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 $r_{DG}$: $r_{DG} = \frac{Tb}{\sqrt{Tb^2 + M_e}}$ with $r_{DG}$ 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 $r_{G}$: $r_{G} = \frac{b}{b+Tb^2/M_e}$ with $b=M/(M+M_e)$, $M$ the number of markers.

Goddard (2009) formula $r_{Go}$: $r_{Go} = \sqrt{1 - \frac{\lambda}{2V_{a}} \log \left(\frac{1+a+2\sqrt{a}}{1+a-2\sqrt{a}}\right)}$ with $\lambda = \frac{M_e}{h^2 \log(2N_e)}$ and $a = 1 + 2\frac{M_e}{Tb^2 \log(2N_e)}$ with $N_e$ the effective size of the population.

Meuwissen et al. (2013) formula $r_{M}$: $r_{M} = \sqrt{\frac{\theta + 1 + \sqrt{(\theta + 1)^2 - 4\theta b}}{2\theta b}}$ with $\theta = T b^2 / M_e$.

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:

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

$M_{e5} = 4N_e L$: Stam (1980).

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

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

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Real data (76 cases)</th>
<th>Simulated data (69 cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Min</td>
</tr>
<tr>
<td>Observed accuracy</td>
<td>0.57</td>
<td>0.17</td>
</tr>
<tr>
<td>h²</td>
<td>0.88</td>
<td>0.58</td>
</tr>
<tr>
<td>T</td>
<td>812</td>
<td>250</td>
</tr>
<tr>
<td>M</td>
<td>29011</td>
<td>18991</td>
</tr>
<tr>
<td>N_e</td>
<td>127</td>
<td>45</td>
</tr>
<tr>
<td>L</td>
<td>31.6</td>
<td>-</td>
</tr>
<tr>
<td>M_e1</td>
<td>822</td>
<td>329</td>
</tr>
<tr>
<td>M_e2</td>
<td>1264</td>
<td>542</td>
</tr>
<tr>
<td>M_e3</td>
<td>1421</td>
<td>625</td>
</tr>
<tr>
<td>M_e4</td>
<td>1621</td>
<td>737</td>
</tr>
<tr>
<td>M_e5</td>
<td>7998</td>
<td>2844</td>
</tr>
</tbody>
</table>

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, h², M and L as covariates.

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, h² or M increase, and decreases when M_e increases. For M_e the maximum is not reached with the 50000 markers corresponding to the more common beadship. T and M induce more important variations of r_G than h² 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.

Results and Discussion

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:

\[ f\left(r_{\text{GG}}\right|T) = \iiint_{M,M_e,h^2} f\left(r_{\text{GG}}\right|T,M,M_e,h^2)p(M)p(M_e)p(h^2)dMdM_edh^2 \]

with \( f\left(r_{\text{GG}}\right|T,M,M_e,h^2) \) 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.
The average accuracy is different for each formula: $r_D(0.31)<r_M(0.39)<r_G(0.44)<r_{Go}(0.51)$. 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 more 1000 animals, +0.01 for each more 100 000 markers, +0.008 for each more Morgan. Predicted accuracy underestimated observed accuracy for large $h^2$: -0.057 by 0.1 $h^2$.

Figure 5 displays the least squares means for differences between observed and predicted accuracies depending on formulas used to compute $r_G$ 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 $M_e$ and was underestimated when using $M_e$ or $M_e$. So according to these results, the best formula to predict accuracy depends on the formula used to compute $M_e$. With $M_2$, $M_3$, or $M_4$ 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.

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**


