Proc. Assoc. Advmt. Anim. Breed. Genet. 22:85-88

ACCURACY OF GENOMIC PREDICTION WHEN QTL EFFECTS ARE NOT NORMALLY DISTRIBUTED

M.E. Goddard[1,2] and T.H.E. Meuwissen[3]

[1] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Victoria, Australia
[2] AgriBio, Department of Economic Development, Jobs, Transport and Resources, Victoria, Australia
[3] Norwegian Life Sciences University, Norway

SUMMARY

In this paper we examine, using simulation and an analytical method, the factors that control the accuracy of genomic prediction when the effects of chromosome segments are not normally distributed, for instance, because many chromosome segments do not contain a QTL. In this situation non-linear methods of analysis give higher accuracy than GBLUP but the advantage is small unless the distribution of chromosome segment effects departs markedly from a normal distribution and the distribution assumed by the method of analysis also departs markedly from a normal distribution. The effect of sample size on accuracy of non-linear methods is similar to that with GBLUP but the advantage of non-linear methods over GBLUP increases with sample size when accuracy is low.

INTRODUCTION

Before implementing genomic prediction of breeding values (genomic selection), it would be useful to be able to predict the accuracy that might be achieved or at least to understand the factors controlling accuracy so that the optimum combination could be used. If genomic estimated breeding values (GEBVs) are estimated using GBLUP (Meuwissen et al 2001), there is good theory to predict the accuracy (Daetwyler et al 2008, Goddard 2009). In this case, the accuracy or correlation between EBV and true breeding value (r) is approximately given by MacLeod et al (2014)

$$r^2 = \theta c/(1+\theta - h^2 r^2) \quad (1)$$

where c = the proportion of genetic variance explained by markers
h^2 = heritability
$\theta = Nh^2c/M_e$
N= number of records in the training population
M_e = effective number of independent chromosome segments in the genome.

This is not an explicit formula for r^2 because r^2 appears on both sides of the equation. However, we choose to present the formula in this way because it makes clear the way in which increasing accuracy decreases the unexplained variance and so further increases accuracy. If the causal variants or QTL have similar properties to the markers, then $c = M/(M + M_e)$ where M is the number of markers. However, c is often less than this presumably because the QTL have lower linkage disequilibrium (LD) with the markers than the markers do amongst themselves.

Estimation of breeding values using GBLUP, as above, is a Bayesian prediction if it is assumed that the effects of the markers are all drawn from a normal distribution with mean zero and constant variance. That is, a model in which the genomic relationships between the animals is estimated from the markers (GBLUP) is equivalent to a model in which SNP effects are assumed to be normally distributed (SNP-BLUP). Other assumptions about the distribution of marker effects lead to other methods of estimation of which some have been called Bayes A, B, C or R. Although BLUP is a linear estimate in the phenotypic values (y), these other Bayesian methods are non-linear in y. These non-linear Bayesian methods give higher accuracy than BLUP in some

cases (MacLeod e al 2014) but there is no theory that predicts how much more accurate and in what circumstances. As well as the parameters that affect GBLUP accuracy, the accuracy of non-linear methods could be affected by the true distribution of marker effects and the distribution assumed by the method of analysis. The aim of this paper is illustrate how these parameters affect the accuracy of non-linear Bayesian methods of predicting breeding value. We use simulation and a simplified analytical model.

MATERIALS AND METHODS

Analytical method. Here we assume that the markers and QTL are identical and there are M_e independent QTL so that the accuracy of estimating a single QTL effect (r) is equal to the accuracy with which the combined value of all QTL is estimated. This can then be calculated using numerical integration. That is, $r^2 = V(\hat{q})/V(q)$ and $V(\hat{q}) = \int f(q)E(\hat{q}|q)^2\, dq$, where q is the effect of a QTL assumed to have a mean of zero, f(q) is the distribution of QTL effects, $E(\hat{q}|q)$ is the expectation of the estimate of q ($\hat{q}$) *g*iven q.

Simulation. We simulated a genome of length 1M in a population of N_e = 1000 until it reached mutation-drift equilibrium. At this point there were approximately 33,000 SNPs segregating of which between 3 and 290 were designated as QTL and their effect sampled from a distribution that was either exponential or gamma (shape parameter = 0.09) or t-distribution (degrees of freedom = 4.1 or 4.2). The scale of the effects was adjusted so that a fixed heritability was reached after adding normally distributed environmental effects. The linkage disequilibrium among the markers means that the effective number of chromosome segments (M_e) is approximately 300. The simulated data on 200 animals were analysed with BLUP, Bayes A, Bayes B (Meuwissen et al 2001) and Bayes R (Erbe et al 2012) and the correlation between true breeding value and EBV calculated in an independent set of animals. Because the results depend to θ, the simulation approximately corresponds to a genome of 30 M but with a sample size of 30 * 200 = 6000.

RESULTS AND DISCUSSION

Simulation results. Table 1 lists the accuracy achieved when h^2 = 0.5 and the all 33,000 markers were used so that all genetic variance is explained by the markers (c=1 in equation 1).

Table 1. Effect of distribution of QTL and distribution assumed by the method of analysis on accuracy (%) of EBVs

For Bayes R Sim. = simulation results, anal. = analytic approximation, all other results are from simulation.

No. QTL	Distribution	Method of analysis				
		GBLUP	Bayes B	Bayes R		Bayes A
				sim.	anal.	
3	exponential	51	97	95	98	67
30	exponential	49	83	82	85	54
30	gamma	48	88	89	96	65
30	t (df = 4.105)	54	81	82	81	57
290	t (df = 4.225)	52	57	55	61	51

When GBLUP is used, assuming a normal distribution of marker effects, the accuracy is nearly the same (~0.5) regardless of the true distribution of QTL effects. Although there are 33,000 SNPs, there are only about 300 effective independent chromosome segments. Therefore the last

distribution in table 1 with 290 QTL with effects drawn from a t distribution does not differ greatly from a distribution in which all chromosome segments have an effect drawn from a normal distribution. Consequently the Bayesian methods have little advantage over GBLUP. When there are less than or equal to 30 QTL, many chromosome segments have zero effect and the distribution differs more markedly from a normal distribution. In these cases Bayes B and Bayes R have similar accuracy and an advantage over GBLUP. Bayes B and Bayes R assume a distribution of marker effects in which some effects are zero and this agrees with the true distribution in the first 4 cases in table 1. Bayes A assumes no effects are zero but all SNP effects follow a t-distribution. The accuracy it achieves is in between that of GBLUP and Bayes B or R.

The accuracy of the non-linear methods (e.g. Bayes B and R) depends in part on the kurtosis of the distribution of effects of chromosome segments. If many segments have zero effect (i.e. no QTL in the segment) the kurtosis is increased. However, the kurtosis is not the only parameter of the distribution that affects the accuracy of EBVs. In table 1 the gamma distribution with 30 QTL and the exponential distribution with 3 QTL have similar kurtosis but the exponential distribution leads to higher accuracy. This is because the gamma distribution with shape parameter of 0.094 has some large effects but also many very small effects that are hard to estimate accurately.

The results in table 1 can be summarised by

- the true distribution must differ greatly from a normal before non-linear methods have an advantage over GBLUP,
- it is not worthwhile to use a non-linear method of analysis unless it assumes a distribution of marker effects that differ greatly from a normal distribution.

Analytical method. Here we calculated the accuracy of estimating the effect of a single QTL assuming that the method of analysis used the same distribution of QTL effects as used to generate true QTL effects. Table 1 shows that the analytical method overestimates the accuracy found by simulation. This is expected. The analytical method assumes there is only one marker per effective chromosome segment, whereas in the simulation there are approximately 100. The GBLUP analysis shrinks estimates of marker effects but the amount of shrinkage is not effected by the size of the estimated effect. Consequently, the effect of a chromosome segment can be shared among several markers with little loss of accuracy. But the non-linear methods shrinks apparently large effects less than small effects (Figure 2) and so, if the effect of a single QTL is shared among several markers, the effect is shrunk too much and this reduces the accuracy.

Apart from this over prediction of accuracy, the analytical method does predict the differences in accuracy between distributions (Table 1) and, although not shown here, it also predicts the effect of changing θ reasonably well. In figure 1, we use the analytical method to examine the effect of θ on accuracy. The y-axis of the graph is $T = r^2/(1-r^2)$. For GBLUP analysis this is almost equal to θ but differs from it due to the $-h^2r^2$ term in equation 1. This term corrects for the reduction in error variance when estimating the effect of one marker due to the simultaneous prediction of the effects of all other markers (Daetwyler et al 2008). Consequently, T is slightly greater than θ for GBLUP and this disparity increases slightly with θ. For the non-linear methods, T increases faster than linear in θ and the advantage over GBLUP increases with θ at first and then reaches a constant ratio.

In real data within one breed, the distribution of QTL effects may be most similar to the t-distribution with 290 QTL in 300 effective chromosome segments corresponding to 8100 QTL in a 30M genome. This would explain why non-linear methods enjoy only a small advantage over BLUP in many cases. The advantage of non-linear methods would be expected to increase if multiple breeds were analysed or the population had a high effective population size e.g. in humans.

REFERENCES

Daetwyler, H.D., Villanueva, B. & Woolliams, J.A. 2008 Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLoS ONE* **3**, e3395.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. and Goddard M.E. (2012) *J. Dairy Sci.* **95**: 4114.

Goddard, M. 2009 Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-257.

MacLeod IM, Hayes BJ, & Goddard ME (2014) The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. *Genetics* 198(4):1671-1684.

Meuwissen, T., Hayes, B. and Goddard, M. 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819 - 1829.

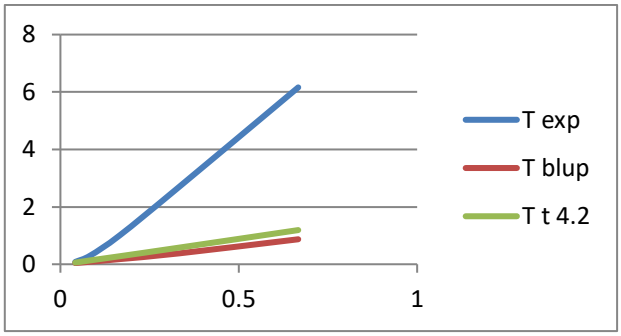

Figure 1. The effect of θ on $T = r^2 / (1-r^2)$. The graphs show the effect of θ on accuracy from the analytical method for the exponential distribution of 30 QTL effects (T exp), the normal distribution of 300 QTL effects (T blup) and the t-distribution with degrees of freedom = 4.225 of 290 QTL effects

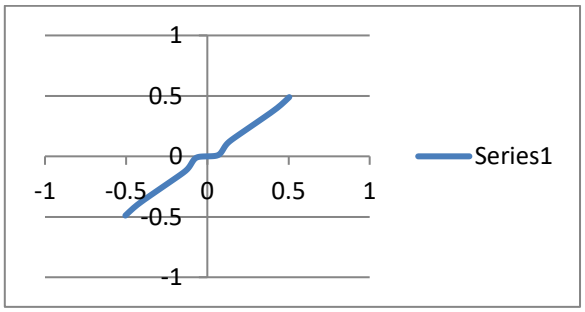

Figure 2. Estimated QTL effect size vs true QTL effect size from the analytical method under the exponential distribution of 30 QTL in 300 effective chromosomal segments (arbitary scale of effect sizes)