Estimation of Breeding Values for Haploid Chromosomes

M.P.L. Calus[*], H.A. Mulder[*], and R.F. Veerkamp[*]

Introduction

Estimation of breeding values using SNP information is becoming common practice in animal breeding to enable genomic selection. Different methods have been proposed to parameterize the model with SNP or haplotype effects (Calus *et al.* (2008); Meuwissen *et al.* (2001)). The simplest model proposed is where each SNP is assumed to have the same variance, for example by calculating a genomic relationship matrix directly from the SNP data. This model is obviously unrealistic when QTLs with large effect are known. On the other hand, models such as BayesB are proposed, that derive whether a SNP has an effect on the trait of interest, or no effect at all. Another option is to replace the SNPs in the model by haplotypes (Calus *et al.* (2008); Villumsen *et al.* (2009)), so that effectively breeding values for chromosome segments are estimated. Models with individual loci or haplotypes explicitly modeled are difficult to implement in routine mixed model evaluation and software, and therefore typically solved using Gibbs sampling. We investigated an alternative method that estimates breeding values and variances specific per chromosome using mixed model technology, and demonstrate its predictive ability in real data compared to three established models.

Material and methods

Data. The analyses are based on genotypes of 516 cows having records for fat percentage measured as an average across the first 15 wk of lactation. Quality control steps applied to the SNP data included, that the SNP are positioned on one of the 29 autosomes or the X chromosome, a call rate for each SNP of over 90%, a GenCall score >0.2 and a GenTrain score >0.55, a minor allele frequency of >2.5% and a lack of deviation from Hardy Weinberg equilibrium, χ^2<600 (for more details, see Verbyla *et al.* (in press)). After these steps, 41,272 out of 54,001 SNPs remained. FastPHASE (Scheet and Stephens (2006)) was used to phase the genotype data, and to impute missing genotypes.

Models. Breeding values were predicted using four different models. The first two models are described as:

$$y_{ijk} = \mu + age_j + ys_k + animal_i + e_{ijk}$$

where y_{ijk} is the phenotypic record of animal i, μ is the overall mean, age_j is effect of age class j, ys_k is the fixed effect of year-season k, $animal_i$ is the random polygenic effect of

[*] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, P.O. Box 65, 8200AB Lelystad, The Netherlands

animal *i*, and e_{ijk} is a random residual for animal *i*. In the first model (A), the genetic variance was estimated as $\mathbf{A}\sigma_a^2$, where **A** is the additive relationship matrix. In the second model (G), the genetic variance was estimated as $\mathbf{G}\sigma_a^2$, where **G** is the genomic relationship matrix calculated as $\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$ (VanRaden (2008)). ASReml (Gilmour *et al.* (2006)) was used to run the first and second model.

The third model (BayesC) is described as:

$$y_{ijklm} = \mu + age_j + ys_k + \sum_{l=1}^{41,272}\sum_{m=1}^{2} SNP_{ilm} + e_{ijklm}$$

where SNP_{ilm} is a random effect for allele *m* at locus *l* of animal *i.* Gibbs sampling was used to sample the SNP effects from two distributions. One distribution resembles SNPs associated with a QTL; the other distribution resembles SNPs with no association with a QTL. For further details, see Calus *et al.* (2008) and Meuwissen and Goddard (2004).

The fourth model (CHROM) is described as:

$$y_{ijklm} = \mu + age_j + ys_k + \sum_{l=1}^{30}\sum_{m=1}^{2} CHROM_{ilm} + e_{ijklm}$$

where $CHROM_{ilm}$ is a random effect for haploid copy *m* of chromosome *l* of animal *i.* In $CHROM_{ilm}$ the similarity between all phased haploid chromosomes was calculated as follows. First, the formula described by (Eding and Meuwissen (2001)) was used to calculate the similarity for each locus. Those similarities were then averaged across all loci on a chromosome to obtain a chromosome specific similarity. So for each chromosome *l*, a matrix with similarities ($\mathbf{C}_l$) was constructed, where each animal had two entries, one for each chromosome. All matrices $\mathbf{C}_l$ were checked for negative eigenvalues and bended whenever necessary. The variance of chromosome *l* was modeled as $\mathbf{C}_l\sigma_{c_l}^2$. The total genetic variance was calculated as twice the sum of all chromosome variances. Model CHROM was run using ASReml (Gilmour *et al.* (2006)).

All models were first run using all data to estimate the variance components. In addition, CHROM model was also run for each chromosome separately.

Cross-validation. A tenfold cross-validation was performed. Only for model CHROM, the breeding values in the cross-validation were estimated in a BLUP run using the estimated variance components of the full run. In the other models, the variances were re-estimated every time. In each cross-validation, the phenotype of one of every ten animals was omitted, such that all animals had their breeding value predicted once while not having phenotypic information. The accuracy of predicting the phenotypes was calculated as the correlation between adjusted phenotypes and estimated breeding values. Adjusted phenotypes were calculated for animal *i* as $y_{ijk}^{*} = y_{ijk} - a\hat{g}e_j - y\hat{s}_k$.

Results and discussion

Residual variances were similar across models (Table 1). Estimated genetic variance was similar for models A and G, but substantially higher for CHROM. The genetic variance was not calculated from the results of BayesC, since no straightforward procedure is described for that purpose. Estimated genetic variances apply to the base generation. Since the G matrix was calculated using allele frequencies in the current population, it assumes that the current generation is the base generation. Model A, however, places the base generation further back in the past. In the CHROM model, inbreeding is modeled as the similarity between haploid chromosomes within an animal. This value was on average 0.64, implying a high level of inbreeding, and a base generation placed even further back in time. The estimates for the additive genetic variance across the models are in line with the respective order of their base generations. The base generations can be standardized by adjusting the matrices **G** and $\mathbf{C}_l$ but this is not straightforward given the different nature of the models.

Table 1: Heritability of fat percentage, and accuracy of prediction, estimated using different models.

Model	$\hat{\sigma}_e^2$	$\hat{\sigma}_a^2$	h^2	se	Accuracy of prediction	
					Phenotype included	Phenotype excluded
A	0.019	0.148	0.886	0.096	0.996	0.425
G	0.037	0.120	0.764	0.078	0.984	0.463
CHROM	0.027	0.368	0.933	0.028	0.985	0.597
BayesC	0.019				0.998	0.781

The heritability of CHROM was higher than that of A, while the heritability of the model G was lower than that of A (Table 1). The results show that modeling the relationships more precisely, leads to a decrease in the standard error of the heritability. The heritability for model CHROM was calculated for all chromosomes separately, when only one chromosome was fitted or when all chromosomes were fitted simultaneously (Figure 1). The results show that the heritabilities are largely overestimated when only one chromosome was included in the model, i.e. across chromosomes they sum to 10.8. This overestimation was cured by fitting all chromosomes simultaneously, or including a polygenic component in the model. When including a polygenic component, the total heritability had a maximum value of 0.91 (results not shown).

The accuracy of predicting the phenotype for animals with phenotypes included in the analysis was >0.98 for all models (Table 2). The accuracy for animals with phenotypes excluded was lowest for model A, but only slightly higher for model G. The accuracy of A may be relatively high due to the structure of the data; many animals had (half-)sibs, and 121 mother-daughter pairs were present. In the present dataset the additional benefit of G compared to A was therefore limited. The model BayesC showed the highest accuracy for animals with phenotypes excluded, because it captured the effect of DGAT optimally. The accuracy of CHROM was intermediate between models G and BayesC, because it allows to use a chromosome specific variance (Figure 1), but is not as good able as BayesC to separate out the effect of an individual locus with a large effect.

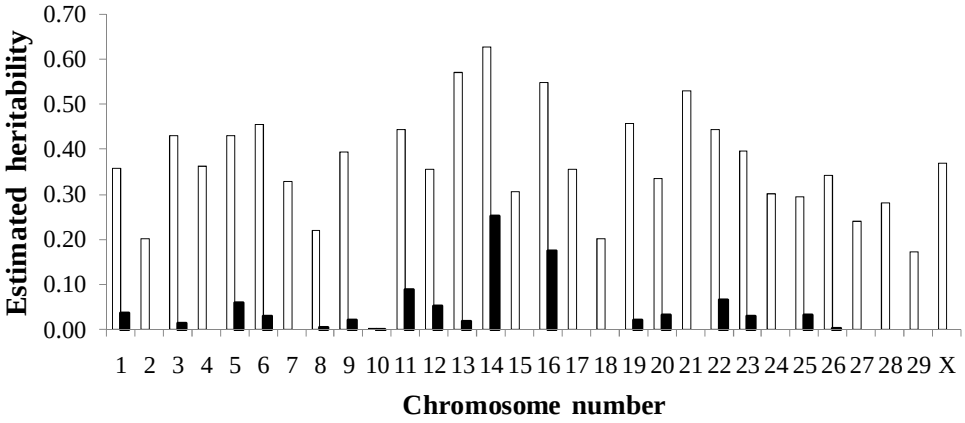

Figure 1: Heritability for fat% for all chromosomes estimated using model CHROM. □ (■) indicates the effects when one (all) chromosome(s) is (are) included in the model.

Conclusion

The presented model CHROM allows straightforward estimation of chromosome specific variance components and breeding values with available software packages. Standardization of base generations to allow proper comparison of estimated variance components across models needs to be resolved. For fat percentage, with one gene with large effect, accuracy of prediction of model CHROM was intermediate to the accuracies of models G and BayesC.

Acknowledgements

Part of this work was carried out as part of the RobustMilk project that is financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. The content of this paper is the sole responsibility of the authors, and it does not necessarily represent the views of the Commission or its services.

References

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

Eding, H. and Meuwissen, T.H.E. (2001). *J. Anim. Breed. Genet.* 118:141-159.

Gilmour, A., Cullis, B., Welham, S. *et al.* 2006. ASReml User Guide (Release 2).

Meuwissen, T.H.E. and Goddard, M.E. (2004). *Genet. Sel. Evol.* 36:261-279.

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

Scheet, P. and Stephens, M. (2006). *Am. J. Hum. Genet.* 78:629-644.

VanRaden, P.M. (2008). *J. Dairy Sci.* 91:4414-4423.

Verbyla, K.L., Calus, M.P.L., Mulder, H.A. *et al.* (in press). *J. Dairy Sci.*

Villumsen, T.M., Janss, L., and Lund, M.S. (2009). *J. Anim. Breed. Genet.* 126:3-13.