Domestic and Interbull information in the single step genomic evaluation of Holstein milk production

J. Přibyl, J. Bauer, P. Pešek, J. Přibylová, L. Vostrý, L. Zavadilová

Institute of Animal Science, Prague-Uhříněves, Czech Republic

ABSTRACT: Estimated breeding values and genomic enhanced breeding values for milk production of young genotyped Holstein bulls were predicted using a conventional animal model, ridge regression genomic prediction procedure, genomic best linear unbiased prediction, single-step genomic best linear unbiased prediction, and one-step blending procedures. For prediction, the nation-wide database of domestic Czech production records was combined with deregressed proofs from Interbull files through 2008, which had been transformed by multiple across country evaluation to reflect domestic production conditions. 1259 genotyped bulls had already been proven in 2008. Analyses were run that used Interbull values only for these genotyped bulls and used Interbull values for all available sires. Predictions were validated by comparing correlations of breeding value predictions with estimated breeding values and daughter-yield-deviations after progeny test in 2012 of 140 young genotyped bulls and their associated reliabilities. Combining domestic data with Interbull estimated breeding values improved prediction of both estimated breeding values and genomic enhanced breeding values. Prediction by animal model (traditional estimated breeding values) using only the domestic database had 0.29 validated reliability of prediction; whereas combining the nation-wide domestic database with all available deregressed proofs for genotyped and non-genotyped sires from Interbull resulted in reliability of 0.34, compared to 0.36 when using Interbull data only. The highest reliabilities were for predictions from the single-step genomic best linear unbiased prediction procedure using combined data, or with all available deregressed proofs from Interbull only (one-step blending approach), which reached validated reliabilities for genomic enhanced breeding values predictions 0.53 and 0.54, respectively.

Keywords: genomic breeding value; single-step prediction; animal model; validated reliability

List of abbreviations: EBV = estimated breeding value, GEBV = genomic enhanced breeding value, BLUP = best linear unbiased prediction, RRBLUP = ridge regression genomic prediction procedure, GBLUP = genomic best linear unbiased prediction, ssGBLUP = single-step genomic best linear unbiased prediction, MACE = multiple across country evaluation (Interbull breeding values), DYD = daughter yield deviations, DRP = deregressed proofs (one-step blending approach), DGV = direct genetic values, PA = parent average, YD = yield deviation, ERC = effective record contributions, MAF = minor allele frequency, **A** = pedigree relationship matrix, **G** = genomic relationship matrix, HYS = herd-year-season effect

INTRODUCTION

In small Holstein populations, a substantial proportion of matings are often to imported bulls or semen. In such cases, sires have low and only indirect genetic relationship to the domestic population. Interbull multiple across country evaluation (MACE) genetic correlations of the Czech Republic with other countries are approximately 0.85, resulting in reduced reliability of estimated breeding values (EBV) of foreign sires after imports to about 72% of reliabilities in the country of origin. To some countries with a different climate and production conditions (e.g. Ireland, Israel, New Zealand), genetic correlations are even lower, about 0.75. These circumstances negatively influence national genetic evaluations of animals and also international MACE comparisons. Typically,

Supported by the Ministry of Agriculture of the Czech Republic (Projects No. MZE0002701404 and No. QI111A167).

however, the criterion for selection is the production and rank of animals under domestic management and environmental conditions.

Genomic enhanced breeding value (GEBV) is used mainly for the evaluation of young animals without own performance. Values of young animals are predicted according to the relation to other animals with phenotype performance records. This relation is on the basis of relationship according to common ancestors, and/or on the basis of common segments of genome. All available sources of information are used to achieve the highest possible reliability of prediction, but with the attention to avoid double counting of information sources. Some methodical aspects connected with EBV and GEBV predictions are described by Pribyl et al. (2010).

Two-stage approaches work initially with parts of genome and followed by summation, or all genetic information simultaneously is used for genomic evaluation. Second group of procedures is more accurate and sometimes named as a single-stage approach (Schulz-Streeck et al. 2013). Parts of cattle genome suitable for genetic evaluation of animals were analyzed by Szyda et al. (2013). For simultaneous evaluation, the multi-step procedures are using a variety of regression-based methodologies (Meuwissen et al. 2001), including the ridge regression genomic prediction procedure (RRBLUP), and Bayesian procedures, and the genomic best linear unbiased prediction (GBLUP) method using a genomic relationship matrix (VanRaden 2008). Pseudo-phenotypes, daughter yield deviations (DYD) or deregressed proofs (DRP) are bases for direct genetic values (DGV) calculation. Genetic markers do not explain all genetic variability of analyzed traits (Liu et al. 2011), therefore DGV are then combined with residual polygenic effect (or parent average; PA) in a selection index to produce GEBV. Misztal et al. (2009), Aguilar et al. (2010), Christensen and Lund (2010), and Legarra and Ducrocq (2012) developed a single-step genomic best linear unbiased prediction (ssGBLUP) which effectively combines nation-wide production records and pedigree databases with genomic information, and produces directly GEBV. This method overcomes several critical assumptions required by multi-step procedures, and allows common rank of all genotyped and ungenotyped animals in a population.

Pribyl et al. (2012) used this methodology for the genetic evaluation of the Czech Holstein population. Despite using a small number of proven reference bulls, genotyping of proven and young bulls led to an increased correlation of the GEBVs of young bulls with their EBV prediction after progeny test. As mentioned, imported sires typically have a low genetic relationship to the domestic population. Therefore using information from global Interbull EBVs could improve accuracy of prediction.

Gao et al. (2012) and Su et al. (2012) used DRP of sires as input data instead of national production records in ssGBLUP, naming this approach one-step blending approach. Pribyl et al. (2013) combined in ssGBLUP nation-wide databases of production with all available Interbull DRPs. Implanting the Interbull file converted by MACE into a scale reflecting Czech production conditions improved accuracy of prediction. To demonstrate possible benefits of combining data sources, the relatively new issue from Interbull (year 2011) was used. Improvement depended on correlation of young genotyped bulls with Interbull database (unpublished results). The newer the data, and the younger the Interbull bulls, the higher the improvement in accuracy of prediction of young bulls under domestic conditions.

The aim of this study was to compare methods of genetic prediction for young bulls by GEBV using both domestic and global Interbull data from the year 2008. Predictions were verified according to domestic production data until the calving year 2012.

MATERIAL AND METHODS

Datasets description. Production records from first lactations of Czech Holstein cows, Interbull milk yield EBVs of bulls, and pedigree databases were used.

Two overlapping datasets of domestic milk production performance data and a dataset that included Interbull breeding values (MACE) were used:

(1) Domestic: 969 269 1st lactations, calving years 1991–2008, and 1 762 905 animals in the pedigree file;

(2) Domestic: 1 185 225 1st lactations, calving years 1991–2012, and 1 958 139 animals in the pedigree file;

(3) Interbull: 98 037 EBVs through year 2008, average reliability 0.70 (> 0.28), converted by MACE on a national scale, and 268 451 animals in the pedigree file.

Values were modified in order that variability of EBV of domestic proven bulls and of Interbull EBVs was similar. Estimated breeding values were deregressed (Rozzi et al. 1990; Schaeffer 1994) and pseudo-data yield deviation (YD) and effective record contributions (ERC) were calculated, considering sire as an animal with its own production:

$$ERC = ((1 - h^2)/h^2) \times (rel/(1 - rel))$$

where:
h = heritability
rel = reliability of estimated breeding values (EBV)

Bulls were genotyped by Illumina BovineSNP50 BeadChip V2 (Illumina Inc., San Diego, USA). To eliminate possible input errors, data were edited for: minor allele frequency (MAF) < 0.05, Gscore < 0.60, number of loci per bull < 90%, number of bulls per locus < 90%, substantial error of prediction of old proven bulls in the training set – absolute difference of input DRP with predicted DGV > 708 kg, large discrepancy of part of relationship matrix **A22** to genomic relationship matrix **G** – absolute difference in relationship to others > 3 animals > 0.30, and proportion of Holstein genes < 85%.

After editing, 39 904 loci for 1605 bulls out of which 1259 were already proven in 2008 (training set), 140 young with 0 daughters in 2008 and > 50 daughters (average 67) in 2012, and 206 other bulls with a small number of daughters were used.

Data were evaluated by weighted analysis using several statistical procedures. Because ERC was used as the weight for individual records, for all domestic production records ERC was set equal to 1.

Involved methods of evaluation

(1) Ridge regression genomic prediction procedure (RRBLUP) was performed according to the following model:

$$y_j = \mu + \Sigma\delta_i \cdot g_{ij} + e_j$$

where
y_j = deregressed proofs (DRP) of milk production for bull j
μ = common constant (contemporary group) as a fixed effect
δ_i = regression coefficient for locus i (this effect is considered as random with covariance matrix equal to identity matrix multiplied by a constant σ^2_a/m where σ^2_a = total genetic variance and m = number of loci)
g_{ij} = value of alleles in locus i < 0, 1, 2 > for bull j
e_j = random error

Estimated marker effects are used to predict direct genetic values (DGV) of young animals:

$$DGV_j = \mu + \Sigma\delta_i \cdot g_{ij}$$

(2) Genomic best linear unbiased prediction (GBLUP) was done by the model:

$$y_j = \mu + an_j + e_j$$

where
μ = common constant (contemporary group) as a fixed effect
an_j = direct genetic values (DGV) of animal j with genomic relationship matrix **G** for all genotyped animals
e_j = random error

(3) Best linear unbiased prediction (BLUP) and single-step genomic best linear unbiased prediction (ssGBLUP) were performed according to the animal model:

$$y_{ij} = HYS_i + \beta_1 \cdot ca_j + \beta_2 \cdot ca_j^2 + \beta_3 \cdot do_j + \beta_4 \cdot do_j^2 + an_j + e_{ij}$$

where
y_{ij} = first lactation milk yield of cow, or deregressed proofs (DRP) of milk production for bull
HYSi = contemporary group within a herd in a 3-month calving period (fixed effect)
β_1, β_2, β_3, β_4 = regression coefficients
ca_j, ca_j^2 = parameters for curvilinear regressions on calving age (fixed effect)
do_j, do_j^2 = parameters for curvilinear regressions on days open (fixed effect)
an_j = estimated breeding values (EBV) or genomic enhanced breeding value (GEBV) of animal j with pedigree additive relationship matrix **A** in BLUP, or matrix **H** in ssGBLUP

H is the pedigree additive relationship matrix **A** augmented by genomic relationship matrix **G**. Weights of 80% **G** and 20% additive pedigree relationship matrix only for genotyped animals $\mathbf{A_{22}}$ were used for incorporation into **H**.

Matrix **G** was constructed according to deviations from the averages of observed allele frequencies and was standardized by division by the average value of the diagonal of **G** (Forni et al. 2011), then shifted, so that the elements of the **A22** and ele-

ments of **G** would have the same average (Vitezica et al. 2011).

Deregressed proofs (DRPs) processed from MACE values are free from influence of systematic environmental effects and all of them are on the same scale. For inclusion into BLUP calculations, DRPs are therefore assigned to an additional class of herd-year-season effect (HYS), and for independent variables (calving age, days open) average values from domestic population are used.

The programme G-matrix (Version 2.0, 2011; http://dmu.agrsci.dk) was used for construction of the **G** relationship matrix, and the DMU (Version 6, release 5.0., 2010; http://dmu.agrsci.dk) software package was used for genetic prediction. Data files were handled with help of SAS (Statistical Analysis System, Version 9.4, 2012).

Procedures for the various models for genetic prediction are summarized in Table 1.

Domestic production records were used in BLUP and ssGBLUP genetic prediction procedures (Table 2), whereas in RRBLUP and GBLUP analyses, Interbull DRPs from 1259 referenced bulls were utilized, which represented a total of 57 864 ERCs. These values were combined in an index with EBV estimates according to pedigree information from the domestic Holstein population. Of all DRPs available from Interbull, a total of 98 037 were used in BLUP and ssGBLUP procedures, and this database represents 785 276 ERCs. This method corresponds to "one-step blending approach". The combination of both domestic and Interbull databases identified 1 064 912 records (1 632 668 ERCs) that were analyzed by BLUP and ssGBLUP procedures. In these analyses, Interbull DRPs were used only when sires did not have daughters in the domestic population.

Procedures were validated by calculating correlations among predictors of genetic merit for 140 young bulls that had no daughter records in 2008 but > 50 daughter records in 2012, that is, their EBVs and DYDs after progeny test (Szyda et al. 2008, 2011). Average validated reliabilities (*VRel*) (Gao et al. 2012; Su et al. 2012) were computed from correlation of prediction with DYD by the following formula:

$$VRel = r^2_{P,DYD}/rel_{DYD}$$

where

$r_{P,DYD}$ = correlation of predicted method with daughter yield deviations (DYD) after progeny test

rel_{DYD} = reliability as affected by number of progeny, corresponding with daughter yield deviations (DYD)

RESULTS AND DISCUSSION

Merging domestic production records with Interbull files notably increased volume of input data for genetic evaluation (Table 2).

Results were expressed as deviations from a base population of 2116 proven sires, each having at least 60 daughters in 2008. Average value (EBV/GEBV) of prediction of young bulls had deviation in a case of evaluating the domestic population equal to 657, and 672 kg of milk for BLUP and ssGBLUP methods, respectively (Table 3), whereas from combined data these averages were 651 and 640 kg, respectively. Average values of prediction

Table 1. Prediction procedures

Method	Calculated value	Sources of production 2008		
		Domestic (D)	**I**nterbull (I)	**D + I**
BLUP	EBV	D-EBV	I-EBV	D+I-EBV
RRBLUP	DGV		rI-DGV	
	GEBV*			rI-GEBV
GBLUP	DGV		gI-DGV	
	GEBV*			gI-GEBV
ssGBLUP**	GEBV	D-GEBV	I-GEBV***	D+I-GEBV

EBV = estimated breeding value, DGV = direct genetic values, GEBV = genomic enhanced breeding value, BLUP = best linear unbiased prediction, RRBLUP = ridge regression genomic prediction procedure, GBLUP = genomic best linear unbiased prediction, ssGBLUP = single-step genomic best linear unbiased prediction, g = genomic, r = ridge regression

*GEBV = 0.8 DGV + 0.2 D-EBV

genomic relationship **G is weighted 80% and pedigree relationship **A22** 20%

***one-step blending approach

Table 2. Size of data for predictions 2008

	Records	Weights ERC	Procedure
Domestic (D)	969 269	969 269	D-EBV D-GEBV
Interbull (I) for genotyped bulls	1 259	57 864	rI-DGV
	970 528**	240 145	rI-GEBV
	1 259	57 864	gI-DGV
	970 528**	240 145	gI-GEBV
Interbull (I) for all bulls	98 037	785 276	I-EBV I-GEBV
D + I all*	1 064 912	1 632 668	DI-EBV DI-GEBV

ERC = effective record contributions, EBV = estimated breeding value, GEBV = genomic enhanced breeding value, DGV = direct genetic values, g = genomic, r = ridge regression
*from Interbull file, only bulls with no domestic daughters
**including pedigree information from domestic population

were in a good agreement with results based upon progeny test, in which average EBV for this group of young bulls was 629 kg of milk. In average, predictions of breeding values of young bulls in all methods were overestimated by about 1.7–6.8%.

Correlations of predictions with EBV after progeny test (EBV12) were noticeably higher than with DYD (DYD12) (Table 3). D-EBV were EBVs of young animals, reflecting the response of pedigree of (imported) young bulls in a domestic condition. Predictions of young bulls according to this "common" BLUP-Animal Model analysis were correlated with EBV12 by 0.59 and with DYD12 by 0.47. Corresponding validated reliability (*VRel*) was 0.29. Predictions derived from ssGBLUP of domestic data reached *VRel* of 0.48.

Predictions with DGV by RRBLUP and GBLUP, which were according to Interbull DRP for genotyped bulls only, were correlated to EBV12 by 0.60 and 0.59 respectively, and correlated to DYD12 by 0.57. Differences in correlations to EBV12 and to DYD12 were much lower than when using BLUP and ssGBLUP and domestic databases. Corresponding *VRel* were 0.42 and 0.41, respectively. After combination with pedigree values, reliabilities reached *VRel* of 0.47. This is close to the value obtained on the domestic population using ssGBLUP.

Predictions of EBV using the BLUP method including all Interbull DRP (one-step blending approach) versus using combined data reached *VRel* of 0.36 and 0.34, respectively, which were notably higher than from domestic population data only (Table 3). Predictions by GEBV with ssGBLUP from Interbull and combined data had *VRel* values of 0.54 and 0.53, respectively. Values achieved by using solely Interbull data and by combined data were similar.

Table 3. Average genetic predictions for 140 young bulls, correlations of predictions with results after progeny test, and validated reliabilities

Data 2008	Mean milk (kg)*	Method	EBV 2012	DYD 2012	Validated reliability
Domestic (D)	657	D-EBV	0.59	0.47	0.29
	672	D-GEBV	0.70	0.61	0.48
Interbull (I) for genotyped bulls		rI-DGV	0.60	0.57	0.42
		rI-GEBV	0.67	0.61	0.47
		gI-DGV	0.59	0.57	0.41
		gI-GEBV	0.66	0.61	0.47
Interbull (I) for all bulls		I-EBV	0.62	0.53	0.36
		I-GEBV	0.70	0.65	0.54
D + I all	651	D+I-EBV	0.63	0.51	0.34
	640	D+I-GEBV	0.73**	0.64	0.53
Data 2012	629	D-EBV			

EBV = estimated breeding value, DYD = daughter yield deviations, GEBV = genomic enhanced breeding value, DGV = direct genetic values, g = genomic, r = ridge regression
*difference of EBV/GEBV from basis of 2116 proven sires each with at least 60 daughters in 2008
**when using for response variable GEBV12 means GEBV by single-step genomic best linear unbiased prediction (ssGBLUP) in year 2012, the highest correlation is for DI-GEBV with value 0.75

In combined data, only Interbull sires that did not have domestic daughters were used for BLUP and ssGBLUP procedures. In methods I-EBV and I-GEBV (one-step blending approach), all available data from Interbull were used, including contributions from the Czech population. Therefore sources of information were similar in both cases. The Interbull database contained 785 276 ERC (Table 2) connected directly to sires, which had substantial predicative ability, greater than a population of cows of similar size. On the other hand, the Interbull database was generated under production conditions not closely similar to those of the Czech domestic herds.

The resulting reliabilities are within the range of values achieved by the literature cited in this study.

CONCLUSION

Combining genetic evaluation of all domestic records with all available Interbull EBVs, both for genotyped and ungenotyped sires, and transformed by MACE into domestic production conditions improved prediction both of EBV and GEBV.

The ssGBLUP method enabled using daughter's production records and/or DRPs both for genotyped and ungenotyped sires in joint genetic evaluation.

Generally, the most reliable genetic predictions, according to repeated calculations, were produced by the ssGBLUP procedure utilizing combined data. Differences in accuracy of prediction between ssGBLUP in combined data and ssGBLUP using only Interbull data (one-step blending approach) were small.

Acknowledgement. We thank the Czech Moravian Breeding Corporation (Prague, Czech Republic) and the Holstein Cattle Breeders Association of the Czech Republic (Hradištko, Czech Republic) for supply of data files. We are grateful to Dr. Per Madsen, University of Aarhus, Denmark, and to Prof. Ignacy Misztal, University of Georgia, USA, for methodical support and for provision of computer programs. We gratefully acknowledge the helpful comments of anonymous reviewers.

REFERENCES

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.

Christensen O.F., Lund M.S. (2010): Genomic prediction when some animals are not genotyped. Genetics Selection Evolution, 42, 2.

Forni S., Aguilar I., Misztal I. (2011): Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution, 43, 1.

Gao H., Christensen O.F., Madsen P., Nielsen U.S., Zhang Y., Lund M.S., Su G. (2012): Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genetics Selection Evolution, 44, 8.

Legarra A., Ducrocq V. (2012): Computation strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. Journal of Dairy Science, 95, 4629–4654.

Liu Z.T., Seefried F.R., Reinhardt F., Rensing S., Thaller G., Reents R. (2011): Impact of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genetics Selection Evolution, 43, 19.

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

Misztal I., Legarra A., Aguilar I. (2009): Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science, 92, 4648–4655.

Pribyl J., Rehout V., Citek J., Pribylova J. (2010): Genetic evaluation of dairy cattle using a simple heritable genetic ground**.** Journal of the Science of Food and Agriculture, 90, 1765–1773.

Pribyl J., Haman J., Kott T., Pribylova J., Simeckova M., Vostry L., Zavadilova L., Cermak V., Ruzicka Z., Splichal J., Verner M., Motycka J., Vondrasek L. (2012): Single-step prediction of genomic breeding value in a small dairy cattle population with strong import of foreign genes. Czech Journal of Animal Science, 57, 151–159.

Pribyl J., Madsen P., Bauer J., Pribylova J., Simeckova M., Vostry L., Zavadilova L. (2013): Contribution of domestic production records, Interbull estimated breeding values, and single nucleotide polymorphism genetic markers to the single-step genomic evaluation of milk production. Journal of Dairy Science, 96, 1865–1873.

Rozzi P., Schaeffer L.R., Burnside E.B., Schlote W. (1990): International evaluation of Holstein-Friesian dairy sires from three countries. Livestock Production Science, 24, 15–28.

Schaeffer L. (1994): Multiple-country comparison of dairy sire. Journal of Dairy Science, 77, 2671–2678.

Schulz-Streeck T., Ogutu J.O., Piepho H.P. (2013): Comparisons of single-stage and two-stage approaches to genomic selection. Theoretical and Applied Genetics, 126, 69–82.

Su G., Madsen P., Nielsen U.S., Mantysaari E.A., Aamand G.P., Christensen O.F., Lund M.S. (2012): Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy Science, 95, 909–917.

Szyda J., Ptak E., Komisarek J., Zarnecki A. (2008): Practical application of daughter yield deviations in dairy cattle breeding. Journal of Applied Genetics, 49, 183–191.

Szyda J., Zarnecki A., Suchocki T., Kaminski S. (2011): Fitting and validating the genomic evaluation model to Polish Holstein-Friesian cattle. Journal of Applied Genetics, 52, 363–366.

Szyda J., Zukowski K., Kaminski S., Zarnecki A. (2013): Testing different single nucleotide polymorphism selection strategies for prediction of genomic breeding values in dairy cattle based on low density panels. Czech Journal of Animal Science, 58, 136–145.

VanRaden P.M. (2008): Efficient methods to compute genomic predictions. Journal of Dairy Science, 91, 4414–4423.

Vitezica Z.G., Aguilar I., Misztal I., Legarra A. (2011): Bias in genomic predictions for populations under selection. Genetics Research, 93, 357–366.

Received: 2013–11–13
Accepted after corrections: 2014–03–07

Corresponding Author

Ing. Ludmila Zavadilová, CSc., Institute of Animal Science, Přátelství 815, 104 00 Prague 10-Uhříněves, Czech Republic
Phone: +420 267 009 608, e-mail: zavadilova.ludmila@vuzv.cz