Single-step genomic predictions of a minor breed, concurrently with the national genomic evaluations of main breeds.

***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), Animal genetics, Myllytie 1, FI-31600 Jokioinen, Finland
[2] Nordic Cattle Genetic Evaluation, Agro Food Park 15, DK-8200 Aarhus N, Denmark

Abstract

Farmers, maintaining indigenous cattle breeds, are typically lacking accurate (G)EBV for their animals. Finnish local cattle breed (Finncattle, FIC) in 2019 presented less than 1% of milk recorded cows of the country. Breeding values for FIC are calculated jointly with Red Dairy Cattle (RDC) of Finland, Denmark, and Sweden (DFS). To perform joint GEBV prediction for FIC and RDC breeds, we propose a single-step GTBLUP approach with metafounders (MF). We used genomic data from 917 FIC and 168 476 RDC animals and test-day milk records from FIC and RDC cows. Originally assigned 137 unknown parent groups were replaced by 137 MF with and the meta-founder relationships were derived using co-variance functions.

Key words: ssGTBLUP, metafounders, co-variance function, Finncattle, Red Dairy cattle

Introduction

Genetic and genomic prediction is common in large dairy cattle breeds (e.g., Holstein, Red Dairy cattle, or Jersey). The situation is the opposite for small breeds with limited genomic data available. Insufficient phenotypic and genomic information can be considered a synonym for bias, overprediction, and low accuracy (Andonov et al., 2017). An approach to overcome the limitation is to perform genomic prediction of a small breed jointly with a genetically related large breed.

Finncattle is the indigenous breed of Finland presented by less than 1% of milk recorded cows in the country (Soini et al., 2019). However, the current breed is expected to share genetics with Finnish and Swedish Ayrshire and Friesian cattle due to an open herd book, fourth generation cross is accepted as a pure breed. The routine genetic evaluation of the breed is performed by NAV (Nordic Cattle Genetic Evaluation, Denmark) jointly with Nordic (Denmark, Finland, and Sweden) Red Dairy cattle (RDC) and Finnish Holstein (HOL) since 2006. However, genomic prediction for FIC is not yet available as FIC breeders and farmers just recently started to collect genotypes from FIC cows and bulls. Current genotyping has focused on the western subpopulation which are the major part of FIC animals.

Genomic prediction in Nordic dairy cattle is currently performed as a two-step approach (VanRaden, 2008), but many efforts have been done to move to the single-step approach (Mäntysaari et al., 2020). Theoretically, a single-step method (Aguilar et al., 2010; Christensen and Lund, 2010) and, recently proposed, metafouders (Legarra et al, 2015) would allow joint FIC and RDC genomic prediction in a sophisticated way. However, some worry about a possible drop in the quality of genomic prediction in RDC due to presence of FIC genotypes in reference population exist. The original MF approach should also be adjusted in a way to include as many MF as unknown parent groups (UPG).

Aim of the current study was to: investigate ways to extend the number of MF to the same as UPG; perform single-step genomic prediction using RDC and FIC phenotypes and genotypes simultaneously; and see if the inclusion of FIC genotypes has negative impacts on quality of RDC genomic prediction.

Materials and Methods

Data

Phenotypic data and pedigree were obtained from the August 2020 NAV RDC production traits evaluation. Protein and milk yield test-day records were available from 3.6 million RDC, 0.86 million HOL, and 30 thousand FIC cows. The reduced data set was created by omitting the records collected in 2017-2020 in order to assess prediction ability of the models. All the breeding values were estimated using the official NAV test day model with 27 traits: milk, fat, protein x 3 lactations x 3 countries.

Full pedigree included 4.6 million RDC, 1 million HOL, and 34.6 thousand FIC cows and 76.6 thousand RDC, 22.5 thousand HOL, and 1.5 thousand FIC bulls. Truncated pedigree was created for estimation of base population allele frequencies (AF) by keeping only genotyped individuals and one generation of their parents. The genomic dataset included 168 476 RDC and 917 FIC animals with 46 914 markers per genotype available. Imputation and quality control of the genotypes were done by NAV.

UPG and MF

In the truncated pedigree, unknown parents were replaced by a set of 20 groups. The set was only used in estimation of base population allele frequencies and to compute the “regular” gamma matrix ($\boldsymbol{\Gamma}_{\mathbf{20}}$) required in the metafounder approach. The groups were formed as country × breed (or just breed) by time intervals: Finnish, Swedish, and Danish RDC (FIN RDC, SWE RDC, and DNK RDC in <1990,1990-2000,and >2000) = 9 groups; RDC from other countries (RDC OTHER in <2000 and ≥2000) = 2 groups; FIC (FIC in <1980,1980-1990, and >1990) = 3 groups; other breeds (OTHER <2000 and ≥2000) = 2 groups; HOL (HOL <1960,1960-1980,1981-2000, and >2000) = 4 groups.

In the full pedigree the UPG defined by NAV were replaced by set of 137 groups. The 137 UPG were formed based on breed, country, selection path, and sex and birth decade. In the genomic prediction the set were considered as either UPGs or MFs.

Gamma matrix

The base population AF for RDC, FIC, and OTHER groups were computed using RDC and FIC genotypes in BPOP program using GLS model (Strandén and Mäntysaari, 2020). HOL AFs were estimated using HOL genotypes by M. Koivula (personal communication). Markers with minor allele frequency ≤ 0.05 by breed were deleted, and only common markers for HOL, RDC, and FIC breeds were selected. Obtained marker set of 40,536 markers was used to estimate $\boldsymbol{\Gamma}_{\mathbf{20}}$. The $\boldsymbol{\Gamma}_{\mathbf{20}}$ matrix was computed as $8 * cov(\boldsymbol{P})$, where $\boldsymbol{P}$ is m by n matrix of AF with m = number of SNPs and n = number of base populations (groups). The $\boldsymbol{\Gamma}_{\mathbf{20}}$ was needed to predict $\boldsymbol{\Gamma}_{\mathbf{137}}$.

The estimation of $\boldsymbol{\Gamma}_{\mathbf{137}}$ was done using co-variance function described in Tijani et al. (1999): $\boldsymbol{\Gamma}_{\mathbf{137}} = \Phi_{137}\mathbf{K}\Phi'_{137}$, where Φ_{137} is the model matrix describing the groups and $\mathbf{K}$ is a matrix of co-variance function coefficients estimated as $\mathbf{K} = (\Phi'_{20}\ \Phi_{20})^{-1} * \Phi'_{20}\boldsymbol{\Gamma}_{\mathbf{20}}\Phi_{20} * (\Phi'_{20}\ \Phi_{20})^{-1}$. Here Φ_{20} is the model matrix functions proposed for given (20) MF.

Statistical model

Genetic prediction was done using the following models: 1) single-step GTBLUP (Mäntysaari et al. 2017) with 137 UPGs; 2) single-step GTBLUP with 137 MFs; and 3) original TD BLUP animal model (Lidauer et al., 2006).

ssGTBLUP UPG model included genomic relationship matrix (**G**) built with residual polygenic effect = 30% and a base population allele frequency = 0.5. Diagonal of **G** was scaled by trace($\mathbf{A}_{\mathbf{22}}$)/trace(**G**). Inbreeding was accounted in the inverse of pedigree relationship matrix ($\mathbf{A}^{-1}$) and submatrix of genotyped animals ($\mathbf{A}_{22}$), and a full QP transformation for UPG was used (Matilainen et al., 2018).

ssGTBLUP MF model included **G** matrix build with the same assumptions as in *ssGTBLUP*

UPG, except the scaling. The inverses of $\mathbf{A}$ and $\mathbf{A}_{22}$ matrices were build using $\boldsymbol{\Gamma}_{\mathbf{137}}$ and MF inbreeding was accounted.

TD BLUP model had 137 UPGs and inbreeding accounted in the inverse of pedigree relationship matrix ($\mathbf{A}^{-1}$).

Computations were done with MiX99 software (Strandén & Lidauer, 1999). Validation of genomic prediction was performed by regression of genomic estimated breeding values (GEBVs) obtained using the full data on the corresponding GEBVs from the reduced data (Legarra and Reverter, 2018). Criteria for selection of validation bulls was >20 daughters with records in the full and no daughters in the reduced data. The set of validation cows included cows with at least one record in the full and no records in the reduced data sets.

Results & Discussion

Gamma and relationship matrices.

Figure 1 has the Γ_{20} matrix as a heatmap plot. The dull red color implies fair kinship between the MF, in opposite the bright red color implies high kinship. FIC MFs were barely related to modern HOL and OTHER MFs. The relationships between FIC and RDC MFs were alike across all time intervals. Expectedly HOL MFs had the lowest kinship with the other groups.

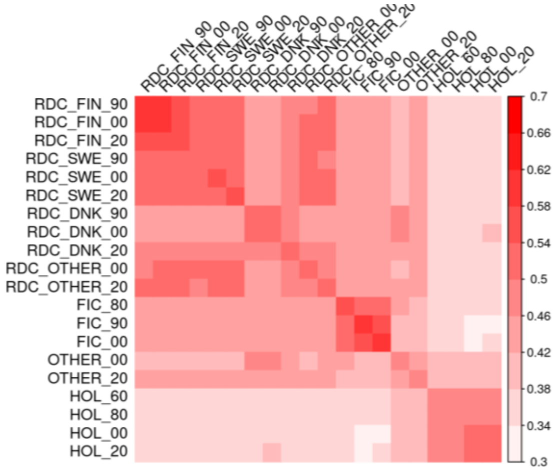

Figure 1. Heatmap plot of the Γ_{20} matrix. Diagonal = self-relationship of the MFs; off-diagonals are relationship between MFs.

The $\mathbf{K}$ matrix used to compute the $\boldsymbol{\Gamma}_{\mathbf{137}}$ matrix is in Appendix I. The $\boldsymbol{\Gamma}_{\mathbf{137}}$ matrix (Figure 2) had a structure replicating, to some extent, the structure of the $\boldsymbol{\Gamma}_{\mathbf{20}}$ matrix.

Average diagonal elements of the $\mathbf{A}_{22}$, $\mathbf{A}_{22}^{\Gamma_{137}}$, and $\mathbf{G}_{05}$ by birth year of genotyped animals are presented in Figure 3. Use of the $\boldsymbol{\Gamma}_{137}$ matrix lifted the diagonal elements of the $\mathbf{A}_{22}$ matrix closer to those of $\mathbf{G}_{05}$. Correlation between the diagonal elements of $\mathbf{G}_{05}$ and $\mathbf{A}_{22}$ increased from 0.51 to 0.71 after augmentation by the $\boldsymbol{\Gamma}_{137}$ matrix. In FIC, correlation between the off-diagonal elements of $\mathbf{G}_{05}$ and $\mathbf{A}_{22}$ increased from 0.63 to 0.67 after augmentation by the $\boldsymbol{\Gamma}_{137}$ matrix, but between the diagonal elements remained unchanged.

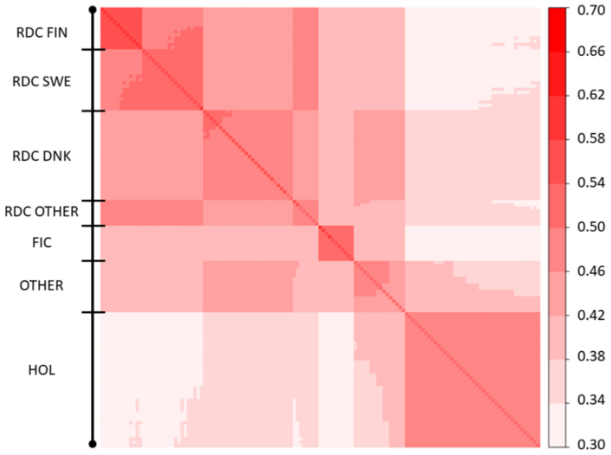

Figure 2. Heatmap plot of the Γ_{137} matrix. Diagonal = self-relationship of the MFs; off-diagonals are relationship between MFs.

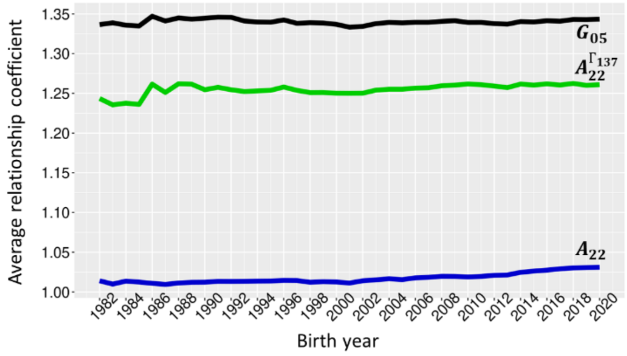

Figure 3. Average diagonal element of the relationship matrices ($\mathbf{A}_{22}$, $\mathbf{A}_{22}^{\Gamma_{137}}$, and $\mathbf{G}_{05}$) by birth year of the genotyped animals.

Genomic prediction and validation.

Table 1 has validation results for genomic prediction in FIC bulls for protein and milk yields. Model predictive values (R^2) were the same for the *ssGTBLUP UPG* and the *ssGTBLUP MF* models. The highest $b_1$value in the protein was obtained for *ssGTBLUP UPG.* Surprisingly high R^2 and b_1 were attained for the *TD BLUP* model, especially in milk. This may be due to the low number of a candidate bulls used to perform the regression analysis.

In the cows set (Table 2), the highest coefficient of determination in protein and milk was obtained for the *ssGTBLUP MF* model.

Table 1. Validation results in the 21 FIC bulls for protein and milk yield (G)EBV predicted using *ssGTBLUP UPG*, *ssGTBLUP MF*, and *TD BLUP* models ($GEBV_{UPG}$, $GEBV_{MF}$, and EBV_{UPG}).

	Model	MDiff	b_1(±SE)	R^2
Protein	$GEBV_{UPG}$	-4.2	0.90 (±0.2)	0.66
	$GEBV_{MF}$	-4.0	0.79 (±0.1)	0.66
	EBV_{UPG}	-3.1	0.82 (±0.2)	0.53
Milk	$GEBV_{UPG}$	-195	0.80 (±0.2)	0.60
	$GEBV_{MF}$	-177	0.92 (±0.2)	0.60
	EBV_{UPG}	-203	0.93 (±0.1)	0.69

MDiff = mean (G)EBV from full data minus (G)EBV from reduced data;
$\boldsymbol{b_1}$ = the regression coefficient;
$\boldsymbol{R^2}$ = the coefficient of determination of LR-model (Legarra and Reverter 2018).

Table 2. Validation results in the 109 FIC cows for protein and milk yield (G)EBV predicted using *ssGTBLUP UPG*, *MF*, and *TD BLUP* models ($GEBV_{UPG}$, $GEBV_{MF}$, and EBV_{UPG}).

	Model	MDiff	b_1(±SE)	R^2
Protein	$GEBV_{UPG}$	3.1	0.89 (±0.1)	0.48
	$GEBV_{MF}$	3.5	0.83 (±0.1)	0.50
	EBV_{UPG}	4.3	0.79 (±0.2)	0.32
Milk	$GEBV_{UPG}$	66	1.04 (±0.1)	0.59
	$GEBV_{MF}$	76	0.99 (±0.1)	0.61
	EBV_{UPG}	90	0.94 (±0.1)	0.48

MDiff = mean (G)EBV from full data minus (G)EBV from reduced data;
$\boldsymbol{b_1}$ = the regression coefficient;
$\boldsymbol{R^2}$ = the coefficient of determination.

Because performance comparison of genomic prediction in RDC breed was not among the aims of the current study, we have not presented validation results for those. However, to describe how much the presence of FIC genotypes in joint evaluations affects the RDC GEBVs, *ssGTBLUP UPG* model was run also without the FIC genotypes. The correlation of GEBVs from both the models for RDC AI bulls was >0.999 (Figure 4). As was expected, FIC genotypes did not bias RDC evaluations. The proportion of genotyped RDC animals would always be hundred times higher than FIC. Thus, RDC animals would not be much affected even if the number of FIC genotypes increases. Joint evaluation of RDC and FIC leads to high impact of RDC genomic information which may cause false-positive overprediction in FIC.

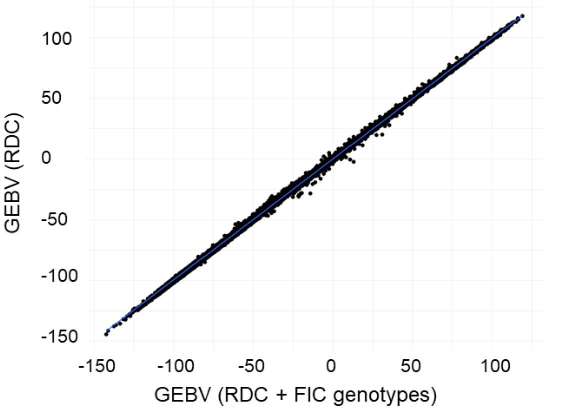

Figure 4. Scatter plot of RDC bulls milk GEBVs predicted using *ssGTBLUP UPG* model with RDC and FIC genotypes and *ssGTBLUP UPG* model with RDC genotypes only.

The current study showed that co-variance function allows to use the same number of MF as UPG. Thus, the same groups can be used in MF as UPG in a routine evaluation. The MF approach gave slightly higher validation reliability then UPG with full QP transformation. However, correlation between diagonals of $\mathbf{G}_{05}$ and $\mathbf{A}_{22}$ after use of Γ_{137} increased for RDC but not for FIC animals. The allele frequency change in time dictated by Φ_{20} matrix presumably assigned overly strict time trend to FIC animals. Further work for the better approach to estimate $\mathbf{\Gamma}_{20}$ matrix seems justified.

Conclusions

The results showed that the use of co-variance functions to get same amount of MFs as UPGs is feasible. The MF approach showed slightly higher R^2 than the UPG approach. Influence of FIC genotypes on RDC GEBVs was not detected. This suggests that presence of FIC genotypes would not harm RDC single-step evaluations.

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

Andonov, S., Lourenco, D.A.L., Fragomeni, B.O., Masuda, Y., Pocrnic, I., Tsuruta, S., and Misztal, I. 2017. Accuracy of breeding values in small genotyped populations using different sources of external information—A simulation study. Journal of Dairy Science, 100(1): 395-401. https://doi.org/10.3168/jds.2016-11335.

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and 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(2), 743-752. https://doi.org/10.3168/jds.2009-2730.

Christensen, O. F., and Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. Genetics Selection Evolution, 42, 2. https://doi.org/10.1186/1297-9686-42-2.

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., and Misztal, I. (2015). Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. Genetics, 200 (2), 455–468. https://doi.org/10.1534/genetics.115.177014

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

Lidauer, M., Pedersen, J., Pösö, J., Mäntysaari, E.A., Strandén, I., Madsen, P., Nielsen, U.S., Eriksson, J.-Ä., Johansson, K., and Aamand, G.P. 2006. Interbull Bulletin 35, 103-108.

Matilainen, K., Strandén, I., Aamand, G. P., and Mäntysaari, E. A. 2018. Single step genomic evaluation for female fertility in Nordic Red dairy cattle. Journal of Animal Breeding and Genetics,135(5),337–348. https://doi.org/10.1111/jbg.12353

Mäntysaari, E.A., Evans R.D., and Strandén, I. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals, Journal of Animal Science, Volume 95, Issue 11, November 2017, Pages 4728–4737, https://doi.org/10.2527/jas2017.1912

Mäntysaari, E. A, Koivula, M., and Strandén, I. (2020). Symposium review: Single-step genomic evaluations in dairy cattle. Journal of Dairy Science, 103(6), 5314-5326. https://doi.org/10.3168/jds.2019-17754

Soini, K., Pouta, E., Latvala T., and Lilja, T. 2019. Agrobiodiversity Products in Alternative Food System: Case of Finnish Native Cattle Breeds. Sustainability, 11(12), 3408. https://doi.org/10.3390/su11123408

Strandén, I., and Lidauer, M. 1999. Solving large mixed models using preconditioned conjugate gradient iteration. Journal of Dairy Science, 82:2779–2787.

Strandén, I., and Mäntysaari, E. A. 2020. Bpop: an efficient program for estimating base population allele frequencies in single and multiple group structured populations. Agricultural and Food Science, 29(3), 166–176. https://doi.org/10.23986/afsci.90955

Tijani, A., Wiggans, G. R., Van Tassell, C. P., Philpot, J. C., and Gengler N. 1999. Use of (co) variance functions to describe (co)variances for test day yield. J. Dairy Sci. 82.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91, 4414-4423.

Appendix I. The matrix of co-variance function coefficients estimated as $\mathbf{K} = (\Phi_{20}' \, \Phi_{20})^{-1} * \Phi_{20}' \boldsymbol{\Gamma}_{\mathbf{20}} \Phi_{20} * (\Phi_{20}' \, \Phi_{20})^{-1}$

Time	RDC	RDC	RDC	RDC	FIC	OTHER	HOL
trend	FIC	SWE	DNK	OTHER			
0.0297	-0.0035	-0.0031	-0.0213	-0.0108	-0.0150	-0.0142	-0.0003
	0.5697	0.5201	0.4528	0.5091	0.4377	0.4274	0.3416
		0.5328	0.4618	0.5047	0.4298	0.4296	0.3411
			0.5161	0.4609	0.4376	0.4667	0.3719
				0.5046	0.4345	0.4379	0.3543
					0.5504	0.4243	0.3430
						0.4730	0.3918