Modeling unknown parent groups or metafounders in single step genomic BLUP – results of a simulation study

J. Himmelbauer[1,2], H. Schwarzenbacher[1], C. Fuerst[1], B. Fuerst-Waltl[2]

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

[2]BOKU University, Vienna, Gregor-Mendel Str. 33, 1180 Vienna, Austria

Corresponding author: himmelbauer@zuchtdata.at

Abstract

The concepts considering for unknown parents are crucial in improving genetic evaluations in animal breeding by accounting for genetic differences within base populations. This study builds on a previous simulation study for the German-Austrian-Czech Fleckvieh population, presenting results that compare metafounders (MF) and unknown parent groups (UPG) for single-step genomic best linear unbiased prediction, and includes detailed analyses for scaling variance components when using MF. The results show that in both settings with complete and incomplete pedigree, evaluations using MF show the best bias and dispersion results, with minimal impact from incomplete pedigree information. In contrast, evaluations without UPG or MF and evaluations where UPG were incorporated via Quaas-Pollak-transformation in the pedigree-based and genomic relationship matrix (UPG_fullQP) exhibit substantial overestimation and overdispersion, emphasizing the importance of accurate relationship modeling in genetic evaluations. This study found that estimating variance components using MF and scaling variance components lead to the same heritability. However, using adapted variance components results in moderate overestimation and slight overdispersion of GEBV. The validation method based on the linear regression method could not detect the significant overestimation and overdispersion in UPG_fullQP. This means that commonly used validation methods tend to underestimate the advantages of MF in populations with numerous unknown pedigrees, highlighting challenges in model optimization for handling unknown parents.

Key words: ssGBLUP, unknown parents, metafounder, simulation, dairy cattle

Introduction

Thompson (1979) and Quaas (1988) published the concept of unknown parent groups (UPG) to account for genetic differences within subgroups of base populations, incorporating animals with missing parents and diverse genetic backgrounds into genetic evaluations. UPGs can have non-zero means but are assumed to be non-inbred and unrelated, similar to the base population. For single-step genomic best linear unbiased prediction (ssGBLUP) Legarra et al. (2015) extended this concept and introduced metafounder (MF), which can model relationships within and across subpopulations.

ssGBLUP uses an integrated relationship matrix (H), combining the pedigree-based (A) and genomic (G) relationship matrices. Ideally, both matrices should refer to the same base population (Christensen, 2012), though this is often not the case in cattle populations without adjustments. Methods to align G with A include those by VanRaden (2008), Vitezica et al. (2011), and Christensen (2012). MF is addressing this alignment by adapting A to match G.

In the German-Austrian-Czech Fleckvieh population, the first ssGBLUP genomic evaluation was published in April 2021 (Himmelbauer et al., 2021), using 15 UPGs for most fitness traits. MF is considered the gold

standard for ssGBLUP implementations (Meyer et al., 2018). Therefore, it is likely to be one of the next development steps in the national genomic evaluation system.

A small preliminary study for the case without unknown parents has already been published in Himmelbauer et al. (2023a). The detailed results based on a simulated cattle population based on two base populations, several scenarios and different pedigree settings were published in Himmelbauer et al. (2024). The results presented in this paper are in part a small selection from Himmelbauer et al. (2024) supplemented by more detailed analyses for scaling the variance components.

Materials and Methods

Simulating metafounders

The fundamental methodology employed for simulating the population is analogous to that described in Himmelbauer et al. (2024). The procedure begins by dividing the founder population into two subpopulations, with each subsequently selected independently. The populations are then reunited, forming the basis of the pedigree. From this point onwards, the pedigree is recorded, while the heritability (h^2) for the trait under selection is set to 0.3. Subsequently, a period of 30 years was simulated with selection based on PBLUP, followed by an 8-year period of selection based on ssGBLUP. Figure 1 provides a schematic representation of the simulation process.

Dataset

The entire simulation documented all pedigree information, phenotypes, genotypes, and TBV for all animals across all years. This data was used to create the study's dataset, described in detail by Himmelbauer et al. (2024) as "low pedigree completeness." In the simulation, all females with offspring had phenotypes. To mimic routine datasets, 90% of the phenotypes from the first 15 generations were randomly deleted. Additionally, 75% of males and 30% of females born in the last eight years were randomly genotyped. The final dataset includes approximately 154,500 phenotypes, 143,400 genotypes, and a total of around 1,105,500 animals in the pedigree.

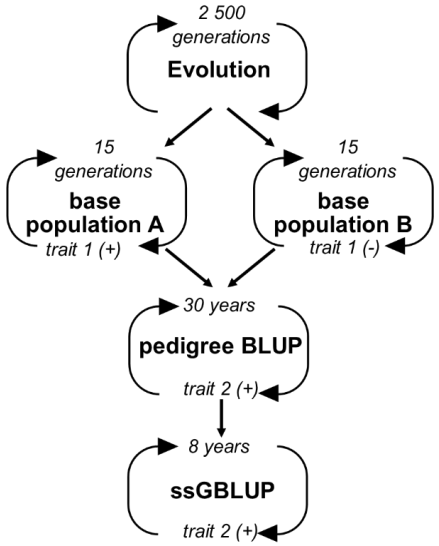

Figure 1. Schematic overview of the simulation process.

A reduced dataset was created for validation, using the same animals and genotypes but excluding the phenotypes from the last three years. Specifically, the phenotypes of all females born in years 32, 33, and 34 were excluded, resulting in 133,500 phenotypes in the reduced dataset.

For most analyses, some animals are assumed to have unknown sires and/or dams. The proportions of missing parents are 7.5% for sires and 10% for dams and are consistent across all birth years. Animals with unknown parents are randomly selected, but the potential for genomic parentage verification was considered, such that parents that can be identified with certainty (genotyped sires and dams of genotyped animals) or with a high probability (e.g., genotyped dam's sires of genotyped animals) are not deleted. This approach reflects practical scenarios and prevents double counting in genetic evaluations (Pimentel et al., 2022).

Pedigree settings

Two pedigree settings were tested, resulting in different classifications of unknown parents as UPG or MF.

Full pedigree

This setting uses the complete pedigree with no missing parents, except for animals born in year 0, forming the pedigree base. Base animals are assigned to their true subpopulations (purebred A or B), forming two UPG or MF.

True missing pedigree

This setting simulates unknown parents according to the previously described procedure. UPG or MF classification is based on subpopulation (purebred A, B, or crossbred AB), sex (missing sire or dam), and year of birth (grouped in five-year intervals). Since the full pedigree is known, true subpopulation and year of birth for missing parents are used.

Genetic evaluations

In order to test different methods of accounting for unknown parents, a series of genetic evaluations were conducted for the two pedigree settings. To calculate the estimated linear regression validation statistics (LR) (Legarra and Reverter, 2018), all evaluations were also computed for the truncated datasets. Except for the evaluation with scaled variances, we used the simulated genetic variance ($\sigma^2_{unrelated} = 0.3$) and for all evaluations, true base allele frequencies were used to construct the genomic relationship matrix.

All evaluations were conducted using MiX99 (MiX99 Development Team, 2019). The G matrix for ssGBLUP was prepared as in Himmelbauer et al. (2023b) using the HGINV program (Strandén and Mäntysaari, 2020), based on VanRaden's method 1 (VanRaden, 2008) and the approach for Proven and Young (Misztal et al., 2014a). For evaluations using the MF approach, base allele frequencies were set to 0.5, as outlined by Legarra et al. (2015).

1) ssGBLUP without UPG (no_UPG):

An ssGBLUP was used to estimate GEBV without UPGs. All unknown parents were set to 0, assigning them to a single base population.

2) ssGBLUP with UPG in A (UPG_alteredQP):

This ssGBLUP used UPGs in the pedigree, modeled as random. UPGs were included in the inverse pedigree relationship matrix (A^{-1}) and the inverse pedigree relationship matrix for genotyped animals (A_{22}^{-1}), but not in the inverse genomic relationship matrix (G^{-1}). This approach follows Masuda et al. (2018, 2022) and Strandén et al. (2022).

3) ssGBLUP with UPG in H (UPG_fullQP):

This method also used UPG and QP transformation was applied to A^{-1}, A_{22}^{-1} and G^{-1} as described in (Misztal et al., 2013).

4) ssGBLUP with MF and true Γ (MF_true):

In this ssGBLUP, unknown parents were represented by MF, with relationships defined by the true Γ. The variance-covariance matrix for breeding values was

$$var(u) = H_\Gamma \cdot \sigma^2_{unrelated},$$

where $\sigma^2_{unrelated}$ is 0.3 and H_Γ is the combined relationship matrix as described in Legarra et al. (2015).

5) ssGBLUP with MF, true Γ and scaled variances (MF_sc):

This evaluation is similar to MF_true, but with scaled variance components according to Legarra et al. (2015). The additive genetic variance was scaled using:

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

(Legarra et al., 2015). The variance-covariance matrix for breeding values was then

$$var(u) = H_\Gamma \cdot \sigma^2_{related}.$$

Estimation of variance components

The variance components were estimated using AIREML (Misztal et al., 2014b). The data used

correspond to those used for the "full pedigree" pedigree setting, i.e., a complete pedigree with two base populations and all phenotypes that were also used for all test runs analyzed. Genotypes were not used in the variance component estimation.

Two different approaches were tested. On the one hand, the variance components were estimated in the case that the relationships between the two base populations (Γ) were not taken into account, and on the other hand with consideration of (Γ) in the creation of A. In addition, the results from the first approach were scaled using the scaling method of Legarra et al. (2015) and compared with those results from the second approach.

Analyzing results

All comparisons are based on 10 repetitions of the previously described simulation.

True validation statistics:

Two measures, bias and dispersion, are used to compare the different evaluations. These are calculated using the youngest animals with genotypes born in the last year of the simulation, totaling 14,672 animals.

Bias, the mean difference between (G)EBV and TBV, is calculated as

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

Positive bias values indicate overestimation. Given that the genetic standard deviation for the trait is 1, the bias can be interpreted as genetic standard deviations.

Dispersion is measured by the regression coefficient b_1 from the regression:

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

where b_0 is the intercept, b_1 the regression coefficient and e the residuals.

Estimated validation statistics using linear regression (LR) method:

To obtain validation statistics, (G)EBV for certain validation animals based on a full dataset are compared with those from a reduced dataset. Two validation groups were defined: a male group (around 530 genotyped bulls born between years 30-32) and a female group (around 12,400 genotyped females born between years 32-34). Bulls in the male group have no daughters with records in the reduced dataset but at least 20 daughters in the full dataset. Cows in the female group have no phenotypes in the reduced dataset but have records in the full dataset.

Based on Himmelbauer et al. (2023b), the LR method accurately estimates bias, dispersion, and validation reliability (Legarra and Reverter, 2018; Macedo et al., 2020). Bias, the mean difference of GEBV between reduced and full datasets, is calculated as:

$$b = \overline{GEBV_r} - \overline{GEBV_f}$$

A bias of 0 indicates unbiased (G)EBV. Positive values indicate overestimation, and negative values indicate underestimation.

Dispersion is calculated as:

$$b_1 = \frac{\mathrm{cov}(GEBV_f, GEBV_r)}{\mathrm{var}(GEBV_r)}.$$

If $b_1 = 1$, there is no over- or underdispersion, $b_1 < 1$ indicates overdispersion, and $b_1 > 1$ indicates underdispersion.

Reliability is calculated as:

$$r^2 = \frac{\mathrm{cov}(GEBV_f, GEBV_r)}{\sigma_g^2},$$

where σ_g^2 is the true genetic variance in the validation group.

These statistics were calculated for both male and female animals across all pedigree settings, and genetic evaluations, and are based on 10 replicates.

Results & Discussion

Bias and dispersion

Figure 2 presents bias and dispersion results for the two pedigree settings and different evaluations. Regarding bias in the full pedigree setting, no_UPG, UPG_alteredQP, and UPG_fullQP show a slight overestimation of around 0.04 genetic standard deviations, with UPG_fullQP slightly higher at 0.07. MF with

true Γ underestimates by approximately 0.02 genetic standard deviations, while scaled variance components overestimate by 0.03 genetic standard deviations. In the true missing pedigree setting, no_UPG and UPG_fullQP exhibit substantial overestimation of 0.24 and 0.40 genetic standard deviations, respectively. MF evaluations show slight underestimation, with minimal impact from incomplete pedigree information.

Similar trends apply to dispersion (Figure 2, second row). In the full pedigree setting evaluations no_UPG, UPG_alteredQP, and UPG_fullQP show similar results regarding dispersion of around 0.96. MF_true and MF_est perform best with a regression coefficient of 1.00. However, scaling variance components with MF slightly worsens dispersion. In the true missing pedigree setting UPG_alteredQP maintains a similar dispersion level of 0.96, while no_UPG decreases to 0.93 due to incomplete pedigree data. UPG_fullQP shows the most significant impact, decreasing dispersion coefficients from 0.96 (full pedigree) to 0.74 (true missing pedigree). MF evaluations show consistent dispersion values, with MF_true at 1.00 and MF_sc at 0.97 for the true missing pedigree setting.

The evaluation using MF and true Γ shows the best bias and dispersion results in both pedigree settings, aligning with Bradford et al. (2019). Clear differences are observed in evaluations with and without UPG. The upward bias in no_UPG, along with overdispersion, arises because relationships in A, which only considers known relationships, do not match those in G, where all genomic relationships are fully considered.

UPG_fullQP exhibits significant bias and overdispersion due to double-counting when UPG is considered in G, despite G's complete genomic relationships. Similar issues have been identified in other studies (Bradford et al., 2019; Masuda et al., 2021; Meyer, 2021). UPG_alteredQP yields results similar to those with a complete pedigree because G accurately accounts for relationships and is unaffected by incomplete pedigrees.

All results presented here are also part of the study published by Himmelbauer et al. (2024). Beside a more detailed discussion of the presented results, in Himmelbauer et al. (2024) a comparison of these results with the results of a second scenario with less unknown parents, two additional pedigree settings and two additional genetic evaluations are presented.

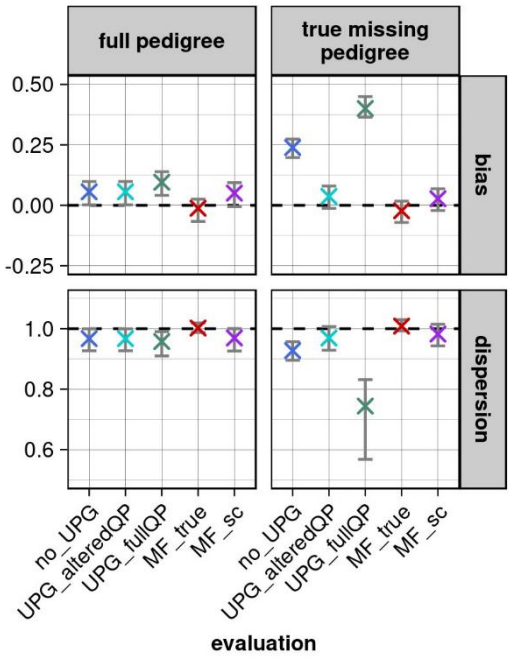

Figure 2. Comparison of true validation statistics (bias, dispersion) for 2 pedigree settings and 5 evaluation methods. The error bars in the plot show the range from minimum to maximum and the "x" show the means over 10 repetitions.

Scaling or estimating variance components

This study demonstrated that scaled variance components, compared to non-scaled ones, tend to result in a moderate overestimation rather than slight underestimation of GEBV. In terms of dispersion, scaled variance components have a negative effect, causing slight overdispersion. Similar effects were observed in a scenario with a complete pedigree and only two MF (Himmelbauer et al., 2023a). The effects of scaled variance components in this study are comparable, but to a lesser extent, with those reported by Himmelbauer et al. (2023b) in a

scenario with excessively too high heritability. This suggests that scaling may lead to a slightly too high heritability estimate.

Variance component estimation was performed to analyze this aspect in detail. Using A (pedigree relatedness without considering metafounder) it was possible to accurately estimate the simulated heritability (h^2) (Table 1 without MF/Γ). However, using A_Γ (pedigree relationship matrix considering MF relationships) resulted in an average h^2 of 0.3887, significantly higher than the simulated h^2. Yet, this h^2 corresponds closely to the scaled value calculated using the formula from Legarra et al. (2015), shown in Table 1 (scaled) as 0.3854. These analyses confirm that the scaling method of Legarra et al. (2015) works well and yields nearly identical results as estimating variance components with MF relationships. However, why the scaled h^2 provides slightly poorer validation results compared to unscaled h^2 remains unresolved.

Additional analyses in Himmelbauer et al., (2024) indicated that scaled variance components appear to influence the GEBV of the animals themselves but not the estimation of the MF effects.

Table 1: Results of variance component estimation with and without using MF and scaling variances.

	without MF/Γ		
	genetic variance	residual variance	h^2
mean	1.0663	2.3891	0.3088
min	0.9625	2.1878	0.2708
max	1.1142	2.6268	0.3224
	with MF/Γ		
	genetic variance	residual variance	h^2
mean	1.5212	2.3941	0.3887
min	1.3761	2.1918	0.3435
max	1.5941	2.6303	0.4045
	scaled		
	genetic variance	residual variance	h^2
mean	1.4970	2.3891	0.3854
min	1.3565	2.1878	0.3405
max	1.5644	2.6268	0.4006

Estimated validation statistics using LR method

Figure 3 displays the results of validation using the LR method for both pedigree settings and all genetic evaluations. Reliability shows minor variations among genetic evaluations, with those using MF performing slightly less effectively compared to other ssGBLUP evaluations. Notably, the validation does neither detect the full extent of the significant overestimation of no_UPG and UPG_fullQP, nor the extent of the pronounced overdispersion of UPG_fullQP in settings with incomplete pedigree. Validation statistics reveal no substantial differences between full and missing pedigree settings. In terms of bias, MF_true

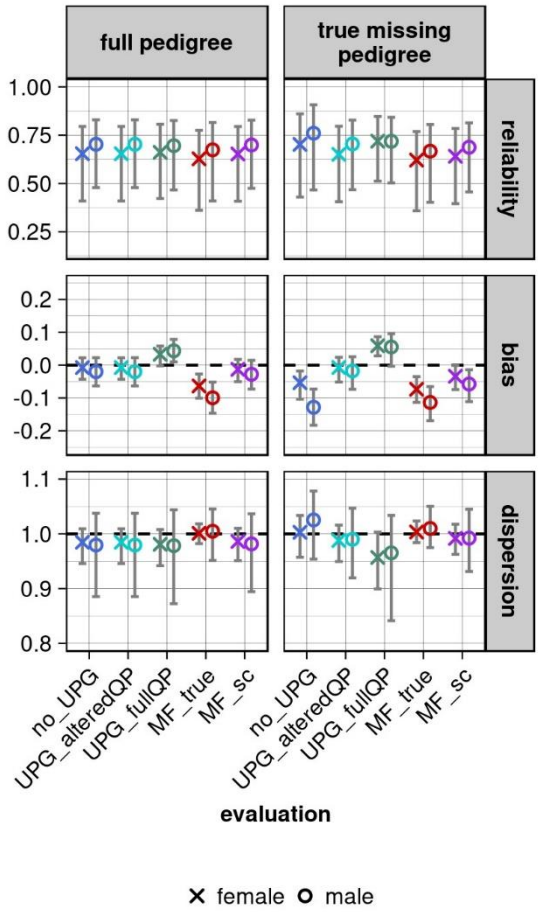

Figure 3. Comparison of estimated validation statistics (reliability, bias, dispersion) based on the LR method for 2 pedigree settings and 5 evaluation methods. The error bars in the plot show the range from minimum to maximum and the "x" and "o" show the means over 10 repetitions.

exhibits slight downward bias compared to other evaluations. However, regarding dispersion, both MF_true and MF_sc show a regression coefficient close to 1.00 across both pedigree settings, while no_UPG and UPG_alteredQP demonstrate slightly worse but still quite good results based on this validation.

The main conclusion is that also with this validation method evaluations using MF generally perform very well or at least better compared to other evaluation methods. However, differences in validation statistics are notably smaller than with true validation statistics. It is important to note that the significant bias and dispersion observed with UPG_fullQP is not detected by LR validation statistics. Such methods can only detect bias and dispersion if these issues are corrected in evaluations using complete datasets (Himmelbauer et al., 2023b). This is not the case here, as GEBV from UPG_fullQP appear nearly unbiased.

In summary, in populations with numerous unknown pedigrees, commonly used validation methods tend to underestimate the advantages of MF compared to other evaluations. This emphasizes the challenge of identifying an optimal model for dealing with unknown parents in practice.

Conclusions

In conclusion, the findings of this study indicate that MF has a positive effect on reducing bias and dispersion. The study highlights the potential of significant bias and dispersion when UPG is considered in an incorrect manner. Furthermore, the scaling of variance components was found to have a small detrimental effect on true validation statistics, rather than an enhancing one. The study also shows that the method of scaling variance components proposed by Legarra et al. (2015) leads to similar results as those obtained by estimating variance components using metafounder. Finally, the study identified limitations in the use of the LR method for assessing the effectiveness of MF in this context. These findings emphasize some of the challenges and outcomes associated with implementing MF in dairy cattle populations.

References

Bradford, H.L., Y. Masuda, P.M. VanRaden, A. Legarra, and I. Misztal. 2019. Modeling missing pedigree in single-step genomic BLUP. *Journal of Dairy Science* 102:2336–2346. https://doi.org/10.3168/jds.2018-15434.

Christensen, O.F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genetics Selection Evolution* 44:1–10. https://doi.org/10.1186/1297-9686-46-20.

Himmelbauer, J., H. Schwarzenbacher, and C. Fuerst. 2021. Implementation of single-step evaluations for fitness traits in the German and Austrian Fleckvieh and Brown Swiss populations. *Interbull Bulletin* 56:82–89. https://journal.interbull.org/index.php/ib/article/view/79/79.

Himmelbauer, J., H. Schwarzenbacher, C. Fuerst, and B. Fuerst-Waltl. 2023a. Investigation on the Metafounder Concept in ssGBLUP Based on a Simulated Cattle Population. *Interbull Bulletin* 59:124–131. https://journal.interbull.org/index.php/ib/article/view/1878/1882.

Himmelbauer, J., H. Schwarzenbacher, C. Fuerst, and B. Fuerst-Waltl. 2023b. Comparison of different validation methods for single-step genomic evaluations based on a simulated cattle population. *Journal of Dairy Science* 106:9026–9043. https://doi.org/10.3168/jds.2023-23575.

Himmelbauer, J., H. Schwarzenbacher, C. Fuerst, and B. Fuerst-Waltl. 2024. Exploring unknown parent groups and metafounders in single-step genomic BLUP: Insights from a simulated cattle population. *Journal of*

Dairy Science (uncorrected proof). https://doi.org/10.3168/jds.2024-24891.

Legarra, A., O.F. Christensen, Z.G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral Relationships using Metafounders: Finite Ancestral Populations and Across Population Relationships. *Genetics* 200:455–468. https://doi.org/10.1534/genetics.115.177014.

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution* 50:1–18. https://doi.org/10.1186/s12711-018-0426-6.

Macedo, F.L., A. Reverter, and A. Legarra. 2020. Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. *Journal of Dairy Science* 103:529–544. https://doi.org/10.3168/jds.2019-16603.

Masuda, Y., I. Misztal, P. VanRaden, and T. Lawlor. 2018. Genomic predictability of single-step GBLUP for production traits in US Holstein. Page 182 in Abstracts of the 2018 American Dairy Science. Journal of Dairy Science, Knoxville, Tennessee. URL: https://www.adsa.org/Portals/0/SiteContent/Docs/Meetings/PastMeetings/Annual/2018/ADSA2018_full_abstracts_book.pdf.

Masuda, Y., S. Tsuruta, M. Bermann, H.L. Bradford, and I. Misztal. 2021. Comparison of models for missing pedigree in single-step genomic prediction. *Journal of Animal Science* 99:1–10. https://doi.org/10.1093/jas/skab019.

Masuda, Y., P.M. VanRaden, S. Tsuruta, D.A.L. Lourenco, and I. Misztal. 2022. Invited review: Unknown-parent groups and metafounders in single-step genomic BLUP. *Journal of Dairy Science* 105:923–939. https://doi.org/10.3168/jds.2021-20293.

Meyer, K. 2021. Impact of missing pedigrees in single-step genomic evaluation. *Animal Production Science* 61:1760–1773. https://doi.org/10.1071/AN21045.

Meyer, K., B. Tier, and A. Swan. 2018. Estimates of genetic trend for single-step genomic evaluations. *Genetics Selection Evolution* 50:1–11. https://doi.org/10.1186/s12711-018-0410-1.

Misztal, I., Z.G. Vitezica, A. Legarra, I. Aguilar, and A.A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. *Journal of Animal Breeding and Genetics* 130:252–258. https://doi.org/10.1111/jbg.12025.

Misztal, I., A. Legarra, and I. Aguilar. 2014a. Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science* 97:3943–3952. https://doi.org/10.3168/jds.2013-7752.

Misztal, I., S. Tsuruta, D.A.L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs.

MiX99 Development Team. 2019. MiX99: A software package for solving large mixed model equations. url: http://www.luke.fi/mix99.

Pimentel, E., C. Edel, R. Emmerling, and K.-U. Götz. 2022. Effects of missing or wrong pedigree records on Single-Step predictions. Pages 1302–1305 in Proceedings of the World Congress on Genetics Applied to Livestock Production WCGALP, Rotterdam. https://doi.org/10.3920/978-90-8686-940-4.

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

Strandén, I., G.P. Aamand, and E.A. Mäntysaari. 2022. Single-step genomic BLUP with genetic groups and automatic adjustment for allele coding. *Genetics Selection Evolution* 54:1–12. https://doi.org/10.1186/s12711-022-00721-x.

Strandén, I., and E. Mäntysaari. 2020. HGINV program. Version 0.91. Natural Resources Institute Finland (Luke). Jokioinen, Finland.

Thompson, R. 1979. Sire evaluation. BIOMETRICS 35:339–353. https://doi.org/10.2307/2529955.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Vitezica, Z.G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. *Genetics Research* 93:357–366. https://doi.org/10.1017/S001667231100022X.