Single-step genomic BLUP with many metafounders

Andrei A. Kudinov1*, Minna Koivula1, Gert P. Aamand2, Ismo Strandén1 and Esa A. Mäntysaari1

1Natural Resources Institute Finland (Luke), Jokioinen, Finland, 2Nordic Cattle Genetic Evaluation, Aarhus, Denmark

Single-step genomic BLUP (ssGBLUP) model for routine genomic prediction of breeding values is developed intensively for many dairy cattle populations. Compatibility between the genomic (G) and the pedigree (A) relationship matrices remains an important challenge required in ssGBLUP. The compatibility relates to the amount of missing pedigree information. There are two prevailing approaches to account for the incomplete pedigree information: unknown parent groups (UPG) and metafounders (MF). Unknown parent groups have been used routinely in pedigree-based evaluations to account for the differences in genetic level between groups of animals with missing parents. The MF approach is an extension of the UPG approach. The MF approach defines MF which are related pseudo-individuals. The MF approach needs a Γ matrix of the size number of MF to describe relationships between MF. The UPG and MF can be the same. However, the challenge in the MF approach is the estimation of Γ having many MF, typically needed in dairy cattle. In our study, we present an approach to fit the same amount of MF as UPG in ssGBLUP with Woodbury matrix identity (ssGTBLUP). We used 305-day milk, protein, and fat yield data from the DFS (Denmark, Finland, Sweden) Red Dairy cattle population. The pedigree had more than 6 million animals of which 207,475 were genotyped. We constructed the preliminary Gamma matrix (Γpre) with 29 MF which was expanded to 148 MF by a covariance function (Γ148). The quality of the extrapolation of the Γpre matrix was studied by comparing average off-diagonal elements between breed groups. On average relationships among MF in Γ148 were 1.8% higher than in Γpre. The use of Γ148 increased the correlation between the G and A matrices by 0.13 and 0.11 for the diagonal and off-diagonal elements, respectively. [GIEBV were predicted using the ssGTBLUP and Pedigree-BLUP models with the MF and UPG. The prediction reliabilities were slightly higher for the ssGTBLUP model using MF than UPG. The ssGBLUP MF model showed less overprediction compared to other models.

KEYWORDS

genetic groups, genomic evaluation, red dairy cattle, finncattle, co-variance function
1 Introduction

Genomic prediction in dairy cattle started in 2009 for US Holsteins, Jersey, and Brown Swiss (Wiggans et al., 2017). Since then, most dairy populations publish genomic estimated breeding values (GEBV) using a multi-step approach (Masuda et al., 2022). The term “multi” stands for a cascade of steps used to obtain GEBV: calculation of pseudo-observations for genotyped proven bulls and cows, estimation of SNP effects, prediction of direct genomic values, and blending of genomic values with pedigree index (Wiggans et al., 2011). In contrast, a single-step genomic BLUP (ssGBLUP) model accounts for pedigree, phenotypic, and genomic data simultaneously to obtain GEBVs for all animals (Legarra et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). Despite the preselection bias in the multi-step GEBV (Patry and Ducrocq, 2011) and the benefits of the single-step model (Legarra et al., 2014), the latter is used only for a few dairy populations (Mäntysaari et al., 2020; Misztal et al., 2020; Masuda et al., 2022). High computational load, compatibility challenges for the genomic and the pedigree relationship matrices, and improper accounting of unknown parents impede the wide implementation of the single-step approach (Mäntysaari et al., 2020). In dairy cattle, these problems can be expected to be amplified due to the many generations in pedigrees, intensive selection, and the vast exchange of breeding material between populations.

The original ssGBLUP requires the inverse matrices of \(A_{22}\) and \(G\) in the inverted joint relationship matrix \(H^{-1}\) (Aguilar et al., 2010; Christensen and Lund, 2010) where \(A_{22}\) is the pedigree relationship matrix of the genotyped animals and \(G\) is the genomic relationship matrix. When the number of genotyped animals in the \(G\) matrix \((n)\) exceeds the number of markers \((m)\), direct inversion of the \(G\) matrix is not possible without regularization such as adding a small value to the diagonal or a residual polygenic matrix (Mäntysaari et al., 2017). When \(n \gg m\), the single-step method becomes computationally challenging. Several computational approaches have been proposed for the computation of \(G^{-1}\) to allow feasible application of ssGBLUP for large datasets (see review by Misztal et al., 2020). For instance, the method called ssGTBLUP (Mäntysaari et al., 2017) uses the relationship matrix of genotyped animals \((A_{22})\) as the regularization matrix to avoid singularity, the Woodbury matrix identity for the \(G\) inverse, and a sparse presentation of the \(A_{22}^{-1}\) to solve the computational challenges. In the data set with 178K genotyped animals to obtain GEBV, the ssGTBLUP model used 33% of the memory and 55% of the wall-clock time needed by the original ssGBLUP (Koivula et al., 2021a).

The difference in average off- and diagonal elements of \(A_{22}\) and \(G\) matrices is known as a single-step compatibility issue (Vitezica et al., 2011). To balance the matrices implies adjusting either the pedigree or the genomic relationship matrix to make the matrices more similar. The concept of the \(A\) adjustment was suggested by Christensen (2012) and further developed into the metafounder (MF) approach (Legarra et al., 2015). Metafounders are related inbred pseudo-individuals that are used as unknown parents in the pedigree. Relationships between MF are described by a covariance matrix (\(\Gamma\)), which is used to build a relationship matrix \(A^\Gamma\). Estimation of \(\Gamma\) can be based on estimates of base allele frequencies (AF) for each MF (García-Baccino et al., 2017). An important assumption of the MF approach is that the \(G\) matrix is constructed with all AF equal to 0.5 (Legarra et al., 2015). Applicability of MF has been shown in livestock (Koivula et al., 2021b, 2022), sheep (Granado-Tajada et al., 2020), and pig (Xiang et al., 2017) data sets. The MF approach was also reported as a perfect choice for multi-breed evaluations in case computation of accurate \(\Gamma\) is possible (Poulsen et al., 2022).

The number of MF in the reported studies on ssGBLUP in the large dairy cattle breeds has nearly always been less than the number of unknown parent groups (UPG). Allocation of few MF by breed or by breed by time help to achieve an accurate estimation of \(\Gamma\) due to the even distribution of MF across genotyped animals (Kudinov et al., 2020; Masuda et al., 2021). When MF are used in ssGBLUP, it would be natural to use the same number of MF as there are UPG in the pedigree-based animal model (PBLUP). However, accurate estimation of an unstructured \(\Gamma\) matrix of large size is difficult, especially if some of the UPG groups have no descendants among the genotyped animals or genotyped individuals are several generations away.

The aim of this study was to propose an approach to construct \(\Gamma\) with the same number of MF as routinely defined UPG. The proposed approach was applied to the Red Dairy Cattle 305-day data and pedigree used for the milk production evaluation in Nordic countries (Denmark, Finland, and Sweden). Both PBLUP and ssGTBLUP models were used. The predictions used either UPG or MF in equal numbers. Thus, the predictive performance of four models was investigated.

2 Materials and methods

2.1 Data

Data were 305-day milk, protein, and fat yield records from three lactations of Nordic Red Dairy Cattle (RDC), Finnish Holstein (HOL), and Finncattle (FIC) cows. Records were from January 1988 to June 2021. The total number of records by trait were: 9.45, 8.99, and 8.98 million for milk, protein, and fat, respectively (Table 1). Pedigree included 6.05 million cows and 118,363 bulls, of which 8,427 were RDC and 278 were FIC proven bulls. Genetic groups were defined as breed x country x five- or 10-year period for RDC, and as breed x five- or 10-year period for HOL, FIC, and other breeds. In total, there were 148 groups: 61 RDC, 45 HOL, 16 FIC, and 26 for breed group OTHER. The group OTHER included 23 breeds majorly beef cattle.

Frontiers in Genetics 02
### TABLE 1 Number of records by lactation, trait, and breed in 305-day Nordic (Denmark, Finland, Sweden) Red Dairy cattle production data.

| Lactation | RDC  | HOL  | FIC  |
|-----------|------|------|------|
|           | I    | II   | III  | I    | II   | III  | I    | II   | III  |
| Milk      | 3,468,211 | 2,516,689 | 1,546,056 | 837,905 | 628,763 | 394,169 | 27,620 | 19,227 | 12,103 |
| Protein   | 3,362,027 | 2,414,026 | 1,466,707 | 771,209 | 567,914 | 349,911 | 24,999 | 17,381 | 10,896 |
| Fat       | 3,361,935 | 2,413,956 | 1,466,662 | 771,214 | 567,913 | 349,911 | 24,999 | 17,382 | 10,896 |

*Red Dairy cattle.*

*Finnish Holstein.*

*Finncattle.*

### TABLE 2 Groups used to compute preliminary Γ matrix.

| Breed  | Origin  | Birth years of the animals descending from the group (MF) | Abbreviation | Genotype set  | Proportion of genotyped animals tied to the group (%) |
|--------|---------|--------------------------------------------------------|--------------|---------------|-----------------------------------------------------|
| RDC    | FIN     | <1970                                                  | FIN70        | RDC           | 3.62                                                |
| RDC    | FIN     | 1971–1980                                             | FIN80        | RDC           | 7.35                                                |
| RDC    | FIN     | 1981–1990                                             | FIN90        | RDC           | 7.91                                                |
| RDC    | FIN     | 1991–2000                                             | FIN90        | RDC           | 6.17                                                |
| RDC    | FIN     | 2001–2010                                             | FIN10        | RDC           | 7.29                                                |
| RDC    | FIN     | 2011–2020                                             | FIN20        | RDC           | 1.51                                                |
| RDC    | SWE     | <1970                                                  | SWE70        | RDC           | 4.20                                                |
| RDC    | SWE     | 1971–1980                                             | SWE80        | RDC           | 6.85                                                |
| RDC    | SWE     | 1981–1990                                             | SWE90        | RDC           | 11.00                                               |
| RDC    | SWE     | 1991–2000                                             | SWE10        | RDC           | 8.68                                                |
| RDC    | SWE     | 2001–2010                                             | SWE20        | RDC           | 1.14                                                |
| RDC    | DNK     | <1980                                                  | DNK80        | RDC           | 2.17                                                |
| RDC    | DNK     | 1981–1990                                             | DNK90        | RDC           | 3.76                                                |
| RDC    | DNK     | 1991–2000                                             | DNK10        | RDC           | 4.26                                                |
| RDC    | DNK     | 2001–2010                                             | DNK20        | RDC           | 4.61                                                |
| RDC    | NOR     | <2000                                                  | NOR20        | RDC           | 0.80                                                |
| RDC    | NOR     | 2001–2020                                             | NOR20        | RDC           | 4.51                                                |
| RDC    | ANY     | <2000                                                  | RDC00        | RDC           | 0.24                                                |
| RDC    | ANY     | 2001–2020                                             | RDC20        | RDC           | 5.93                                                |
| FIC    | FIN     | <1990                                                  | FIC90        | RDC           | 1.20                                                |
| FIC    | FIN     | 1991–2000                                             | FIC10        | RDC           | 0.37                                                |
| FIC    | FIN     | 2000–2020                                             | FIC20        | RDC           | 0.15                                                |
| OTHER  | ANY     | <2020                                                  |OTH20         | RDC           | 2.60                                                |
| HOL    | ANY     | <1970                                                  | HOL70        | HOL           | 13.42                                               |
| HOL    | ANY     | 1971–1990                                             | HOL90        | HOL           | 6.35                                                |
| HOL    | ANY     | 1991–2010                                             | HOL10        | HOL           | 3.06                                                |
| HOL    | ANY     | 2010–2020                                             | HOL20        | HOL           | 1.41                                                |

*Breedes were RDC, Read Dairy Cattle; FIC, Finncattle; HOL, Holstein, and OTHER, other than listed.*

*Countries of origin: FIN, Finland; SWE, Sweden; DNK, Denmark; NOR, Norway, and ANY, any than specially specified.*

*RDC, and HOL, genotypes set include 46,914 and 46,342 markers, respectively. RDC, set included FIC, genotypes imputed along with RDC, genotypes.*

*Proportion of genotypes tied to the group in a particular genotype set.*
Genomic data were used from 206,140 RDC animals (6,018 proven bulls and 85,142 cows with records) and 1,335 FIC animals (160 proven bulls and 845 cows with records). Before 2019 the bulls were genotyped with Illumina Bovine SNP50 array and most cows with Illumina Bovine LD array (Illumina, San Diego, CA, USA). Since 2019 both bulls and cows were genotyped with Eurogenomics EG MD array (https://www.eurogenomics.com/). Quality control and imputation of genotypes to 46,914 SNPs were performed by NAV (Nordic Genetic Evaluation, Denmark). Genomic markers were not filtered on minor allele frequency and no edits were done concerning across and within breeds polymorphism. HOL genotypes were not presented in the current study.

## 2.2 Statistical models

Four prediction models were investigated using a multi-trait multi-lactation model: single-step GTBLUP with UPG in $H^{-1}$ (ssUPG), single-step GTBLUP with MF (ssMF), pedigree-based BLUP with UPG in $A^{-1}$ (pUPG), and pedigree-based BLUP with MF (pMF). The traits were milk, protein, and fat yield in three lactations i.e. - nine traits total. The linear mixed effects model was:

$$y = Xb + Zu + e,$$

where $y$ is the vector of phenotypes, $X$ is the design matrix relating fixed effects to the phenotypes, $b$ is the vector of fixed effects, $Z$ is the design matrix relating the breeding values to the phenotypes, $u \sim N(0, A\sigma^2_u)$ is the vector of random animal breeding values, and $e \sim N(0, I\sigma^2_e)$ is the residual vector. Matrix $A$ is the pedigree-based relationship matrix, $I$ is an identity matrix, $\sigma^2_u$ and $\sigma^2_e$ are genetic and residual variances, respectively. Fixed effects in $b$ were calving year by season, calving age, herd by year, and calving age by breed. Calving age by breed effect consists of linear ($\alpha$), quadratic ($\alpha^2$), and cubic ($\alpha^3$) regression coefficients of calving age multiplied by pedigree-based breed proportions of an animal (Lidauer et al., 2010; Christensen and Lund, 2010) so that the general level of breed remained to be modeled by $u$. The regression coefficients were centered over all data to zero according to mean calving age as $\alpha = (\text{calving age} - \text{calving age})/365$.

### 2.2.1 Single-step GTBLUP

The mixed model equations (MME) of the original ssGBLUP model require the inverse of a joint relationship matrix $H^{-1}$ (Aguilar et al., 2010; Christensen and Lund, 2010):

$$H^{-1} = A^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{pmatrix},$$

where $A_{22}$ is the part in $A$ for the genotyped animals, and $G$ is the genomic relationship matrix (VanRaden, 2008). Regularization matrix $C = wA_{22}$ was added to the marker-based matrix $G$, where $w$ is the residual polygenic proportion, i.e., the genomic relationship matrix was $G_{c,w} = (1-w)G + wA_{22}$ (Mäntysaari et al., 2017). We used $w$ equal to 30% to keep the comparability to the studies by Koivula et al. (2021b, 2022). The $G_{c,w}$ matrix was constructed with the assumption that AF of all markers was equal to 0. Thus, $G_{c,w} = (1-w)Z_{101}Z_{101}^T/k + wA_{22}$ where $k = m/2$ is the scaling factor, $m$ is the number markers, and $Z_{101}$ is the matrix of genotype counts with values of 0 for the heterozygote and values -1 and +1 for homozygotes. The inverse genomic relationship matrix can be expressed as (Mäntysaari et al., 2017) $G_{c,w}^{-1} = 1/k^2A_{22}^{-1} - T_u^T T_w$ where $T_u = \frac{1}{k}Z_{101} Z_{101}^T A_{22}^{-1} Z_{101}$ and $L_w$ is the Cholesky decomposition of $\frac{1}{k^2}Z_{101} Z_{101}^T A_{22}^{-1} Z_{101}^T + \frac{1}{k}I$.

#### 2.2.2 Single-step GTBLUP with UPG

The joint relationship matrix augmented by UPG (Quaas and Pollak, 1981; Misztal et al., 2013; Matilainen et al., 2018) was computed as shown in Koivula et al. (2021a):

$$H^{-1} = A_{\text{UPG}}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & B_{11} B_{12} \\ 0 & B_{11} B_{22} \end{pmatrix},$$

where

$$A_{\text{UPG}}^{-1} = \begin{pmatrix} A_{11} & A_{12} & -A_{11} Q_1 + A_{12} Q_2 \\ A_{12} & A_{22} & -A_{12} Q_1 + A_{22} Q_2 \\ -A_{11} Q_1 + A_{12} Q_2 & -A_{12} Q_1 + A_{22} Q_2 & Q A^T A^{-1} Q \end{pmatrix},$$

and

$$B = \frac{1 - w}{w} \begin{pmatrix} A_{11}^{-1} & -A_{11}^{-1} Q_1 \\ -Q A_{22}^{-1} & Q A_{22}^{-1} \end{pmatrix} = \begin{pmatrix} T_u^T T_w & -T_u^T T_w Q_1 \\ -Q T_u^T T_w & Q T_u^T T_w Q_1 \end{pmatrix}.$$

The $Q$ matrix has proportions of genes contributed from each UPG according to pedigree information. The subscripts 1 and 2 in $Q$ pertain to genotyped and non-genotyped animals. Subscripts 1 and 2 in $B$ pertain to genotyped animals and UPGs, respectively. The UPGs were modeled as random effect. Inbreeding coefficients were accounted in both pedigree-based relationship matrices.

### 2.2.3 Single-step GTBLUP with MF

In the MF approach (Christensen, 2012; Legarra et al., 2015), the $H^{-1}$ matrix was replaced by:

$$(H^*)^{-1} = (A^*)^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G_{c,w}^{-1} - A_{22}^{-1} \end{pmatrix},$$

where $G_{c,w} = (1-w)G + wA_{22}$, $A^*$ is pedigree relationship matrix formed with a $\Gamma$ matrix, $A_{22}^*$ is the submatrix of $A^*$ for the genotyped animals, and $\Gamma$ was variance covariance matrix of the MF. Inbreeding coefficients estimated using the $\Gamma$ matrix were used in the inverses of $(A^*)^{-1}$ and $(A_{22}^*)^{-1}$.

Frontiers in Genetics 04

frontiersin.org
2.2.4 Pedigree BLUP

The pedigree-based models (pUPG and pMF) were similar to their corresponding single-step models, except that the genomic data was excluded from the prediction. In pUPG model UPGs were accounted in \( (A_{\text{UPG}}^{-1} \) Quas and Pollak, 1981). In pMF model the \( A^{-1} \) was replaced by \((A')^{-1}\).

2.5 Estimation of the \( \Gamma \) matrix

Let the number of MF be \( r \) such that the \( \Gamma \) matrix has size \( r \times r \). In the MF approach, the \( \Gamma \) matrix describes the variance-covariance structure of MF. It can be estimated by \( \text{8Gov}(P) \), where \( P \) is an \( m \times r \) matrix of estimated base population AF for each marker and MF (Legarra et al., 2015; Garcia-Baccino et al., 2017). In the studied data set, 44% of the UPG were not linked to the following general steps were used:

a) Estimate allele frequencies for a set of base groups;
b) From estimated allele frequencies, calculate the preliminary \( \Gamma \) matrix \( (\Gamma_{\text{pre}}) \) for the base groups;
c) Solve the matrix \( K \) in the covariance function \( \Gamma_{\text{pre}} \Phi_{\text{pre}} K \Phi_{\text{pre}}' \) \( E \) using \( \Gamma_{\text{pre}} \) and the model matrix \( \Phi_{\text{pre}} \); The matrix \( E \) is null if row rank of \( \Phi_{\text{pre}} \) is equal to dimension of \( \Gamma_{\text{pre}} \) and, if not \( E \) represents least squares errors of estimation.
d) Compute the \( \Gamma \) for the large number of groups as \( \Phi_{\Gamma}K\Phi_{\Gamma}' \).

The model matrices \( \Phi_{\text{pre}} \) and \( \Phi_{\Gamma} \) define linear model by group and time for the set of base MF and for all MF, respectively.

The technical detailed steps used to compute \( \Gamma \) for 148 groups \( (\Gamma_{148}) \) were:

a) The pedigree was pruned to include only one ancestor generation of genotyped animals as in Kudinov et al. (2020). Truncation of the pedigree helped to achieve equal distribution of the genomic information over UPGs. Missing parents in the truncated pedigree were replaced by 26 groups formed by breed, country, and time interval (Table 2). All HOL ancestors were assigned to the same group regardless of country and time. Estimation of base population AF for each of the groups \( P_{\text{RDC}} \) was performed using the GLS method (McPeck et al., 2004; Garcia-Baccino et al., 2017). The HOL group estimated from RDC genotypes was dropped from \( P_{\text{RDC}} \). The HOL AF \( (P_{\text{HOL}}) \) were the same as used in Kudinov et al. (2020)—calculated using Holstein genotypes (M. Koivula, personal communication). The joint \( P_{\text{RDC,HOL}} \) matrix of size 29 by 45,823 was created by merging compatible SNPs in \( P_{\text{RDC}} \) and \( P_{\text{HOL}} \). Number of SNPs dropped from \( P_{\text{RDC}} \) and \( P_{\text{HOL}} \) where 1,091 and 519, respectively.

b) Three \( \Gamma \) matrices \( (\Gamma_{\text{RDC,HOL}}, \Gamma_{\text{RDC}}, \text{ and } \Gamma_{\text{HOL}}) \) were computed using \( P_{\text{RDC,HOL}}, P_{\text{RDC}}, \text{ and } P_{\text{HOL}} \). A pre-\( \Gamma \) matrix \( (\Gamma_{\text{pre}} \text{ Figure 1}) \) was created by replacing the diagonal elements of \( \Gamma_{\text{RDC,HOL}} \) by diagonal elements of \( \Gamma_{\text{RDC}} \) and \( \Gamma_{\text{HOL}} \) at corresponding places. The diagonal values in \( \Gamma_{\text{pre}} \) were larger than in \( \Gamma_{\text{RDC,HOL}} \).

c) Structure of \( \Gamma_{\text{pre}} \) was computed with covariance function \( \Phi_{\text{pre}} K \Phi_{\text{pre}}' \) (Kirkpatrick et al., 1990), where \( \Phi_{\text{pre}} \) was a model matrix having standardized year of the MF (Appendix 1) and \( K \) was a matrix of co-variance function coefficients. Year standardization was done using formula \( (\text{year}_{\text{max}} - \text{year}_{\text{min}}) / 2 \), where \( \text{year}_{\text{max}} \) is a year of the MF, \( \text{year}_{\text{min}} \) and year_{max} are the first (1950) and the last (2021) year points among the 148 groups in the pedigree.

The matrix \( K \) was estimated as (Tijani et al., 1999):

\[
\mathbf{K} = (\Phi_{\text{pre}})^{-1} \mathbf{P}_{\text{pre}} \mathbf{P}_{\text{pre}} \gamma_{\text{pre}} (\Phi_{\text{pre}})^{-1} - 1
\]

leading to estimate

\[
\mathbf{K} = \begin{bmatrix}
0.0108 & 0.0093 & 0.0083 & -0.0042 & -0.0004 & 0.0003 & -0.0058 & -0.0054 & 0.0046 \\
0.6563 & 0.6203 & 0.5551 & 0.5961 & 0.5850 & 0.5511 & 0.5343 & 0.5158 \\
0.6349 & 0.5621 & 0.6094 & 0.5795 & 0.5434 & 0.5145 & 0.5164 & 0.5022 \\
0.6098 & 0.5544 & 0.5644 & 0.5467 & 0.5778 & 0.5358 & 0.5358 & 0.5159 \\
0.6353 & 0.6578 & 0.5441 & 0.5375 & 0.5358 & 0.5358 & 0.5358 & 0.5358 \\
0.5899 & 0.5442 & 0.5458 & 0.5325 & 0.5325 & 0.5325 & 0.5325 & 0.5325 \\
0.6868 & 0.5530 & 0.5121 & 0.5285 & 0.5285 & 0.5285 & 0.5285 & 0.5285 \\
0.6946 & 0.5285 & 0.5285 & 0.5285 & 0.5285 & 0.5285 & 0.5285 & 0.5285 \\
0.6653
\end{bmatrix}
\]

d) Finally, the \( \Gamma_{148} \) was estimated as \( \Phi_{148} \Phi_{148}' \) where the matrix \( \Phi_{148} \) was designed same way as \( \Phi_{\text{pre}} \) but using all the MF. Rank of the \( \Gamma_{148} \) matrix is only 9. To avoid \( A' \) matrix singularity we reduced the off-diagonal values of \( \Gamma_{148} \) by 2.5% and increased the diagonal values by 2.5%

2.6 Validation of model fit

Validation of the prediction models was done using modified forward prediction (Mäntysaari et al., 2010). For the validation a reduced phenotypic data set was constructed by removing records from the last 4 years of data, i.e., June 2017 to June 2021. Daughter yield deviations (DYD) for bulls and yield deviations (YD) for cows were computed using the full data set using the same model which was applied to reduced data. Bias of evaluation was estimated by the linear regression coefficient \( (h_i) \) from the weighted regression of DYD/YD on the corresponding [G]EBV predicted with the reduced data. The weight of DYD for bull \( i \) was \( \text{EDC}_i / (\text{EDC}_i + \lambda_b) \), where \( \lambda_b \) is \((4-h^2)/h^2\), \( h^2 \) is heritability of the trait, and EDC\( _i \) is the effective daughter contributions of bull \( i \) computed as in Taskinen et al. (2014). Weight for cow YD\( _i \) was computed as \( \text{ERC}_i / (\text{ERC}_i + \lambda_c) \), where \( \lambda_c \) is \((1-h^2)/h^2\) and ERC\( _i \) is the effective record contribution of cow \( j \) (Přibyl et al., 2013). Adjusted validation reliability was attained by dividing...
the coefficient of determination from the regression model ($R^2$) by the average weight of DYD ($R^2_{DYD}$) and YD ($R^2_{YD}$) for bulls and cows, respectively. Average genetic trends were plotted using the trait specific combined [G]EBVs computed as 

$$[G]\text{EBV}_{\text{parity1}} \times 0.30 + [G]\text{EBV}_{\text{parity2}} \times 0.25 + [G]\text{EBV}_{\text{parity3}} \times 0.45$$

(https://nordicebv.info/wp-content/uploads/2021/10/NAV-routine-genetic-evaluation_EDITYSS-08102021.pdf).

2.7 Software

Pedigree truncation and estimation of the inbreeding coefficients was done using RelaX2 v.1.95 software. The AF were estimated using Bpop v. 0.98 program (Strandén and Vuori, 2006; Strandén and Mäntysaari, 2020), T matrix and its diagonal needed in the ssGTBLUP model were computed using hgtinv v.0.83 program. The computation of [G]EBV predictions and the estimation of EDC/ERC used MiX99 software (Strandén and Lidauer, 1999). MiX99 software uses preconditioned conjugate gradient (PCG) iteration. The PCG method was assumed to be converged when convergency criteria $< 1e-6$ was achieved. Convergency criteria was defined as a Euclidean norm of the difference between the right-hand side (RHS) of the MME and the one predicted by the current solutions relative to the norm of RHS. The matrices ($A_{22}^{-1}$) and ($A_{12}^{-1}$) used by MiX99 and hginv were constructed using the given pedigree and inbreeding files, and in case of MF, by file with the $\Gamma^{-1}$ matrix.

3 Results and discussion

3.1 Relationship matrices

Elements of $\Gamma_{pre}$ ranged from 0.59 to 0.74 and from 0.51 to 0.69 for the diagonal and off-diagonal elements, respectively. The lowest and highest diagonal values (self-relationship, Legarra et al., 2015) were in groups HOL 1960 and RDC FIN 1990, respectively. In the $\Gamma_{idh}$ matrix, diagonal elements were in a range
from 0.61 to 0.73 (Figure 2). The lowest and highest self-relationships were in HOL SWE 1970 and OTHER 1960 groups, respectively. The off-diagonal elements of $\Gamma^{148}$ ranged from 0.48 to 0.69. The highest average relationships were observed between the FIN and SWE RDC groups, as expected. Relationships between HOL and RDC DNK were higher than with the other RDC groups due to the larger proportion of HOL sires in the RDC DNK pedigree. Similarly, the FIC groups were genetically closer to RDC FIN than to the other groups due to historical crossbreeding. Relationship coefficients between the RDC subgroups in our study ranged from 0.54 to 0.65 which was much higher than the range 0.09–0.18 presented between the biological types of Montana cattle breed (Kluska et al., 2021). Average relationships between RDC and HOL breed (0.52) was close to presented between HOL and Jersey breeds (0.48, Legarra et al., 2015).

Because $\Gamma^{148}$ is an extrapolated matrix of $\Gamma^{pre}$ we expect these to be alike. The difference between the two matrices was assessed...
values in $\mathbf{A}_{22}$ in our study was higher than presented for HOL in Koivula et al. (2022). Average relationship coefficients of $\mathbf{A}_{22}$, $\mathbf{A}_{22}^\Gamma$, and $\mathbf{G}_{05}$ in Koivula et al. (2022) had a steady increase by animal’s birth year. However, similar behavior in our study was observed only for the $\mathbf{A}_{22}$ matrix. A slight decrease in the average relationship coefficient of $\mathbf{G}_{05}$ and $\mathbf{A}_{22}^\Gamma$ was observed after year 2000. This might be caused by the establishment of the joint Nordic RDC evaluation and admixture of the breeds in the three populations. The total increase in the average relationships in the 40-year period were 2.81%, 0.97%, and 0.72% for $\mathbf{A}_{22}$, $\mathbf{A}_{22}^\Gamma$, and $\mathbf{G}_{05}$, respectively. The MF approach is beneficial in an admixed population such as Nordic RDC, as it helps to balance the $\mathbf{G}_{05}$ and $\mathbf{A}_{22}$ matrices. However, $\mathbf{A}_{22}^\Gamma$ and $\mathbf{G}_{05}$ were not on exactly on the same scale as the mean diagonal and off-diagonal elements in $\mathbf{A}_{22}$, were still somewhat lower than in $\mathbf{G}_{05}$ (Table 4). So, some compatibility issues between the pedigree- and genomic-based relationship matrices remained.

In addition to $\mathbf{G}_{\text{pre}}^{\text{RDC-HOL}}$ was tested as source for $\mathbf{G}_{148}^{\text{pre}}$ and corresponding $\mathbf{A}_{22}^{\text{pre}}$ estimation. The mean difference between $\mathbf{G}_{\text{pre}}^{\text{RDC-HOL}}$ and $\mathbf{G}_{\text{pre}}^{\text{RDC-HOL}}$ was 0.03. Diagonal elements in $\mathbf{A}_{22}^{\text{pre}}$ constructed using extrapolated $\mathbf{G}_{\text{pre}}^{\text{RDC-HOL}}$ were on average 0.02 lower than in $\mathbf{A}_{22}$ used for genomic prediction. Even though construction of $\mathbf{G}_{\text{pre}}$ with $\mathbf{G}_{\text{RDC}}$ and $\mathbf{G}_{\text{HOL}}$ diagonals helped to lift $\mathbf{A}_{22}$ closer to $\mathbf{G}_{05}$, this step was not vital and $\mathbf{G}_{\text{RDC-HOL}}$ might have been used as is.

Filtering of the SNPs by minor allele frequency (MAF) for the $\mathbf{G}$ matrix estimation was elaborated in our previous study (Kudinov et al., 2020), and indirectly performed in Legarra et al. (2015). In the current study, we avoided MAF filtering of SNPs used to compute $\mathbf{G}_{\text{pre}}$. That helped to compute a $\mathbf{A}_{22}^{\text{pre}}$ matrix closer to $\mathbf{G}_{05}$, as the same set of markers was used to construct $\mathbf{G}_{05}$. We observed that if selection of SNPs is used, it should be applied to both $\mathbf{G}$ and $\mathbf{G}_{05}$, i.e., the same set of markers should be used consistently. It is reasonable to keep the set of markers used in $\mathbf{G}_{05}$ and $\mathbf{G}$ as compatible as possible.

The presented approach in our study allows to fit the same number of MF as UPG and define MF for base population groups not linked to the genotypes. However, approach requires several arbitrary steps that need to be customized for each population. For instance, the definition of the groups will be different in $\mathbf{G}_{\text{pre}}$. We defined the groups in $\mathbf{G}_{\text{pre}}$ by breed, country, and time. If any of defined groups had less than 0.1% of genotyped animals, we have had to combine it. In our study, the definition of the time variable in the base populations used to compute AF was the last year of the time interval, another way is to use mean, median or the first year. Because the year definition in each of the groups is used in the model matrix $\Phi$ and resulting covariance function, average diagonal of $\mathbf{A}_{22}^{\text{pre}}$ would expectedly decrease. The standardized year of the MF in $\Phi$ was computed with the same formula. However, this can be adjusted for specific breed or country. Use of a covariance function in routine genomic prediction need re-estimation of the $\mathbf{G}$ matrix when new genetic groups are defined, but not re-estimation of base AF.

### Table 4 Mean, minimum (Min), and maximum (Max) element values of $\mathbf{A}_{22}$, $\mathbf{A}_{22}^\Gamma$, and $\mathbf{G}_{05}$ from diagonal and off-diagonal.

| Elements | Matrix | Mean | Min | Max |
|----------|--------|------|-----|-----|
|          | $\mathbf{A}_{22}$ | 1.03 | 1.00 | 1.29 |
|          | $\mathbf{A}_{22}^\Gamma$ | 1.31 | 1.25 | 1.50 |
|          | $\mathbf{G}_{05}$ | 1.34 | 1.00 | 1.58 |
| Off-diagonal | $\mathbf{A}_{22}$ | 0.06 | 0.00 | 0.82 |
|          | $\mathbf{A}_{22}^\Gamma$ | 0.61 | 0.50 | 1.17 |
|          | $\mathbf{G}_{05}$ | 0.68 | 0.47 | 1.40 |

$\mathbf{A}_{22}$ - the pedigree relationship matrix of genotyped animals; $\mathbf{A}_{22}^\Gamma$ - the pedigree relationship matrix of genotyped animals augmented by the $\mathbf{G}_{05}$ - the genomic relationship matrix with allele frequencies equal to 0.5.

Using percentage deviation from the mean off-diagonal values in breed groups (Table 3). The average off-diagonal value of $\mathbf{G}_{148}$ was 1.8% higher than in $\mathbf{G}_{\text{pre}}$. For instance, the average relationships between the RDC FIN and FIC groups were 0.54 and 0.56 in $\mathbf{G}_{\text{pre}}$ and $\mathbf{G}_{148}$, respectively. Thus, the covariance function allows to extrapolate the $\mathbf{G}$ matrix for the MF approach in order to have the same number of MF as UPG.

Application of the $\mathbf{G}_{148}$ matrix to the pedigree-based relationship matrix lifted the average diagonal elements of $\mathbf{A}_{22}$ closer to $\mathbf{G}_{05}$ (Figure 3). The smallest diagonal and off-diagonal values of $\mathbf{A}_{22}$ increased by 0.25 (from one to 1.25) and by 0.50 (from 0 to 0.50), respectively, by using $\mathbf{G}_{148}$ as the basis for $\mathbf{A}_{22}^\Gamma$ (Table 4). The increase was close to that in Kudinov et al. (2020) - 0.27 and 0.48 for the diagonal and off-diagonal elements, respectively. The correlations in the diagonal and off-diagonal elements were higher between $\mathbf{G}_{05}$ and $\mathbf{A}_{22}^\Gamma$ (0.70 and 0.88) than between $\mathbf{G}_{05}$ and $\mathbf{A}_{22}$ (0.57 and 0.77). The overall magnitude of
### Table 5: [G]EBV validation test regression coefficients ($b_1$) and weighted validation reliabilities ($R^2_{EDC}$) for RDC validation bulls.

| Model* | ssUPG | ssMF | pUPG | pMF |
|--------|--------|------|------|-----|
| Trait - parity | Number of bulls | $b_1$ | $R^2_{EDC}$ | $b_1$ | $R^2_{EDC}$ | $b_1$ | $R^2_{EDC}$ | $b_1$ | $R^2_{EDC}$ |
| Milk 1 | 284 | 0.80 | 0.45 | 0.82 | 0.48 | 0.82 | 0.20 | 0.87 | 0.21 |
| Milk 2 | 243 | 0.77 | 0.37 | 0.81 | 0.39 | 0.89 | 0.22 | 0.93 | 0.22 |
| Milk 3 | 180 | 0.76 | 0.21 | 0.79 | 0.22 | 0.91 | 0.13 | 0.96 | 0.13 |
| Protein 1 | 287 | 0.64 | 0.34 | 0.66 | 0.36 | 0.54 | 0.11 | 0.59 | 0.11 |
| Protein 2 | 240 | 0.62 | 0.30 | 0.70 | 0.31 | 0.66 | 0.14 | 0.70 | 0.14 |
| Protein 3 | 181 | 0.66 | 0.18 | 0.68 | 0.19 | 0.74 | 0.10 | 0.80 | 0.10 |
| Fat 1 | 284 | 0.68 | 0.40 | 0.71 | 0.41 | 0.63 | 0.17 | 0.67 | 0.17 |
| Fat 2 | 241 | 0.73 | 0.39 | 0.78 | 0.40 | 0.70 | 0.16 | 0.73 | 0.16 |
| Fat 3 | 181 | 0.72 | 0.21 | 0.75 | 0.20 | 0.79 | 0.12 | 0.83 | 0.12 |

*Models are: ssUPG, single-step GTBLUP, with UPG, accounted in $H^1$; ssMF, single-step GTBLUP, with MF; pUPG-pedigree BLUP, with UPG, accounted in $A^1$; and pMF, pedigree BLUP, with MF.

*bRegression coefficient $b_1$ in DYD = $b_0 + b_1[G]EBV$ equation has been multiplied by two.

*cValidation reliability was obtained from coefficient of determination of the model ($R_{model}^2$), after correcting it by the average weight of DYDs ($R^2_{EDC}$).

### Table 6: [G]EBV validation test regression coefficients ($b_1$) and weighted validation reliabilities ($R^2_{ERC}$) for RDC validation cows.

| Model* | ssUPG | ssMF | pUPG | pMF |
|--------|--------|------|------|-----|
| Trait - parity | Number of cows | $b_1$ | $R^2_{ERC}$ | $b_1$ | $R^2_{ERC}$ | $b_1$ | $R^2_{ERC}$ | $b_1$ | $R^2_{ERC}$ |
| Milk 1 | 32,133 | 1.03 | 0.50 | 1.04 | 0.51 | 0.91 | 0.17 | 0.97 | 0.16 |
| Milk 2 | 27,097 | 0.96 | 0.41 | 0.97 | 0.42 | 0.93 | 0.15 | 0.98 | 0.15 |
| Milk 3 | 10,923 | 0.91 | 0.34 | 0.92 | 0.34 | 0.85 | 0.13 | 0.90 | 0.13 |
| Protein 1 | 33,298 | 0.87 | 0.36 | 0.89 | 0.36 | 0.80 | 0.14 | 0.84 | 0.14 |
| Protein 2 | 25,491 | 0.82 | 0.33 | 0.84 | 0.33 | 0.83 | 0.13 | 0.88 | 0.13 |
| Protein 3 | 10,387 | 0.83 | 0.29 | 0.84 | 0.29 | 0.78 | 0.11 | 0.83 | 0.11 |
| Fat 1 | 34,129 | 0.88 | 0.38 | 0.90 | 0.38 | 0.84 | 0.15 | 0.89 | 0.15 |
| Fat 2 | 25,987 | 0.83 | 0.32 | 0.85 | 0.32 | 0.85 | 0.14 | 0.90 | 0.14 |
| Fat 3 | 10,494 | 0.86 | 0.31 | 0.86 | 0.31 | 0.82 | 0.13 | 0.87 | 0.13 |

*Models are: ssUPG, single-step GTBLUP, with UPG, accounted in $H^1$, ssMF, single-step GTBLUP, with MF; pUPG-pedigree BLUP, with UPG, accounted in $A^1$, and pMF, pedigree BLUP, with MF.

*Validation reliability was obtained from coefficient of determination of the model ($R_{model}^2$), after correcting it by the average weight of DYDs ($R^2_{ERC}$).

### 3.2 Model runs and validation

The ssMF and ssUPG models converged in 1,388 and 2,802 iterations. The wall-clock time per iteration was similar for ssMF and ssUPG models. The mix99 runtime for ssMF model in Intel Xeon 2.8 Ghz machine with four cores was 11 h 19 min.

Table 5 presents bull validation results for the nine traits. For all traits, the highest prediction reliability was obtained by the ssMF model. The regression slopes ($b_1$) obtained by ssMF were slightly higher than by ssUPG. Prediction reliability by the pUPG and pMF models were the same. However, the slopes ($b_1$) with pMF were closer to one than with pUPG. In all traits and models, quality of prediction decreased from lactation one to 3. For the validation cows (Table 6), the ssMF model gave slightly better validation reliability in milk than the ssUPG model. For protein and fat, the same...
prediction reliability was estimated in both single-step models. The slope in ssMF was closer to one than in the other models. However, in the first parity milk trait, the slope was slightly above one in ssUPG and ssMF. The MF approach improved quality of genomic prediction in the studied population similarly as reported in Bradford et al. (2019), Masuda et al. (2021), and Koivula et al. (2022) The bias of prediction in the ssMF model was lower as reported for dairy sheep (Macedo et al., 2020). Nonetheless bias in our study remain significant in all single-step models.

Genetic trends for combined milk, fat, and protein GEBVs are presented for the genotyped bulls with at least 50 daughters and all RDC cows in Figures 4, 5, respectively. Average GEBV were centered using the mean GEBV of RDC cows born in 2007. Both genomic and non-genomic models had similar shape in the UPG and MF instance. The average GEBV levels were higher than the
average EBV levels. Similar difference has been observed in other single-step studies (Ma et al., 2015; Silva et al., 2019; Koivula et al., 2022). Overprediction in single-step with MF was reduced in our study similar to reported in Masuda et al. (2021) and Koivula et al. (2022).

Reduction of heritability by additive variance scaling was suggested by Legarra et al. (2015) when MF are used for genomic prediction. Base populations in models with UPGs are assumed unrelated, which is contrary to MF. In order to solve that problem additive variance was suggested to be scaled by \((1 + tr(Γ)/(2n)) − 1\), where \(tr(Γ)\) is the sum of diagonal elements of the \(Γ\) matrix (Legarra et al., 2015). However, this is based on assumption that the current population is a homogenous mixture of all the base populations that the MF will present. In reality base populations have influence unequally to the studied population, and thus we kept the same genetic variances in the UPG and MF models.

4 Conclusion

We presented a method to utilize the same number of MF as UPG in single-step GBLUP. The Covariance functions allowed smooth extrapolation of the \(Γ\) matrix with 29 metafounders to 148 in the pedigree of all animals. Use of \(Γ_{148}\) increased correlation between the elements of pedigree and genomic relationship matrices. The \(Γ_{148}\) matrix was tested in the ssGTBLUP approach and compared with UPG based ssGTBLUP. Results showed a slight improvement in prediction reliability and overprediction in the MF model over the UPG model.

Data availability statement

The data analyzed in this study was obtained from Finncattle Foundation (Finland), Finnish Breeder Association (FABA, Finland), Swedish Cattle Farmers Association (Växa), Landbrug and Fødevarer F.m.b.A (L and F), Nordic Cattle Genetic Evaluation (NAV, Denmark), Viking Genetics (Denmark), and corresponding farmers. Requests to access these datasets should be directed to the Director of Nordic Cattle Genetic Evaluation, Gert P. Aamand, gap@lf.dk.

Author contributions

AK executed the analysis and wrote the manuscript. EM did the study design. MK, GA, IS, and EM provided support during analysis execution, reviewed, and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Project “Genomic evaluations for Western Finncattle” financed by Luke (Finland) and Finncattle Foundation (Finland).

Acknowledgments

We acknowledge Finncattle Foundation (Finland), Finnish Breeder Association (Finland), Nordic Cattle Genetic Evaluation (Denmark), and Viking Genetics (Denmark). We kindly acknowledge two reviewers participated in production of the current manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93, 743–752. doi:10.3168/jds.2009-2730

Bradford, H. L., Masuda, Y., VanRaden, P. M., Legarra, A., and Misztal, I. (2019). Modeling missing pedigree in single-step genomic BLUP. J. Dairy Sci. 102, 2336–2346. doi:10.3168/jds.2018-15434

Christensen, O. F. (2012). Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet. Sel. Evol. 44, 37. doi:10.1186/1297-9686-44-37

Christensen, O. F., and Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42, 2. doi:10.1186/1297-9686-42-2

Garcia-Baccino, C. A., Legarra, A., Christensen, O. F., Misztal, I., Pocrnic, I., Vitezica, Z. V., et al. (2017). Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genet. Sel. Evol. 49, 34. doi:10.1186/s12711-017-0309-2

Granado-Tajada, I., Legarra, A., and Ugarte, E. (2020). Exploring the inclusion of genomic information and metafounders in Latxa dairy sheep genetic evaluations. J. Dairy Sci. 103 (7), 6346–6353. doi:10.3168/jds.2019-18033
random regression test-day model. Model for genomic relationships of metafounders applied on large scale single-step Woodbury matrix identity-based genomic relationship inverse. SSGBLUP with metafounders and unknown parent groups. Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and genomic model improved reliability and reduced the bias of genomic predictions in J. Dairy Sci. 104 (9), 10049–10058. doi:10.3168/jds.2020-19821

Kudinov, A. A., Mäntysaari, E. A., Aamand, G. P., and Strandén, I. (2020). Metafounder approach for single-step genomic evaluations of Red dairy cattle. J. Dairy Sci. 103 (7), 6299–6310. doi:10.3168/jds.2019-17483

Legarra, A., Aguilar, I., and Mårtal, I. (2009). A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92, 4665–4663. doi:10.3168/jds.2009-2061

Legarra, A., Christensen, O. F., Aguilar, I., and Mårtal, I. (2014). Single Step, a general approach for genomic selection. Livestock Sci. 166, 54–65. doi:10.1016/j.livsci.2014.04.029

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

Lidauer, M. H., Pöös, J., Pedersen, J., Lassen, J., Madsen, P., Mäntysaari, E. A., et al. (2015). Across-country test-day model evaluations for Holstein, nordic red cattle, and Jersey. J. Dairy Sci. 98, 1256–1299. doi:10.3168/jds.2014-8307

Ma, P., Lund, M. S., Nielsen, U. S., Aamand, G. P., and Su, G. (2015). Single-step genomic model improved reliability and reduced the bias of genomic predictions in Danish Jersey. J. Dairy Sci. 98, 9026–9034. doi:10.3168/jds.2015-9703

Macedo, E. L., Christensen, O. F., Astruc, J. M., Aguilar, I., Masuda, Y., and Legarra, A. (2020). Bias and accuracy of dairy sheep evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. Genet. Sel. Evol. 52 (47), 47. doi:10.1186/s12711-020-00567-1

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

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

Mäntysaari, E. A., Evans, R., and Strandén, I. (2017) Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. J. Anim. Sci. 95, 4728–4737. doi:10.2527/jas.2017.19112

Masuda, Y., Tsuruta, S., Bermann, M., Bradford, H. L., and Mårtal, I. (2021). Comparison of models for missing pedigree in single-step genomic prediction. J. Anim. Sci. 99 (2), sak019. doi:10.1093/jas/skab019

Masuda, Y., VanRaden, P. M., Shogo, T., Lourenço, D. A. L., and Mårtal, I. (2022). Invited review: Unknown-parent groups and metafounders in single-step genomic BLUP. J. Dairy Sci. 105 (2), 923–939. doi:10.3168/jds.2021-20293

Matilainen, K., Strandén, I., Aamand, G. P., and Mäntysaari, E. A. (2018). Single step genomic evaluation for female fertility in Nordic Red dairy cattle. J. Anim. Breed. Genet. 135, 337–348. doi:10.1111/jbg.12353

McPeek, M. S., Xiaodong, W., and Ober, C. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60, 359–367. doi:10.1111/j.0006-341X.2004.00180.x

Mårtal, I., Lourenço, D. A. L., and Legarra, A. L. (2020). Current status of genomic evaluation. J. Anim. Sci. 98, skaa101. doi:10.1093/jas/skaa101

Mårtal, I., Vitezica, Z. G., Legarra, A., Aguilar, I., and Swan, A. A. (2013) Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130, 252–258. doi:10.1111/jbg.12025

Patry, C., and Ducrocq, V. (2011). Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J. Dairy Sci. 94, 1011–1020. doi:10.3168/jds.2010-3884

Poulsen, B. G., Osterlind, T., Nielsen, B., and Christensen, O. F. (2022). Predictive performances of animal models using different multibreed relationship matrices in systems with rotational crossbreeding. Genet. Sel. Evol. 54 (1), 25–17. doi:10.1186/s12711-022-00714-w

Pribyl, J., Madsen, P., Bauer, J., Pribylova, J., Simečková, M., Vostřá, L., et al. (2013). Contribution of domestic production records, Interbull estimated breeding values, and single nucleotide polymorphisms to the single-step genomic evaluation of milk production. J. Dairy Sci. 96 (3), 1865–1873. doi:10.3168/jds.2012-6157

Quaas, R. L., and Pollak, E. J. (1991). Modified equations for sire models with groups. J. Dairy Sci. 64, 1868–1872. doi:10.3168/jds.1991.64-2777-8

Silva, A. A., Silva, D. A., Silva, F. F., Costa, N. C., Lopes, P. S., Caetano, A. R., et al. (2019). Autoregressive single-step test-day model for genomic evaluations of Portuguese Holstein cattle. J. Dairy Sci. 102 (7), 6530–6639. doi:10.3168/jds.2018-15191

Strandén, I., and Lidauer, M. (1999). Solving large mixed linear models using preconditioned conjugate gradient iteration. J. Dairy Sci. 82, 2779–2787. doi:10.3168/jds.2002-0302(99)57535-9

Strandén, I., and Mäntysaari, E. A. (2020). Bpog: An efficient program for estimating base population allele frequencies in single and multiple group structured populations. AFSci. 29 (3), 166–176. doi:10.23986/afsci.90955

Strandén, I., and Vuurik, K. (2006). “RelaX2: Pedigree analysis program,” in Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, 13–18 August 2006 (Belo Horizonte, MG, Brazil: Instituto Prociência), 27–30.

Taskinen, M., Mäntysaari, E. A., Aamand, G. P., and Strandén, I. (2014). “Comparison of breeding values from single-step and bivariate blending methods,” in Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, August 2014 (Vancouver, BC, Canada: WCGALP), 17–22.

Tajani, A., Wiggans, G. R., Van Tassell, C. P., Philpot, J. C., and Gengler, N. (1999). Use of (co) variance functions to describe (co)variances for test day yield. J. Dairy Sci. 82 (1), 22610–22614. doi:10.3168/jds.2007-0980

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423. doi:10.3168/jds.2007-09980

Vitezica, Z. G., Aguilar, I., Mårtal, I., and Legarra, A. (2011). Bias in genomic predictions for populations under selection. Genet. Res. 93, 357–366. doi:10.1017/S001667231100022X

Wiggens, G. R., Cole, J. B., Hubbard, S. M., and Sonstegard, T. S. (2017). Genomic selection in dairy cattle: The USDA experience. Annu. Rev. Anim. Biosci. 5 (1), 309–327. doi:10.1146/annurev-animal-021815-111442

Wiggens, G. R., VanRaden, P. M., and Cooper, T. A. (2011). The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94 (6), 3202–3211. doi:10.3168/jds.2010-3866

Xiang, T., Christensen, O. F., and Legarra, A. (2017). Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. J. Anim. Sci. 95, 1472–1480. doi:10.2527/jas.2016.1155
Appendix 1: Representation of the $\Phi_{\text{pre}}$ matrix used in formula 

$$(\Phi_{\text{pre}} \Phi_{\text{pre}})^{-1} \Phi_{\text{pre}} \Gamma_{\text{pre}} \Phi_{\text{pre}} (\Phi_{\text{pre}} \Phi_{\text{pre}})^{-1}$$

to compute the $K$ matrix.

| BirthYear | STY | FIN | SWE | DNK | NOR | RDC | FIC | OTHER | HOL |
|-----------|-----|-----|-----|-----|-----|-----|-----|-------|-----|
| 1970      | -0.437 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1980      | -0.155 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1990      | 0.127 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2000      | 0.489 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2010      | 0.690 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2021      | 0.972 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1970      | -0.437 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1980      | -0.155 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1990      | 0.127 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2000      | 0.489 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2010      | 0.690 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2021      | 0.972 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1980      | -0.155 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1990      | 0.127 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2000      | 0.489 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2010      | 0.690 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2021      | 0.972 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1980      | -0.155 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1990      | 0.127 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2000      | 0.489 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2010      | 0.690 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2021      | 0.972 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1980      | -0.155 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1990      | 0.127 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2000      | 0.489 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2010      | 0.690 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2021      | 0.972 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

$^1$ STY - birth year of the group standardized as 

$$2 \left(\frac{\text{year}_{\text{max}} - \text{year}_{\text{min}}}{\text{year}_{\text{max}} - \text{year}_{\text{min}}} - 1\right)$$

where $\text{year}_{\text{MF}}$ is the year of the MF, 

$\text{year}_{\text{min}} = 1950$ and $\text{year}_{\text{max}} = 2021$. 

By Kudinov et al. 10.3389/fgene.2022.1012205