Exploring unknown parent groups and metafounders in single-step genomic BLUP: Insights from a simulated cattle population

Judith Himmelbauer,1,2* Hermann Schwarzenbacher,1 Christian Fuerst,1 and Birgit Fuerst-Waltl2
1ZuchtData EDV-Dienstleistungen GmbH, Dresdner Straße 89/B1/18, 1200 Vienna, Austria
2University of Natural Resources and Life Sciences, Vienna, Gregor-Mendel Str. 33, 1180 Vienna, Austria

ABSTRACT

This study explores how the metafounder (MF) concept enhances genetic evaluations in dairy cattle populations using single-step genomic best linear unbiased prediction (ssGBLUP). By improving the consideration of relationships among founder populations, MF ensures accurate alignment of pedigree and genomic relationships. The research aims to propose a method for grouping MF based on genotypic information, assess different approaches for estimating the gamma matrix, and compare unknown parent groups (UPG) and MF methodologies across various scenarios, including those with low and high pedigree completeness based on a simulated dairy cattle population.

In the scenario where unknown ancestors are rare, the impact of UPG or MF on breeding values is minimal but MF still performs slightly better compared with UPG. The scenario with lower genotyping rates and more unknown parents shows significant differences in evaluations with and without UPG and also compared with MF. The study shows that ssGBLUP evaluations where UPG are considered via Quaas-Pollak-transformation in the pedigree-based and genomic relationship matrix (UPG_fullQP) results in double counting and subsequently in a pronounced bias and overdispersion.

Another focus is on the estimation of the gamma matrix, emphasizing the importance of crossbred genotypes for accuracy. Challenges emerge in classifying animals into subpopulations and further into MF or UPG, but the method used in this study, which is based on genotypes, results in predictions which are comparable to those obtained using the true subpopulations for the assignment.

Estimated validation results using the linear regression method confirm the superior performance of MF evaluations, although differences compared with true validations are smaller. Notably, UPG_fullQP's extreme bias is less evident in routine validation statistics.

Key Words: genetic evaluation, single-step GBLUP, metafounder, UPG

INTRODUCTION

The single-step genomic best linear unbiased prediction (ssGBLUP) method has become established as the gold standard in genetic evaluation in recent years. As described in Legarra et al. (2009), this method combines the pedigree-based relationship matrix (A) and the genomic relationship matrix (G) into an integrated relationship matrix (H). It is crucial for both matrices to pertain to the same base population, as highlighted by Christensen (2012). However, achieving this assumption is often challenging in routine dairy cattle populations. Various methodologies, as proposed by Christensen (2012), VanRaden (2008), and Vitezica et al. (2011) exist to match G to A.

Legarra et al. (2015) introduced the concept of metafounders (MF) to address this issue by adapting A to G. The fundamental approach involves using allele frequencies of 0.5 for all SNPs in the computation of G and assigning unknown parents in the pedigree to pseudo-individuals, called MF. Accounting for unknown parents using unknown parent groups (UPG) was initially introduced by Thompson (1979) and Quaas (1988) and accommodates genetic differences within subgroups of the base populations. UPG, also known as genetic groups or phantom parents, are widely used in animal breeding to incorporate animals with missing parents and diverse genetic backgrounds in genetic evaluations. Although the mean genetic level of UPG may differ from zero, they are presumed to be non-inbred and unrelated, similar to the base population.

After ssGBLUP gained increasing application in routine genetic evaluation, it became necessary to consider UPG in ssGBLUP. Initially, pedigree based UPG in A were taken into account. Publications on this topic report both positive effects (Koivula et al., 2015), but also...
convergence issues (Matilainen et al., 2016), especially for traits with low heritability. Misztal et al. (2013) further developed this topic and published an approach to integrate UPG in $H$ (i.e., using Quas-Pollak ($QP$)-transformation). Some studies comparing both variants reported improved convergence behavior (Matilainen et al., 2016) and reduced bias (Misztal et al., 2013, 2017), but there are also studies reporting better results for UPG in $A$ (Masuda et al., 2018). Additionally, another method of modeling UPG in ssGBLUP was developed, where UPG are only considered in the construction of inverse of modeling UPG in ssGBLUP was developed, where matrix for genotyped animals) but not in inverse $A$, leading to very good results regarding bias and dispersion (Masuda et al., 2018, 2019; Koivula et al., 2021; Masuda et al., 2021; Belay et al., 2022; Koivula et al., 2022a; Strandén et al., 2022). However, few studies have specifically addressed the underlying assumptions and their impact on genomic estimated breeding values ($GEBV$). One of them, Masuda et al. (2021), is a comprehensive study examining various approaches to integrate UPG in $H$ and comparing their effects on $GEBV$. The commonly used approach with $QP$-transformation in $H$, however, showed distorted evaluation results in this study, attributed to unmet assumptions. Similar results were also found in Bradford et al. (2019).

Legarra et al. (2015) extended the concept of UPG through MF, introducing relationships within and across different base populations. Building upon this concept, MF cannot only balance base populations of $A$ and $G$ but also address differences in genetic levels of subpopulations and relationships both within and between these base populations.

There are already some studies that compare evaluations with UPGs and MF (Meyer et al., 2018; Bradford et al., 2019; Kudinov et al., 2020; Macedo et al., 2020a; Masuda et al., 2021; Meyer, 2021). While the results and conclusions always depend on the data used, it can be summarized that MF seems to be the theoretically most accurate method to account for unknown pedigrees in ssGBLUP implementations.

However, not all details that are necessary for the implementation of MF into routine genetic evaluations are fully clarified. One aspect is, for instance, the definition of UPG or MF, which involves determining how unknown parents can be meaningfully classified into subpopulations and consequently into MF or UPG. The commonly applied strategy is to categorize animals based on year of birth, sex and sometimes also based on their country of origin. This is also used in most publications on this topic (Westell et al., 1988; Interbull, 2001; Kudinov et al., 2020; Meyer, 2021). However, this method is quite arbitrary. In contrast to UPG, the additional desired effect of MF is, as mentioned above, adjusting $A$ to match $G$. This requires that the genetic structure present in $G$ is also present in $A$. In other words, instead of trusting pedigree or year of birth, one would assign animals to UPG or MF based on actual genotypes. To our knowledge, such a method for grouping unknown parents and examining its effect in different genetic evaluations based on UPG or MF has not yet been published. The development and testing of such a grouping strategy will thus be a part of the present study.

The estimation of the gamma matrix ($\Gamma$), which defines the relationship structure within and between the MF, is a current area of research. Legarra et al. (2015) suggested different possible methods for estimating $\Gamma$. Different estimation methods were also derived and tested in Garcia-Baccino et al. (2017), one of which using estimated base allele frequencies. Several extensions and adaptations of this method have been published (Kudinov et al., 2021; Macedo et al., 2021; Legarra et al., 2023). These studies usually include a theoretical derivation of the method used, the application in a routine data set and/or a comparison with the true $\Gamma$ in a simulated situation. The aim of the present study is also to compare the estimated $\Gamma$, computed based on different data sets, with the true $\Gamma$, but also to examine the effect of the estimated $\Gamma$ compared with the true $\Gamma$ on the results of the genetic evaluation.

The scaling of variance components when using MF is a topic with relatively scarce research. Legarra et al. (2015) theoretically derived that the variance components of a model without MF differ from those in a model with MF. The authors suggest re-estimating the variance components taking MF into account or using a theoretically derived scaling factor that assumes a complex mixture of all metafounders. However, only a few studies reported that they used this scaling in their MF implementation (Macedo et al., 2020a; Meyer, 2021). Other authors do not mention this scaling in their publications (Bradford et al., 2019; Masuda et al., 2021), maybe because some software packages do the scaling by default, when MF are modeled. Kudinov et al. (2022) even report that the scaling was not suitable for their application. In this study we examine the effect of scaled and unscaled variance components on the validation statistics of breeding values.

Pruning of pedigrees is a common method to reduce the gap between the pedigree base and the first animals with phenotype or genotype and might lead to higher accuracy in the estimation of $\Gamma$ and MF effects. This strategy has been proposed and implemented in several publications (Macedo et al., 2021; Napel et al., 2022), and for example, Macedo et al. (2021) reported less biased results in a routine data set for sheep.

To sum up, this study aims to present an idea of grouping MF based on genotypes, explore different methods.
for $\Gamma$ estimation and compare genetic evaluations with UPG and MF for a simulated population with 2 base populations and different scenarios and settings with and without unknown pedigrees.

**MATERIALS AND METHODS**

Because no human or animal subjects were used, this analysis did not require approval by an Institutional Animal Care and Use Committee or Institutional Review Board.

**Simulating metafounders**

The R package AlphaSimR (Gaynor et al., 2021) was used for the simulation. The aim was to mimic a dairy cattle population which is based on 2 related and inbred base populations (MF) and crossbreeding takes place between these 2 base populations. Overall, the genetic parameters and the basic procedure are the same as those already used in (Himmelbauer et al., 2023a; b).

The genome comprises 30 chromosomes, each spanning a genetic length of 1 Morgan. For every chromosome, 1,660 single nucleotide polymorphisms (SNPs) and 30 quantitative trait loci (QTL) were uniformly generated, resulting in a total of 49,800 SNPs and 1,800 QTL (900 for each of the 2 traits) across the genome. The mutation rate was set at $2.5 \times 10^{-8}$ per base pair.

Initially, a simulation encompassed 2,500 generations of evolution, yielding 84,000 founder animals, evenly split between males and females. The effective population size ($Ne$) was set at 150. The historical $Ne$ throughout these 2,500 generations was chosen to mimic the linkage disequilibrium structure observed in the current German, Austrian and Czech Fleckvieh population, as documented by Pausch et al. (2013).

Following this evolution-step, the founder population is split into 2 subpopulations. Subsequently, both resultant subpopulations undergo a 15-generation selection based on the true breeding value (TBV) for trait 1. In this process, subpopulation A is selected for high values, while subpopulation B is selected for low values of trait 1, creating a genetic distance between the 2 populations. Subsequently, the 2 subpopulations are combined again, and a second trait (trait 2) is introduced, characterized by a heritability of 0.3 and a random genetic correlation with trait 1 ranging from 0.3 to 0.5. The trait is sex limited, where only females with progeny have phenotypes (e.g., milk). There is only one phenotype per animal and no additional fixed effects or contemporary groups are simulated. Following this, a comprehensive selection strategy as used in Himmelbauer et al. (2023) was performed, including 30 years of pedigree BLUP (PB-LUP) and 8 years of genomic selection by ssGBLUP. It is noteworthy that the selection is based only on trait 2, there are no unknown parents in the pedigree, and the 2 different base populations are treated as one pedigree base in all genetic evaluations. Some adjustments were made compared with the simulation described in Himmelbauer et al. (2023): To guarantee the availability of phenotypes and genotypes for both purebred populations (A and B) and the crossbred population (AB) at the end of the selection process, animals are independently selected based on their subpopulations. In simple terms, female and male animals are separated into different age groups to simulate a realistic age structure and overlapping generations. Within each age group of female animals (f0-f6), there are separate groups for purebred animals from population A, purebred animals from population B, and crossbred animals from populations AB. Similarly, male animals are divided into age groups m0-m4, with separate groups for purebred animals from population A and population B.

The mating process is carefully managed, ensuring that females from subpopulations A, B, and AB are mated with males from purebred populations A and B. This is implemented in such a way that each possible combination of male and female subpopulations occurs with equal frequency in each simulated year. Table 1 provides information on the potential combinations and the corresponding frequencies at which these combinations are generated. A visual representation of the simulation approach is shown in Figure 1.

**Data sets**

Throughout the entire simulation, all pedigree information, phenotypes, and genotypes for all animals across all years were recorded. Additionally, all TBV were collected. This was then used to create 2 data sets for 2 different scenarios:

**High pedigree completeness.** The first scenario is intended to represent the observed situation in the breeding population of the German-Austrian-Czech Simmental population as realistically as possible. This data set contains all animals and all pedigree information. All female animals with offspring have a phenotype, whereby 90% of the phenotypes of animals born in the first 15 years of the simulation are randomly deleted. The reason for this is that in real data sets it is not common to have the complete phenotype information back to the pedigree base. The genotype of all males and a random 60% of the females born in the last 8 years of the simulation (years 31 to 38) are used for genetic evaluation in this scenario. In addition, all sires and 1% of unselected males born in years 16 to 30 are genotyped and all dams and 1% of unselected females born in years 28 to 30 are genotyped. In total, this data set therefore comprises around 154,500
phenotypes, 204,900 genotypes and a total of 1,105,500 animals in the pedigree.

In addition, a reduced data set, which is used for validation, was created using the same animals and the same genotypes. However, the last 3 years of phenotypes are excluded, i.e., the phenotypes of all females born in years 32, 33 and 34. This results in 133,500 phenotypes remaining in the reduced data set.

For most of the following studies, it is necessary that some animals have unknown sires and/or dams. The number of unknown dams and sires per year was adjusted as closely as possible to the actual gaps in the pedigree. The proportions of missing parents for the current German, Austrian and Czech Fleckvieh population and for this scenario can be seen in Figure 2 (top and center). Older animals have missing parents more often than younger animals, with the proportion of missing sires in the data set falling from 7.5% in the oldest cohort to 0.2% in the youngest cohort. The proportion of missing dams in the data set is between 10% and 0.1%. In principle, the animals whose parents are set to unknown in the pedigree are selected at random. However, the possibilities of genomic parentage verification were considered, so that parents that can be found with certainty (genotyped sires and dams of genotyped animals) or with high probability (e.g., genotyped dam’s sires of genotyped animals) are not deleted. On the one hand, this makes the situation with missing pedigrees more realistic, because in practice the pedigrees are completed for such animals and, on the other hand, such a strategy is also necessary to avoid double counting in the genetic evaluation (Pimentel et al., 2022). Figure 3 shows the resulting probabilities at which the respective ancestor is unknown in the pedigree of the genotyped animals born in the last year of the simulation, are shown.

As described for the first scenario, a reduced data set for validation was also created for the second scenario using the same procedure as explained above. The validation data set also contains all animals and genotypes, and the number of phenotypes is reduced to 133,500.

**Low pedigree completeness.** In a second scenario, a somewhat less favorable situation was simulated with more unknown parents in the pedigree and fewer genotyped animals. As in the first data set, all animals are included. The selection of phenotypes and genotypes is identical to the first scenario, with the difference that only 75% of the males and 30% of the females, born in the last 8 years, are genotyped. In total, this data set also includes 1,105,500 animals in the pedigree, 154,500 phenotypes but only 143,400 genotypes. The strategy to create gaps in the pedigree is also the same as in the first scenario, but at higher frequencies. The proportions of missing parents in this scenario are shown in Figure 2 (bottom) and are constant across all cohorts at 7.5% for sires and at 10% for dams, which are the initial values from the first scenario. In Figure 3 (second line, italics) the resulting probabilities, at which the respective ancestor is unknown in the pedigree of the genotyped animals born in the last year of the simulation, are shown.

**Classification of MF and UPG**

Several variants with different pedigree settings were tested in this study. This results in different classifica-
tions of unknown parents to UPG or to MF based on the respective pedigree settings:

**Full pedigree.** In this setting, the true full pedigree without missing parents was used. The parents of the animals born in year 0 are unknown and thus form the pedigree base. Depending on the evaluation test run (see below), these base animals are assigned to their respective true subpopulation (purebred A or purebred B) forming 2 UPG or MF.

**True missing pedigree.** In this setting, unknown parents in the pedigree were simulated according to the procedure described above. Depending on the evaluation method applied, the unknown parents are classified as UPG or MF. The classification was based on subpopulation (purebred A, purebred B or crossbred AB), sex (missing sire or missing dam) and year of birth (5 years were combined into one group). Since the complete pedigree is known, both the respective true subpopulation and the respective true year of birth of the missing parent were used.

**Estimated missing pedigree.** This pedigree setting is intended to simulate a realistic situation, because true year of birth and true subpopulation of the missing parents used in the previous pedigree setting is unknown in real data sets. Therefore, it was necessary to develop a procedure to estimate the assignment to different subpopulations. This was done based on the genomic breed composition (GBC). The method used works as follows:

First, a representative set of 10,000 genotyped animals was randomly selected for analysis. Principal Component Analysis (PCA) on the genotypes was applied to capture the underlying genetic structure (Wilmot et al., 2023), followed by a K-means cluster analysis (Hartigan and Wong, 1979) utilizing the first 2 principal components. The clustering was tested 10 times, where the number of clusters (n) varied between 1 and 10. The optimal number of clusters (n = 2) was determined by evaluating the within cluster sum of squares, which is the sum of the squared deviations from each observation and the cluster center. Cluster centers were then defined, and allele frequencies for both clusters were calculated for animals within these centers.

Subsequently, the GBC was computed for all genotyped animals from the allele frequencies for the 2 subpopulations and the genotypes of the animals using the following linear regression model as outlined by He et al. (2018):

\[ y = 1\mu + Fb + e, \]  

where \( y \) is a \( M \times 1 \) vector containing the genotype of an animal and \( M \) is the number of SNPs, \( F = \{ f_{mt}\} \) is a \( M \times T \) matrix, with \( f_{mt} \) being the allele frequency of SNP \( m \) in population \( t \) and \( T \) is the number of populations, \( \mu \) is the general mean, \( e \) is a vector of residuals and \( b \) is a \( T \times 1 \) vector containing the regression coefficients for all breeds for the respective animal. As described in He et al. (2018), after setting negative coefficients to 0 and scaling the coefficients such that the sum of all coefficients within each animal is 100, these regression coefficients are an estimate for the GBC of the respective genotyped animals. Animals with one GBC greater than 75 were classified as purebred for this subpopulation and all other

![Figure 2. Observed proportions of missing parents for the current German, Austrian and Czech Fleckvieh population (observed pedigree completeness) and simulated proportions (high and low pedigree completeness) based on the year of birth for both scenarios.](image)
animals were classified as crossbred (AB). The threshold value of 75 was chosen to maximize specificity based on accordance with true subpopulation assignment.

To obtain GBC also for non-genotyped animals, an ssGBLUP without UPG and MF was performed, as it is used to estimate breed base representation in the US cattle population (VanRaden, 2011). This ssGBLUP considered 2 traits (GBC for subpopulation 1 and subpopulation 2) with a heritability of 0.999. The calculated GBC for genotyped animals were used as phenotypes. The GEBV were used as estimated GBC for subpopulation 1 and 2 for the respective animals. As the estimated GBC do not always sum up to 100% for all animals, some post-processing steps were implemented. For genotyped animals, the GBC derived from genotype calculations above were used. Proportional scaling was applied to animals with GBC estimates for both subpopulations greater than 0, so that the sum is 100%. In cases where one of the estimates is lower than 0 and the sum exceeded 90%, GBC were set to 0% for the subpopulation where the estimate is lower than 0 and 100% for the other one. For sums below 90%, a uniform constant of 50 (100%/n, where n is the number of subpopulations) was added to both estimates, followed by proportional scaling to maintain a sum of 100%. Animals were classified into subpopulation A, subpopulation B, and crossbred population AB, designating purebred (A, B) for one of the GBC greater than 75 (maximum specificity) and AB for others.

Based on this classification the unknown parents are assigned to UPG or MF. The procedure for the assignment is the same as described above for “true missing pedigree,” with the difference that in this case not the true but the estimated assignment to the subpopulations (purebred A, purebred B and crossbred AB) is used. As the true year of birth of the missing parents is also unknown, the known year of birth of the animal whose parent is unknown is used to approximate the year of birth of the missing parents.

**Pruned estimated missing pedigree.** This pedigree setting is very similar to “estimated missing pedigree” with the difference that the pedigree has been pruned. Phenotypes of animals born in the first 14 years are removed from the data set. Therefore, only 139,800 phenotypes remain in the data set while the number of genotypes used remains the same. The pedigree was pruned to ≤ 3 generations before animals with genotype and/or phenotype. This reduces the number of animals in the pedigree to 698,548. This pruned pedigree was then used to estimate the GBC of all animals in the pedigree as described above and the animals were then classified into subpopulation A, subpopulation B and crossbred population AB. The assignment of unknown parents to UPG or MF is the same as described above, based on estimated subpopulation, sex, and year of birth.

To ensure that the UPG or MF is sufficiently represented by animals, phenotypes and genotypes, groups that occur less than 10 times in the pedigree are merged with a neighboring group of birth years of the same subpopulation and the same sex. To measure the information available to estimate UGP or MF, genetic contributions from the MF or UGP were calculated using RelaX2 (Strandén and Vuori, 2006). The genetic contributions for one animal always sum up to 1 and the different contributions give the proportion of genes that can be traced back to the respective UPG or MF. The genetic contributions of the individual groups are calculated for all animals with phenotype. The contributions within each group are then summed over all animals with phenotype. This sum more or less represents the phenotype equivalents that contribute to the estimation of the respective group.

**Figure 3.** Illustration of the 3-generation pedigree of a validation animal showing the probability that the respective ancestor is unknown in the pedigree for the 2 different scenarios with high (first line) and low (italics, second line) pedigree completeness.
The same is done for all animals with genotype. Groups whose phenotype and/or genotype equivalents are less than 100 are also merged with neighboring group of birth years of the same breed and sex. Therefore, the number of UPG or MF in each replicate varies between 26 and 29 for estimated pedigree and between 34 and 38 for true pedigree due to the more balanced distribution of animals in the subpopulations.

**Gamma matrix**

The true $\Gamma$ was computed using true allele frequencies calculated based on the genotypes of all animals which are replaced by the different MF in the pedigree and the following formula proposed in Garcia-Baccino et al. (2017)

$$\Gamma = 8 \cdot \text{cov}(p_1, p_2, \ldots, p_T), \quad [2]$$

where $T$ is the number of MF and $p_1, p_2, \ldots, p_T$ are vectors containing true allele frequencies of MF. In addition to the true $\Gamma$, 2 alternative estimations were performed. Both estimation methods rely on calculated allele frequencies and employ Equation 2. The software Bpop (Strandén and Mäntysaari, 2020b) was used for estimating allele frequencies and $\Gamma$. This software employs a generalized least square method for allele frequency estimation (Gengler et al., 2007). For the first method (BFQ pure), only genotypes of animals classified as purebred for one of the 2 subpopulations were considered to estimate $\Gamma$. Conversely, in the second method (BFQ all), all genotypes, including those of both purebred and crossbred animals in the final data set, were utilized for estimating $\Gamma$.

It is important to check whether $\Gamma$ is positive definite. Therefore, the eigenvalues for each (true and also estimated) $\Gamma$ were calculated. If the smallest absolute eigenvalue is close to 0, a constant of 0.01 was added to the diagonal of the matrix as a singularity prevention.

**Genetic evaluation methods**

To assess the impact of incorporating MF, for a given data set and pedigree setting various genetic evaluations were examined. A total of 7 evaluation methods were done for both data sets/scenarios and for all pedigree settings respectively. To calculate estimated linear regression ($LR$) validation statistics (Legarra and Reverter, 2018), all evaluations were also computed for the truncated data sets. For all evaluations except the evaluation with scaled variances, the simulated genetic variance, $\sigma^2_{\text{unrelated}} = 0.3$, is used in the evaluation as well as the true base allele frequencies are used to set up the genomic relationship matrix. All evaluations were computed using the commercial software package MiX99 (MiX99 Development Team, 2019). Basically, G for ssGBLUP was computed as described in Himmelbauer et al. (2023a). Preparing G was done using the program HGINV (Strandén and Mäntysaari, 2020a) based on VanRaden's method I (VanRaden, 2008) and the approach for Proven and Young (Miszta et al., 2014). For all evaluations using MF, base allele frequencies equal to 0.5 were used instead of the true base allele frequencies as described in Legarra et al. (2015).

1) **PBLUP with UPG (PED)**. A simple pedigree BLUP was performed. The UPG in the pedigree were taken into account and modeled as random with $\text{var}(g) = I \cdot \sigma^2_{\text{unrelated}}$, where I is an identity matrix.

2) **ssGBLUP without UPG (no UP.G)**. An ssGBLUP with no UPG in the pedigree was used to estimate GEBV. For this purpose, all unknown parents in the pedigree are set to 0 and therefore assigned to one single base population.

3) **ssGBLUP with UPG in A (UPG alteredQP)**. This evaluation is also an ssGBLUP but in this case the UPG in the pedigree are used. UPG are modeled as random and are only considered in the creation of the inverse pedigree relationship matrix ($A^{-1}$), the inverse pedigree relationship matrix for genotyped animals ($A_{22}^{-1}$) but not in the inverse genomic relationship matrix ($G^{-1}$) as described among others in Masuda et al. (2018), (2022) and Strandén et al. (2022).

4) **ssGBLUP with UPG in H (UPG fullQP)**. This method also considers UPG and QP transformation was applied to $A^{-1}$, $A_{22}^{-1}$, and $G^{-1}$ as used in Misztal et al., (2013).

5) **ssGBLUP with MF and true $\Gamma$ (MF_true)**: In this case the unknown parents are modeled as MF in an ssGBLUP evaluation. The relationships between MF are defined by the true $\Gamma$. This results in a variance-covariance-matrix for breeding values of $\text{var}(u) = H_{\Gamma} \cdot \sigma^2_{\text{unrelated}}$, where $\sigma^2_{\text{unrelated}}$ is the true simulated variance of 0.3 and $H_{\Gamma}$ is the combined relationship matrix as described in Legarra et al. (2015).

6) **ssGBLUP with MF and estimated $\Gamma$ (MF_est)**: This is another ssGBLUP with MF, but here the relationships between MF are estimated using the strategy BFQ_all, described above. The variance-covariance-matrix for breeding values can be calculated as for MF_true but using the estimated $\Gamma$ to set up $H_{\Gamma}$.

7) **ssGBLUP with MF, true $\Gamma$ and scaled variances (MF_sc)**: This method is the same as MF_true, but in this case, the variance components are scaled as proposed by Legarra et al. (2015). The additive genetic variance was scaled using the following equation:
Analyzing results

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

**Gamma matrix.** To assess the efficiency of the 2 different methods for estimating \( \Gamma \), the diagonal and off-diagonal values of the estimated \( \Gamma \) are compared against their corresponding values in the true \( \Gamma \). The mean difference for both the diagonal and off-diagonal values is computed for method comparison.

**True validation statistics.** Three validation measures, namely correlation, bias, and dispersion, are used to compare the different evaluations. All measures are calculated based on the youngest animals with genotypes born in the last year of the simulation. The size of this validation group varies for the different scenarios. For the “high pedigree completeness” scenario, the validation group consists of 22,375 animals, whereas for the “low pedigree completeness” scenario, it only consists of 14,672 animals, as fewer animals are genotyped in the second scenario.

The correlation is calculated between the (G)EBV and TBV. The bias is the mean difference between (G)EBV and TBV, calculated using the following formula:

\[
\text{bias} = \text{EBV} - \text{TBV}. \quad [4]
\]

It follows that positive values for bias indicate overestimation. As the genetic standard deviation for the trait of interest is 1, the bias resulting from Equation 4 can be interpreted as genetic standard deviations. The regression coefficient \( b_1 \) of the following regression is used as a measurement of the dispersion:

\[
\text{TBV} = b_0 + b_1 \cdot \text{EBV} + \mathbf{e}, \quad [5]
\]

where \( b_0 \) is the intercept, \( b_1 \) the regression coefficient and \( \mathbf{e} \) the vector of residuals. Additionally, comparisons are made for the group estimators of the UPG and the MF to assess the differences among the 7 evaluation methods.

**Estimated validation statistics using linear regression (LR) method:** To obtain validation statistics as they can be calculated in real data sets, the (G)EBV for certain validation animals based on the full data sets are compared with the (G)EBV for these animals based on the reduced data sets. Two validation groups were defined, a male and a female validation group. The male validation group contains approximately 530 genotyped bulls per replicate that were born between year 30 and year 32 in the simulation. These bulls have no daughters with own records in the reduced data set but have at least 20 daughters with records in the full data set. The female validation group contains around 12,400 genotyped females per replicate from birth years 32 to 34, which have no phenotypes in the reduced data set but have records in the full data set.

Based on the findings from Himmelbauer et al. (2023), the LR method is a simple but quite accurate method to estimate bias, dispersion and validation reliability (Legarra and Reverter, 2018; Macedo et al., 2020b). The bias is the mean difference of (G)EBV from the full and the reduced data set and was calculated with the following formula:

\[
b = \text{GEBV}_r - \text{GEBV}_f, \quad [6]
\]

The expected value for \( b \) when (G)EBV are unbiased is 0, positive values for bias indicate overestimation, whereas negative values indicate underestimation. As the genetic standard deviation for the trait of interest is 1, the bias resulting from Equation 6 can be interpreted as genetic standard deviations. The following formula was used to calculate the dispersion:

\[
b_1 = \frac{\text{cov}(\text{GEBV}_f, \text{GEBV}_r)}{\text{var}(\text{GEBV}_r)}. \quad [7]
\]

For the case \( b_1 = 1 \), there is no under- or overdispersion present in the (G)EBV, \( b_1 < 1 \) indicate overdispersion, meaning high (G)EBV are overestimated and low (G)EBV are underestimated and \( b_1 > 1 \) indicate underdispersion. The reliability was calculated as follows:

\[
r^2 = \frac{\text{cov}(\text{GEBV}_f, \text{GEBV}_r)}{\sigma_g^2}, \quad [8]
\]

where \( \sigma_g^2 \) is the true genetic variance in the validation group.

These statistics were calculated for male and female animals for all scenarios, pedigree settings, and all different genetic evaluations. The results are also based on 10 replicates.

**UPG and MF effects.** To assess the accuracy of the estimated effects for MF and UPG, they were compared...
against their true values. To calculate the true effects, for each MF or UPG, those unknown parent animals that were replaced by the respective MF or UPG were searched in the full pedigree, and the weighted average of their TBV was calculated.

RESULTS

Estimation of Pedigree

The within cluster sum of squares plot for number of clusters \( n \) between 1 and 10 for one repetition is shown in Figure 4 and looks very similar for all repetitions. There is a huge improvement going from one to 2 clusters, a smaller improvement for \( n = 3 \) and almost no improvement for more than 3 clusters.

The result of the clustering and the cluster centers for \( n = 2 \) is shown in Figure 5, where all the 10,000 genotyped animals are plotted using their first 2 principal components.

To assess the estimated GBC for genotyped animals the estimated subpopulation is compared with the true subpopulation in Table 2. Almost all purebred animals are classified correctly to the true purebred subpopulation. About 20% of the true crossbred animals are classified wrongly to subpopulation A.

The comparison of the true subpopulations and the estimated subpopulations based on their estimated GBC are only important for sires and dams that are missing somewhere in the pedigree. In Table 3 the true and estimated classification of missing sires is shown. For dams, the same information is shown in Table 4. Only a small proportion of purebred animals are correctly assigned to the true subpopulation while the majority is classified as crossbred.

Gamma matrix

Figure 6 and Figure 7 show the comparison of true and estimated \( \Gamma \) for different pedigree settings and for both scenarios. For the scenario with high pedigree completeness, the results for the comparison of \( \Gamma \) show that both variants overestimate the values on the diagonal, with the method that also considers the genotypes of crossbred animals showing significantly fewer estimation errors (Figure 6, top). The off-diagonal values are, on average, relatively well estimated by both methods, with only slight tendencies toward underestimation. Again, the BFQ_all method provides slightly better results compared with BFQ_pure (Figure 6, bottom).

For the scenario with low pedigree completeness, it can be observed that \( \Gamma \) can be generally estimated more accurately regardless of the method, and the differences between the 2 methods are smaller. However, the BFQ_all method, especially for settings with estimated pedigree, provides better results than BFQ_pure for both the diagonal and off-diagonal elements (Figure 7).

For both scenarios, an increase in inbreeding, i.e., the diagonal elements, is observed from full pedigree to

![Figure 4. Within cluster sum of squares for the K-means clustering based on the first 2 principal components for number of clusters from 1 to 10 for the first repetition.](image)

![Figure 5. Clustering of the 10,000 selected genotypes into 2 subpopulations and corresponding true subpopulations. The red circles mark the calculated cluster centers.](image)

<table>
<thead>
<tr>
<th></th>
<th>estim. A</th>
<th>estim. AB</th>
<th>estim. B</th>
</tr>
</thead>
<tbody>
<tr>
<td>true A</td>
<td>99.64%</td>
<td>0.36%</td>
<td>0%</td>
</tr>
<tr>
<td>true AB</td>
<td>19.71%</td>
<td>78.59%</td>
<td>1.69%</td>
</tr>
<tr>
<td>true B</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2. Proportions of estimated subpopulations based on the genotype for animals of the three different true subpopulations
missing pedigree. The relationships between the MF also increase when the pedigree is incomplete.

The scaling factor for variance components is calculated according to Equation (3) using $\Gamma$. This results in scaling factors ranging from 0.706 to 0.719 for the settings with full pedigree and from 0.745 to 0.777 for settings with missing pedigrees, resulting in scaled $h^2$ between 0.355 and 0.378 instead of 0.30.

Estimates for UPG and MF

Figure 8 illustrates estimated effects of UPG or MF across different pedigree settings and genetic evaluations considering high pedigree completeness. Groups are analyzed separately by breed and sex. In the “full pedigree” setting, only one estimated value per breed is shown for males and females, as MF or UPG occur only at the pedigree base, with male and female animals of the same subpopulation combined. Results from various genetic evaluations yield similar outcomes, consistent with true effects (Figure 8, first column).

In settings with incomplete pedigrees (Figure 8, second and third column), estimated genetic trends within subpopulations and sexes are obtained by assigning respective MF or UPG to years when unknown parent animals were born. MF_true and MF_sc show very similar results across all scenarios and pedigree settings and closely capture the true genetic trend. One exception is the subpopulation B for the setting with true missing pedigree, where the results underestimate the true effects from around year 20 to the youngest groups. The results for MF_est are similar to MF_true and MF_sc for the setting with estimated missing pedigree. Only in some cases, the estimates for older groups deviate noticeably from the true effects (e.g., the oldest group in the male subpopulation A). For the “true estimated pedigree” setting in some cases MF_est underestimated the youngest groups slightly.

Results from evaluations with UPG, starting from approximately year 20, show a significant underestimation of true effects in both pedigree settings and additionally tend to overestimate older groups in the setting with es-

---

Table 3. Confusion matrix showing the true vs. estimated subpopulations for missing sires

<table>
<thead>
<tr>
<th>Missing sires</th>
<th>estim. A</th>
<th>estim. AB</th>
<th>estim. B</th>
</tr>
</thead>
<tbody>
<tr>
<td>true A</td>
<td>329</td>
<td>47.303</td>
<td>20</td>
</tr>
<tr>
<td>true B</td>
<td>2</td>
<td>52.248</td>
<td>359</td>
</tr>
</tbody>
</table>

Table 4. Confusion matrix showing the true vs. estimated subpopulations for missing dams

<table>
<thead>
<tr>
<th>Missing dams</th>
<th>estim. A</th>
<th>estim. AB</th>
<th>estim. B</th>
</tr>
</thead>
<tbody>
<tr>
<td>true A</td>
<td>670</td>
<td>68.729</td>
<td>366</td>
</tr>
<tr>
<td>true AB</td>
<td>115</td>
<td>4.328</td>
<td>171</td>
</tr>
<tr>
<td>true B</td>
<td>140</td>
<td>67.934</td>
<td>1.661</td>
</tr>
</tbody>
</table>

---

Figure 6. Comparison of true (green) and estimated gamma values (diagonal values on the top and off-diagonal values in the bottom) for the 4 different pedigree settings based on the scenario with high pedigree completeness. BFQ_pure (dark blue) was computed based on estimated allele frequencies from purebred animals. BFQ_all (light blue) was computed based on estimated allele frequencies from purebred and crossbred animals. 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.
hestimated missing pedigrees. PED generally demonstrates good agreement with true effects but slightly underestimates the youngest female group for the “true missing pedigree” setting, tend to underestimate younger groups and overestimate older groups in the “estimated missing pedigree” setting. Female groups in subpopulation AB for the “estimated missing pedigree” setting demonstrate varying patterns, with PED and UPG_alteredQP showing strong overestimates for all groups.

As there is no noticeable effect of pedigree pruning on the effects of UPG or MF, the results for the “pruned estimated pedigree” setting are not shown.

Figure 9 shows the same results for the scenario with low pedigree completeness. For the setting with a full pedigree (Figure 9, first column), the results are, as expected, comparable to those from the scenario with high pedigree completeness. With true missing pedigree (Figure 9, second column), it is interesting to see that all evaluation methods except UPG_fullQP provide very similar results that also align well with the true effects, which was not the case for the scenario with high pedigree completeness. As in the previous scenario, the effects from UPG_fullQP show an underestimation of the groups from the year 20, but to a smaller extent. This pattern is the same for all subpopulations and sexes in this setting.

For the setting with estimated pedigree (Figure 9, third column), the results are again very similar to those from the scenario with high completeness. All evaluations with MF provide similar results, more or less precisely matching the true effects and fully capturing the trend of these effects. The other 3 evaluations with UPG cannot capture this trend and tend to overestimate the older groups and underestimate the younger groups. Again, pedigree pruning has no effect on the estimates for UPG and MF and therefore the results are not shown.

Correlation, bias, and dispersion

In Figure 10, the results for correlation, bias, and dispersion for different pedigree settings and evaluations are presented for the scenario with high pedigree completeness. As expected, the average correlation for all PED evaluations is lower than for the ssGBLUP evaluations. Additionally, the variability between repetitions for PED is significantly higher than for all other genetic evaluations. Except for PED, no substantial differences are observed regarding correlation, neither within nor between the pedigree settings, with an average correlation of about 0.83 for all ssGBLUP evaluations and 0.59 for PED (Figure 10, first row).

Regarding bias and the setting with a complete pedigree, PED again shows the most extreme results with a downward bias of around 0.62 genetic standard deviations. The evaluations no_UPG, UPG_alteredQP, and UPG_fullQP result in a slight overestimation of around 0.04 genetic standard deviations. The overestimation of UPG_fullQP is slightly higher at 0.07. The bias for
the 2 evaluations with MF and true and estimated $\Gamma$ is approximately the same, showing an underestimation of around 0.02 genetic standard deviations, while scaling the variance components leads to an overestimation of 0.03 genetic standard deviations. Surprisingly, the results for bias for all other settings with incomplete pedigree are very similar to those with a complete pedigree (Figure 10, second row).

The same applies to dispersion (Figure 10, third row). Again, the results for different pedigree settings are
nearly identical. PED shows an average overdispersion of around 0.88 with a notably higher variability. The results for evaluations no_UPG, UPG_alteredQP, and UPG_fullQP are very similar and are around 0.95. The best results are obtained by MF_true and MF_est with a regression coefficient of 1.00. However, scaling the variance components when using MF worsens the dispersion slightly.

Figure 11 displays results for the scenario with low pedigree completeness. Due to the increased frequency

**Figure 9.** Comparison of true (light green) and estimated effects for unknown parent groups (UPG) and metafounder (MF) for the scenario with low pedigree completeness, 3 pedigree settings and for 6 evaluation methods. The time trends of the effects are presented in genetic standard deviations and compared separately based on subpopulation and sex. PED: pedigree BLUP with UPG, UPG_alteredQP: single-step genomic BLUP with UPG in A and A\textsubscript{22}, UPG_fullQP: single-step genomic BLUP with UPG in A, A\textsubscript{22} and G, MF\_true: single-step genomic BLUP with MF and true gamma matrix, MF\_est: single-step genomic BLUP with MF and estimated gamma matrix, MF\_sc: single-step genomic BLUP with MF and true gamma matrix and scaled variance components.
of unknown parents, more differences between pedigree settings are evident. Correlation (Figure 11, first row) again shows lower values and higher variability for PED across all pedigree settings. For the full pedigree setting, ssGBLUP evaluations maintain a consistent average correlation of around 0.83, except UPG_fullQP, which results in lower average correlations ranging from 0.73 to 0.80 for incomplete pedigree settings. Regarding bias (Figure 11, second row), significant differences emerge between pedigree settings. PED exhibits a strong underestimation averaging 0.63 genetic standard deviations for all pedigree settings. Patterns observed in the “full pedigree” setting are consistent with the previous scenario, with varying degrees of overestimation for other evaluations. In the setting with true missing pedigree no_UPG and UPG_fullQP show substantial overestimation of 0.24 and 0.40 genetic standard deviations, respectively. Similar results are also observed for the 2 settings with estimated pedigree, with the extreme bias of UPG_fullQP reducing to approximately 0.3. MF evaluations demonstrate slight underestimation, with minimal changes due to incomplete pedigrees.

Dispersion results (Figure 11, third row) reveal consistent overdispersion of around 0.89 for PED across

---

**Figure 10.** Comparison of true validation statistics (correlation, bias, dispersion) for the scenario with high pedigree completeness, 4 pedigree settings and 7 evaluation methods. The error bars in the plot show the range from minimum to maximum and the “x” show the means over 10 repetitions. PED: pedigree BLUP with UPG, no_UPG: single-step genomic BLUP without UPG or MF, UPG_alteredQP: single-step genomic BLUP with UPG in A and A22, UPG_fullQP: single-step genomic BLUP with UPG in A, A22 and G, MF_true: single-step genomic BLUP with MF and true gamma matrix, MF_est: single-step genomic BLUP with MF and estimated gamma matrix, MF_sc: single-step genomic BLUP with MF and true gamma matrix and scaled variance components. correlation: correlation between estimated and true breeding values, bias: mean difference of estimated minus true breeding values given in genetic standard deviations, dispersion: slope of the regression of true breeding values on estimated breeding values.
all pedigree settings. UPG_alteredQP maintains similar dispersion levels (0.96), while the dispersion of no_UPG changes from 0.97 in full pedigree to 0.93 for the settings with incomplete pedigree. UPG_fullQP experiences the most significant impact, with notable decreases in dispersion coefficients from 0.96 (full pedigree) to 0.74 (true missing pedigree), 0.80 (estimated missing pedigree), and 0.83 (pruned estimated pedigree). MF evaluations show no noticeable changes in dispersion, maintaining consistent values of 1.00 for MF_true and MF_est and 0.97 for MF_sc across pedigree settings.

**Estimated validation statistics using LR method**

The results of the validation using LR method of the different pedigree settings and genetic evaluations are shown in Figure 12 for the scenario with low pedigree completeness. The results for the scenario with high pedigree completeness show even less differences between the pedigree settings and evaluations (except for PED) and do not detect any significant bias or dispersion for ssGBLUP evaluations. Therefore, these results are not presented here.

For the scenario with low pedigree completeness the LR method results in clear differences compared with the true validation statistics. For instance, the reliability shows larger discrepancies between the different genetic evaluations, with evaluations using MF performing relatively worse than the other ssGBLUP evaluations. However, the more striking difference is that this validation does not recognize the strong overestimation of no_UPG and UPG_fullQP and the significant overdispersion of UPG_fullQP in the settings with incomplete pedigree. The validation statistics show no substantial differences between the settings with complete and incomplete pedigree. Regarding bias, MF_true and MF_est are slightly more downwards biased compared with the other evaluations. However, concerning dispersion, both MF_true and MF_est again show a regression coefficient very close to 1.00 for all pedigree settings, but also no_UPG and UPG_alteredQP achieve relatively good results based on this validation.

**DISCUSSION**

**Comparison of scenarios**

The first scenario with high pedigree completeness reflects the actual situation present in many current dairy cattle populations in terms of the number of unknown parents in the pedigree. Looking at the frequencies of unknown ancestors in the group of the youngest animals used for validation, these frequencies are already very low (Figure 3, first row). As a result, the effect of UPG or MF on the (G)EBV of these animals is hardly noticeable. Therefore, there are hardly any differences between the results for complete and incomplete pedigrees in terms of correlation, bias, and dispersion for this scenario. Due to the high genotyping rate in male and increasingly in female animals, and the almost complete recording of the pedigree, such populations seem to be relatively robust regarding the consideration of unknown parents. However, in routine populations, there can be subpopulations from other countries or breeds with more unknown parents in their pedigree. For example, Tsuruta et al. (2014) reported 14% missing sires and 19% missing dams for non-US bulls used in the genetic evaluation for US-Holstein cattle. Looking at other animal species, such as sheep for instance, Macedo et al. (2021) reported around 8% unknown parents for a Lacaune dairy sheep population. Therefore, a second scenario with significantly lower genotyping rates and more unknown parents was examined in this study. In Figure 3, the unknown ancestor probabilities for the second scenario are shown in italics in the second row. The proportions of unknown ancestors in the pedigree are clearly higher in this scenario. It should be noted that the observed positive or negative effects of different methods strongly depend on the underlying data structure and must be considered in the interpretation of the results. This is also the reason, why different studies might come to various and sometimes contradictory conclusions for the same method (Bradford et al., 2019).

**Effects of different evaluations for settings with missing pedigrees**

Basically, the analyzed group of animals includes individuals from 3 subpopulations (purebred A, B, and crossbred AB). Since these subpopulations have slightly different genetic levels, a separate analysis of the subpopulations could potentially provide more accurate results. However, as the separate analysis yielded essentially the same results as the joint analysis it was not included in this study.

The strong downward bias and significant overdispersion for PED in both scenarios and all pedigree settings are attributed to genomic preselection. The causes of this preselection bias have been extensively described in several studies (e.g., Patry and Duroc, 2009, 2011; Jibrila et al., 2020, 2021; Himmelbauer et al., 2023b) and are present in this study since animals from the last 8 generations were selected based on GEBV.

It is remarkable that the scenario with high pedigree completeness in the setting with true incomplete pedigree does not show significant differences between estimates with and without UPG. In the case without UPG, neither the genetic trend within a subpopulation nor the genetic
differences between subpopulations are considered, but there are still no significant effects on validation statistics. This leads to the conclusion that the effects of UPG, explaining differences in the levels of (G)EBV of unknown parents, have little to no impact on the animals of the last generation. The reason for this has been described above and is related to the high genotyping rate and the low proportions of animals with unknown ancestors. To investigate the effects of genetic evaluations with and without UPG, or how UPG are considered in evaluations, one must refer to the scenario with low pedigree completeness, where the UPG proportions for validation animals are larger, as more recent pedigrees are unknown.

At the same time, it seems surprising that evaluations with MF show an effect on the GEBV of validation animals in the scenario with high pedigree completeness. Since these positive effects, as seen in the comparison of no_UPG and UPG_alteredQP, cannot be attributed to the different levels of the UPG, it is presumed that the superiority of evaluations with MF comes from the better alignment of $A$ and $G$, as they refer to the same base and the relationships between and within the subpopulations are considered. For instance, Vitezica et al. (2011) and
Christensen et al. (2012) have found that poor alignment can lead to biases. Also, in the scenario with low pedigree completeness, evaluations with MF provide the best results in terms of bias and dispersion, similar to the results for the complete pedigree. This aligns with the findings in Bradford et al. (2019). However, in contrast to the previous scenario, there are now clear differences in evaluations with and without UPG. The upwards bias of no_UPG together with overdispersion is due to the fact that the relationships in A, where only known relationships are considered, and those in G, where all genomic relationships are fully considered, do not match. The genotyped animals from the last generation receive the full contribution of their (in the pedigree known or unknown) ancestors through G, and additionally, all animals are corrected for the genetic level in the base population through A. Since more pedigrees in this scenario have unknown ancestors, the level of the base population is overestimated, leading to bias. Interestingly, Bradford et al. (2019) found a downward bias and slight underdispersion in their simulation study. However, it was also noted in that study that the
impact of different UPG modeling strongly depends on the underlying data and population structure (Bradford et al., 2019).

The large bias and overdispersion of UPG_fullQP can be explained by double-counting which occurs when UPG is considered in G, even though genomic relationships in G are complete. Similar phenomena have been found in several comparable studies (Bradford et al., 2019; Masuda et al., 2021; Meyer, 2021). In comparison, the evaluation UPG_alteredQP in this scenario provides results similar to those with a complete pedigree. The reason is that G correctly considers relationships and is not affected by incomplete pedigrees. The relationships in A and A_{22} are now corrected by modeling through UPG, and the different levels or trends in unknown parents can be accurately considered. Good results using this method were also found in other studies for routine populations (Koivula et al., 2021; Belay et al., 2022; Koivula et al., 2022a).

**Gamma matrix**

Regarding the estimation of \( \Gamma \), several studies have already compared different methods. Garcia-Baccino et al. (2017) demonstrate in their simulation study that estimating \( \Gamma \) through the covariance of base allele frequencies for one MF provides unbiased results. Himmelbauer et al. (2023b) previously compared this method to other methods, for example the one proposed by Legarra et al. (2015), for the case of 2 MF. The authors concluded that, again estimating from base allele frequencies best approximates the true \( \Gamma \). Therefore, the present study, which additionally examines the case for more than 2 MF, focuses on the effect of including genotypes from a crossbred population on the estimation of base allele frequencies and thus on \( \Gamma \). It was demonstrated that, in all scenarios and for all pedigree settings, crossbred genotypes are an important source of information, and both inbreeding and relatedness between the MF can be estimated with less bias. Similar results were also published by Bradford et al. (2019) for a simulated population. Nevertheless, the results show a slight upward bias for the diagonal in the scenario with high pedigree completeness compared with the true \( \Gamma \). Several authors suggested adjustments to the method developed by Garcia-Baccino et al. (2017), especially for cases with many MF. For instance, Kudinov et al. (2021) used a covariance function kernel to extend \( \Gamma \) to multiple MF. Macedo et al. (2021) used the change in inbreeding over time to estimate \( \Gamma \) for different groups of birth years of a population and tested a second method based on single-locus iterative peeling, which, however, led to less optimal results. Legarra et al. (2023) proposed a method based on a partial EM maximum likelihood algorithm to estimate relationships between subpopulations or breeds. Taskinen et al. (2023) showed that interpolating or extrapolating \( \Gamma \) yields good results for the application in a multi-breed ssGBLUP evaluation for Finnish beef cattle.

In the present study, however, more complex methods were avoided since GEBV properties for evaluations using estimated and true \( \Gamma \) revealed no significant differences across tested scenarios and pedigree settings. It seems, on the one hand, that the slight tendency to overestimate the diagonal has little impact on GEBV and, on the other hand, that for the given data structure in this study, estimating \( \Gamma \) for the approximately 30 to 40 MF via the simple covariance of estimated base allele frequencies is sufficiently accurate. However this is not true in all situations, for example, when there are genotyped animals with very deep pedigrees, and there is only little information to estimate recent MF. Unfortunately, there are no known studies that have previously investigated the effects of differently estimated \( \Gamma \) compared with the use of true \( \Gamma \).

In this study, the testing of the positive definiteness of the \( \Gamma \) has proven to be particularly important regarding convergence behavior and also for the corresponding GEBV properties. In various test runs, it was found that even a \( \Gamma \) whose smallest eigenvalues are still slightly greater than 0, and therefore mathematically positive definite, can lead to negative effects on the estimation of MF effects and thus to even more extremely biased GEBV and poorer convergence (results not shown). In such cases, a singularity prevention by adding a constant to the diagonal has proven effective. In the present study, 0.01 was added to the diagonal for this purpose. To our knowledge this aspect has never been mentioned in any previous study on MF, even though it might be an important issue for the implementation of MF. However, similar singularity preventions like this are common for the genomic relationship matrix (e.g., Koivula et al., 2022b).

**Estimation of pedigree and classification of groups**

In livestock, the underlying genetic population structures are often not entirely clear making the assignment of unknown parents to MF or UPG a difficult task. The strategy of detecting subpopulations based on genotypes and PCA has already been tested in another study for identifying subpopulations in cattle populations (Wilmot et al., 2023). The subsequent cluster analysis allows for an automated allocation to subpopulations based on the first 2 principal components. A crucial point is, of course, the choice of the number of clusters. In this study, the so-called elbow method, based on the within-cluster sum of squares, was used. This is a widely used but heuristic method to find the optimal number of clusters. The number of clusters is chosen where an additional cluster does
not significantly improve the modeling of the data. The principle is that a model should be as complex as necessary but as simple as possible (Occam's razor, Good, 1977). In this study, it is not entirely clear whether the optimal number of clusters is 2 or 3 (Figure 4). Since the improvement achieved by having 3 clusters appears relatively small compared with the improvement from one to 2 clusters, 2 was chosen as the optimal number for the presented results. However, in test runs with 3 clusters, very similar results regarding the agreement of the classification with the true pedigree were achieved, and subsequently, there were no significant differences in the validation statistics for the (G)EBV (results not shown). The allocation of all genotyped animals based on allele frequencies works with very high accuracy, as shown in He et al. (2018).

The greatest challenge and at the same time the major weakness of the applied method is the transfer of this information to non-genotyped animals and the classification of non-genotyped animals into subpopulations. Many animals are assigned to incorrect subpopulations through the estimation of the GBC. In the course of developing this approach, different adaptations and ideas were tested, such as imputing, different methods for classifying genotypes or different post-processing adaptations. Of all the strategies tested, the one described here proved to be the most optimal even though the specificity of the approach is not perfect. For example, imputing the genotype for a part of the non-genotyped animals using the genotypes of ancestors and progeny and thus calculate the GBC based on allele frequencies was only possible for very few animals. The reason is that imputation is only relevant for animals with at least one unknown parent, and these are also the animals that less frequently have genotyped descendants. This arises from the assumption of which animals can have unknown parents at all. However, since this also applies to routine populations, it is presumed that the effect of imputing is limited there as well.

Nevertheless, the comparison of the validation statistics of genetic evaluations with true incomplete pedigree and estimated incomplete pedigree showed that for this study the classification reflects the essential population structures and has no negative effects on (G)EBV compared with the evaluations with true missing pedigree. This applies to both the scenario with high and low pedigree completeness. At this point it should be mentioned that the accuracy of the assignment of non-genotyped animals and the impact on the final (G)EBV depends of course on the underlying data structure and might be different for other population structures.

### Pedigree pruning

In general, shortening the pedigree and phenotypes has the effect of reducing the number of animals and the amount of data. However, it also reduces the distance between the pedigree base and those animals providing the majority of genotype and phenotype information. This increases the effective numbers of genotypes and phenotypes contributing to the estimation of the effects of UPG or MF on GEBVs of related animals, and consequently simplifies their estimation (Napel et al., 2022). However, in this study, no significant effect of pedigree pruning was observed. One possible explanation could be that too few generations were removed through pruning. Unlike many routine populations, the original pedigree in this simulation only extends back 38 years, and more extensive pruning as applied in this study without losing substantial numbers of phenotypes and genotypes was not feasible for the underlying data. At the same time, this implies that even in the unpruned pedigree, the distance from the pedigree base to the generations with genotype and phenotype information is probably smaller than in many real populations, which may be another explanation for the limited effects of pedigree pruning.

Nevertheless, one advantage of pruning in this study is that the estimation of the GBC for ungenotyped animals is somewhat easier, as many of the old ungenotyped animals are removed, and the GBC does not need to be estimated as far back to old generations. As a result, more animals are correctly assigned to their respective subpopulations compared with the scenario with an unpruned pedigree. However, this has no positive impact on the final (G)EBV.

### Scaling of variance components

The scaling of variance components using MF was derived by Legarra et al. (2015) based on theoretical aspects. However, only a few studies explicitly mention and apply the scaling (Macedo et al., 2020a; Meyer, 2021). Kudinov et al. (2022) explain their decision not to use the scaling factor by stating that the assumption for using the scaling factor is that all subpopulations described by the MF equally contribute to the base population. This assumption is not met in their data set.

In this study, it was shown that scaled variance components, compared with non-scaled variance components, instead of slight underestimation tend to result in a moderate overestimation of GEBV. Regarding dispersion, scaled variance components have a negative effect and lead to a slight overdispersion. These effects were already described for a scenario with a complete pedigree and only 2 MF in Himmelbauer et al. (2023b). The effects of scaled variance components in this study are...
comparable, although to a smaller extent, with those found in Himmelbauer et al. (2023a) for the scenario with too high heritability. This might be an indication that scaling ends up with a slightly too high heritability. Interestingly, in some test runs based on data sets with a complete pedigree and 2 MF for the 2 base populations, it was found that the results of a variance component estimation considering the 2 MF match with the scaled variance components (results not shown).

It is also noteworthy that in this study it seems that scaled variance components only affect the GEBV of the animals themselves, not the estimation of the effects of the MF. This can be concluded from the fact that the estimated effects of the MF for MF_true and MF_sc are exactly the same (Figures Figure 8 and Figure 9), with the only difference between these 2 evaluations being the scaled variance components.

Currently, the authors of the present study are not aware of any other publications describing the effect of scaling variance components on the GEBV. Based on the results found here, it seems appropriate to avoid scaling, as it does not lead to improvement but rather a deterioration in the validation results.

**Estimates for UPG and MF**

The comparison of true and estimated effects for MF and UPG reveals that the evaluations with MF predict the true effects very well. This was also found in Bradford et al. (2019). An exception are estimates of MF_est for groups close to the pedigree base, which are partly overestimated in both scenarios, especially in settings with estimated pedigree. The likely reason for this is that in the estimation of Γ, the relationships between this old MF and other MF, or the inbreeding within this MF, were incorrectly estimated. The comparison of estimated and true base allele frequencies has shown that the base allele frequencies in groups, which are distant from young cohorts where routine genotyping takes place, are estimated less precisely, which of course affects Γ.

Another phenomenon are the estimated trends from MF_est which appear to be downward biased in the scenario with high pedigree completeness and true incomplete pedigree for all subpopulations and males as well as females. Interestingly, in this scenario and setting, MF_true and MF_sc also underestimated the trend for subpopulation B. This might indicate a slightly insufficient coverage of MF with genotypes and/or phenotypes, meaning that there may not be enough information to estimate the MF effects correctly, which therefore leads to a biased genetic trend of UPG (Tsursuta et al., 2014).

However, for almost all scenarios and pedigree settings, evaluation methods PED, UPG_alteredQP, and UPG_fullQP tend to significantly underestimate the genetic trend. Similar results were shown in the study by Bradford et al. (2019). Masuda et al. (2021) also compared results for the trend in UPG and MF, with similar findings. The only exception is the scenario with low pedigree completeness and true incomplete pedigree. Here, only UPG_fullQP shows a clear underestimation of the trend. The reason for this could be that, compared with the scenario with high pedigree completeness, UPGs are more frequent in the pedigree in this scenario, providing more genotype and phenotype information for estimation.

**Estimated validation results using LR method**

The main conclusion is that evaluations with MF still show a very well or even better performance in most scenarios and pedigree settings compared with other evaluations. However, the differences in validation statistics, especially in the scenario with low pedigree completeness, are significantly smaller than for true validation statistics. It is noteworthy that the extreme bias and dispersion of UPG_fullQP are not reflected in the real validation statistics. Such validation methods can only detect bias and dispersion if bias and dispersion are corrected in the evaluation with the complete data set and there is no or at least less bias and over- or underdispersion in the (G)EBV from the complete data set (Himmelbauer et al., 2023b). This does not seem to be the case here, resulting in the GEBV from UPG_fullQP appearing almost unbiased.

Somewhat surprising is that MF_true and MF_est show a small but noticeable bias in the scenario with low pedigree completeness, which, according to the validation with TBV, is not present. It seems that there is a slight overestimation when using the full data set. The lower reliability of MF estimates is also somewhat surprising.

In summary, it can be stated that especially in populations with many unknown pedigrees, routinely applicable validation methods underestimate the superiority of MF compared with other evaluations, especially UPG_fullQP. This makes it even more challenging in practice to find an optimal model for considering unknown parents. Stochastic simulations such as the one used in this study can help to some extent and provide important insights, but they can never reflect the full complexity of the real data.

**CONCLUSION**

In summary, this study shows that already in a scenario with 2 base populations and incomplete pedigree, using MF in ssGBLUP significantly improves bias and dispersion in the youngest animals compared with evaluations with UPG. This applies to populations with high and low pedigree completeness. Additionally, a method for classi-
fying animals into MF was tested, showing that genotype information is helpful to define UPG or MF in this study. Computing $\Gamma$ from allele frequencies and incorporating genotypes from purebred and crossbred animals, leads to similar results as using the true $\Gamma$. Scaling of variance components did not improve the validation results in this study, but even worsened them in some cases. Although evaluations with MF have been found to be particularly advantageous in terms of unbiased GEBV and accurate estimation of MF effects, this study also shows that validation methods used in routine evaluations likely underestimate the advantages.

ACKNOWLEDGMENTS

This study was partly funded by the project “FoKUHs,” a project of Cattle Breeders Austria (Vienna), aimed to establish systematic herd genotyping in farms with complete recording of health data and finally change the genomic evaluation to the single-step method in Austria. The authors would like to acknowledge all farmers and breeding associations for their participation in the project. Gratitude is expressed to the Austrian Federal Ministry of Agriculture, Forestry, Regions, and Water Management (Vienna) and the Federal Provinces of Austria for their generous financial support of this project.

The authors have not stated any conflicts of interest.

REFERENCES


Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future


VanRaden, P.M. 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss 94.


