Single-step genomic prediction in small breeds: Finncattle case.

A.A. Kudinov[1*], M. Koivula[1], I. Strandén[1], G.P. Aamand[2], and E.A. Mäntysaari[1]

[1] Natural Resources Institute Finland (Luke), 31600, Jokioinen, Finland; [2] Nordic Cattle Genetic Evaluation (NAV), DK-8200, Aarhus, Denmark;
[*]andrei.kudinov@luke.fi

Abstract

Genomic prediction is of great interest in small breeds. When breed-wise populations cannot exchange data, the establishment of a multi-breed evaluation is a tempting option. Single-step genomic prediction is a useful tool when limited data is available. In the current study, we developed a single-step genomic prediction for the small indigenous Finncattle breed with or without data from national major breed Nordic Red Dairy cattle (RDC). The reliability of genomic prediction was assessed using modified Interbull and Legarra & Reverter validation approaches. Inclusion of the RDC data to Finncattle evaluations helped to improve the quality of prediction in traditional and genomic models. The highest prediction reliability for milk and protein was obtained using the multi-breed single-step model.

Introduction

Genomic prediction has been widely used as a two-step procedure in large dairy cattle populations since 2009. The use of DNA markers allows predicting the breeding value of a candidate animal with high reliability at an early age (VanRaden, 2020). However, in small breeds, the use of the two-step genomic prediction approach might be challenging due to the limited data available. Thus, the single-step method (Christensen & Lund, 2010) would be the best way to perform genomic prediction with a limited number of genotyped animals and phenotypic records (Song et al., 2019). A widely reported procedure to increase the power of genomic prediction is the exchange of genomic, phenotypic, or MACE (Multi-trait Across-Country Evaluations, Interbull, Sweden) information between countries (Kudinov et al., 2021a). Naturally, the exchange can only be done with the common breeds and cannot be used when the breed is unique. In such a situation multi-breed evaluation might be the solution. Usually when multi-breed evaluations are used, the solutions for SNP effects are typically estimated by breed as the two-step procedure. The applicability of the multi-breed single-step genomic prediction model has not been studied extensively.

Finncattle (FIC) is the local Finnish breed. Breeding values for FIC have been evaluated by Nordic Cattle Genetic Evaluation (NAV) in the same evaluation with Nordic Red Dairy Cattle (RDC) and Finnish Holstein (Fin HOL, Kudinov et al., 2021b). Multi-breed data is used because many of the herds raise cows from the three breeds in the same facilities and feeding. The current genomic evaluation for RDC animals uses the two-step approach with pure RDC animals. Because genotyping of FIC breed has started only recently, the development of FIC genomic prediction models has become topical. Assuming the benefits and willingness of Nordic countries to run genomic prediction using a single-step model, the development of a multi-breed single-step model was started.

The focus of the current study was to assess the accuracy of genomic prediction for FIC cows with and without RDC phenotypic and genomic information. We performed validation of the model fit by comparing results from approaches suggested by Legarra & Reverter (2018) and modified Mäntysaari et al. (2010).

Materials & Methods

Data. Pedigree and phenotypic records were obtained from the August 2020 NAV RDC milk and protein evaluations. Pedigree included 4,741,559 RDC, 1,047,697 Fin HOL, 36,133 FIC, and 101,144 other breed animals. There were 137 unknown parent groups (UPG) formed using breed by country by time classes. Phenotypic data was presented by 305-day milk and protein first lactation records from 3,468,530 RDC, 811,363 FIN HOL, and 27,810 FIC cows. Genomic data include genotypes from 187 FIC and 35,729 RDC bulls, and 756 FIC and 133,107 RDC cows. Raw genotypes were imputed and edited by NAV and included 46,914 SNP markers that are used in the current national multi-step genomic evaluations.

Statistical models. Single-step GBLUP and pedigree-based BLUP (PBLUP) were used to estimate breeding values. All models used the same pedigree and UPG, but the phenotypic and genomic data differed. Genomic prediction was performed using a single-step GTBLUP (Mäntysaari et al., 2017; Koivula et al., 2021) model with two different datasets: $\text{ssGTBLUP}_{\text{FIC+RDC}}$ used FIC and RDC data; $\text{ssGTBLUP}_{\text{FIC}}$ used FIC data only. Corresponding PBLUP predictions used either FIC and RDC data – $\text{PBLUP}_{\text{FIC+RDC}}$ or FIC data only – $\text{PBLUP}_{\text{FIC}}$. In the single-step models, UPG were substituted by the same number of metafounders (MF, Legarra et al., 2015). The relationships between MF in the pedigree were modeled using a gamma matrix (**Γ**) of size 137 as presented in Kudinov et al. (2021b). The inverses of pedigree relationship matrix (**A**) and submatrix for genotyped animals ($\mathbf{A}_{22}$) were built with MF inbreeding coefficients. Genomic relationship matrix (**G**) was built assuming a base population allele frequency = 0.5 and residual polygenic effect = 30%. Model runs were done with MiX99 software (Strandén & Lidauer, 1999).

The prediction power of the genetic evaluation models was examined using forward prediction validation. In the reduced data, records from 2017 - 2020 were omitted. The validation used 73 genotyped cows that had no record in the reduced milk and protein data but had records in the full data. Validation reliability of predictions was assessed by two methods: I) production prediction validation (PPV) - regression of yield deviations (YD) of cows computed with full genomic and phenotypic data on (genomic) estimated breeding values ([G]EBV) computed using truncated data ($[\text{G}]\text{EBV}_{\text{red}}$; $\text{YD} - \mu_{[\text{G}]\text{EBVred}} = b_0 + b_1 ([\text{G}]\text{EBV}_{\text{red}} - \mu_{[\text{G}]\text{EBVred}})$, where $\mu_{[\text{G}]\text{EBVred}}$ is mean of $[\text{G}]\text{EBV}_{\text{red}}$). Similar to the Interbull GEBV validation test (Mäntysaari et al., 2010), the YD of cow *j* was weighted by w_j computed as $w_j = \text{ERC}_j / (\text{ERC}_j + \lambda)$, where $\lambda = (1\text{-}h^2)/h^2$;$h^2$ is heritability of the trait and ERC_j is the effective record contribution (Přibyl et al., 2013), the coefficient of determination (R^2) from the regression equation was adjusted by mean of the weights (R^2_w); II) Linear-regression validation (LRV, Legarra and Reverter, 2018) of [G]EBV computed using the full data on $[\text{G}]\text{EBV}_{\text{red}}$ using formulae $[\text{G}]\text{EBV} - \mu_{[\text{G}]\text{EBVred}} = b_0 + b_1([\text{G}]\text{EBV}_{\text{red}} - \mu_{[\text{G}]\text{EBVred}})$.

Results and Discussion

Table 1 presents validation results for the first lactation milk yield in 73 FIC cows obtained using the PPV and LRV approaches. In the PPV approach, the highest validation reliability (R^2_w) was obtained for the $\text{ssGTBLUP}_{\text{FIC+RDC}}$ model (0.52), as expected. Elimination of the RDC data from the single-step model ($\text{ssGTBLUP}_{\text{FIC}}$) led to a 0.12 decrease in R^2_w (0.40). In both data scenarios, the single-step models showed higher validation reliabilities than their corresponding animal models. The increase was 0.12 and 0.07 for RDC+FIC and FIC data, respectively. The regression coefficient (b_1), indicating the bias, was over 1.0 for the $\text{ssGTBLUP}_{\text{FIC+RDC}}$ model due to underprediction of GEBV differences by the reduced data. In the LRV approach, the b_1 coefficient behaved like in PPV and the value was over one (1.05) in the $\text{ssGTBLUP}_{\text{FIC+RDC}}$ model. The highest correlation between the full and the reduced data predictions was observed in ssGTBLUP$_{\text{FIC+RDC}}$ suggesting a higher predictive ability of the model compared to the other methods. For protein (Table 2), in general, the R_w^2 values were higher than in milk. The b$_1$ coefficients were over one in PPV for both genomic models (1.07 and 1.03). In the LRV approach, b$_1$ was close to one in all models and the highest coefficient of correlations (cor) were in models with genomic data (0.83 and 0.78). The cor values for milk and protein were alike.

Table 1. Results from validation test in 73 FIC cows for 305d first lactation milk.

	PPV[1]			LRV[2]		
Model	b_0[3] (kg)	b_1[4]	R_w^2[5]	b_0 (kg)	b_1	cor[6]
ssGTBLUP$_{\text{FIC+RDC}}$	-60	1.07	0.52	1.2	1.05	0.84
ssGTBLUP$_{\text{FIC}}$	-34	1.01	0.40	-21	0.99	0.78
PBLUP$_{\text{FIC+RDC}}$	-52	0.99	0.40	2.2	0.99	0.73
PBLUP$_{\text{FIC}}$	-20	0.92	0.33	-13	0.96	0.70

[1] Prediction of full data YD by reduced data [G]EBV using linear model YD – $\mu_{\text{[G]EBVred}}$ = b$_0$ + b$_1$([G]EBVred - $\mu_{\text{[G]EBVred}}$)
[2] Prediction of full data [G]EBV by reduced data [G]EBV using linear model [G]EBV – $\mu_{\text{[G]EBVred}}$ = b$_0$ + b$_1$([G]EBVred - $\mu_{\text{[G]EBVred}}$)
[3] b_0 = the bias
[4] b_1 = the regression coefficient
[5] R_w^2 = the coefficient of determination from PPV adjusted by the ERC
[6] cor = coefficient of correlation

The addition of RDC phenotypic data to FIC evaluations improved the prediction reliability of the FIC traditional model. Further augmentation of the model by FIC and RDC genotypes enhanced prediction even further.

Table 2. Results from validation test in 73 FIC cows for 305d first lactation protein.

	PPV[1]			LRV[2]		
Model	b_0[3](kg)	b_1[4]	R_w^2[5]	b_0 (kg)	b_1	cor[6]
ssGTBLUP$_{\text{FIC+RDC}}$	-1.91	1.07	0.61	-0.03	0.99	0.83
ssGTBLUP$_{\text{FIC}}$	-0.11	1.03	0.49	0.14	0.98	0.78
PBLUP$_{\text{FIC+RDC}}$	-2.09	1.00	0.44	-0.23	0.98	0.72
PBLUP$_{\text{FIC}}$	0.19	0.98	0.42	0.32	0.97	0.71

[1] Prediction of full data YD by reduced data [G]EBV using linear model YD – $\mu_{\text{[G]EBVred}}$ = b$_0$ + b$_1$([G]EBVred - $\mu_{\text{[G]EBVred}}$)
[2] Prediction of full data [G]EBV by reduced data [G]EBV using linear model [G]EBV – $\mu_{\text{[G]EBVred}}$ = b$_0$ + b$_1$([G]EBVred - $\mu_{\text{[G]EBVred}}$)
[3] b_0 = the bias
[4] b_1 = the regression coefficient
[5] R_w^2 = the coefficient of determination from PPV adjusted by the ERC
[6] cor = coefficient of correlation

Both PPV and LRV approaches gave similar results for the analyzed data. The unfit of the PPV for complex data sets (low heritability, indirect genetic values) discussed in Legarra and Reverter (2018) was not applicable for our research. However, the performance of the LRV approach is computationally less challenging compared to PPV and could be prioritized. Implementation of the ssGTBLUP model for FIC and RDC Test-Day data is the next development step. In the analysis of the test day data, the LRV approach allows a straightforward way to perform prediction validation.

Conclusions

A multi-breed single-step model for FIC and RDC was successfully used in genomic prediction using 305-day first lactation milk and protein data. The genomic model with FIC and RDC data showed higher prediction reliability than using FIC data only.

Acknowledgements

We acknowledge Project “Genomic evaluations for Western Finncattle” financed by Luke (Finland) and Finncattle Foundation (Finland). Data providers: Finnish Breeder Association (FABA, Finland), Nordic Cattle Genetic Evaluation (NAV, Denmark), and VikingGenetics (Denmark).

References

Christensen O.F., and Lund M. S. (2010) Genetics Selection Evolution, 42, 2. https://doi.org/10.1186/1297-9686-42-2

Koivula M., Strandén I., Aamand G.P., and Mäntysaari E.A. (2021) J. Dairy Sci. 104(9):10049-10058. https://doi.org/*10.3168/jds.2020-1982*

Kudinov A.A., Koivula M., Strandén I., Aamand G.P., and Mäntysaari E.A. (2021) Interbull Bulletin, 56, 174-179.

Kudinov A.A., Mäntysaari E.A., Pitkänen T.J., Saksa E.I., Aamand G.P. et al. (2021) J. Animal Breed. Genet. 00, 1– 12. https://doi.org/10.1111/jbg.12660

Legarra A., Christensen O.F., Vitezica Z.G., Aguilar I., and Misztal I. (2015). Genetics, 200(2), 455–468. https://doi.org/10.1534/genetics.115.177014

Legarra A., and Reverter A. (2018). Genetics Selection Evolution 50:53. https://doi.org/10.1186/s12711-018-0426-6

Mäntysaari E.A., Liu Z., and VanRaden P.M. (2010) Interbull Bulletin, 41,17–22.

Mäntysaari, E.A., Evans R.D., and Strandén, I. (2017) J. Animal Sci. 95:4728–4737, https://doi.org/10.2527/jas2017.1912

Přibyl J., Madsen P., Bauer J., Přibylová J., Šimečková M. et al. (2013) J. Dairy Sci. 96:1865-1873. https://doi.org/10.3168/jds.2012-6157.

Song H., Zhang J., Zhang Q., and Ding X. (2019) Frontiers in Genetics, 9, 730. https://doi.org/10.3389/fgene.2018.00730

Strandén I., and Lidauer M. (1999) J. Dairy Sci. 82:2779–2787. https://doi.org/10.3168/jds.S0022-0302(99)75535-9

VanRaden P.M. (2020) J. Dairy Sci. 103:5291–5301. https://doi.org/10.3168/jds.2019-17684