Considering Dominance In Genomic Selection

R. Wellmann, J. Bennewitz [1]

Introduction

The prediction of breeding values on the basis of genome wide dense marker maps starts to become common in practice. No dominance is considered so far. The inclusion of dominance effects into models for the prediction of genomic breeding values could increase the accuracy of the predictions if the data structure allows it. Moreover, the predicted dominance effects could be used to choose mating pairs with good combining ability by recovering inbreeding depression and utilizing possible overdominance.

The genetic evaluation method for populations with dominance and inbreeding, proposed by de Boer and Hoeschele (1992) assumes that locus effects are fixed and randomness stems from random sampling of genotypes. But in most models for the prediction of genomic breeding values, the marker effects are at random and the genotypes of the individuals are fixed. This randomness arises either from uncertainty about the marker effects, or from random sampling of the population from all hypothetical populations. Daetwyler *et al.* (2008) and Goddard (2008) derived approximations for the accuracy r_{BV} of predicted breeding values. Summarized by Meuwissen (2009) we have

$$r_{BV} = \sqrt{\frac{Nh^2}{Nh^2 + Q_e}} = \sqrt{\frac{ah^2}{1 + ah^2}} \tag{1}$$

approximately, where h^2 is the heritability and $a = \frac{N}{Q_e}$ can be calculated from the number of training records N and the effective number of QTL loci Q_e which depends on the effective population size N_e and the chromosome sizes in Morgan. But the accuracy also depends on whether the model assumptions are satisfied by the data and on the method that is chosen for the prediction (here BLUP). One situation that leads to a violation of model assumptions is the presence of dominance and inbreeding depression. The aim of this study is to assess the effects of different methods and of different amounts of dominance variance on the accuracy of predicted breeding values and dominance effects.

Simulation

A Fisher-Wright diploid population with population size $N = 1000$ was simulated by sampling with replacement for 5000 generations. Thereafter, the population size decreased for 400 generations from 1000 to 100 according to $N_{e,t-400} = 100+900\frac{1-e^{0.005t-2}}{1-e^{-2}}$ in order to reproduce the LD-pattern that is observed in cattle (compare Villa-Angulo *et al.*, 2009). Starting with generation 1, the population was maintained without selection for 5 generations with $N_{e,t} = 100$ but $N_t = 1000$. This was achieved by unequal

[1]Universität Hohenheim, Institut für Tierhaltung und Tierzüchtung, D-70599 Stuttgart, Germany

numbers of males and females. Mating was random with replacement, except in generation 0, where a factorial mating was carried out. The marker effects were predicted in generation 1. These predictions were used to predict breeding values and dominance values in generations 1 to 5.

An infinite-sites mutation model was assumed. The genome had only one chromosome which equals 1 Morgan, making use of the scaling argument of Meuwissen (2009). From about 12000 SNPs 1666 markers were identified based on the MAF and on the distance to neighbouring markers. Each SNP became a QTL with a probability of 1%. This resulted in 120 QTL on average. The distribution of the QTL effects a_i with mean 0 and variance σ_a^2 was a mixture of a double exponential distribution and a normal distribution, i.e.

$$a_i \sim 0.95 \cdot \mathcal{L}(0, u^2) + 0.05 \cdot \mathcal{N}(0, (5u)^2)$$

with appropriate $u > 0$. This resulted in many QTLs with small effects and few QTLs with large effect. Dominance degrees h_i were normally distributed with mean μ_h, variance σ_h^2, and they were independent from the additive effects. Dominance effects $d_i = h_i|a_i|$ and the additive effects a_i were dependent, although their covariance is 0. According to Bennewitz and Meuwissen (2010), $\mu_h = 0.193$ and $\sigma_h = 0.312$ would be realistic, although these parameters would depend on the simulated trait. The small genome size and QTLs with large effect however resulted in a large sampling error, that is, realized additive variance and dominance variance deviated from their expectations. Instead to compare different scenarios with designed additive and dominance variance, we therefore considered σ_a^2, σ_h^2 and μ_h as random variables, i.e. different randomly chosen values were assumed for each replicate. This resulted in a wide range of the realized additive variance and dominance variance of the simulated populations. 50 traits with independent QTL effects were simulated for 10 replicated populations (i.e. 500 data sets) in order to speed computation time.

Statistical model

The statistical model used for the prediction of breeding values and dominance values was

$$Y_i = \mu + \beta \widehat{F}_i + \sum_{n \in \mathcal{M}} \tilde{A}_n (v_{ni} + m_{ni}) + \sum_{n \in \mathcal{M}} (\tilde{D}_n - \mu_D)(v_{ni} - m_{ni})^2 + E_i.$$

Thereby, μ and β are fixed parameters and $\widehat{F}_i$ is the estimated inbreeding coefficient of individual i. The paternal and maternal alleles $v_{ni}, m_{ni} \in \{0, 1\}$ at the nth marker of individual i were fixed explanatory variables. The additive and dominance effects of the markers $\tilde{A}_n$ and $\tilde{D}_n$ were assumed to be normal distributed and independent with $E(\tilde{A}_n) = 0$ and $E(\tilde{D}_n) = \mu_D$. The variances of the marker effects were calculated from additive variance, dominance variance and from inbreeding depression of the respective trait by assuming independence and equal variances for all additive effects (resp. dominance effects) and that the markers explain the full additive variance and dominance variance. Genomic breeding values EBV_i and dominance values EDV_i were

calculated from predicted marker effects as (Falconer and Mackay, 1996):

$$EBV_i = \sum_{n \in \mathcal{M}} \big(\widehat{A}_n + (\widehat{D}_n + \widehat{\mu}_D)(q_n - p_n)\big)(v_{ni} + m_{ni}), \qquad EDV_i = \sum_{n \in \mathcal{M}} -2(\widehat{D}_n + \widehat{\mu}_D)(v_{ni} - p_n)(m_{ni} - p_n),$$

where p_n and $q_n = 1 - p_n$ are the frequencies of the alleles of the nth marker in the respective generation and $\widehat{\mu}_D$ is calculated from $\widehat{\beta}$.

Results and Discussion

Solving Equation (1) for ah^2 shows that $\frac{r^2_{BV}}{1-r^2_{BV}} = ah^2$. Therefore, we assumed a linear model with the heritability h^2, the fraction of dominance variance d^2 and the inbreeding depression $Inbr$ as explanatory variables and $\frac{r^2_{BV}}{1-r^2_{BV}}$ as the dependent variable in order to predict the accuracies. The equation for trait $k = 1, ..., 50$ in population $j = 1, ..., 10$ is

$$\frac{r^2_{BV,jk}}{1 - r^2_{BV,jk}} = a_1 h^2_{jk} + a_2 d^2_{jk} + a_3 |100 Inbr_{jk}| + e_{jk},$$

where $r_{BV,jk}$ is the mean accuracy in generations $1 - 5$, e_{jk} is the error, and $Inbr$ is the change of the trait value when the inbreeding coefficient increases by 1%. The same model was assumed to predict the accuracy of the dominance value r_{DV}. We write $a_{BV} = (a_1, a_2, a_3)^T$ or $a_{DV} = (a_1, a_2, a_3)^T$, depending on whether r_{BV} or r_{DV} was predicted. For the BLUP model without dominance, we obtained the least squares estimate $\widehat{a}_{BV} = (9.0, 3.2, -0.3)^T$ and for the BLUP model with dominance, we obtained $\widehat{a}_{BV} = (9.0, 4.4, -0.3)^T$ and $\widehat{a}_{DV} = (0.2, 2.4, 1.0)^T$.

The interpretation of the results is complicated by the fact that covariates were empirically correlated. Inbreeding depression and dominance variance are correlated because inbreeding depression contributes to dominance variance. Heritability and dominance variance are positively correlated because it can be shown that σ^2_a affects additive and dominance variance in the same way. Therefore, the effects of some covariates could be captured by others. Average values were $d^2 = 0.035, Inbr = -0.0043$, and $h^2 = 0.25$. For BLUP without dominance, the apparently positive effect of dominance variance ($a_2 = 3.2$) on r_{BV} was on average overcompensated by the negative effect of inbreeding depression ($a_3 = -0.3$), but not so for BLUP with dominance. The effect of heritability on r_{BV} is smaller than predicted from Equation (1) for several reasons. At first, N_e decreased with $N_e = 100$ only in the last generations. Secondly, the model assumptions were not satisfied by the data. Moreover, the accuracy was averaged over 5 generations and decreases over time. The inclusion of dominance increased the accuracy of the breeding values only little. Inbreeding depression decreased r_{BV} even if the model accounts for it. This possibly arised because the estimates of the inbreeding coefficients were not accurate enough due to the limited number of markers. Nevertheless, inbreeding depression could well be utilized to predict dominance values.

Conclusion

The BLUP Model has the strong advantage that computations are feasible even for relatively large numbers of markers and that the model is robust to the true distribution of the marker effects as stated by Goddard (2008). But it has the disadvantage that several model assumptions are often not satisfied by the data. Modifications of the BLUP procedure that account for the violated assumptions could therefore increase the accuracies of predicted dominance values and breeding values. Heuristics show that the accuracy of predicted dominance values that are obtained by BLUP is indeed below optimum (unpublished results). However, modifications could make the procedure less robust. A violated assumption is that additive and dominance effects are not independent, although their covariance was 0 in the simulations. One approach to account for this is to predict the dominance effects conditional on the predicted additive effects and vice versa. Although QTL effects were not normally distributed, the assumption of normally distributed marker effects may be not so bad since the number of QTLs was quite large and it is sufficient for the markers to collectively predict the effects of the haplotypes to which they belong. But breeding values and dominance values depend on the frequencies of the QTL and these frequencies do not need to coincide with the frequencies of the markers that capture the QTL effects. Attempts should be made to regress those marker effects back that are not needed for the prediction, e.g. when the markers are not in linkage with a QTL or when they are redundant. This could be done by the use of different methods (e.g. LASSO) or by assigning different variances to different parts of the chromosomes, e.g. depending on the effects that are predicted for other traits. A non-normal distribution of the marker effects may be of little importance if prior knowledge about large QTL is available so that markers that capture their effects have expectations different from zero. These general issues are also important when dominance is present.

References

Bennewitz, J., Meuwissen, T. H. E. (2010). *J. Anim. Breed. Genet.*, **127**:171-179

de Boer, I. J. M., Hoeschele, I. (1992). *Theor. Appl. Genet.*, **86**:245-258

Daetwyler, H. D., Villanuvea, B., Woolliams, J.A. (2008). *PLoS ONE*, **3**:e3395

Falconer, D. S., Mackay, T. F. C. (1996). Introduction to quantitative genetics. London, UK: Longman

Goddard, M. (2008). *Genetica*, **136**:245-257

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

Meuwissen, T. H. E. (2009). *Gen. Sel. Evol.*, **41**:35

Villa-Angulo, R., Matukumalli, L. K., Gill et al. (2009). *BMC Genetics*, **10**:19.