STØTTET AF
mælkeafgiftsfonden

Single step evaluations using haplotype segments

M. L. Makgahlela[1, 2], T. Knürr[2], G.P. Aamand[3], I. Strandén[2] and E. A. Mäntysaari[2]

[1]*Department of Agricultural Sciences, University of Helsinki, Finland*
[2]*MTT Agrifood Research Finland, 31600 Jokioinen, Finland*
[3]*NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark*

Abstract

The low reliability of genomic evaluations in some breeds may be improved by regression on ancestral haplotypes instead of all available markers. The aim of this paper was to examine the use of haplotypes in constructing genomic relationships for single step evaluations. BayesB model was fitted to all markers simultaneously to estimate marker effects, using deregressed proofs (DRP) of reference bulls as data. Phased 5-marker haplo-blocks were constructed from the markers with the highest absolute effect size and 4 adjacent markers. Haplo-blocks and their estimated variances were then used to compute genomic relationships. Evaluations used cow DRP of milk and protein as data, weighted by effective record contribution. Estimated genomic breeding values (GEBV) were validated from regression coefficients in a linear model. The validation reliabilities for milk over marker-based methods were improved by up to 4% with 1500 segments and 40% weight on the pedigree. Reliabilities were smaller for protein but comparable when the weight on pedigree was 40%. Inflation levels for both traits were higher with haplo-blocks than individual markers. Haplo-blocks appeared to be beneficial and indicate a need to further assess the optimal number of haplotypes and the weight on pedigree for single step evaluations.

Keywords: haplotype segments, genomic relationships, genomic breeding values

Introduction

Genomic evaluations in many dairy cattle populations are based on regression on a large number of individual markers (Hayes *et al*., 2009, Kearney *et al*., 2009; Reinhardt *et al*., 2009). The procedure for estimating GEBV was demonstrated by VanRaden *et al*. (2009). However, the variance of the quantitative trait loci (QTL) explained, amongst others, depends on 1) the extent of linkage disequilibrium (LD) between single nucleotide polymorphism (SNP) markers and QTL (Daetwyler *et al*., 2008) 2) the additive genetic relationships between individuals (Habier *et al*., 2007). Consequently, the validation reliabilities of genomic evaluations in populations such as the Nordic Red dairy cattle (RDC) have been limited by primarily, insufficient LD due to admixture.

An alternative approach for improving the observed reliabilities may be to construct haplotype segments surrounding the putative QTLs. Genomic selection methods, as originally proposed by Meuwissen *et al*. (2001) were based on haplotype segments of two adjacent multi-allelic markers. Several methods of grouping or constructing haplotypes have since been described (e.g., Hayes *et al*., 2007; Calus *et al*., 2008; Edriss *et al*., 2013). The limitation that has discouraged the uptake of haplo-blocks in evaluations is that many effects need to be estimated depending on the number of segments, which would require more data (Hayes *et al*., 2007). In addition, haplotyping requires phasing of genotypes, which can be performed using free and considerably reliable software. Of particular interest, studies using simulated and real data have shown that evaluations with haplo-blocks are more reliable than individual markers, especially when the marker density is low (Calus *et al*., 2008; de Roos *et al*., 2011). This is because ancestral haplotype segments capture greater LD with QTL than individual markers. Further, if only moderate number of blocks is needed, the use of haplo-blocks may reduce computing requirements for genomic evaluations. Objective of this study is to

examine the use of genomic relationship matrix (**G**) constructed using haplotype segments in single step evaluations applied on the Nordic RDC population.

Materials and Methods

The genotype data contained 38,194 SNP genotypes for 4,727 bulls, born between 1971 and 2008. These data were already edited to remove uninformative SNP. The full RDC pedigree contained 4,873,448 animals. Phenotypes were DRP of 3,633,481 cows for milk and protein. These data were used to solve bull DRP in MiX99 program. The effective daughter contributions (EDC) for all animals in the pedigree were calculated using ApaX program (Strandén *et al*., 2001). To obtain the reduced data (DRP_R), the original DRP data above (DRP_F) were edited to remove 487,033 cows born after 2005. Thus, DRP_R data only included daughters of 4,208 bulls born between 1971 and 2005, which were defined as the reference population.

The construction of haplotype segments

Firstly, we used DRP_R of reference bulls and BayesB evaluation model to estimate marker effects by fitting all SNP simultaneously. Markers were then ranked and pre-selected based on their absolute effect sizes. The SNP data were phased with Beagle 3.3 (Browning and Browning, 2007). Each 5-SNP haplotype segment included a SNP with the highest absolute effect size and 4 adjacent markers. Finally, we estimated the relative variances for those pre-selected segments.

Statistical Analyses

Instead of performance records, cow DRP_R were used as data to compute GEBV for all animals in the RDC pedigree with MIX99. The single-step approach proposed by Christensen and Lund (2010) and Aguilar *et al.* (2010) was as follows:

$$DRP_{R_{cow}} = \mathbf{1}_n\mu + \boldsymbol{Wa} + \boldsymbol{e}$$

where **W** is the design matrix associating animal effects **a** with appropriate observations. We assumed that $\mathbf{e} \sim N(0, \mathbf{R}\sigma_e^2)$ where **R** is a diagonal matrix of $1/t$ and t is the cow's effective record number. It is assumed that $\mathbf{a} \sim N(0, \mathbf{H}\sigma_a^2)$, where σ_a^2 is the genetic variance and **H** is the unified relationship matrix. The variance parameters with h^2 close to 0.50 for both traits were obtained from the national evaluations. Single-step was implemented in MiX99 as described in Mäntysaari *et al*. (2011), where the pedigree file for all animals is read directly into the program, and additionally, covariance structure of the unified relationship matrix for the genotyped individuals (Aguilar *et al.*, 2010), which is given by:

$$\mathbf{H}^{22} - \mathbf{A}^{22} = \mathbf{G}_w^{-1} - \mathbf{A}_{22}^{-1}$$

where $\mathbf{G}_w = (1-w)\mathbf{G}k + w\mathbf{A}_{22}$, $\mathbf{A}_{22}$ is the pedigree sub-matrix for the genotyped bulls and $k = \text{trace}(\mathbf{A}_{22})/\text{trace}(\mathbf{G})$. The haplo-block **G** was defined as **ZDZ**' where $\mathbf{Z}_{i,j}$ is 0, 1, or 2 copies for the j^{th} haplotype and the matrix **D** is a diagonal of haplo-block variances. The haplo-block **G** was compared with the marker-based relationship matrix of VanRaden (2008) computed as $\mathbf{XX}'/2\sum p_j(1-p_j)$ where $\mathbf{X}_{i,j}$ is $0\text{-}2p_j$, $1\text{-}2p_j$ or $2\text{-}2p_j$ with p_j being the frequency of the 2nd allele. Evaluations compared genomic relationships estimated using 750 haplo-blocks (HAP750), 1500 haplo-blocks (HAP1500) and individual markers (ssGLUP). For each method, the value of the weight *w*, which measures the variance explained by **A**, was varied at 0.10, 0.20 and 0.40.

The validation data were defined as bulls born between 2002 and 2008 with EDC>=20. The DRPs of these bulls were solved from the animal model run with DRP_F. The Interbull GEBV validation test presented by Mäntysaari *et al*. (2010) was used to assess GEBV from different models as:

$$DRP_{F_{bull}} = \boldsymbol{b}_o + \boldsymbol{b_1}\hat{\boldsymbol{a}} + \boldsymbol{e}$$

where $DRP_{F_{bull}}$ has the DRP and $\hat{\boldsymbol{a}}$ has the estimated GEBV of the validation bulls. Validation models were weighted by reliability of the bull defined as $r^2_{DRP} = EDC_i/(EDC_i + \lambda)$, where $\lambda = (4 - h^2)/h^2$ and h^2 for both traits were as mentioned above. The validation reliability was derived by correcting the model R^2 by mean reliability of bulls as $R^2_{GEBV} = \frac{r^2_{(DRP,GEBV)}}{\bar{r}^2_{DRP}}$.

Results and Discussion

Due to differences in scale between the haplo-based and marker-based relationship matrices, the estimated genomic relationships could not be compared directly. However, when examining the distributions of diagonal elements of the $\mathbf{G}_w$ matrices with w=0.1, there was no clear distinction between methods (distributions not shown). Alternative means of assessing differences in **G** between haplo-blocks and individual markers will be studied.

In Figure 1, we illustrate the validation reliabilities of GEBV using different weights (%) of **A** in the unified relationship matrix $\mathbf{H}^{22}$. For all the weights tested, the validation reliabilities of GEBV for milk appeared slightly higher than ssGBLUP using 1500 than 750 ancestral haplotypes. For the HAP1500 model, the reliabilities over ssGBLUP increased by 1, 2 and 4%-units when the weight on **A** matrix increased by 10, 20 and 40%, respectively. Although the patterns were similar for protein, the validation reliabilities were slightly higher with ssGBLUP. The reliabilities for ssGLUP for both traits peaked when the weight on **A** was 20% and tended to decline with increasing weight. Our reliabilities were 2 to 8% higher for milk and -5 to 0% for protein compared to those found by Koivula *et al*. (2012) using ssGBLUP with test day data, and higher for both traits with cow DRP data (Mäntysaari *et al*., 2011). The advantage of haplo-blocks over individual markers was shown for genomic evaluations (Calus *et al*., 2008; de Roos *et al*., 2011; Edriss *et al*., 2013), marker-assisted selection (Hayes *et al*., 2007) and QTL mapping (Grapes *et al*., 2006). While goals are not similar, the common underlying feature here was the ability of haplotypes to improve the LD captured. This agreement between studies was irrespective of the method used to construct haplotypes. The tendency of our reliabilities to increase with increasing weight on **A** may be due to use of few pre-selected markers, which explain most but not all the variance of the QTL. The effect of number of markers on a haplotype was not examined here but was also reported to influence the accuracy of predictions (Grapes *et al*., 2006; Hayes *et al*., 2007).

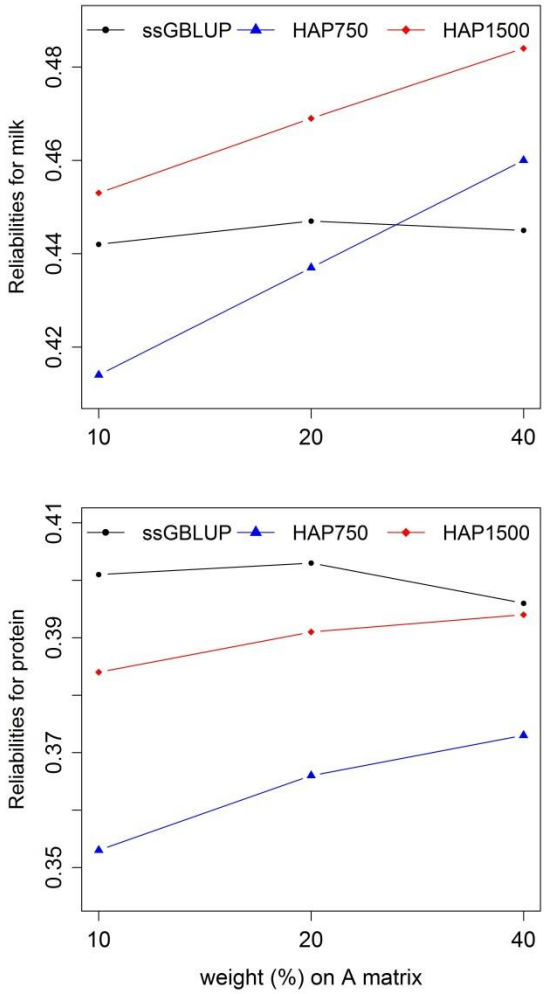

Figure 1. The validation reliabilities for milk and protein from single step with genomic relationship matrix computed as marker-based (ssGBLUP), haplo-block model with 750

(HAP750) and 1500 (HAP1500) segments. The x-axis shows the % weight explained by the pedigree-based matrix (**A**).

The models were also examined on the inflation levels (b_1) of GEBV (Figure 2). The inflation of GEBV was greater with haplo-blocks than ssGBLUP. This was contrary to Edriss *et al*. (2013) who found reduction in bias for some haplo-block models constructed from genealogy information. The difference in b_1 terms between HAP1500 and ssGBLUP was 5 units using 10% weight on **A** but reduced to 1% with 40% weight. This tendency was found for all models and indicates a need for optimal weighting of haplo-blocks and pedigree information for single step evaluations.

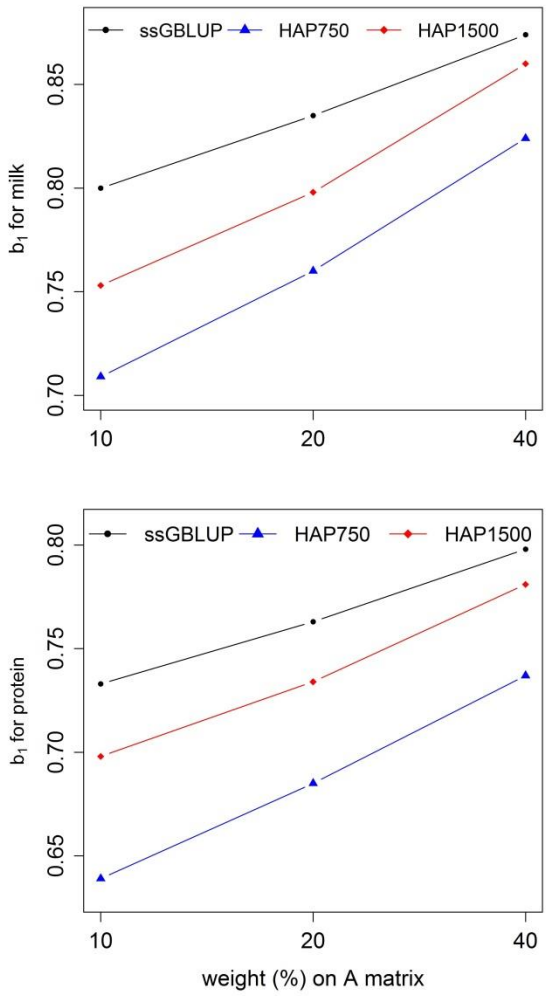

Figure 2. The inflation factors (b_1) for milk and protein from single step with genomic relationship matrix computed as marker-based (ssGBLUP), the haplo-block model with 750 (HAP750) and 1500 (HAP1500) segments. The x-axis shows the % weight explained by the pedigree-based matrix (**A**).

Conclusions

The use of haplotype segments appeared to be beneficial for single step evaluations. For more optimal gain, the results indicate a need for balance between the number of haplotype segments and the weight on pedigree relationships in the construction of a unified relationship matrix. The inflations of GEBV however, were slightly higher with haplo-block but appeared to be improving with more haplotype segments and weight on the pedigree relationship matrix.

References

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J., Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93,743-752.

Browning, S. R., and B. L. Browning. 2009. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084-1097.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178, 553-561.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42, 2.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3, e3395.

de Roos, A. P. W., C. Schrooten, and T. Druet. 2011. Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. J. Dairy Sci. 94, 4708-4714.

Edriss, V., R. L. Fernando, G. Su, M. S. Lund and, B. Guldbrandtsen. 2013. The effect of using genealogy-based haplotypes for genomic prediction. Genet. Sel. Evol. 45, 5

Grapes, L., M. Z. Firat, Dekkers, J.C.M., Rothschild, M. F. and, R. L. Fernando. 2006. Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. Genetics. 172, 1955-1965.

Habier, D., R. L. Fernando and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177, 2389-2397.

Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M.E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res. 89, 215-220.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92, 433-443.

Kearney, F., A. Cromie, and D. P. Berry. 2009. Implementation and uptake of genomic evaluations in Ireland. Interbull Bull. 40, 227-230.

Koivula, M., I. Strandén, J. Pösö, G. P. Aamand, and E. A. Mäntysaari. 2012. Single step genomic evaluations for the Nordic Red dairy cattle test day data. Interbull Bull. 26, 115-120.

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

Mäntysaari, E. A., Z. Liu, and P. VanRaden. 2010. Interbull validation test for genomic evaluations. Interbull Bull. 41, 17–22.

Mäntysaari, E. A., M. Koivula, I. Strandén, J. Pösö, and G. P. Aamand. 2011. Estimation of GEBV using deregressed individual cow breeding values. Interbull Bull. 44, 26-29.

Strandén, I., Lidauer, M., Mäntysaari, E.A. & Pösö, J. 2001. Calculation of Interbull weighting factors for the Finnish test day model. Interbull Bull. 26, 78-81.

Reinhardt, F., Z. Liu, F. Seefried, and G. Thaller. 2009. Implementation of genomic evaluation in German Holsteins. Interbull Bull. 40, 219-226.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91, 4414-4423.

VanRaden, P. M., C. P. Van Tassell, G.R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92, 16-24.