Investigation On the Metafounder Concept in ssGBLUP Based On a Simulated Cattle Population

J. Himmelbauer,[a,b], H. Schwarzenbacher[a], C. Fuerst[a], B. Fuerst-Waltl[b]

[a]*ZuchtData EDV-Dienstleistungen GmbH, Dresdner Straße 89/B1/18, 1200 Vienna, Austria*

[b]*University of Natural Resources and Life Sciences, Vienna, Department of Sustainable Agricultural Systems, Institute of Livestock Sciences, Gregor-Mendel Str. 33, 1180 Vienna, Austria*

Corresponding author: himmelbauer@zuchtdata.at

Abstract

Single-step genomic best linear unbiased prediction (ssGBLUP) has become a popular tool for genetic evaluations in dairy cattle populations. The use of the metafounder (MF) concept allows better consideration of relationships within and between founder populations and ensures correct matching of pedigree and genomic relationships. This study investigates the use of the MF concept in a simulated dairy cattle population where the base population consists of two related and inbred founder populations. The objectives are to compare genetic evaluations with and without MF and to investigate different methods of estimating MF parameters (Γ).Results show that genetic evaluations using MF are less biased and less inflated compared to evaluations using unknown parent groups or not accounting for the different founder populations. However, testing different methods to estimate Γ revealed a tendency to overestimate the relationships within and between the founder populations, leading to an overestimation of pedigree relationships compared to the genomic relationships. In summary, the MF concept in ssGBLUP is superior in this simulated scenario with two founder populations, but care must be taken when estimating Γ to ensure consistency between pedigree and genomic relationships. In general, these findings highlight the importance of considering relationships within and between founder populations in single-step genetic evaluations.

Key words: ssGBLUP, metafounder, simulation, dairy cattle

Introduction

Single-step genomic best linear unbiased prediction (ssGBLUP) uses an integrated relationship matrix (H), which combines the pedigree based relationship matrix (A) and the genomic relationship matrix (G). For this purpose, both matrices are supposed to refer to the same base population (Christensen, 2012). Without dedicated measures, this is usually not the case in cattle populations. In practice, there are several methods to match G to A (Christensen, 2012; VanRaden, 2008; Vitezica et al., 2011). Legarra et al. (2015) published the concept of metafounders (MF), which follows the idea of adapting A to G. The basic ideas are to use allele frequencies equal to 0.5 for all SNPs in the calculation of G and to assign unknown parents in the pedigree to pseudo-individuals (metafounder, MF).

Thompson (1979) and Quaas (1988) introduced the concept of unknown parent groups (UPG), which account for genetic differences within subgroups in the base populations. Since then, UPG, also known as genetic groups or phantom parents, are widely used in animal breeding, because they allow incorporating animals with missing parents and diverse genetic background in the genetic evaluation. UPG may therefore have means different from zero, but are assumed to be non-inbred and unrelated, just as the base population. MF may be seen as an extension to this concept by introducing relationships within and across UPG (Legarra et al., 2015).

For the German-Austrian-Czech Fleckvieh population, the first genomic evaluation using the ssGBLUP approach was published in April 2021 (Himmelbauer et al., 2021). To account for unknown parents, 15 UPG are presently used for most of the fitness traits. MF is the current gold standard for ssGBLUP implementations as shown e.g. by Meyer et al. (2018) and will therefore likely be the next evolution step in the national genomic evaluation system. For reasons discussed above, the aim of this study is to test different methods for gamma estimation and to compare the difference between different genetic evaluations with and without MF for a very simple population structure with two base populations and without any unknown pedigrees.

Materials and Methods

Simulating metafounders

The basic approach for simulating the population is the same as that used and described in detail in Himmelbauer et al. (2023). The main difference, however, is that for this study not only one but two related and inbred base populations (MF) are simulated. To achieve this, the founder population is split after 2 500 generations of evolution. Both subpopulations are then selected for additional 15 generations based on the true breeding value (TBV) for trait 1, with subpopulation A selected for high and subpopulation B selected for low values of trait 1. The two subpopulations are then merged again, and a second trait (trait 2) is created with a heritability of 0.3 and a genetic correlation to trait 1 between 0.3 and 0.5. This is followed by 30 years of selection by pedigree BLUP (PBLUP) and 8 years of selection by ssGBLUP (ignoring the two separated base populations) based on trait 2 as described in Himmelbauer et al. (2023) with small adaptions: To ensure that at the end of the selection process phenotypes and genotypes of both purebred populations (A and B) and the crossbred population (AB) are available, animals are selected separately by subpopulation. Mating is controlled such that females from subpopulations A, B, and AB are mated with males from purebred populations A and B in a way that each possible combination of male and female subpopulations occurs with the same frequency in each simulated year. The schematic overview of the simulation approach is shown in Figure 1.

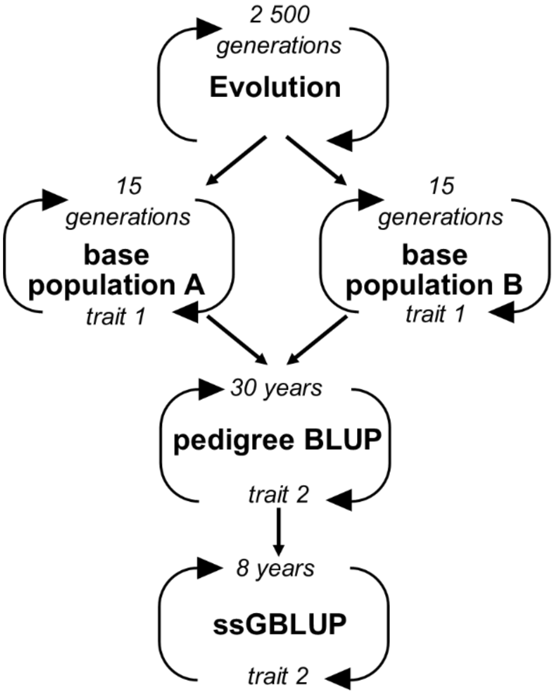

Figure 1. Schematic overview of simulation.

Dataset

The data set from the last year of the simulation serves as input for all further test runs. Basically, all females with offspring have a phenotype in the simulation. In routine datasets, phenotypes are usually not available back to the pedigree base, therefore 90% of the phenotypes from animals of the first 15 generations were randomly deleted. The final dataset consists of about 154 500 phenotypes, 204 900 genotypes and in total of about 1 105 500 animals in the pedigree.

Estimating Gamma Matrix

The true Gamma matrix ($\boldsymbol{\Gamma}$) was calculated using true allele frequencies in the base populations ($\boldsymbol{p_A}$ and $\boldsymbol{p_B}$) in the following formula derived in Garcia-Baccino et al. (2017)

$$\boldsymbol{\Gamma} = \mathbf{8} \cdot \boldsymbol{cov}(\boldsymbol{p_A}, \boldsymbol{p_B}).$$

Additionally, $\boldsymbol{\Gamma}$ was estimated using four different methods. Two methods are based on estimated base allele frequencies and equation (1). Base allele frequencies were estimated using the software Bpop (Strandén &

Mäntysaari, 2020b), which makes use of a generalized least square (GLS) method. For the first method (BFQ_pure), only genotypes of purebred animals of the two subpopulations were used to estimate $\boldsymbol{\Gamma}$. For the second method (BFQ_all), all genotypes of the final dataset, including all crossbred animals, were used. The third method (MM_pure) tested for estimating $\boldsymbol{\Gamma}$ corresponds to the method described in Legarra et al. (2015) as "Method of moments based on summary statistics for multiple pure populations" and again uses only genotypes from the purebred populations. The last method (MM_cross) is equivalent to the "Method of moments based on summary statistics for populations with crosses" and uses crossbred genotypes as described in Legarra et al. (2015).

Genetic evaluations

To evaluate the effect of inclusion of MF, several different genetic evaluations were tested with the same dataset. There are no unknown parents in the pedigree. Only the parents of the pedigree base are unknown and replaced with the true base populations. An exception is the evaluation without UPG, where the parents of the pedigree base are all set to zero.

1) PBLUP with two UPG (PED):

A simple pedigree BLUP, where the UPG were treated as random, was applied on the dataset. The evaluation was done using the commercial software package MiX99 (MiX99 Development Team, 2019).

2) ssGBLUP without UPG (no_UPG):

Breeding values were estimated based on a ssGBLUP with no UPG in the pedigree. All animals in the pedigree were traced back to one single pedigree base population. The preparation of the genomic relationship matrix (G) for ssGBLUP was done with the program HGINV (Strandén & Mäntysaari, 2020a) based on VanRaden's method 1 (VanRaden, 2008) with true base allele frequencies from the founder population and the approach for proven and young (Misztal et al., 2015). Details on the computation of the G-Matrix are the same as in Himmelbauer et al. (2023).

3) ssGBLUP with two UPG (UPG_qp):

This method is the same as no_UPG, described above, with the difference that here the true base populations were used as parents in the pedigree base. The two base populations were modeled as UPG and Quaas and Pollak (QP) transformed UPG were included in inverse G.

4) ssGBLUP with two MF and true $\boldsymbol{\Gamma}$ (MF_true):

The fourth evaluation is a ssGBLUP where the two base populations were modeled as MF. In this case the true $\boldsymbol{\Gamma}$ was used to define the relationships between the MF.

5) ssGBLUP with two MF and estimated $\boldsymbol{\Gamma}$ (MF_est):

This evaluation is equivalent to MF_true, but here an estimated $\boldsymbol{\Gamma}$ was used. The used $\boldsymbol{\Gamma}$ was estimated using strategy BFQ_all, described above.

6) ssGBLUP with two MF, true $\boldsymbol{\Gamma}$ and scaled variances (MF_sc):

This evaluation is the same as MF_true, but in this case, scaled variance components as proposed by Legarra et al. (2015) were used. The additive genetic variance was scaled using the following equation (Legarra et al., 2015):

$$\boldsymbol{\sigma^2_{related}} \approx \frac{\boldsymbol{\sigma^2_{unrelated}}}{\boldsymbol{1 + \frac{\overline{d\imath ag(\Gamma)}}{2} - \bar{\Gamma}}}$$

Analyzing results

All comparisons are based on 10 repetitions of the simulation described above. To evaluate the performance of the different methods to estimate Γ, the diagonal and off-diagonal values of the estimated Γ are compared to the corresponding values of the true Γ.

The comparison of the different evaluations is done using three validation measures based on the youngest animals born in the last year of the simulation. Firstly, the correlation between estimated breeding values (EBVs) and true breeding values (TBVs) is calculated. Secondly, the bias is calculated using the following formula

$$b = \overline{EBV} - \overline{TBV}.$$

Third, the regression coefficient of the following regression is used as a measurement of the dispersion:

$$TBV = b_0 + b_1 \cdot EBV + e.$$

Additionally, the estimates for the group estimators of the UPG and the MF are compared to evaluate the differences between the five evaluation methods. Because the level of the base populations varies across replicates, the estimated difference between the two base populations is compared with the true difference rather than the absolute values.

Results & Discussion

Gamma-matrix

The diagonal of $\boldsymbol{\Gamma}$ is a measure for the inbreeding in the metafounder populations. The true mean diagonal value in this study was 0.631, with values ranging between 0.622 and 0.645. There are also no systematic differences between the two MF within a replicate because both MF populations are the same size and have the same history of evolution.

Basically, all tested methods overestimate the inbreeding of MF, but the two methods based only on genotypes of purebred animals show a significantly higher overestimation (Figure 2, top).

The off-diagonal of Γ represents the relationship between the two MF. In this study, the true value is between 0.566 and 0.585 with an arithmetic mean over 10 repetitions of 0.575. Both methods based on base allele frequencies give a very good estimate of the true value, whereas the other two methods show a clear overestimation (Figure 2, bottom).

In combination, this means that the method BFQ_all is the best at estimating the true Γ in this study where two MF are simulated. This is in line with the results for one MF shown in Garcia-Baccino et al. (2017). An interesting conclusion from the comparison between BFQ_pure and BFQ_all is that genotypes from crossbred animals are very important in the estimation of base allele frequencies in this situation.

Results for UPG/MF

The mean true difference in the genetic level between the two base populations over all repetitions is 0.834 genetic standard deviations, but with a quite high variation between 0.604 and 1.051 genetic standard deviations. All metafounder evaluations slightly underestimate the difference between the base populations by about 0.025 genetic standard deviations, but also with a relatively high error variance

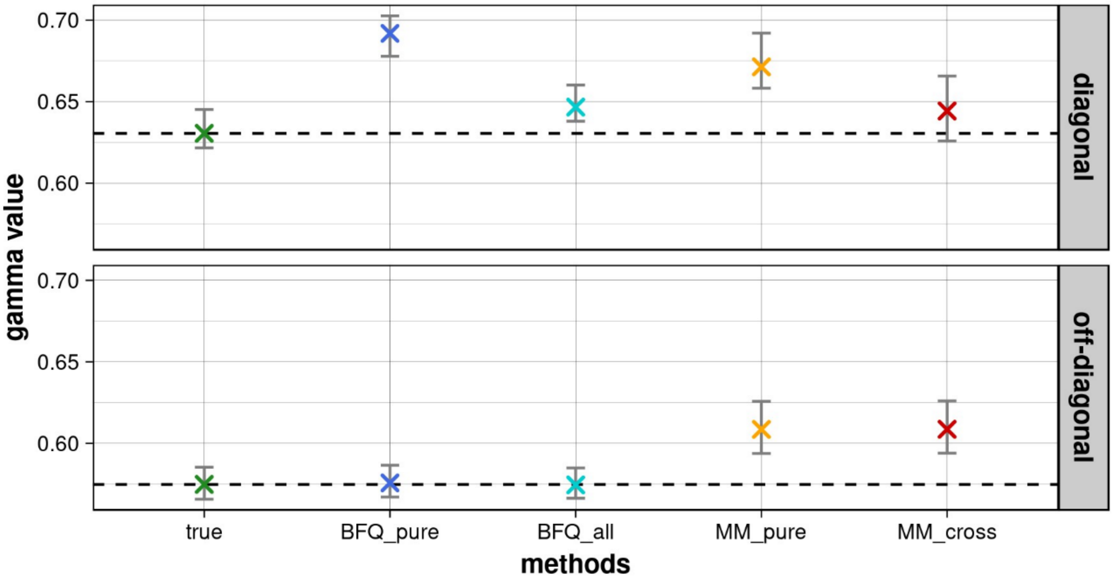

Figure 2. Comparison between true and estimated Γ for diagonal value and off-diagonal value separately. The error bars in the plot show the range from minimum to maximum and the "X" show the means over 10 repetitions. The dashed black lines indicate the true values.

between -0.16 and +0.18 (Figure 3). On average, the estimates from PED and UPG_qp are less biased, and the error variance for UPG_qp is also significantly lower than for the other estimates. This result is somehow surprising that a model with UPG can estimate the level difference of the base populations better than the MF models, one even with the true Γ matrix.

Correlation to true breeding value

The correlation of estimated breeding values (EBV) to true breeding values (TBV) for the youngest animals is more or less the same for all different evaluations (Figure 4, top). Only for breeding values from PED the correlation is substantially lower, as to be expected. Interestingly, there are hardly any differences in the correlation between no_UPG, UPG_qp and MF_true. There are already other studies on the use of MF in simulated and routine datasets and many of them report only small differences between evaluations with and without MF in terms of correlations or R^2 (Garcia-Baccino et al., 2017; Kudinov et al., 2022; Meyer, 2021). But unlike our findings, most studies report at least a slight improvement in correlation when using MF. This discrepancy may arise because our dataset uses MF exclusively at the pedigree base, without UPG or MF further along the pedigree. When MF are used in younger animals, the impact on correlation compared to UPG or not accounting for unknown parents in the final generation maybe becomes more pronounced than observed in our current study.

Bias

Regarding bias, the breeding values from PED show a significant downward bias of 0.627 genetic standard deviations, whereas the EBV from no_UPG and UPG_qp are on average slightly biased upwards by 0.08 and 0.04 genetic standard deviations, respectively (Figure 4, middle). The strong bias of EBV from PBLUP can be explained by the bias due to genomic preselection and was also observed in previous studies (Mäntysaari et al., 2018; Patry & Ducrocq, 2011). The EBV from MF_true and MF_est are mostly unbiased. Less biased results for evaluations with MF were also found in other publications (e.g. Garcia-Baccino et al., 2017). It is interesting to note that the breeding values from the MF model with scaled variance components are also slightly biased upward by about 0.04 genetic standard deviations.

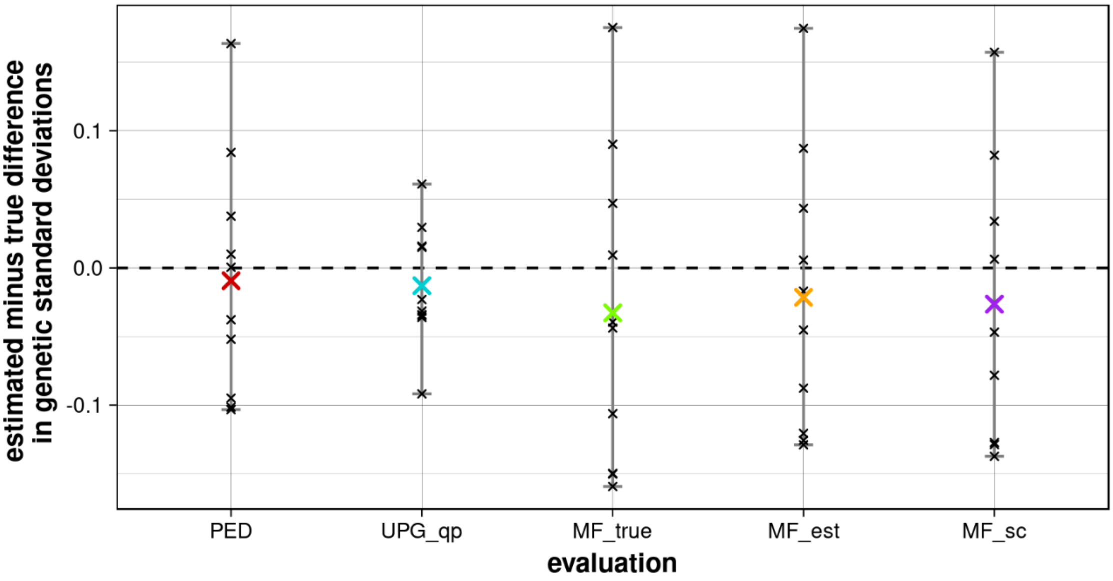

Figure 3. Estimated minus true difference between the genetic levels of the two base populations for different genetic evaluations. Results are given in genetic standard deviations. The error bars in the plot show the range from minimum to maximum and the capital colored "X" show the arithmetic means over 10 repetitions. The small "x" indicate the results for each repetition.

Dispersion

Another effect of genomic preselection is the clear overdispersion of EBV from PBLUP, resulting in a regression coefficient of 0.82. Similar results have also been reported in several publications (Mäntysaari et al., 2018; Patry & Ducrocq, 2009, 2011). EBV from no_UPG and UPG_qp and also MF_sc show an overdispersion with a regression coefficient of around 0.95. There is no difference in the dispersion between EBV from MF_true and MF_est. Both evaluations lead to EBV with a regression coefficient of around 1.01, meaning that there is neither over-, nor a notable underdispersion. Other studies have also shown that the use of MF has a positive effect on dispersion and leads to less inflated breeding values (Garcia-Baccino et al., 2017; Kudinov et al., 2022; Macedo et al., 2021; Meyer, 2021).

Further simulations and analyses (results not shown) have shown that the differences between the estimates depend strongly on the difference in genetic levels between the two base populations. In simulations where the level differences between the two base populations are smaller, the positive effects of the evaluations with MF on dispersion are not so clear. In that case UPG_qp or even no_UPG give comparable or even better results with respect to dispersion than models with MF. One explanation could be that in situations with minimal or no differences in the genetic level of the base groups, MF simulates a difference that is not present at the level of causative loci.

Effects of estimated Gamma-Matrix

As there are hardly any differences in the results for MF, correlation, bias and dispersion between MF_true and MF_est, it can be concluded that the small differences between true and estimated Γ have no notable effects on the validation statistics of the evaluation in this simulated dataset. However, in the present study there are only two MF, and these only used at the pedigree base without any younger unknown parents. In more complex data sets and especially in routine data sets with multiple and also younger MF, the differences between evaluations with estimated and true Γ are likely to be larger.

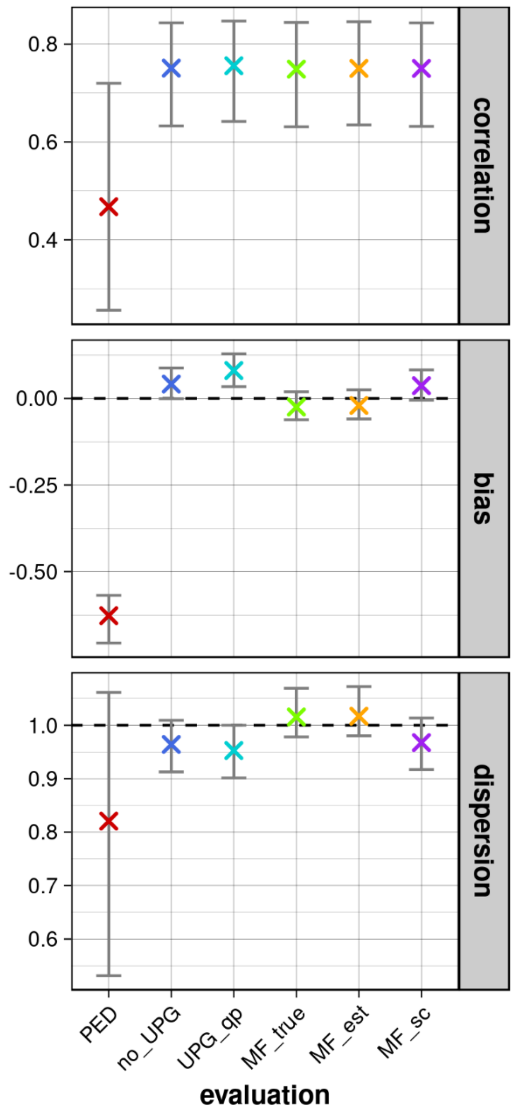

Figure 4. Results for correlation (top), bias (middle) and dispersion (bottom) for the youngest animals for different genetic evaluations. Bias (estimated minus true) is given in genetic standard deviations. The error bars in the plots show the range from minimum to maximum and the capital colored "X" show the arithmetic means over 10 repetitions.

Effects of scaling variance components

Applying the formula published in Legarra et al. (2015) on the true Γ and scaling the true variance components, results in a higher h^2. On average

$$\boldsymbol{\sigma^2_{related} \approx \frac{0.3}{0.713} = 0.421}$$

resulting in $h^2 = \frac{0.421}{0.421+0.7} = 0.376$ instead of 0.3. Using the scaled variance components in the ssGBLUP there are no remarkable differences between MF_true and MF_sc on the estimation of MF and the correlation in the validation group (Figure 3 and Figure 4, top). But compared to MF_true, scaled variance components lead to more bias and overdispersion (Figure 4, middle and bottom). These results are unexpected because it is derived in Legarra et al. (2015) that MF relatedness requires variance components to be adjusted. But there are already other authors reporting no positive or even negative effects of scaling variance components (Kudinov et al., 2022). Overall, the validation results (especially bias and dispersion) of the estimates with scaled variance components tend to show similar results to those found in other studies where the effect of an incorrect h^2 (in this case too high h^2) was investigated (Himmelbauer et al., 2023). This could be interpreted as suggesting that scaling the variance components in this case may lead to a too high h^2.

Conclusion

In summary, this study could show that already in a very simple situation with two base populations and otherwise complete pedigree, ssGBLUP with MF have significant positive effects on bias and dispersion in the youngest animal group compared to UPG. Regarding the estimation of the Γ, the method based on base allele frequencies proved to be the best method, with genotypes of crossbred animals playing an important role in the estimation of base allele frequencies. It is also interesting to note that scaling the variance components in this study did not improve the validation results, but worsened them.

But of course, it should be noted that this study uses very strong simplifications and rather optimal conditions compared to real applications. Therefore, further investigations with more MF and unknown pedigrees are necessary to be able to make statements that are more applicable to routine data.

References

Christensen, O. F. (2012). Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet.Select.Evol*, *44*(27). https://doi.org/10.1186/1297-9686-46-20

Garcia-Baccino, C. A., Legarra, A., Christensen, O. F., Misztal, I., Pocrnic, I., Vitezica, Z. G., & Cantet, R. J. C. (2017). Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. *Genet.Select.Evol*, *49*(34), 1–14. https://doi.org/10.1186/s12711-017-0309-2

Himmelbauer, J., Schwarzenbacher, H., & Fuerst, C. (2021). Implementation of single-step evaluations for fitness traits in the German and Austrian Fleckvieh and Brown Swiss populations. *Interbull Bulletin*, *56*, 82–89.

Himmelbauer, J., Schwarzenbacher, H., Fuerst, C., & Fuerst-Waltl, B. (2023). Comparison of different validation methods for single-step genomic evaluations based on a simulated cattle population. *J.Dairy.Sci*, https://doi.org/10.3168/jds.2023-23575

Kudinov, A. A., Koivula, M., Aamand, G. P., Strandén, I., & Mäntysaari, E. A. (2022). Single-step genomic BLUP with many metafounders. *Front.Genet*, *13*, 1012205. https://doi.org/10.3389/fgene.2022.1012205

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., & 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

Macedo, F. L., Astruc, J. M., Meuwissen, T. H. E., & Legarra, A. (2021). Removing data and using metafounders alleviates biases for all traits in Lacaune dairy sheep predictions. *J.Dairy.Sci*, *105*, 2439–2452. https://doi.org/10.3168/jds.2021-20860

Mäntysaari, E. A., Aamand, G. P., Strandén, I., & Koivula, M. (2018). Solutions for the fixed effects, yield deviations and daughter yield deviations from a data subject to genomic selection. Interbull Meeting 2018, Auckland, New Zealand. https://interbull.org/web/static/web/1500EsaM%C3%A4ntysaari.pdf

Meyer, K. (2021). Impact of missing pedigrees in single-step genomic evaluation. *Anim.Prod.Sci*, *61*(18), 1760–1773. https://doi.org/10.1071/AN21045

Meyer, K., Tier, B., & Swan, A. (2018). Estimates of genetic trend for single-step genomic evaluations. *Genet.Select.Evol*, *50*(39), 1–11. https://doi.org/10.1186/s12711-018-0410-1

Misztal, I., Fragomeni, B. O., Lourenco, D. A., Tsuruta, S., Masuda, Y., & Aguilar, I. (2015). Efficient inversion of genomic relationship matrix by the algorithm for proven and young (APY). *Interbull Bull*, *49*, 111–116.

MiX99 Development Team. (2019). *MiX99: A software package for solving large mixed model equations* (Release XI/2019, version 19.1129) [Computer software]. Natural Resources Institute Finland (Luke). url: http://www.luke.fi/mix99

Patry, C., & Ducrocq, V. (2009). Bias due to genomic selection. *Interbull Bulletin*, *39*, 77–82.

Patry, C., & Ducrocq, V. (2011). Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J.Dairy.Sci*, *94*(2), 1011–1020. https://doi.org/10.3168/jds.2010-3804

Quaas, R. L. (1988). Additive Genetic Model with Groups and Relationships. *J.Dairy.Sci*, *71*(5), 1338–1345. https://doi.org/10.1016/S0022-0302(88)79986-5

Strandén, I., & Mäntysaari, E. (2020a). *HGINV program* (Nov 2020, Version 0.993) [Computer software]. Natural Resources Institute Finland (Luke).

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

Thompson, R. (1979). Sire evaluation. *Biometrics*, *35*(1), 339–353. Scopus. https://doi.org/10.2307/2529955

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *J.Dairy.Sci 91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Vitezica, Z. G., Aguilar, I., Misztal, I., & Legarra, A. (2011). Bias in genomic predictions for populations under selection.*Genet.res*, *93*, 357–366. https://doi.org/10.1017/S001667231100022X