Effect of Cow Reference Group on Validation Accuracy of Genomic Evaluation

M. Koivula*, I. Strandén*, G. P. Aamand# and E. A. Mäntysaari*
*Genetic Research, MTT Agrifood Research Finland, Jokioinen
# NAV Nordic Cattle Genetic Evaluation, Aarhus, Denmark

ABSTRACT: We studied the effect of including genotyped cows in the reference population of the Nordic Red Dairy Cattle on the validation accuracy of genomic breeding values. Deregressed individual cow EBVs (DRP) were used in single-step genomic evaluations. The accuracy of evaluations was calculated after including 0 or 3,111 or 5,593 genotyped cows in the reference population. All evaluations used 4,188 genotyped bulls in the reference. The gain in accuracy was as less than expected, varying from 0.8% to 2.6%-units for the production traits. Still, genotyping cows and subsequent inclusion in the reference population is advantageous and should be increased.

Keywords: Genomic selection; GEBV; single-step GBLUP

Introduction

In genomic evaluation, the reference population consists of individuals with both genotypes and performance records. Accurate genomic evaluations require large reference populations with reliably estimated breeding values (EBV) (Goddard and Hayes (2009)): the larger the reference population the more reliable the genomic evaluations. At beginning, the reference populations consisted only of progeny-tested bulls, and genomic evaluations were based only on averaged performances of bull's daughters. By including genotyped cows in the reference population, the size of reference group could be easily increased (Dassonville et al. (2012), Babts et al. (2012)). For example, in USA cow evaluations have been included in US genomic evaluations since their beginning (Wiggans et al. (2011)).

In the DFS countries (Finland, Denmark and Sweden) the validation accuracies for the genomic evaluations of Red Dairy Cattle (RDC) and Jersey have not been as high as the genomic evaluations of Holstein breed. The reason has been suggested to be the larger effective population size (Goddard (2009)) but naturally smaller populations have not as many accurately evaluated bulls to be included into reference population either. To overcome the problem, the DFS breeding and AI companies have started a cow genotyping project, where a low cost low density chip is offered for the breeders in aim for voluntary genotyping of entire herds.

Inclusion of cow genotypes and phenotypes in the single-step genomic evaluations is straightforward. The single-step genomic evaluation (Aguilar et al. (2010) and Christensen and Lund (2010)) does not divide the population into training group (reference population) and prediction group (validation population), but instead the genomic data is included along the phenotypic data and pedigree relationships information. However, the estimation of benefits from including daughter genotype data into single-step evaluation is more complicated. If the genotyped cows are youngest age class, they cannot be included into evaluation, unless their contemporaries, including the daughters of validation bulls are also included. However, a single-step evaluation can be also computed using a data with animal model deregressed genetic evaluations. This gives a possibility to include records to genotyped females that are of the same age as the validation bulls.

The aim of this paper was to study how much the inclusion of cow genotypes into single-step evaluation based on individual cow deregressed genetic evaluations can improve the accuracy of genomic breeding values.

Materials and Methods

Data. Genotype data contained 46,943 SNPs for 12,928 genotyped Nordic Red Dairy animals (5,467 bulls and 7,461 cows). EBVs and effective record contributions (ERC) of the Nordic NAV evaluations from the January 2014 for milk, protein, and fat were used to represent production trait evaluations. ERCs and deregressed EBVs (DRPs) were obtained for all the 3.7 million RDC cows with records. The variance parameters in ERC approximation were from the average daily TD, and the same values (h^2_{milk}=0.48, $h^2_{protein}$=0.48, and h^2_{fat}=0.49) were used throughout the study. Deregression used the full pedigree of 5.1 million animals in the NAV evaluation. For the genomic evaluations and validation, three different reduced data sets were created from the full cow DRP data. From the DATA-0, DRPs of young genotyped cows and genotyped bull dams were excluded. The dataset contained ~2.8 million cow DRPs. DATA-I had all data in DATA-0 and included DRPs of 3,111 (young) genotyped cows, but not genotyped bull dams. DATA-II was like DATA-0 but in addition had DRPs of 5,593 genotyped cows, including genotyped bull dams (n=45).

Validation bulls were chosen from genotyped bulls born 2005–2010 and having ERC>= 3, corresponding more than 20 daughters in the full cow DRP data. Finally there were 926 validation test bulls. All evaluations used 4,188 genotyped bulls in the reference. Records of daughters of validation bulls were removed from all test data sets. Moreover, non-genotyped daughters of reference bulls born after 2008 were excluded from the data sets. The total number of daughters removed was 319,257.

Model. Single-step approach (ssGBLUP) following, e.g., Aguilar et al. (2010) and Christensen and Lund (2010), was based on model:

$$\mathbf{y}_t = 1\mu + \mathbf{Wa} + \mathbf{e},$$

where $\mathbf{y}_t$ is a vector of DRP of all cows, $\mathbf{a}$ is the vector of additive genetic effects, and $\mathbf{W}$ is incidence matrix relating the breeding values **a** to corresponding observations $\mathbf{y}_t$. The variance var(**a**)=$\mathbf{H}\boldsymbol{\sigma}_a^2$ and var(**e**)=$\mathbf{D}^{-1}\boldsymbol{\sigma}_e^2$, where the diagonal matrix **D** consists of ERC of the animals. For the **H** -matrix, the variance-covariance structure of genotyped animals is the genomic relationship **G** and the relationships of non-genotyped animals are "corrected" with respect to differences in genomic and pedigree based relationships of their genotyped relatives. The inverse of the **H** –matrix is simply

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_w^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where $\mathbf{A}_{22}$ is the sub-matrix of pedigree based numerator relationship matrix **A** for the 12,928 genotyped animals, and the relationship matrix $\mathbf{G}_w$= w**G** + (1-w)$\mathbf{A}_{22}$ is constructed using genomic and pedigree information. The **G** was scaled by dividing it by a scalar in order to have on average the same diagonals as $\mathbf{A}_{22}$ before the matrices **G** and $\mathbf{A}_{22}$ were combined. The constant w is proportion of polygenic effect not accounted by the SNPs. When the MME for single-step is considered, the difference to normal animal model is the matrix block **B**= $\mathbf{G}_w^{-1}$ – $\mathbf{A}_{22}^{-1}$. To improve the properties of the ssGBLUP, we used **B**= $\tau\mathbf{G}_w^{-1}$ – $\omega\mathbf{A}_{22}^{-1}$ (Tsuruta et al. (2011), (2013)). The parameters were w=0.10, τ=1.6 and ω=0.5. These weights for the type of information were found to give least inflation of variance for genomic predictions. Two different **B**-matrices were constructed. One with all genotyped animals included in the **G** (allG), the second with only bull genotypes included in the **G** (bullG).

GEBVs were validated with Interbull GEBV test (Mäntysaari et al. (2010)).

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

where **y** is the daughter yield deviations (DYD) of the test bulls in the full data, and $\hat{\mathbf{a}}$ is the genomic predictions for these bulls from the analysis based on the reduced data. The reliabilities of DYD (r^2_{DYDi} =$EDC_i/(EDC_i + \lambda)$) were used as weights. Effective daughter contributions (EDC) were calculated using Interbull recommendations. The validation reliability of the model was obtained from the R^2 (coefficient of determination) of the model, after correcting it by the average reliability of DYDs of the test bulls i.e. .

$$R^2_{validation} = R^2_{model} / \overline{r^2_{DYD}}$$

First the GEBVs for validation bulls were calculated using DATA-0, I or II. Second, the GEBVs of the validation bulls were used to predict the DYDs of the bulls calculated from the full cow DRP data with the animal model.

Expected reliability. The accuracy of genomic evaluations is known to be related to the size of the reference group, heritability, effective population size and genotyping density. To address the theoretical expectation of accuracy we applied equation D2 in Erbe et al. (2013), as $r^2_{est} = w^2\ N_h/(N_h+M_e)$, with $N_h=N_{bull}r^2_{DRP}+N_{cow}h^2$. With the Holstein milk data Erbe et al. (2013) estimated w=0.875 and M_e=1045, but we assumed RDC to have about 2 times larger effective population size, and, thus, used M_e=2090.

Results and Discussion

For production traits the improvement in R^2 due to 3,111 genotyped reference cows was from 0.5 to 1.4 %-units (Table 1). When 2,482 more genotyped cows were included into the reference population (total 5,593 cows), increase in R^2 was 2.6 %-units for milk, 0.8 %-units for protein and 1.3 % -units for fat. The expected reliability for milk GEBV was approximated to be 0.50, 0.55 and 0.58 GEBV with bulls only, with bulls and 3,111 cows, and with including all 5,593 cows, respectively. The accuracy increase predicted equation by Erbe et al. (2013) would suggest that 5,593 cows would add same information as 2,738 bulls, but this was not realized. Still, the results indicate that, genotyping cows and subsequent inclusion in the reference population was advantageous and is expected to further increase the accuracies.

Table 1. Bull validation results from different DRP data sets. DATA-0=0, DATA-I=3,111 and DATA-II=5,593 genotyped cows in the reference population. Regression coefficients (b_1) and validation reliabilities (R^2) from the conventional parent averages (PA), GEBVs from ssGBLUP with the all genotypes ($GEBV_{allG}$) or only bull genotypes ($GEBV_{bullG}$) included in G matrix.

	Milk		Protein		Fat	
	b_1	R^2	b_1	R^2	b_1	R^2
Data 0						
PA	0.896	0.350	0.788	0.270	0.752	0.275
$GEBV_{allG}$	0.968	0.475	0.870	0.405	0.896	0.447
$GEBV_{BullG}$	0.918	0.471	0.830	0.406	0.868	0.451
Data I						
PA	0.868	0.343	0.74	0.266	0.684	0.265
$GEBV_{allG}$	0.948	0.489	0.836	0.410	0.859	0.456
$GEBV_{BullG}$	0.902	0.464	0.806	0.398	0.840	0.448
Data II						
PA	0.868	0.345	0.744	0.267	0.684	0.265
$GEBV_{allG}$	0.954	0.501	0.840	0.413	0.862	0.460
$GEBV_{BullG}$	0.902	0.466	0.804	0.399	0.838	0.448

While the effect of number of genotyped cows was consistent and positive for the reliability of GEBV, the variance inflation b_1 did not increase same way. Inclusion of genotyped cows seemed to create more bias, but the bias was less when all the cows (DATA-II) were included. It was not quantified if this could be because of inclusion of bull dam records, but this seems unlikely because only 45 bull dams had genotypes, and none of them were dams of validation bulls.

Wiggans et al. (2011) and Dassonville et al. (2012) found that, the inclusion of cow genotypes can result in a decrease in the reliability of bull genomic evaluations. The reason for this decrease was assumed to be in pre-selection of cows, since cows selected for genotyping are based on their high genetic merit or potential for a high genetic evaluation. Thus, potential bull dams have been the first cows to be genotyped. However, in this study we were unable to see any difference in the accuracy when the genotyped bull dams were excluded from the analyses.

If only bull genotypes were used in the block **B**-matrix, validation results of bulls did not gain from inclusion of DRPs of genotyped cows (Table 1). Thus, it seems that if DRPs of genotyped cows are included in the analyses, it is better to in also include the genotypes.

The trends in GEBVs (Figure 1) show no difference whether DRPs of genotyped cows are included in the data or not. Thus, including information of genotyped cows seems not to give rise to any problems in trends either.

Current study was based on cow individual deregressed genetic evaluations. However, this was done only to evaluate the value of cow genotype data. In true single step evaluations, the genotyped cows can be included along with all their contemporaries, and the gain from the information is most likely larger.

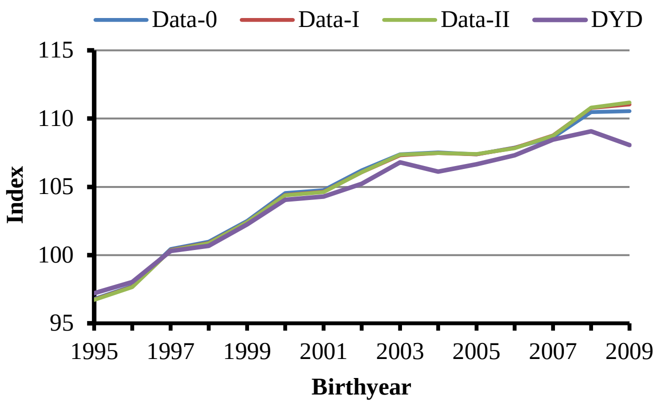

Figure 1. Trendlines for milk GEBVs from ssGBLUP using cow DRPs. Data 0=0, Data I=3,111 and Data II=5,593 genotyped cows in the reference population. DYD=bull daughter yield deviation

Conclusion

We observed consistent increase in GEBV reliability after inclusion of genotyped cows with their records in ssGBLUP reference population, although the gain was less than expected in theory. The number of cows probably is still small to get higher gain for validation accuracy. However, genotyping cows and subsequent inclusion in the reference population is advantageous and number of genotyped cows should be increased.

Literature Cited

Aguilar, I., Misztal, I., Johnson, D.L. et al. (2010). J. Dairy Sci. 93:743-752.

Bapst, B, Baes, C., Seefried, F. R. et al. (2012). Interbull Bull. 47:187-191.

Christensen, O.F. and Lund, M.S. (2010). Genet. Sel. Evol. 42:2.

Dassonneville, R., Baur, A., Fritz, S. et al. (2012). Genet. Sel. Evol. 44:40.

Goddard, M.E. (2009). Genetica 136:245-257.

Goddard, M.E. and Hayes, B.J. (2009). Nat. Rev. Genet. 10:6, 381- 391.

Erbe M, Gredler B, Seefried FR et al. (2013). PLoS ONE 8(12): e81046. doi:10.1371/journal.pone.0081046

Mäntysaari, E.A., Liu, Z. and VanRaden, P. (2010). Interbull Bull. 40: 1-5.

Tsuruta, S., I. Misztal, I. Aguilar et al. (2011). J. Dairy Sci. 94:4198–4204.

Tsuruta, S., I. Misztal, I. and Lawlor, T.J. (2013). J. Dairy Sci. 96 :3332–3335.

Wiggans, G.R., Cooper, T.A., VanRaden, P.M. et al. (2011). J. Dairy Sci. 94, 6188–6193.