Single Step Genomic Evaluations for the Nordic Red Dairy Cattle Test Day Data

***Minna Koivula*[1]*, Ismo Strandén*[1]*, Jukka Pösö*[2]*, Gert Pedersen Aamand*[3] *and Esa A. Mäntysaari*[1]**

[1] Genetic Research, MTT Agrifood Research Finland, Jokioinen

[2] Faba Co, Vantaa, Finland

[3] NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark

Abstract

Most genomic evaluations are currently based on multi step -approach that requires 1) traditional evaluation with an animal model, 2) extraction of pseudo-observations, and 3) the genomic model to predict direct genomic values (DGV) of candidate animals without own records. In the single step analysis the phenotypic records are combined directly with genomic information, and the resulting genomic enhanced breeding value (GEBV) already combine both sources of information optimally. The objectives of this study were to evaluate the feasibility of the TD single step model using phenotypic records of Nordic RDC cows, and to quantify the accuracy when using single step TD model. The results show that the use of phenotypic test-day records in single step analysis is realistic and easy to implement. Moreover, single step TD models give comparable results to original TD models and considerably higher GEBV validation reliabilities and validation regression coefficients. This indicates that inflation is smaller than with DGVs from sire model validations although it still exists. Thus, it provides a good alternative to the current multi step -approach.

Key words: genomic evaluation, single step, TD model, Nordic RDC

Introduction

Most genomic evaluations are currently based on multi step- approach that requires 1) traditional evaluation with an animal model, 2) extraction of pseudo-observations, and 3) the genomic model that is used to predict direct genomic values (DGV) of candidate animals without own records (Hayes *et al.*, 2009; VanRaden, 2008; VanRaden *et al.*, 2009). The accuracies of genomic predictions can be improved by combining genomic information and information from traditional EBV (e.g. VanRaden *et al.*, 2009) yielding genomic enhanced breeding values (GEBV).

In the long run multi step DGVs and GEBVs has an inherent problem. Firstly, the parent averages (PA) of progeny of genomically selected animals do not automatically include genomic information. Secondly, when the animals are selected by their GEBV, the future estimation of unbiased EBVs becomes difficult.

In the single step analysis the phenotypic records are combined directly with genomic information, and the resulting GEBV already combine both sources of information optimally (Aguilar *et al.*, 2010; Christensen and Lund, 2010; Misztal *et al.*, 2009). This kind of single step approach has been rated computationally demanding with large dataset and multi-trait analysis (Su *et al.*, 2012). However, the single step method has been successfully applied e.g. for final scores of over 6 million Holsteins with higher accuracy compared to a multi-step procedure (Aguilar *et al.*, 2010). Thus, despite of high computational requirements, the single step method is suitable for multiple-trait analyses.

A joint random regression test-day (TD) model is currently used for the official Nordic genetic evaluation of production (Lidauer *et al.,* 2006) and udder health traits (Negussie *et al.,* 2010) in Nordic Red Dairy Cattle (RDC). As more selection decisions are made using genomic information, it is becoming essential that all genomic information is included in national evaluations. The objectives of this study were to evaluate the feasibility of the TD single step model using phenotypic records of Nordic RDC cows, and to estimate the accuracy when using single step TD model.

Materials and Methods

Official evaluation data from March 2012 for the RDC were obtained from the Nordic Cattle Genetic Evaluation (NAV). For the production traits the full TD data included 3,538,966 cows with a total of 95.6 million records and 4,774,687 animals in the Nordic RDC pedigree. The full udder health data had 4,400,436 cows with 77.3 million records and the pedigree included 5,437,876 animals. A reduced data were obtained by deleting four years of observations (data cut from February 2008). Thus EBVs and genomic enhanced breeding values (GEBV) were obtained for all animals in the pedigree with a 1) Full data (EBV_F and $GEBV_F$), and 2) Reduced data (EBV_R = parent average, PA, and $GEBV_R$).

Routine 2011 evaluation models were used in analyses. However, the production TD model was run without heterogeneous variance correction. The GEBVs were obtained from the single step TD evaluation models. The implementation of single step in MiX99 constructs the $\mathbf{A}^{-1}$ matrix directly by reading the pedigree file while iterating on data, and reads the $\mathbf{G}^{-1}$- $\mathbf{A}_{22}^{-1}$ block of the $\mathbf{H}^{-1}$ matrix for the genotyped animals (Aguilar *et al.,* 2010) from a separate file during each PCG iteration cycle. The $\mathbf{A}_{22}$ matrix was a relationship matrix of 5,729 genotyped RDC animals. To form the **G** genotypes for 38,194 SNPs were used. First, the method 1 in VanRaden et al. (2008) was applied. Then the raw **G** was scaled by dividing it by a scalar in order to have on average the same diagonals as $\mathbf{A}_{22}$, and finally the matrix was regressed 20% towards $\mathbf{A}_{22}$ (Christensen and Lund, 2010). The regression can be interpreted as a fraction of genetic variance not explained by SNP genotypes.

The effective daughter contributions (EDC) were printed out from the ApaX99 -program (Strandén *et al.,* 2001) for all the animals in the pedigree. The variance parameters in EDC approximation were for the lactation average TD, and the same values (h^2_{milk}=0.40, $h^2_{protein}$=0.28, and h^2_{fat}=0.32, h^2_{SCC}=0.35, h^2_{CM}=0.11) were used throughout the study. For the validation, the deregression of bull EBVs of lactation averages was done using Secant method in option DeRegress (Strandén and Mäntysaari, 2010) in MiX99 package (Strandén and Lidauer, 1999). Bull's EDC were used as weighting factors. Deregressions used the full pedigree in NAV evaluation and EBV_F for the bulls from full data.

Bulls that were born between years 2003-2007 and had EBV_F based on EDC>20 in the full data, but had only PA information in the reduced data were defined as candidate bulls. Validation reliability of predictions was assessed using Interbull validation protocol (Mäntysaari *et al.,* 2010) with

$$\mathbf{y}=\mathbf{1}b_0 + b_1\,\hat{\mathbf{a}} + \mathbf{e}$$

where **y** are the DRP of the candidate bulls in the full data, and $\hat{\mathbf{a}}$ are the genomic prediction for these bulls from the analysis based on the reduced data ($GEBV_R$). The reliabilities of DRP (r^2_{DRPi} =EDC_i/(EDC_i + λ), λ= (4 - h^2)/h^2) were used as weights. The estimate of b_1 was derived from maximum likelihood (ML) estimates of variance components. This was done by fitting a simple random model in SAS PROC MIXED with the option repeated. In

ML all the animals are expected to have DRPs and for the non-genotyped animals the GEBVs are declared missing.

The validation reliability of the model was obtained from the R^2 of the model, after correcting it by the average reliability of DRPs of the candidate bulls, $R^2_{validation}=R^2_{model}/(r^2DRP_{mean})$. In order to estimate the further gain from the genomic information over the traditional PA (Mäntysaari *et al.*, 2010; VanRaden *et al.*, 2009), the same validation test were also applied to PA.

Results & Discussion

When fixed number of 1,500 iteration rounds of PCG were made,no difference in time and convergence statistics were observed among models. Both TD and single step TD models took about 12 hours to run with 4 Intel Xeon® 3.6 GHz processors , and there were no notable differences in convergence. Thus, the only significant extra computations in the single step method were due to the construction of H-matrix block which was done only once. In the udder health evaluations the models were run into the same strict level of convergence ($\text{sqrt}(C_d)=10^{-4}$). Computing times varied from 17 to 31 hours with one 3.6 GHz Intel Xeon® depending on method and data set. Adding the genomic information to the model increased the number of PCG iterations from 2,199 to 3,752. Use of parallel computing also in the udder health models would have reduced computing time considerably.

Within candidate bulls the correlation between full data EBV_F and full data $GEBV_F$ was almost one (varying from 0.99 in production traits and SCC to 0.96 in CM). The correlation between EBV_F and $GEBV_R$ (varying from 0.51 to 0.70) was clearly higher than between EBV_F and PA (varying from 0.40 to 0.56). Also, the correlation between $GEBV_F$ and $GEBV_R$ was higher than between EBV_F and PA. The PA naturally has moderately high correlation (0.7-0.8) with candidate bull $GEBV_R$. Table 1 gives the correlations between different models in milk and CM. In other studied traits correlations were very similar.

Table 1. Correlations among GEBVs and EBVs in candidate bulls. Above diagonal for milk and below diagonal CM.

	PA	EBV_F	$GEBV_R$	$GEBV_F$
PA	1	0.51	0.80	0.51
EBV_F	0.40	1	0.63	0.99
$GEBV_R$	0.70	0.51	1	0.67
$GEBV_F$	0.40	0.96	0.63	1

Table 2. Interbull GEBV test regression coefficients (b_1) and validation reliabilities (R^2) for Nordic RDC bulls. The PA is the parent average. The r^2_{DRP} is the average reliability of DRPs for candidate bulls.

	PA		$GEBV_R$		r^2_{DRP}
	b_1	R^2	b_1	R^2	
Milk	0.82	0.25	0.88	0.40	0.93
Protein	0.81	0.23	0.90	0.40	0.91
Fat	0.78	0.29	0.85	0.50	0.91
SCC	0.86	0.15	0.87	0.31	0.87
CM	0.77	0.13	0.76	0.27	0.80

The model validation results are presented in Table 2. Single step TD model R^2 for milk, protein and fat $GEBV_R$ were 0.40, 0.40 and 0.50, and 0.31 and 0.27 for SCC and CM. The PA based on the same data gave on average 17 % units lower R^2 for milk, protein and fat, and 15.5 % units lower R^2 for SCC and CM. In all the traits, the b_1 values were lower than the expected value of one, indicating that differences among bulls were over evaluated by GEBV. However, the over dispersion seems to be very similar or even higher with the PA (Table 2). This suggests that GEBVs are less biased than PA.

Validation reliabilities from the current study are higher than the validation reliabilities we obtained for RDC with 2-step approach and with sire model single step genomic evaluations (Su *et al.*, 2012), or from GEB based on animal model single step evaluations (Table 3). Moreover, based on the b_1s, the GEBVs from the phenotypic records seem to be less inflated than DGVs from sire model or GEBVs from animal model deregressions, but still the models in current study would fail the Interbull GEBV validation test.

Table 3. Interbull GEBV test results for milk, protein and fat single step evaluations for Nordic RDC bulls. The PA is the animal model parent average, DGV is direct genomic values from 2-step fit, $GEBV_{SM}$ is GEBVs using single step genomic model with sire DRPs (Koivula et al. 2012), and $GEBV_{AM}$ is GEBVs using single step genomic model with animal model DRPs (Mäntysaari et al. 2011). $GEBV_R$ states the results from current study.

	Milk		Protein		Fat	
	b_1	R^2	b_1	R^2	b_1	R^2
PA_{AM}	0.70	0.22	0.89	0.25	0.80	0.28
DGV	0.76	0.30	0.77	0.31	0.85	0.40
$GEBV_{SM}$	0.69	0.32	0.74	0.35	0.80	0.44
$GEBV_{AM}$	0.72	0.35	0.81	0.38	0.79	0.45
$GEBV_R$	0.88	0.40	0.90	0.40	0.85	0.50

In single step evaluation the GEBV of a bull calf will be mainly based on its' genotype, while GEBV of accurately proven bull will be based on daughters. It is important that the evaluations smoothly move from genomic evaluation to progeny test evaluation. To study this, candidate bulls were divided into different categories according to their EDC. In this comparison also young bulls with lower EDC were accepted as candidates. The EDC categories used were 5< EDC< 40, 40< EDC< 100, 100< EDC< 160, and EDC> 160. Figure 1 show the R^2 and b_1 for the different EDC categories. Interestingly there is no clear trend in b_1 and R^2. Generally it seems that with EDC<40 the b_1 values were close to one or larger than one, and were lower for other EDC categories. For most traits, lowest b_1 and R^2 values were found in the candidate bulls with EDC>160. The values in this category were similar to those presented for whole candidate bull group in Table 2. Clearly, when candidate bulls are analysed as one group, the candidate bulls with biggest EDC affect most to the validation seem to behave differently than other traits. This is presumable caused by too low heritability for the trait assumed in TD evaluations.

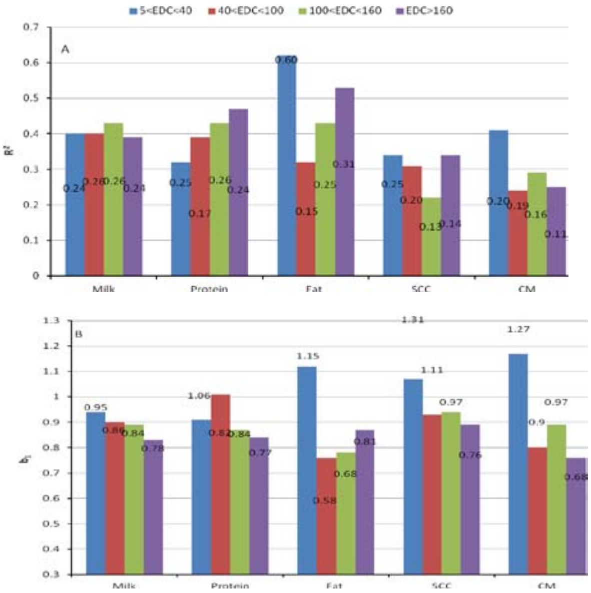

Figure 1. A) Validation reliabilities (R^2) and B) regression coefficients (b_1) by EDC group for candidate bulls. For the comparison the values of R^2 and b_1 for PA are given along the bars.

Comparability of EBV_F and $GEBV_F$ was also assessed by plotting genetic trends for the bulls by year of birth. Figure 2 shows genetic trends for genotyped RDC bulls in A) milk and B) CM. The figure shows that there is no

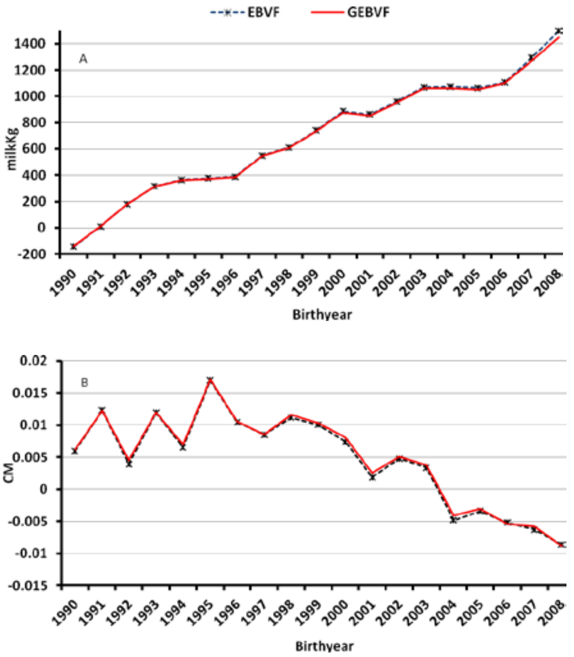

Figure 2. Genetic trends for A) milk and B) CM EBV_F and $GEBV_F$.

difference in the trends from standard TD model and single step TD model. Moreover, when the top 100 bulls with PA or $GEBV_R$ are listed 39 and 49 are the same as in EBV_F top list, respectively. On the other hand, genomic information causes some re-ranking of bulls with daughters as if the bulls were selected based on $GEBV_F$, 93/100 were the same with EBV_F.

The current study shows that single step method is easy to implement straight with the national evaluation model. Phenotypic records are combined directly with genomic information and resulting GEBV directly combines both sources of information. Additional computational costs in the single step approach may be lower than in a two-step G-BLUP approach. Here the number of genotyped bulls was relatively low, and the **G** matrix was easy to invert. For populations with more than 20,000-30,000 genotyped animals, an algorithm suggested by Aguilar *et al.* (2011) can be implemented.

Conclusions

The results show that the use of phenotypic test-day records in single step analysis is feasible. It provides a good alternative to the current multi step approach. The single-step TD model is easy to implement in and it gives comparable results to original models. Moreover use of phenotypic records give higher validation reliabilities compared to earlier validations using sire model or animal model deregressions.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra A. & Tsuruta, S. 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.

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet 128*, 422–428.

Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., du Croz, J., Greenbaum, A., Hammarling, S., McKenney, J.D. & Sorensen, D. 1990. LAPACK: a portable linear algebra library for high-performance computers. *Computer Science Dept. Technical Report CS-90-105*, University of Tennessee, Knoxville, TN, May 1990.

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

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

Koivula, M., Strandén, I., Su, G. & Mäntysaari, E.A. 2012. Different methods to calculate genomic predictions – comparisons of SNP-BLUP, G-BLUP and H-BLUP. *J. Dairy Sci.*, in press

Misztal, I., Legarra, A. & Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci*. *92*, 4648-4655.

Mäntysaari, E., Liu, Z. & VanRaden, P. 2010. Interbull validation test for genomic evaluations. *Interbull Bulletin 40*, 1-5.

Mäntysaari, E.A., Koivula, M., Strandén, I., Pösö, J. & Aamand, G.P. 2011. Estimation of GEBVs using deregressed individual cow breeding values. *Interbull Bulletin 44*, 19-24.

Lidauer, M., Pedersen, J., Pösö, J., Mäntysaari, E.A., Strandén, I., Madsen, P., Nielsen, U.S., Eriksson, J.-Å., Johansson, K. & Aamand, G.P. 2006. *Interbull Bulletin 35*, 103-108.

Negussie, E., Lidauer, M., Mäntysaari, E.A., Strandén, I., Pösö, J., Nielsen, U.S., Johansson, K., Eriksson, J.-Å. & Aamand, G.P. 2010. Combining test day SCS with clinical mastitis and udder type traits: A random regression model for joint genetic evaluation of udder health. *Interbull Bulletin 42*, 25-35.

Strandén, I. & Lidauer, M. 1999. Solving large mixed models using preconditioned conjugate gradient iteration. *J. Dairy Sci*. *82*, 2779-2787.

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

Strandén, I. & Mäntysaari, E.A. 2010. A recipe for multiple trait deregression. *Interbull Bulletin 42*, 21-24.

Su, G., Madsen, P., Nielsen, U.S., Mäntysaari, E.A., Aamand, G.P., Christensen, O.F. & Lund, M.S. 2012. Genomic prediction for the Nordic Red Cattle using one-step and selection index blending approaches. *J. Dairy Sci. 95*, 909–917.

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

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