

*Short Report*

## **Dissecting an interplay between genomic and pedigree sources of information to estimate breeding values for milk yield in Polish Holstein-Friesian dairy cattle in a one-step approach based on a random regression test day model**

**Tomasz Suchocki<sup>1,2\*</sup>, Zengting Liu<sup>3</sup>, Andrzej Żarnecki<sup>2</sup>, Joanna Szyda<sup>1,2</sup>**

<sup>1</sup> Wrocław University of Environmental and Life Sciences, Institute of Genetics, Biostatistic Group, Kozuchowska 7, 51-631 Wrocław, Poland

<sup>2</sup> National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

<sup>3</sup> vit w.V ., Heinrich-Schroeder Weg 1, 27-283 Verden/Aller, Germany

*(Accepted February 17, 2017)*

**In the future an approach incorporating cows' measured phenotypes and marker genotypes of cows and bulls within a single model can be applied. The most important advantage of such a model is the simultaneous use of pedigree and marker-based genomic relationship data. Such a solution allows the use of both genotyped and non-genotyped animals in the prediction procedure. This pilot study is aimed towards implementation of a one-step approach in a random regression test day model context for the Polish Holstein Friesian population, considering various ways of adjusting the relationship matrix. Data consisted of 890 animals (10 genotyped bulls, 100 cows with phenotypic data and 780 ancestors without genotypes or phenotypes). Random regression test day models with a polygenic effect on milk yield modeled by second order Legendre polynomials for the estimation of variance-covariance parameters and were used for prediction of genomically enhanced breeding values (GEBV). In this model, a matrix combining pedigree and marker-based information was used instead of a traditional numerator relationship matrix. In this matrix the proportions of information coming from pedigree and markers were defined by weighting parameters  $w$  and  $1-w$  for pedigree and**

---

\*Corresponding author: tomasz.suchocki@up.wroc.pl

marker-based information matrices, respectively. Various weights of the two sources of information were considered. The accuracy of GEBV both for genotyped bulls and for cows with phenotypes was highest for weighting parameter  $w=0$  and lowest for  $w=1$ . Incorporating genomic information into a conventional genetic evaluation improves reliabilities of breeding value prediction, however, pedigree information is important to maintain the stability of evaluation for non-genotyped animals. Implementation of the single-step approach in a random regression test day model framework is very attractive for genomic prediction in dairy cattle, since it allows to incorporate genomic information directly into a conventional genetic evaluation. However, for accurate predictions it is essential to achieve the right balance between the numerator relationship and markers-based relationship information.

**KEY WORDS:** genomic prediction / reliability / single-step model

Dairy cattle selection based on dense marker panels has become a very important part of breeding programs. Although up to now evaluations based on markers are routinely done as the so-called genomic evaluations, separately from the conventional evaluation based entirely on cow records and pedigree information, it is only a matter of time until cow records, pedigree and marker genotypes are combined into a single model. For this to happen procedures must be developed to efficiently solve the corresponding mixed model equations and to appropriately weight both sources of information. The inverse of the genomic relationship matrix, which is computationally feasible for large data sets used in routine evaluation, was independently proposed by Aguilar *et al.* [2010] and Christensen and Lund [2010]. Only recently Koivula *et al.* [2015] proposed implementation of the model in the context of test day data.

The current study is a pilot for the estimation of a weighting parameter for pedigree and genomic derived relationships for the population of Polish Holstein-Friesian dairy cattle.

## Material and methods

The data set consisted of 890 animals of the Polish Holstein-Friesian breed forming a structure of 10 proven bulls with genotypes, 100 cows with test day records for milk yield at the 1<sup>st</sup> parity, and their 780 ancestors. The cows were daughters of the bulls, with 10 daughters originating from the same herd per bull. At least nine test day records per cow were available, making 987 records in total. Single nucleotide polymorphism (SNP) data genotyped using the Illumina BovineSNP50 BeadChip was available for the 10 bulls. Genotypic data was edited based on minor allele frequency (MAF) by removing SNPs with MAF lower than 0.05, so that 41,085 SNPs were used in the final analysis.

The first step of the analysis was to estimate covariance components between random regression coefficients. It was done using the ASReml software [Gilmour *et al.* 2006], based on the following model:

$$\mathbf{y} = \mathbf{h}\mathbf{t}\mathbf{d} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e} \quad (1)$$

where:

- $\mathbf{y}$  – vector of test day records for milk yield of the 100 cows;
- $\mathbf{htd}$  – vector of fixed herd-test-day effects;
- $\mathbf{a}$  – vector of random regression coefficients of an animal additive polygenic effect;
- $\mathbf{p}$  – vector of random regression coefficients of a permanent environmental effect;
- $\mathbf{e}$  – vector of residuals.

It is assumed that  $\mathbf{a} \sim \mathbf{N}(\mathbf{0}, \mathbf{A} \otimes \mathbf{G}_a)$ ,  $\mathbf{p} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{P})$ , and  $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_n \sigma_e^2)$ , where  $\mathbf{A}$  is a numerator relationship matrix;  $\mathbf{G}_a$  is a covariance matrix for random regression coefficients of an additive polygenic effect;  $\mathbf{P}$  is a covariance matrix for random regression coefficients of a permanent environmental effect;  $\mathbf{I}_n$  and  $\mathbf{I}_N$  are identity matrices of dimension number of cows ( $n$ ) and number of all test day records ( $N$ ), respectively;  $\sigma_e^2$  is a residual variance.  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$  are design matrices for additive polygenic and permanent environmental effects, respectively, with elements corresponding to orthogonal polynomial functions of days in milk (DIM). Both additive polygenic and permanent environmental effects were modeled using Legendre polynomials of order 2.

The next step was to estimate breeding values of all animals (genotyped bulls, cows with phenotypic data and their ancestors) by combining pedigree and genomic sources of information. This so-called single step model was set up as follows:

$$\mathbf{y} = \mathbf{htd} + \mathbf{Z}_1^* \mathbf{a}^* + \mathbf{Z}_2 \mathbf{p} + \mathbf{e}^* \quad (2)$$

where  $\mathbf{a}^*$  is a vector of random regression coefficients of animal additive polygenic effects of all animals and  $\mathbf{Z}_1^*$  is a corresponding matrix with orthogonal polynomial functions of DIM,  $\mathbf{p}$  and  $\mathbf{Z}_2$  define permanent environmental effects and are the same as in model (1). Additive polygenic and permanent environmental effects, the same as in model (1), were modeled using Legendre polynomials of order 2. It was assumed that  $\mathbf{a}^* \sim \mathbf{N}(\mathbf{0}, \mathbf{H} \otimes \hat{\mathbf{G}}_a)$  and  $\mathbf{e}^* \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_N \sigma_e^2)$  where  $\hat{\mathbf{G}}_a$  contain covariances between orthogonal polynomials estimated by model (1).  $\mathbf{H}$  is a relationship matrix defined as:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{G}_w \mathbf{A}_{22}^{-1} \mathbf{A}_{21} + \mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21} & \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{G}_w \\ \mathbf{G}_w \mathbf{A}_{22}^{-1} \mathbf{A}_{21} & \mathbf{G}_w \end{bmatrix}$$

with  $\mathbf{A}_{11}$ ,  $\mathbf{A}_{22}$ , and  $\mathbf{A}_{12}$  representing numerator relationship matrices for non-genotyped, genotyped, and between non-genotyped and genotyped animals, respectively. These four matrices are sub-matrices of the relationship matrix  $\mathbf{A}^*$  between all animals:

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

$\mathbf{G}_w = (\mathbf{1} - w)\mathbf{G}^* + w\mathbf{A}_{22}$  where  $\mathbf{G}^*$  is a SNP-based genomic relationship matrix adjusted for the difference between the original genomic relationship matrix ( $\mathbf{G}$ ) constructed

as proposed by VanRaden [2008] and the  $\mathbf{A}_{22}$  component of the numerator relationship matrix, using two parameters and [Christensen *et al.* 2011]:

$$\mathbf{G}^* = \mathbf{G}\beta + \alpha$$

The parameters  $\alpha$  and  $\beta$  are derived from the following set of equations:

$$\mathbf{avg.diag}(\mathbf{G})\beta + \alpha = \mathbf{avg.diag}(\mathbf{A}_{22})$$

$$\mathbf{avg.offdiag}(\mathbf{G})\beta + \alpha = \mathbf{avg.offdiag}(\mathbf{A}_{22})$$

In the current study various values of weighting parameter  $w$  (within the range from 0 to 1 incremented by 0.1) were tested. Note that the inverse of  $\mathbf{H}$  is given by

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_w^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} + \mathbf{A}^{*-1}, [\text{Aguilar } et al. 2010, \text{Christensen and Lund 2010}].$$

The 305-day lactation based breeding values were calculated for each animal based on random regression coefficients for the additive polygenic effect estimated by model (2):

$$\hat{\mathbf{u}}_i = \sum_{\tau=5}^{305} [\varphi_0(\tau)\hat{\mathbf{a}}_{i0} + \varphi_1(\tau)\hat{\mathbf{a}}_{i1} + \varphi_2(\tau)\hat{\mathbf{a}}_{i2}],$$

where  $\varphi_0(\tau)$ ,  $\varphi_1(\tau)$  and  $\varphi_2(\tau)$  define the Legendre polynomial evaluated at DIM  $\tau$  and  $(\hat{\mathbf{a}}_{i0}, \hat{\mathbf{a}}_{i1}, \hat{\mathbf{a}}_{i2})$  are random regression coefficients for the additive polygenic effect of animal  $i$ . The coefficients of Legendre polynomials have the following form:

$$\varphi_0(\tau) = \sqrt{\frac{1}{2}},$$

$$\varphi_1(\tau) = \sqrt{\frac{3}{2}} \left( \frac{2(\tau - 5)}{300} - 1 \right), \quad \varphi_2(\tau) = \sqrt{\frac{5}{2}} \cdot \frac{1}{2} \left( 3 \left( \frac{2(\tau - 5)}{300} - 1 \right)^2 - 1 \right).$$

The reliability of breeding values for each animal was estimated as:

$$\mathbf{K} = \left[ \sum_{\tau=5}^{305} \varphi_0(\tau), \sum_{\tau=5}^{305} \varphi_1(\tau), \sum_{\tau=5}^{305} \varphi_2(\tau) \right]$$

where

$\mathbf{K} = \left[ \sum_{\tau=5}^{305} \varphi_0(\tau), \sum_{\tau=5}^{305} \varphi_1(\tau), \sum_{\tau=5}^{305} \varphi_2(\tau) \right]$  and  $\mathbf{PEV}_i$  represents the prediction error variance expressed as a diagonal element of and inverse of a random regression coefficients for the additive polygenic effect matrix corresponding to animal  $i$ . Accuracy is defined as a square root of reliability.

The R software [R Development Core Team 2009] was used to create custom-written programs to predict breeding values and their reliabilities.

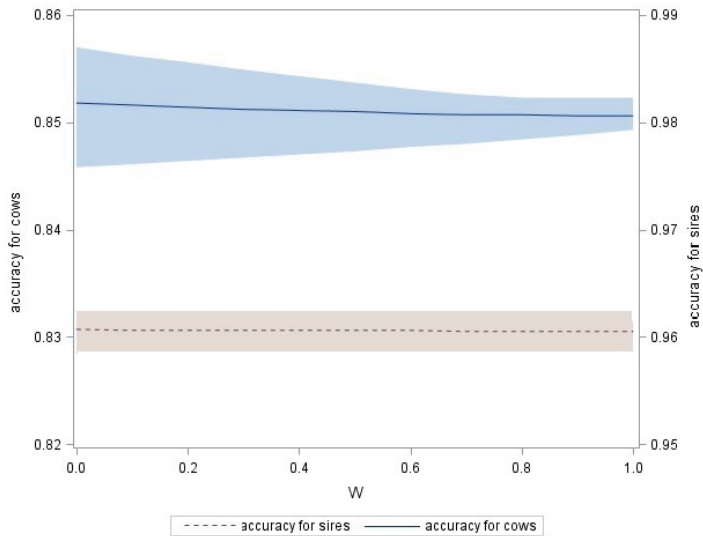


Fig. 1. Average accuracy of GEBV for different values of weighting parameter  $w$  for genotyped bulls and cows with records. Shaded parts of the Figure correspond to the standard deviation of the reliability of breeding values.

## Results and discussion

The average reliability of breeding values was not markedly influenced by weighting parameter  $w$  and ranged between 0.8506 and 0.8518 for the 100 cows with records as well as between 0.9605 and 0.9607 for the genotyped bulls (Fig. 1). Still, a constant trend may be observed, that the highest reliabilities were always obtained for  $w = 0$  corresponding to a pure genomic evaluation, while the lowest reliabilities were always obtained for  $w = 1$  corresponding to a pure pedigree based evaluation. The average accuracy for ancestors amounted to 0.2616 and was independent of  $w$ . The variation of GEBV reliability among individuals, expressed by its standard deviation, did not depend on  $w$  and amounted to 0.002 and 0.092 for bulls and ancestors, respectively. However, for cows the standard deviation of reliability decreased with increasing  $w$ , ranging from 0.006 for  $w = 0$  to 0.002 for  $w = 1$ , indicating that genomic information provides increased reliability of some cows. Both evaluations (for  $w = 0$  and for  $w = 1$ ) resulted in differences between average breeding values in the order of 1.23 genetic standard deviations (540 kg milk) for the genotyped bulls, 0.49 genetic standard deviations (216 kg milk) for cows with records and 0.03 genetic standard deviations (13 kg milk) for ancestors. The ranking of bulls' breeding values was not influenced by  $w$ . Standard deviations of GEBV were independent of parameter  $w$  for genotyped bulls, cows with records and ancestors.

The results of our study indicate that incorporating genomic information into a conventional genetic evaluation improves reliabilities of breeding value prediction.

However, pedigree information is important to maintain the stability of evaluation for non-genotyped animals by reducing the standard deviation of prediction reliability. Although estimated differences were very small due to a small sample size, still the trend of  $w$  and accuracy was consistent.

***Acknowledgements.** The SNP data was obtained in cooperation with the Genomika Polska Consortium.*

#### REFERENCES

1. AGUILAR I., MISZTAL I., JOHNSON D. L., LEGARRA A., TSURUTA S., LAWLOR T. J., 2010 – Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* 93, 743-752.
2. CHRISTENSEN O. F., LUND M. S., 2010 – Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* 42, 2.
3. CHRISTENSEN O., MADSEN P., NIELSEN B., OSTERSEN T., SU G., 2011 – Genomic prediction on pigs using the 338 single-step method. In: 62nd Annual Meeting of the European Federation of Animal 339 Science (EAAP); Stavanger, Norway.
4. GILMOUR A. R., GOGEL B. J., CULLIS B. R., THOMPSON R., 2006 – ASREML User Guide Release 2.0. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.
5. GILMOUR A. R., THOMPSON R., CULLIS B. R., 1995 – Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* 51, 1440-1450.
6. KOIVULA M., STRANDÉN I., PÖSÖ J., AAMAND G. P., MÄNTYSAARI E. A., 2015 – Single-step genomic evaluation using multitrait random regression model and test-day data. *Journal of Dairy Science* 98, 2775-2784.
7. R Development Core Team, 2009 – R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
8. VAN RADEN P. M., 2008 – Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91, 4414-4423.