenotypic level of 0.33, considering bulls with 100 daughters. This value was derived using the formula $r_{IH}^2 = \frac{\frac{1}{4}nh^2}{1+\frac{1}{4}(n-1)h^2}$ (e.g. Mrode, 2005), where r_{IH}^2 is the reliability of selection (in this case the used heritability to simulate the phenotypes of the animals in the reference population), n is the number of daughters and h^2 is the heritability at the phenotypic level. All animals in generations 1 and 2 had one phenotype, while animals in generations 3 and 4 had no phenotypes.

Methods to Calculate Reliability

Reliabilities were estimated based on 1) prediction error variances obtained from the left-hand side of the mixed model equations (REL_LHS), 2) a cross-validation (REL_CV), and 3) as the squared correlation between simulated and estimated breeding values (REL_TRUE). REL_LHS was calculated as

REL_LHS = 1 – (*PEV* / ($G_{i,i}\hat{\sigma}_a^2$)), where $G_{i,i}$ is the diagonal element of animal *i* in the G matrix. REL_CV was calculated as the squared correlation between the phenotype (resembling DYDs with h^2 of 0.90) and the estimated breeding value, adjusting for the h^2 of the phenotype:

REL_CV = $r^2(phen, dgv)/h^2$.

Given the structure of the data, in each generation all animals had an equal amount of information. Therefore, the reliability of the DGVs within generations was expected to be similar for all animals, and REL_LHS and REL_TRUE was calculated within each generation.

The following model was used to estimate the DGVs in ASReml (Gilmour *et al.*, 2006):

$y_i = \mu + DGV_i + e_i$

where DGV and its variance were simultaneously estimated. The DGV were distributed as *N(***0**,**G** $\hat{\sigma}_a^2$ *)*, were **G** is a genomic relationship matrix calculated as $\mathbf{G} = \dfrac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$ (VanRaden, 2008), 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* calculated in the current population.

Based on suggestions during the workshop, additionally the analyses were repeated with **G***, calculated as:

G* = 0.8**G** + 0.2**A**

Results

The results for the different types of reliabilities are presented in Figure 1. REL_TRUE was 0.91 for animals with phenotypes in generations 1 and 2, and decreased thereafter to 0.66 and 0.54 in generations 3 and 4 for animals without phenotypes. REL_CV was highly overestimated in generations 1 and 2 (1.08), and very close to REL_TRUE in generations 3 and 4. REL_LHS was very close to REL_TRUE in generations 1 and 2, and 0.06 and 0.09 higher than REL_TRUE in generations 3 and 4, respectively. Replacing G by G* in the model left REL_CV unchanged, while REL_LHS in generations 3 and 4 decreased.

The variance of the DGV was close to the variance of the TBV in generations 1 and 2, but then dropped in generations 3 and 4 (Table 1). Replacing G by G* did not change the variance of the DGVs. The estimated genetic variance ($\hat{\sigma}_a^2$) was very close to the variance of the TBV.

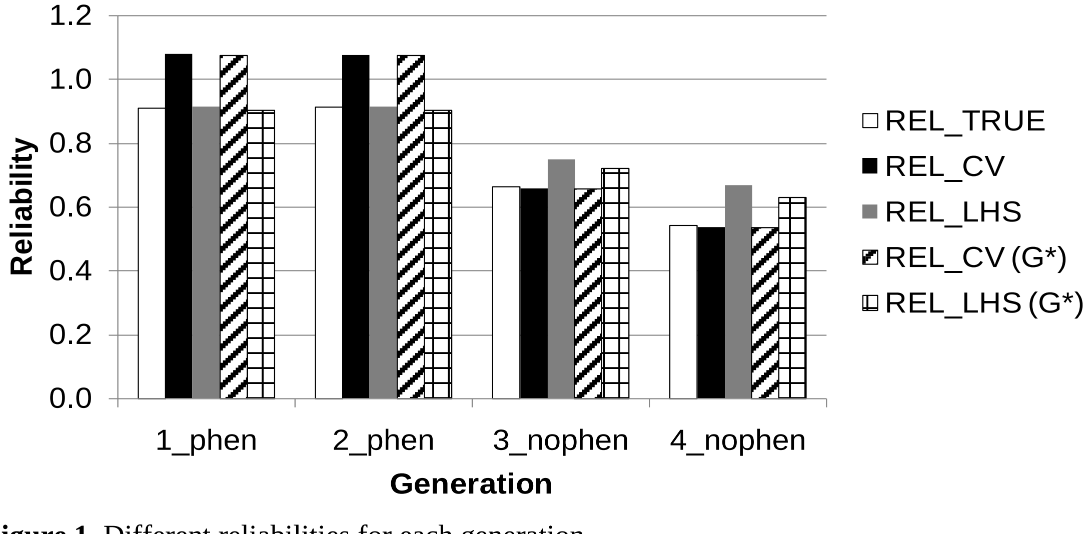

Figure 1. Different reliabilities for each generation.

Table 1. Variances of the true breeding values (TBV), the phenotypes (Phen), direct genomic values (DGV) estimated with G or G*, and the estimated genetic variance in the model ($\hat{\sigma}_a^2$), across the 5 replicates.

			DGV		
Gen	TBV	Phen	G	G*	$\hat{\sigma}_a^2$
1	57.9	64.9	50.0	49.8	57.2[1]
2	56.9	62.0	48.8	48.7	
3	55.4	61.0	39.0	38.8	
4	56.3	62.2	33.3	33.3	

[1]The models yielded only one estimated variance.

Discussion

The objective of this paper was to compare REL_CV and REL_LHS, and to gain more insight in which factors cause differences between them. Since we used simulated data, the true reliability (REL_TRUE) could be estimated. The results clearly showed that REL_LHS accurately estimates the reliability for animals with phenotypes, but overestimates it for animals without phenotypes. REL_CV overestimated the reliability for animals with phenotypes, but accurately estimated the reliability for animals without phenotypes.

REL_CV showed values > 1.0 for animals with phenotypes. Since the DGVs are estimated from the phenotypes, the DGVs will also partly explain the error in the phenotypes. These residual effects that are explained by the DGV incorrectly increase REL_CV.

VanRaden (2009) listed a number of reasons why REL_LHS may differ from REL_CV. The only reason that may play a role in the analysis of our simulated data, is that genetic effects may reside between the markers but are assumed to be located only at the markers. In other words, one reason why REL_LHS overestimated the reliability for animals without phenotypes, may be due to differences in phase between the markers and the QTL in generations 1 and 2 versus 3 and 4.

The phase between the QTL and the neighbouring SNPs, calculated as the correlations between them, were almost identical across the four generations (results not shown). Therefore, the decrease in variance in the estimated DGV and loss in reliability across generations 3 and 4 is not due to a change in phase. In other words, the part of the reliability of the DGV that is decreasing, is not driven by LD between QTL and SNPs. Most likely, this part is explained by pedigree effects that are picked up by the SNPs (Habier *et al.,* 2007), but the accuracy of this part of the prediction decreases rapidly when the distance to the reference population increases. This is however not reflected sufficiently by the model assumptions, leading to the overestimation of REL_LHS. An attempt was made to relax the model assumptions by replacing G by G*, implying that 20% of the genetic variance is explained by a polygenic effect. In this scenario, the overestimation of REL_LHS became smaller, but still was not cured. Further research is needed to better define the weights on G and A in G*, such that the model assumptions fit the data.

Conclusions

REL_LHS overestimated the reliabilities for animals without phenotypes, most likely because the assumption that all explained genetic variance resides at the SNPs is violated. The results show that REL_LHS is the method of choice for animals with phenotypes, while REL_CV is more accurate for animals without phenotypes.

Acknowledgments

NWO – Casimir is acknowledged for financial support.

References

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

Calus, M.P.L., Verbyla, K.L. , Mulder, H.A. & Veerkamp, R.F. 2009. Estimating reliabilities of genomic breeding values. Proc. of the Interbull International Workshop - Genomic Information in

Genetic Evaluations, Barcelona, Spain. *Interbull Bulletin 40,* 199-201.

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

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.

Habier, D., Fernando, R.L. & Dekkers, J.C.M. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics 177,* 2389–2397.

Mrode, R. 2005. *Linear Models for the Prediction of Animal Breeding Values.* 2nd Edition ed. CABI Publishing.

Sigurdsson, A. & Banos, G. 1995. Dependent-variables in international sire evaluations. *Acta Agriculturae Scandinavica Section a-Animal Science. 45(4),* 209-217.

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

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92:1,* 16-24.

VanRaden, P.M. & Wiggans, G.R. 1991. Derivation, calculation, and use of national animal-model information. *J. Dairy Sci. 74:8,* 2737-2746.