Improving single-step genomic prediction reliabilities for clinical mastitis in Nordic Red dairy cattle and Jersey by applying marker-specific weights

A. Chegini[1], I. Strandén [1], E. Karaman[2], T. Iso-Touru[1], J. Pösö[3], G.P. Aamand[4] and M.H. Lidauer[1]

[1] Natural Resources Institute Finland (Luke), 31600 Jokioinen, Finland

[2] Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark

[3] Faba co-op, 01301 Vantaa, Finland

[4] Nordic Cattle Genetic Evaluation, 8200 Aarhus, Denmark

Corresponding author: arash.chegini@luke.fi

Abstract

The standard single-step genomic prediction assumes that all single nucleotide polymorphism (SNP) markers explain an equal amount of genetic variance. The true state may deviate from this assumption, and it has been suggested to consider SNP marker-specific weights when predicting genomic enhanced breeding values (GEBV). We hypothesized that the benefit may be more pronounced in low heritable traits and investigated this hypothesis using the udder health evaluations for Nordic Red (RDC) and Jersey (JER) dairy cattle. In the first step, we develop a standard single-step genomic prediction (ssGBLUP) model based on the currently used multiple-trait evaluation models, and estimated GEBVs. The models included four clinical mastitis (CM) traits, and five correlated traits, namely test-day somatic cell score (SCS) in 1st, 2nd, and 3rd lactations, fore udder attachment and udder depth, and describes all additive genetic effects of an animal by one covariance function. Then, we investigated three alternative approaches, where we applied SNP-marker specific weights. The three approaches for SNP-marker weighting were: 1) a nonlinear method similar to BayesA, 2) the classical formula ($2pq\hat{u}^2$), and 3) the mean of SNP weights for every 20 adjacent SNP markers calculated based on $2pq\hat{u}^2$. To solve the models with SNP marker-specific weights, we applied the single-step SNPBLUP solver implemented in MiX99. We validated the models by forward validation where the last four years of the data were removed. The datasets for RDC and JER included 6.9 and 1.2 million animals of which 5.6 and 0.9 million cows had records, respectively. The number of genotyped animals was 125,789 and 64,777 for RDC and JER, respectively. We found a significant increase in prediction reliability for CM when applying SNP-marker specific weights. For instance, applying the $2pq\hat{u}^2$ weights compared to the standard ssGBLUP for SCS, the prediction reliability increased from 0.58 to 0.64 and from 0.61 to 0.56 for RDC and JER bulls, respectively. We found similar improvements in the prediction reliability for cows. In general, all weighing approaches improved prediction reliability, but the highest improvement was achieved by weighing the SNP-markers by $2pq\hat{u}^2$.

Key words: genomic prediction, SNP marker weights, single-step SNP-BLUP, udder health traits

Introduction

Clinical mastitis (CM) is the costliest disease affecting animal welfare and reducing profitability by lowering milk quality and quantity. Furthermore, it is a lowly heritable trait, which means it will take longer to genetically improve it. Fortunately, studies show that genomic selection can be especially beneficial for traits with high recording costs or traits with low heritability (Meuwissen et al., 2001; Schaeffer, 2006). In addition, it is possible to improve prediction reliability by employing a single-step genomic prediction

(ssGBLUP) model which combines all information from genotyped and non-genotyped animals (Christensen and Lund, 2010).

In a standard ssGBLUP model, the assumption is that all single nucleotide polymorphisms (SNP) are equally important in terms of the amount of genetic variance they explain. This may not be true as some SNPs are in the proximity of influential genes. Results of several studies indicate improvements in prediction reliability by applying SNP marker weights (Wang et al., 2012; Fragomeni et al., 2019). Different formulas have been used to calculate SNP weights ranging from Nonlinear which is a BayesA-like procedure (VanRaden, 2008) to square of marker effect size (Wang et al., 2012). There were some discrepancies between reports which may be due to the differences in the traits, population or breed, and weighing procedures between the studies.

The objective of this study was to investigate the possibility of improving prediction reliability for CM through applying marker weighting in a single-step genomic prediction framework.

Materials and Methods

Data

Records of udder health traits including CM, test-day somatic cell score (SCS) and two udder type traits namely fore udder attachment (UA) and udder depth (UD) from Nordic Red (RDC) and Jersey (JER) dairy cows collected since 1990 in Denmark, Finland and Sweden were used. There were 74.5 and 17.1 million records for 5.6 and 0.9 million RDC and JER, respectively. The number of genotyped animals used in this study was 125,789 and 64,777 for RDC and JER, respectively. The number of SNP markers was 46,914 for RDC and 41,897 for JER.

Observations for CM were grouped into four classes (CM11, CM12, CM2 and CM3) based on the lactation number and the days in milk in which the disease occurred. Also, SCS records were grouped into three classes (SCS1, SCS2 and SCS3) based on the lactation number.

Statistical model

The multi-trait model used in this study is the standard model currently used for the evaluation of udder health traits by Nordic Cattle Genetic Evaluation (NAV) and has been described in detail in Negussie et al. (2010). In brief, the model in matrix notation was:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Tk} + \mathbf{F}_a\mathbf{a} + \mathbf{F}_p\mathbf{p} + \mathbf{e}$$

where **y** is the vector of observations for all nine traits; **b** is the vector of fixed effects; vector **k** contains random herd-year effects for CM, UA and UD and random herd-test-day effects for SCS; vector **a** has the animal additive genetic effects; vector **p** has the random animal non-additive genetic effects and **e** is the random residual. The $\mathbf{F}_a$ and $\mathbf{F}_p$ matrices have the trait-specific covariables from the covariance function. Covariance functions were used to model animal additive and non-additive genetic effects.

Scenarios

First, a standard ssGBLUP was implemented and the results were compared with those of weighted ssGBLUP. In a single-step evaluation, we need a relationship matrix that combines numerator relationship matrix (NRM) with genomic relation matrix (GRM) as follows:

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

where **G** is the GRM and was calculated as $\mathbf{G} = \mathbf{ZZ'} + \mathbf{C}$, where **Z** is a centered and scaled marker matrix and $\mathbf{C} = w\mathbf{A}_{22}$ with w equal to the residual polygenic (RPG) proportion and $\mathbf{A}_{22}$ is the NRM of the genotyped individuals. The amount of RPG proportion was 0.10.

Second scenario was to apply a **Nonlinear** formula (VanRaden, 2008) to weigh the markers as follows:

$$\mathbf{G}_j = \frac{\mathbf{Z}_j\mathbf{W}_j\mathbf{Z}_j'}{\sum_{i=1}^{m} 2p_i(1-p_i)}$$

where m is the number of markers and p_i is allele frequency of marker i. W_j is a diagonal matrix containing the weights for eigenvalue trait j calculated by $1.25^{\frac{|\hat{u}_{ji}|}{sd(\hat{u}_j)}-2}$, where $|\hat{u}_{ji}|$ is the absolute value of the estimated SNP effect for marker i of the eigenvalue trait j and sd($\hat{\mathbf{u}}_j$) is the standard deviation of all estimated SNP effects for eigenvalue trait j.

In the third scenario, markers were weighted using the classical method (Falconer and Mackay, 1996), henceforth referred to as **2pqû²**.

In the last scenario, average weights of every 20 adjacent markers calculated by the classical method were applied (**20SNP_window**).

Validation

To create a reduced dataset for the validation of Legarra and Reverter (2018), the last four years of observations were excluded. Breeding values were predicted using both the reduced and full datasets for each of the scenarios. Combined genomic enhanced breeding values (GEBV) for both CM and SCS were calculated using lactation weights as applied by NAV.

Effective record contribution (ERC) for genotyped animals was calculated. Then, a bull could be a candidate if it had an ERC ≥ 2 using full data and that of zero using reduced data. Corresponding values were 0.9 and zero for cow candidates. All the analyses were implemented using the MiX99 program suite (Pitkänen et al., 2022).

Results & Discussion

Forward validation for CM

Regression of GEBVs using the full dataset on those using the reduced dataset showed slightly lower bias (b_0) for 2pqû² compared to the other scenarios (Table 1). The only exception was the 20SNP_window for RDC bull candidates. The standard ssGBLUP model yielded the lowest dispersion (b_1).

The reliability of predictions using standard ssGBLUP for RDC and JER bull candidates were 0.50 and 0.65, respectively. Corresponding values for RDC and JER cow candidates were 0.74 and 0.72, respectively. All marker weighting scenarios resulted in higher reliabilities (ranging from 0.5% to 13.8%) compared to the standard ssGBLUP, except for 20SNP_window in RDC and JER bulls. The highest prediction reliability was obtained by weighting the markers by the classical formula, i.e., 2pqû².

Table 1. Results of forward validation of bull and cow (within parentheses) candidates for combined clinical mastitis using standard single-step procedure as well as different SNP weighting scenarios for Nordic Red (RDC) and Jersey (JER) dairy cattle.

Breed	Group;n	Model	b_0	b_1	R^2	%gain*
RDC	Bull;86 (Cow;8,440)	standard ssGBLUP	0.002 (0.005)	0.75 (0.87)	0.50 (0.74)	
		Nonlinear	0.001 (0.005)	0.73 (0.85)	0.51 (0.74)	2.0 (1.1)
		2pqû²	0.001 (0.003)	0.68 (0.79)	0.57 (0.78)	13.8 (5.3)
		20SNP_window	0.0004 (0.005)	0.70 (0.85)	0.49 (0.75)	-1.6 (1.8)
JER	Bull;115 (Cow;8,224)	standard ssGBLUP	0.013 (0.010)	0.78 (0.89)	0.65 (0.72)	
		Nonlinear	0.015 (0.012)	0.77 (0.88)	0.66 (0.73)	0.5 (1.9)
		2pqû²	0.010 (0.008)	0.70 (0.79)	0.66 (0.76)	0.9 (5.3)
		20SNP_window	0.012 (0.011)	0.74 (0.87)	0.64 (0.74)	-2.4 (3.1)

* Percent of gain in prediction reliability relative to standard single-step evaluation.

Table 2. Results of forward validation of bull and cow (within parentheses) candidates for combined SCS using standard single-step procedure as well as different SNP weighting scenarios for Nordic Red (RDC) and Jersey (JER) dairy cattle.

Breed	Group;n	Model	b_0	b_1	R^2	%gain*
RDC	Bull;125 (Cow;18,112)	standard ssGBLUP	6.83 (6.11)	0.86 (0.97)	0.58 (0.77)	
		Nonlinear	7.40 (6.84)	0.83 (0.94)	0.60 (0.78)	2.6 (0.6)
		$2pq\hat{u}^2$	7.21 (5.82)	0.77 (0.87)	0.64 (0.79)	11.1 (2.3)
		20SNP_window	6.66 (6.63)	0.82 (0.94)	0.59 (0.78)	2.5 (1.0)
JER	Bull;119 (Cow;6,537)	standard ssGBLUP	8.17 (8.43)	0.81 (0.97)	0.61 (0.79)	
		Nonlinear	7.80 (8.43)	0.80 (0.96)	0.63 (0.80)	2.7 (0.9)
		$2pq\hat{u}^2$	4.06 (5.71)	0.70 (0.87)	0.65 (0.81)	5.4 (2.8)
		20SNP_window	7.55 (7.66)	0.80 (0.95)	0.64 (0.80)	4.0 (1.3)

* Percent of gain in prediction reliability relative to standard single-step evaluation.

The gain in prediction reliability by marker weighting differed by breed and was more advantageous for RDC. This might be due to the differences in the population structure.

Forward validation for SCS

Results of forward validation for SCS are shown in Table 2. Similar to CM, the lowest bias was obtained for the $2pq\hat{u}^2$ approach. Biases were higher for SCS (ranging from 4.06 to 8.43) than for CM.

The lowest and the highest dispersion were for the standard ssGBLUP and $2pq\hat{u}^2$, respectively, which is in line with the results for CM.

The reliability of predictions using standard ssGBLUP for RDC and JER bull candidates were 0.58 and 0.61, respectively. Corresponding values for RDC and JER cow candidates were 0.77 and 0.79, respectively. The amount of improvement in prediction reliability by applying marker weighting ranged from 0.6% to 11.1% for RDC and 0.9% to 5.4% in JER. Similarly, the $2pq\hat{u}^2$ approach resulted in the highest gain in prediction reliability in both breeds compared to the other scenarios. Prediction reliability was on average higher for SCS than for CM. This was expected as the heritability of SCS was higher than that of CM.

Conclusions

This study was conducted to compare predicted breeding values by the standard single-step genomic model with weighted approaches by using records of udder health traits in two Nordic dairy breed populations. Results indicated that marker weighting is beneficial as improvements in bias and prediction reliability were observed for clinical mastitis and somatic cell score. The classical formula to weigh the markers resulted in the highest gain in prediction reliability and the lowest bias. However, the highest dispersion was obtained by applying this approach. It seems that by marker weighting we accept slightly lower precision in exchange for higher accuracy.

Acknowledgments

This study has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815668.

References

Christensen, O.F., and Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:2. https://doi.org/10.1186/1297-9686-42-2.

Falconer, D.S., and Mackay, T.F.C. 1996. Introduction to Quantitative Genetics. New York: Longman. 464 p.

Fragomeni, B.O., Lourenco, D.A.L., Legarra, A., VanRaden, P.M., and Misztal, I. 2019. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. *J. Dairy Sci.* 102:10012–10019. https://doi:10.3168/jds.2019-16262.

Legarra, A., and and Reverter, A. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol.* 50:53. https://doi:10.1186/s12711-018-0426-6.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829. https://doi:10.1093/genetics/157.4.1819.

Negussie, E., Lidauer, M., Mäntysaari, E.A., Strandén, I., Pösö, J., Nielsen, U.S., Johansson, K., Eriksson, J.–Å., and 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 in Denmark, Finland and Sweden. *Interbull Bulletin.* 42:25–31.

Pitkänen, T.J., Gao, H., Kudinov, A., Taskinen, M., Mäntysaari, E.A., Lidauer, M.H., and Strandén, I. 2002. From data to genomic breeding values with the MiX99 software suite. In Proc. of the 12th world congress on genetics
applied to livestock production: 3–8 July, Rotterdam.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. *J. Anim Breed Genet.* 123:218-223.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. https://doi:10.3168/jds.2007-0980.

Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W.M. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res.* 94:73–83. https://doi.org/10.1017/s0016672312000274