J Appl Genetics (2011) 52:363–366
DOI 10.1007/s13353-011-0047-z

ANIMAL GENETICS · SHORT COMMUNICATION

Fitting and validating the genomic evaluation model to Polish Holstein-Friesian cattle

Joanna Szyda · Andrzej Żarnecki · Tomasz Suchocki · Stanisław Kamiński

Received: 30 September 2010 / Revised: 5 April 2011 / Accepted: 6 April 2011 / Published online: 7 May 2011

Abstract The aim of the study was to fit the genomic evaluation model to Polish Holstein-Friesian dairy cattle. A training data set for the estimation of additive effects of single nucleotide polymorphisms (SNPs) consisted of 1227 Polish Holstein-Friesian bulls. Genotypes were obtained by the use of Illumina BovineSNP50 Genotyping BeadChip. Altogether 29 traits were considered: milk-, fat- and protein- yields, somatic cell score, four female fertility traits, and 21 traits describing conformation. The prediction of direct genomic values was based on a mixed model containing deregressed national proofs as a dependent variable and random SNP effects as independent variables. The correlations between direct genomic values and conventional estimated breeding values estimated for the whole data set were overall very high and varied between 0.98 for production traits and 0.78 for non return rates for cows. For the validation data set of 232 bulls the corresponding correlations were 0.38 for milk-, 0.37 for protein-, and 0.32 for fat yields, while the correlations between genomic enhanced breeding values and conventional estimated breeding values for the four traits were: 0.43, 0.44, 0.31, and 0.35. This model was able to pass the interbull validation criteria for genomic selection, which indicates that it is realistic to implement genomic selection in Polish Holstein-Friesian cattle.

Keywords Dairy cattle · Genomic selection · Model validation · Single nucleotide polymorphism

Electronic supplementary material The online version of this article (doi:10.1007/s13353-011-0047-z) contains supplementary material, which is available to authorized users.

J. Szyda (✉) · T. Suchocki
Department of Animal Genetics,
Wrocław University of Environmental and Life Sciences,
Kożuchowska 7,
51–631 Wrocław, Poland
e-mail: joanna.szyda@up.wroc.pl

A. Żarnecki
Institute of Animal Breeding and Genetics,
National Research Institute of Animal Production,
Krakowska 1,
32–083 Kraków, Poland

S. Kamiński
Department of Animal Genetics,
University of Warmia and Mazury,
Oczapowskiego 5,
10–718 Olsztyn, Poland

Recently many countries have incorporated the genomic information, in a form of thousands of single nucleotide polymorphism (SNP) genotypes originating from a microarray technology, into their genetic evaluation systems (Hayes et al. 2009, VanRaden, 2008). It has become evident that the genomic information is now an important part of a routine evaluation of genetic merit in dairy cattle (Liu, 2010). In this paper we describe the results of fitting and validating the genomic selection model to the population of Polish Holstein-Friesian dairy cattle.

The data set used as a training data set for the estimation of additive effects of SNPs consisted of 1227 Polish Holstein-Friesian bulls. The selection of bulls for genotyping was based on two major criteria: on the accuracy of their conventionally estimated breeding values and on the representativeness, in terms of genetic merit, of the selected bulls for the population of all dairy bulls active in Poland. The first criterion was quantified through the number of the effective daughter contribution (EDC) associated with the estimated breeding value (EBV) for milk yield of each bull. Traits were represented by EBVs, which were deregressed

using the method of Jairath et al. (1998) based on the national proofs corresponding to the release from February 2010. Altogether 29 traits were considered, comprising three production traits and a somatic cell score - originating from a random regression test day model as well as four female fertility traits and 21 traits describing type and conformation - originating from an animal model. The traits are listed in online resource 1. Genotypes were generated by the use of Illumina BovineSNP50 Genotyping BeadChip, which consists of 54001 SNPs. The applied SNP selection criteria comprised polymorphism, expressed by the minor allele frequency (MAF), with the minimum MAF of 0.01, and technical quality of a SNP, expressed by the minimum call rate of 90% within the analyzed sample of bulls. Average call rate obtained for our data was high and amounted to 99.66% and 99.75% for all SNPs and for selected SNPs, respectively. For DGV estimation 46267 SNPs were selected, yielding 56502470 bull-SNP genotypes in total for milk yield. For the other traits the total number of bull-SNP genotypes was lower since not all of the genotyped bulls had EBVs available.

The following mixed model was used to estimate the additive effects of the selected N_{snp}=46267 SNPs for up to N_a=1227 bulls with genotypes: $\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$, where $\mathbf{y}$ [N_a] represents a vector of deregressed EBVs (dEBVs), $\mathbf{X}$ is a [N_axN_b] design matrix for fixed effects, $\mathbf{b}$ [N_b] is a vector of N_b fixed effects, which in the current model comprise only a general mean (N_b=1), $\mathbf{Z}$ is a [N_axN_{snp}] design matrix for SNP genotypes, which is parameterized as −1, 0, or 1 for a homozygous, a heterozygous, and an alternative homozygous SNP genotype respectively, $\mathbf{g}$ is a [N_{snp}] vector of random additive SNP effects, and $\mathbf{e}$ is a [N_a] vector of residuals with $\mathbf{e} \sim N\left(0, \mathbf{D}\widehat{\sigma}_e^2\right)$ with $\mathbf{D}$ being a diagonal matrix containing the reciprocal of EDC on the diagonal. The covariance structure of $\mathbf{g}$ was assumed to be $\mathbf{g} \sim N\left(0, \mathbf{I}\frac{\widehat{\sigma}_a^2}{N_{snp}}\right)$, with $\mathbf{I}$ being an identity matrix and $\widehat{\sigma}_a^2$ representing the additive genetic variance of a given trait. The estimation of parameters of the above models was based on solving the mixed model equations: $\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$ (Henderson, 1984), with $\mathbf{R}$ represented by $\mathbf{D}\widehat{\sigma}_e^2$ and $\mathbf{G}$ represented by $\frac{\widehat{\sigma}_a^2}{N_{snp}}$. The iteration on data technique was based on Gauss-Seidel algorithm with residuals update (Legarra and Misztal, 2008). Consequently, the variance of $\mathbf{y}$ is given by $\mathbf{ZGZ}^{\mathrm{T}}+\mathbf{R}$. Note, that the additive genetic variance component $\left(\widehat{\sigma}_a^2\right)$ of this model was not estimated, but was assumed as known, based on the estimates used in the Polish national genetic evaluation model for a corresponding trait.

DGV is defined as the sum of additive effects of SNPs estimated from the above model: $\hat{\mathbf{a}} = \mathbf{X}\hat{\mathbf{b}} + \mathbf{Z}\hat{\mathbf{g}}$. The genomic enhanced breeding values (GEBV) were calculated as a combination of genomic information coming through DGV and the parental information coming through the parent average (PA) using a selection index approach: $GEBV = [REL_{DGV} \quad REL_{PA}] \begin{bmatrix} REL_{DGV} & REL_{DGV}REL_{PA} \\ REL_{DGV}REL_{PA} & REL_{PA} \end{bmatrix}^{-1} \begin{bmatrix} DGV \\ PA \end{bmatrix}$, where REL_{DGV} is reliability of individual's DGV, calculated as explained below, and REL_{PA} is individual's PA reliability originating from the national genetic evaluation. The reliability of DGV was estimated following the approach of Strandén and Garrick (2009), based on the following model: $\mathbf{y}=\mathbf{Xb}+\mathbf{Z}^*\mathbf{a}+\mathbf{e}$, where, $\mathbf{Z}^*$ represents a design matrix for **DGV** - a [N_a] vector of random direct genomic value effects for bulls distributed as $\sim N\left(0, \mathbf{A}_g\widehat{\sigma}_a^2\right)$ with $\boldsymbol{A}_g$ defined as $\mathbf{ZZ}^{\mathrm{T}}\frac{1}{p_{het}^b}$, with p_{het}^b representing the sum over all SNPs of heterozygous genotype frequencies in the base population estimated following (VanRaden, 2008). The reliabilities of bulls' DGVs are given by: $\mathbf{REL}_{\mathrm{DGV}} = diag\left\{\left(\mathbf{A}_g - \frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_a^2}\mathbf{C}^{22}\right)\mathbf{A}_g^{-1}\right\}$, where $\mathbf{C}^{22}$ represents the inverse of the coefficient matrix from the MME corresponding to **DGV**: $\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}^* \\ \mathbf{Z}^{*T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{*T}\mathbf{R}^{-1}\mathbf{Z}^* + \mathbf{A}_g^{-1}\frac{\widehat{\sigma}_e^2}{\widehat{\sigma}_a^2} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}$.

Descriptive statistics regarding the analyzed traits were summarized in online resource 1, which shows that for each of the analyzed traits DGV had similar, but somewhat lower standard deviations than EBV, which was expected since EBV were used as a dependent variable in the SNP effect estimation model. For the training data set estimated correlations between EVB and DGV, were very high and varied between 0.98 for milk yield, 0.78 and 0.81 for non return rates at 56 days for cows and heifers, respectively - traits with the lowest heritability of 0.02.

The highest positive correlations between SNP estimates were observed for interval from calving to first insemination and days open (0.89), size and stature (0.80), as well as between milk and protein yields (0.76), the negative correlations were highest between overall feet and leg score and real leg set (−0.35), between body depth and udder depth (−0.26) and between rear leg rear view and rear leg set (−0.21). Most of the values (except the correlation between body depth and udder depth) well correspond with the estimates obtained for the Polish Holstein-Friesian breed based on conventional, multivariate models (Żarnecki et al., 2003). Manhattan plots of SNP effect estimates for milk and fat yields along the genome were presented in online resource 2. In order to enable comparison of SNP effects, their estimates were transformed to a standard normal distribution and were presented as absolute values.

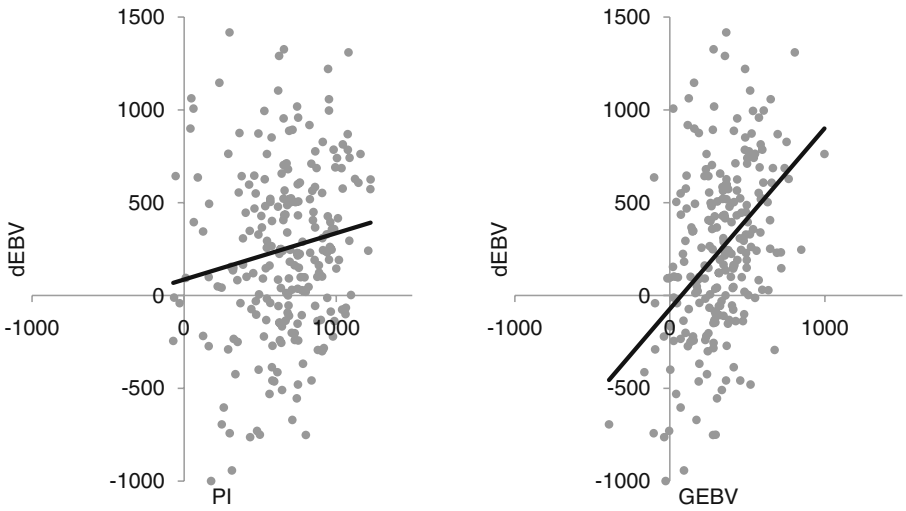

Fig. 1 Predictive ability for PA and GEBV expressed as a linear regression for 232 bulls from the validation data set

The highest SNP estimate for milk yield amounted to 3.67 kg, for fat yield 0.20 kg, and 0.0002 day for non return rate at 56 days of heifers. The main goal of genetic evaluation is not to identify particular loci with considerable effects on a trait, but to assess the sum of all possible additive effects across the genome. However, from the geneticists' perspective, a closer examination of effects if particular SNPs and their links to bovine genomic features are of great interest. Estimates of the effect of SNP on milk and fat yield on BTA14 in a proximity of DGAT1 - a gene having very strong effect on both traits (Grisart et al., 2002) were shown on online resource 3. Our result confirmed that DGAT1 locus has a large effect on milk and fat yields and provides empirical evidence of the validity of SNP effect estimation procedure.

In order to formally validate the genomic selection model the procedure recommended by Interbull (Mäntysaari et al. 2010) was followed. For this purpose the original, training data set was partitioned into an estimation data set consisting of older bulls and a validation data set consisting of younger bulls. The validation data set consisted of 232 bulls, while the remaining 984 bulls were used for the estimation of SNP effects. Validation was done for milk, fat and protein yields. The linear regression coefficients for regression of dEBV on PA and GEBV for the three traits were summarised in online resource 4. In general, models involving PA had much lower slopes than models using GEBV as an independent variable, indicating that the latter models had better predictive ability (Fig. 1). The best prediction, indicated by the slope of 0.96 which is closest to the expected value of 1.00, was estimated for regression of dEBV on GEBV for milk yield, and the worst, with a slope of 0.26 was obtained for regression of dEBV on PA for fat yield. The correlations with EBV (Table 1) were lowest for PA (from 0.14 to 0.26), middle for DGV (from 0.32 to 0.38), and generally the highest when both sources of information were combined into GEBV (from 0.31 to 43). One exception was fat yield, for which the highest correlation was obtained using DGV.

Many simulated as well as real data sets have been analysed in order to compare predictive ability of various models used for the estimation of SNP effects (Clark et al. 2010; Konstantinov and Hayes 2010; Mrode et al. 2010; Shepherd et al. 2010). Summarising the results of those studies one can conclude that no marked differences in predictive abilities can be observed between models. Instead factors related to the trait genetic background (heritability, number of loci with large effects) as well as the structure of the training data set play a key role in determining correlations between the predicted and true genetic merits (Calus, 2010). Results obtained in our study clearly show that a much better accuracy of prediction for

Table 1 Pearson correlation coefficients between EBV from 2010 and PA/DGV/GEBV together with the reliability of DGV and GEBV, calculated based on daughter information from 2004 for the validation data set. N_v is the number of bulls in the validation data set

		Correlation with EBV_{2010}			Reliability	
Trait	N_v	DGV_{2004}	PA_{2004}	$GEBV_{2004}$	DGV	GEBV
Milk yield	232	0.38	0.16	0.43	0.16	0.20
Protein yield	231	0.37	0.26	0.44	0.15	0.21
Fat yield	231	0.32	0.14	0.31	0.12	0.11

selection candidates can be achieved by using a combined information from SNP genotypes (through DGV) and parental EBVs (through PA) instead of the conventional approach based entirely on the EBVs of ancestors.

In our study a low reliability of DGV was obtained for the young selection candidates. It is much lower than values reported for production traits by Hayes et al. (2009), Lund and Su (2009), and VanRaden et al. (2009), which vary between 0.45 and 0.73. The main reason for low values obtained in our study was, as indicated by Hayes et al. (2009) and Habier et al. (2010), a relatively small training data set and corresponding low genetic relatedness between the training and the selection candidate data sets (only 59% of bulls from the validation data set had sires in a training data set). Still, the obtained accuracy of DGV and GEBV was much higher than the accuracy of PA. Moreover, based on the results for protein yield, the predictive ability of the genomic model described here was positively validated by the International Bull Evaluation Service (Interbull and International Bull Evaluation 2010) in August 2010. Consequently, the model presented in this study has been recognised within European Union states by the Directorate of Animal Health and Welfare of the European Commission as a valid procedure for genomic evaluation.

Acknowledgements The project was carried out within the framework of MASinBULL consortium which is supported financially by the Animal Breeding and Insemination Center in Bydgoszcz, Poland.

References

Calus MPL (2010) Genomic breeding value prediction: methods and procedures. Animal 4:157–164

Clark SA, Hickey JM, van der Werf JHJ (2010) How Would Different Models of Genetic Variation Affect Genomic Selection? Proceedings of the 9th WCGALP, Leipzig, Germany

Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelman R, Georges M, Snell R (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res 12:222–231

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42:5

Henderson CR (1984) Applications of Linear Models in Animal Breeding, University of Guelph

Interbull, International Bull Evaluation Service (2010) http://www.interbull.org

Jairath L, Dekkers JCM, Schaeffer LR, Liu Z, Burnside EB, Kolstad B (1998) Genetic Evaluation for Herd Life in Canada. J Dairy Sci 81:550–562

Konstantinov KV, Hayes BJ (2010) Comparison of BLUP and Reproducing kernel Hilbert spaces methods for genomic prediction of breeding values in Australian Holstein Friesian cattle. Proceedings of the 9th WCGALP, Leipzig, Germany

Legarra A, Misztal I (2008) Technical Note: Computing Strategies in Genome-Wide Selection. J Dairy Sci 91:360–366

Liu Z (2010) Dairy cattle genetic evaluation enhanced with genomic information. Proceedings of the 9th WCGALP, Leipzig, Germany

Lund MS, Su G (2009) Genomic selection in the Nordic countries. Interbull 39:29–42

Mäntysaari E, Liu Z, VanRaden P (2010) Interbull Validation Test for Genomic Evaluations. Interbull, Bulletin, 41

Mrode R, Coffey MP, Strandén I, Meuwissen THE, van Kaam JBCHM, Kearney JF, Berry DP (2010) A Comparison Of Various Methods For The Computation Of Genomic Breeding Values Of Dairy Bulls Using Software At Genomicselection.net. Proceedings of the 9th WCGALP, Leipzig, Germany

Shepherd R. Meuwissen THE, Woolliams J (2010) A Fast EM Algorithm For Genomic Selection. Proceedings of the 9th WCGALP, Leipzig, Germany

Strandén I, Garrick DJ (2009) Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92:2971–2975

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited Review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24

Żarnecki A, Morek-Kopeć M, Jagusiak W (2003) Genetic parameters of linearly scored conformation traits of Polish Black-and-White cows. J Anim Feed Sci 12:689–696