Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Is the use of formulas a reliable way to predict the accuracy of genomic selection?

S. Brard *, **A. Ricard**†‡.
*UMR INRA / INPT ENSAT / INPT ENVT, Génétique, Physiologie et Systèmes d'Elevage, INRA, F-31326 Castanet-Tolosan, France, †INRA, UMR 1313, 78352 Jouy-en-Josas, France, ‡IFCE, Recherche et Innovation, 61310 Exmes, France

ABSTRACT: We studied 4 formulas for prediction of accuracy of genomic selection. Our objectives were: to study the variation of accuracy depending on the parameters, and to compare observed accuracy in 13 references (145 values) to the accuracy given by formulas. The marginal distribution of accuracy was studied according to each parameter (range of values defined by the references). Then we compared accuracies predicted using formulas and observed accuracies (got by cross-validation for real data or correlation for simulations). We proved that the size of the reference population and the number of effective segments M_e have a major weight and that there were large differences between formulas. The formula which gave the best prediction of observed accuracy depended on M_e estimation given the effective size of population and so no generalization can be done, and no use of formulas to decide genomic plans.
Keywords: Genomic selection; Accuracy; Effective number of segments

Introduction

Genomic selection is now used in many countries for genetic improvement of animals. Genotypes are used to estimate the breeding values of animals before they get own performances. The accuracy of genomic selection is the correlation between the true breeding value and the breeding value estimated using genotypes. In real data the accuracy of genomic selection is calculated posterior to genomic selection, using cross-validation. This method requires genotyped and phenotyped animals. But formulas have been developed to predict accuracy before genotyping. They use parameters that describe the potential data (animals, trait, and markers). Accuracy predicted with these formulas can encourage the start of genomic selection in a breed or discourage it. However, to our knowledge the weight of the parameters used in the formulas has never been explored in detail, and up to now the results of the formulas have been checked only on data chosen specifically for the purpose of testing the formulas. Our objectives were: (i) to study to what extent variations in parameters induce variations of accuracy calculated using formulas, and (ii) to compare accuracies from literature (obtained mainly in dairy cattle genomic evaluation with simulated or real data) to accuracies predicted using formulas.

Materials and Methods

Four formulas for prediction of accuracy of genomic selection. This work focused on four recently developed formulas for prediction of accuracy of genomic selection.

Daetwyler et al. (2008) formula r_D: $r_{g\hat{g}} = \sqrt{\frac{Th^2}{Th^2+M_e}}$ with $r_{g\hat{g}}$ the accuracy of genomic selection, T the size of the reference population, h^2 the heritability and M_e the number of effective segments.

Goddard et al. (2011) formula r_G: $r_{g\hat{g}} = \sqrt{b\frac{Tbh^2/M_e}{1+Tbh^2/M_e}}$ with b=M/(M+Me), M the number of markers.

Goddard (2009) formula r_{Go}:

$r_{g\hat{g}} = \sqrt{1-\frac{\lambda}{2T\sqrt{a}}}\,Log\left(\frac{1+a+2\sqrt{a}}{1+a-2\sqrt{a}}\right)$ with $\lambda = \frac{M_e}{h^2Log(2N_e)}$ and $a = 1+2\frac{M_e}{Th^2Log(2N_e)}$ with Ne the effective size of the population.

Meuwissen et al. (2013) formula r_M:

$r_{g\hat{g}} = \sqrt{\frac{\theta+1+\sqrt{(\theta+1)^2-4h^2\theta b}}{2h^2}}$ with $\theta = Tbh^2/M_e$.

Formulas for effective number of segments M_e. The effective number of chromosome segment M_e is also called "effective number of loci" or "number of independent chromosome segments". Goddard (2009) assumed that every potential QTL is tagged by a marker. Linkage, by causing linkage disequilibrium, limits the number of markers required to M_e. M_e can be estimated with the formulas:

$M_{e1} = \frac{2N_eL}{Log(4N_eL)}$, where L is the length of the genome in Morgan (Goddard 2009).

$M_{e5} = 4N_eL$ Stam (1980).

$M_{e2} = \frac{2N_eL}{Log(2N_eL)}$, $M_{e3} = \frac{2N_eL}{Log(N_el)}$, where l is the average length of one chromosome (Goddard et al. 2011).

$M_{e4} = 2N_eL$ Hayes et al. (2009b). Formula M_{e1} gives the lowest value of M_e and M_{e5} the highest for the same N_e.

Thirteen publications as source of data. Thirteen publications dealing with accuracy of genomic selection (mainly in dairy cattle) were used to investigate the formulas. They used simulated data (Meuwissen et al. 2001, Habier et al. 2007, Calus et al. 2008, Calus et al. 2009, Habier et al. 2009, Pszczola et al. 2011, Baastiansen et al. 2012, Brito et al. 2012) or real data (Hayes et al. 2009a, Luan et al. 2009, Verbyla et al. 2009, Moser et al. 2010, Habier et al. 2010). The ranges of values found for parameters and accuracies are reported in Table 1. From this point accuracies from publications will be the "observed accuracies" whereas accuracies got by with the formulas will be the "predicted accuracies".

Table1. Range of values found in publications for accuracy of genomic selection and parameters, in real data or simulated data.

	Real data (76 cases)			Simulated data (69 cases)		
	Mean	Min	Max	Mean	Min	Max
Observed accuracy	0.57	0.17	0.78	0.5	0.11	0.9
h^2	0.88	0.58	0.97	0.45	0.1	0.94
T	812	250	2096	880	480	2200
M	29011	18991	42576	41236	100	800000
N_e	127	45	167	184	95	400
L	31.6	-	-	8.77	3	23.33
M_{e1}	822	329	1060	432	81	1646
M_{e2}	1264	542	1610	601	95	2440
M_{e3}	1421	625	1800	669	108	2707
M_{e4}	1621	737	2042	756	124	3038
M_{e5}	7998	2844	10554	4097	570	17190

Variation of accuracy predicted with formulas. To analyze the variation of accuracy, a range of variation was defined for parameters according to their values observed in the articles (Table 2). The marginal probability density function of accuracy was computed for each parameter, by integration over the other parameters, for example for T:

$$f(r_{g\hat{g}}|T) = \iiint_{M,M_e,h^2e\Delta} f(r_{g\hat{g}}|T,M,M_e,h^2)p(M)p(M_e)p(h^2)dMdM_edh^2$$

with $f(r_{g\hat{g}}|T,M,M_e,h^2)$ the 4 previous formulas for accuracy, and p(M), $p(M_e)$, $p(h^2)$ and p(T) the density of each parameter. For each parameter, the density function was chosen to attribute a similar probability to the most common values: M_e and h^2 had a uniform distribution, and M and T had a uniform log distribution.

Table 2. Range of values of parameters chosen for the analysis of variance.

Parameter	Minimum	Maximum
h^2	0.1	0.98
T	250	20000
M	3000	800000
N_e	45	400
M_e	250	20000

Comparison between observed accuracies and predicted accuracies. The 13 publications gave 145 observed accuracies. Predicted accuracies were calculated with the corresponding parameters. An analysis of variance was performed on the difference between observed and predicted accuracies, with sources of variation: formula for accuracy*formula for M_e (20 levels), type (real or simulated data) and T, h^2, M and L as covariates.

Results and Discussion

Analysis of marginal distribution of accuracy. Figures 1 to 4 show the variation of accuracy according to the parameters for the 4 formulas. For r_{Go} only the 2 extreme assumptions about the relationship between M_e and N_e are shown (all other between them). Accuracy increases when T, h^2 or M increase, and decreases when M_e increases. For M, the maximum is not reached with the 50000 markers corresponding to the more common beadship. T and M_e induce more important variations of $r_{g\hat{g}}$ than h^2 or M. The major weight found for M_e is an issue because for a same N_e, M_e varies a lot depending on the formula used for estimation.

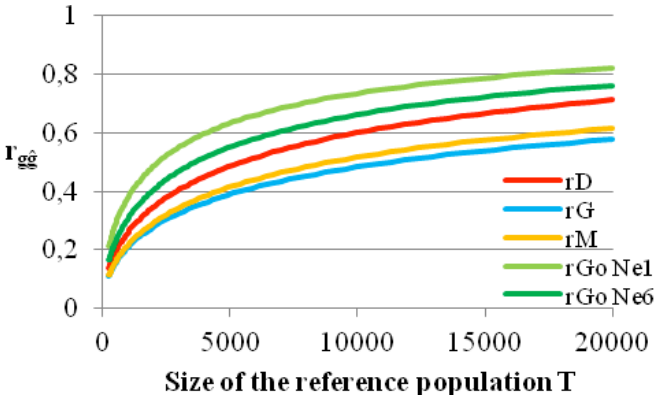

Figure 1. Marginal distribution of accuracy in function of the size of the reference population T.

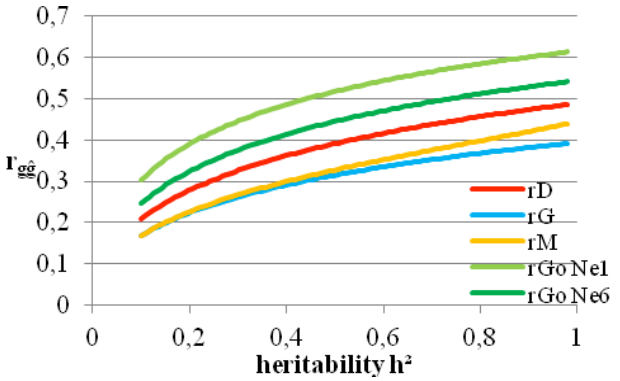

Figure 2. Marginal distribution of accuracy in function of the heritability h².

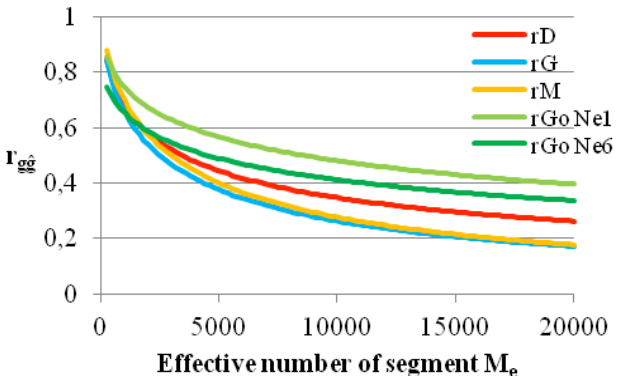

Figure 3. Marginal distribution of accuracy in function of the number of effective segment M_e.

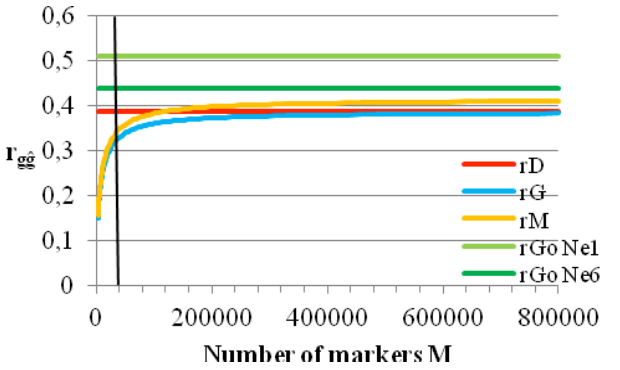

Figure 4. Marginal distribution of accuracy in function of the number of markers M.

The average accuracy is different for each formula: $r_G(0.31)<r_M(0.33)< r_D(0.39)< r_{Go}\ N_{e6}(0.44)< r_{Go}\ N_{e1}(0.51)$. These accuracies are higher than the ones got for the prediction of own performances with an intermediate heritability, but lower than accuracies got by progeny testing. But depending on the values of the parameters, accuracy got with formulas can be much lower and unfavorable (<0.20) or much higher and favorable (>0.70).

Comparison between observed and predicted accuracies. In the variance analysis the type of data was not significant, whereas all other effects were. Predicted accuracy overestimated observed accuracy when T, M and L were large: +0.06 for each more1000 animals, +0.01 for each more 100 000 markers, +0.008 for each more Morgan. Predicted accuracy underestimated observed accuracy for large h²: -0.057 by 0.1 h².

Figure 5 displays the least squares means for difference between observed and predicted accuracies depending on formulas used to compute $r_{g\hat{g}}$ and M_e. Differences between observed and predicted accuracy increased as M_e increased, depending on the method used to get this parameter. Accuracy was overestimated when using M_{e1} and was underestimated when using M_{e4} or M_{e5}. So according to these results, the best formula to predict accuracy depends on the formula used to compute M_e. With M_{e2}, M_{e3} or M_{e4} the formulas that give the best predictions are r_G, r_M and r_{Go} respectively. These results evidenced once again the importance of M_e estimation as, depending on M_e, the prediction of accuracy was over or underestimated.

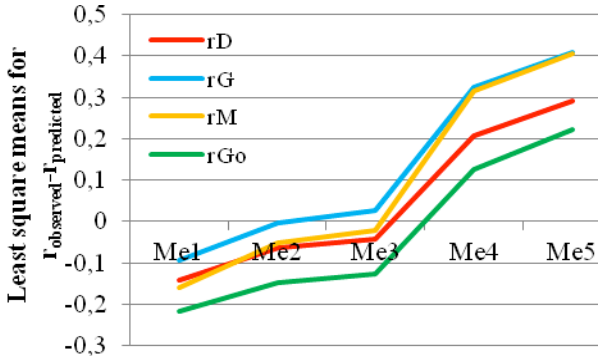

Figure 5. Least square means for the difference between the observed and the predicted accuracies depending on the formulas for accuracy and on the number of effective segment M_e.

Conclusion

So far, the main problem evidenced is that the uncertainty on the appropriate method for M_e estimation prevents to check if formulas for prediction of accuracy actually work, and to determine which one is the better one. The recommendation we make to people wanting to plan genomic selection using formulas would be to be careful with the parameters: we proved that formulas overestimate or underestimate accuracy for extreme values of parameters. Particular attention should be paid to M_e: this parameter has a huge weight and its value is completely different depending on the method for estimation.

Literature Cited

Bastiaansen, J. W. M., Coster, A., Calus, M. P. L. et al. (2012). *Genet. Sel. Evol*., 44, 3.

Brito, F. V., Neto, J. B., Sargolzaei, M., et al. (2012). *BMC Genetics*, 12, 80.

Calus, M. P. L., Meuwissen, T. H. E., de Roos, A. P. W., et al. (2008). *Genetics*, 178, 553-561.

Calus, M. P. L., Meuwissen, T. H. E., Windig, J. J. et al. (2009). *Genet. Sel. Evol*., 41, 11.

Daetwyler, H. D., Villanueva, B., Wooliams, J. A. (2008). *PLoS ONE* 3(10): e3395.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., et al. (2010). *Genetics*, 185, 1021-1031.

Goddard, M. (2009). *Genetica*, 136, 245-257.

Goddard, M. E., Hayes, B. J., Meuwissen, T. H. E. (2011). *J. Anim. Breed. Genet*., 128, 409-421.

Habier, D., Fernando, R. L., Dekkers, J. C. M. (2007). *Genetics*, 177, 2389-2397.

Habier, D., Fernando, R. L., Dekkers, J. C. M. (2009). *Genetics*, 182, 343-353.

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

Habier, D., Fernando, R. L., Garrick, D. J. (2013). *Genetics*, 194, 597-607.

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., et al. (2009a). *Genet. Sel. Evol*., 41, 51.

Hayes, B. J., Visscher, P. M., Goddard, M. E. (2009b). *Genet. Res. (Camb)*, 91, 47-60.

Luan, T., Wooliams, J. A., Lien, S., et al. (2009). *Genetics*, 183, 1119-1126.

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

Meuwissen, T., Hayes, B., Goddard, M. (2013). *Annu. Rev. Anim. Biosci.*, 1, 221-237.

Moser, G., Khatkar, M. S., Hayes, B. J., et al. (2010). *Genet. Sel. Evol*., 42, 37.

Pszczola, M., Mulder, A. H., Calus, M. P. L. (2011). *J. Dairy Sci*., 94, 731-441.

De Roos, A. P. W., Hayes, B. J., Spelman, R. J., et al. (2008). *Genetics*, 179, 1503-1512.

Stam, P. (1980). *Genet. Res*., 35, 131-155.

Verbyla, K. L., Hayes, B. J., Bowman, P. J., et al. (2009). *Genet. Res. (Camb)*, 91, 307-311.