Journal of Animal Science, 2021, Vol. 99, No. 2, 1–10

doi:10.1093/jas/skab019
Advance Access publication January 25, 2021
Received: 20 October 2020 and **Accepted:** 20 January 2021
Animal Genetics and Genomics

ANIMAL GENETICS AND GENOMICS

Comparison of models for missing pedigree in single-step genomic prediction

Yutaka Masuda,[†,1] Shogo Tsuruta,[†] Matias Bermann,[†] Heather L. Bradford,[‡] and Ignacy Misztal[†]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA, [‡]Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

[1]Corresponding author: yutaka@uga.edu

ORCiD numbers: 0000-0002-3428-6284 (Y. Masuda); 0000-0002-6897-6363 (S. Tsuruta); 0000-0002-5374-0710 (M. Bermann); 0000-0001-5404-3872 (H. L. Bradford); 0000-0002-0382-1897 (I. Misztal).

Abstract

Pedigree information is often missing for some animals in a breeding program. Unknown-parent groups (UPGs) are assigned to the missing parents to avoid biased genetic evaluations. Although the use of UPGs is well established for the pedigree model, it is unclear how UPGs are integrated into the inverse of the unified relationship matrix (H-inverse) required for single-step genomic best linear unbiased prediction. A generalization of the UPG model is the metafounder (MF) model. The objectives of this study were to derive 3 H-inverses and to compare genetic trends among models with UPG and MF H-inverses using a simulated purebred population. All inverses were derived using the joint density function of the random breeding values and genetic groups. The breeding values of genotyped animals ($\mathbf{u}_2$) were assumed to be adjusted for UPG effects ($\mathbf{g}$) using matrix $\mathbf{Q}_2$ as $\mathbf{u}_2^* = \mathbf{u}_2 + \mathbf{Q}_2\mathbf{g}$ before incorporating genomic information. The Quaas–Pollak-transformed (QP) H-inverse was derived using a joint density function of $\mathbf{u}_2^*$ and $\mathbf{g}$ updated with genomic information and assuming nonzero $\text{cov}(\mathbf{u}_2^*, \mathbf{g}')$. The modified QP (altered) H-inverse also assumes that the genomic information updates $\mathbf{u}_2^*$ and $\mathbf{g}$, but $\text{cov}(\mathbf{u}_2^*, \mathbf{g}') = \mathbf{0}$. The UPG-encapsulated (EUPG) H-inverse assumed genomic information updates the distribution of $\mathbf{u}_2^*$. The EUPG H-inverse had the same structure as the MF H-inverse. Fifty percent of the genotyped females in the simulation had a missing dam, and missing parents were replaced with UPGs by generation. The simulation study indicated that $\mathbf{u}_2^*$ and $\mathbf{g}$ in models using the QP and altered H-inverses may be inseparable leading to potential biases in genetic trends. Models using the EUPG and MF H-inverses showed no genetic trend biases. These 2 H-inverses yielded the same genomic EBV (GEBV). The predictive ability and inflation of GEBVs from young genotyped animals were nearly identical among models using the QP, altered, EUPG, and MF H-inverses. Although the choice of H-inverse in real applications with enough data may not result in biased genetic trends, the EUPG and MF H-inverses are to be preferred because of theoretical justification and possibility to reduce biases.

Key words: genomic selection, metafounder, relationship matrix, simulation, unknown-parent group

Abbreviations:

EBV	estimated breeding value
EUPG H-inverse	unknown-parent-groups-encapsulated H-inverse
GEBV	genomic estimated breeding value
MF	metafounder
MME	mixed model equations
PCG	preconditioned conjugate gradient
QP H-inverse	Quaas–Pollak-transformed
ssGBLUP	single-step genomic best linear unbiased prediction
TBV	true breeding value
UPG	unknown-parent group

Introduction

The theory of single-step genomic best linear unbiased prediction (ssGBLUP) (Aguilar et al., 2010; Christensen and Lund, 2010) was developed assuming a complete pedigree. However, pedigree information is frequently missing for farm animals making the numerator relationship matrix (**A**) incomplete. Missing parents are assumed to be members of an unselected base-population, an incorrect assumption considering that most livestock populations are selected. Therefore, an incomplete relationship matrix fails to account for selection which may result in biased genetic trends (Kennedy et al., 1988). Unknown-parent groups (UPGs) can be used to model non-zero breeding values for missing parents (Graser et al., 1987; Westell et al., 1988). The breeding value of each animal is adjusted with genetic group effects weighted by the expected fraction of an animal's genes originating from each group (Quaas, 1988). Although the UPG model is well defined for $\mathbf{A}^{-1}$, it is unclear how UPGs can be incorporated into the inverse of the unified relationship matrix ($\mathbf{H}^{-1}$) in ssGBLUP. We will refer to such an inverse as "H-inverse."

Several H-inverses with UPGs are available (Tsuruta et al., 2011; Misztal et al., 2013; Masuda et al., 2019a). Some studies showed similar genetic trends among H-inverses (Matilainen et al., 2018; Masuda et al., 2019a), whereas others reported biased UPG solutions and genomic predictions (Bradford et al., 2019b; Tsuruta et al., 2019). No theoretical justification currently exists for using a particular H-inverse with UPGs.

A generalization of a UPG model resulted in the metafounder (MF) model (Christensen, 2012; Legarra et al., 2015). A limited number of reports on comparisons between UPGs and MFs (Bradford et al., 2019b; Kudinov et al., 2020; Macedo et al., 2020) are available. However, despite a comprehensive explanation of MFs by Legarra et al. (2015), it is unclear how the MF model differs from the UPG model. Thus, the objectives of this study were to present formal derivations of 3 H-inverses with UPGs and to compare genetic trends among models with UPG and MF H-inverses using a simulated purebred population.

Materials and Methods

Animal Care and Use Committee approval was not requested for this study because data were simulated in a computer.

Derivation of H-inverses

The H-inverses were derived using the joint density-function approach (Quaas, 1988; Aguilar et al., 2010). The density function $p(\mathbf{u}|\mathbf{A}, \sigma_u^2)$, where $\mathbf{u}$ is a vector of breeding values, $\mathbf{A}$ is the numerator relationship matrix, and σ_u^2 is the additive genetic variance, was modified by incorporating information from UPGs and genomic markers. The H-inverses were obtained using the joint density function of $\mathbf{u}$ and $\mathbf{g}$ after UPG and genomic information were incorporated into the density function. Figure 1 shows pathways for the derivation each H-inverse. A brief description of the derivation of each H-inverse is presented below, and additional details are given in Supplementary Appendixes A1, A2, and A3.

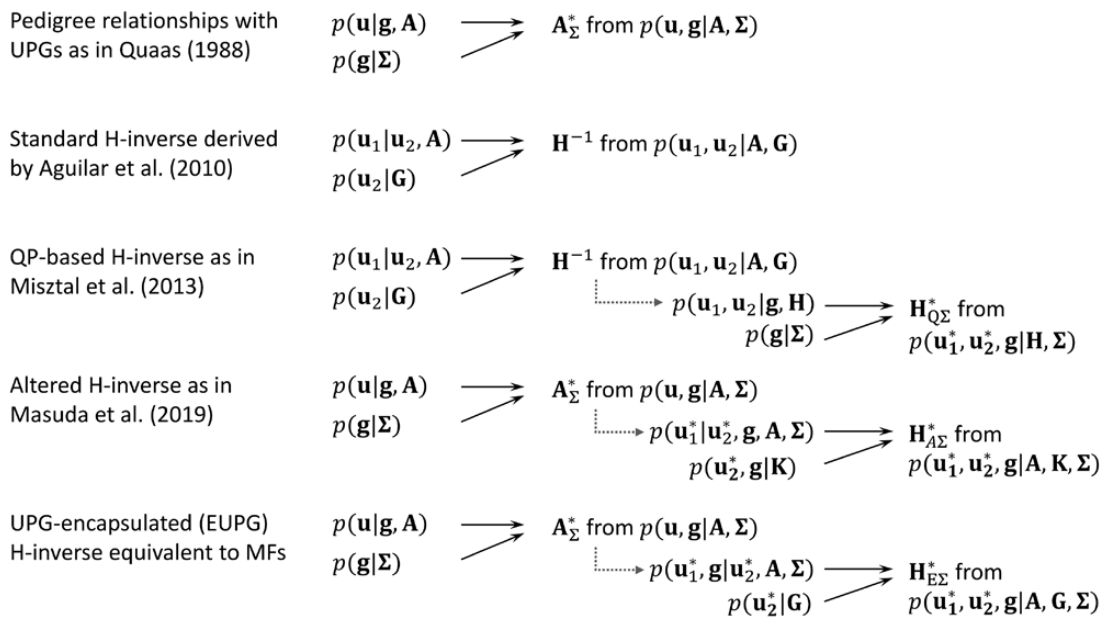

Figure 1. Pathways for the derivation of the inverse of a relationship matrix and H-inverses with UPGs based on joint-density functions. For example, $\mathbf{A}^*_{\Sigma}$ is derived from the joint density $p(\mathbf{u},\mathbf{g}|\mathbf{A},\boldsymbol{\Sigma}) = p(\mathbf{u}|\mathbf{g},\mathbf{A})\,p(\mathbf{g}|\boldsymbol{\Sigma})$. The genetic variance is omitted from the formulas. Symbols include $\mathbf{u}$ = a vector of breeding values, $\mathbf{g}$ = a vector of random UPGs, $\mathbf{u}^* = \mathbf{u} + \mathbf{Q}\mathbf{g}$, where $\mathbf{Q}$ is a matrix relating animals to UPGs, $\mathbf{A}$ = numerator relationship matrix, $\mathbf{G}$ = genomic relationship matrix, $\mathbf{H}$ = unified relationship matrix with pedigree and genomic information, $\boldsymbol{\Sigma} = \text{var}(\mathbf{g})$, $\mathbf{K} = \text{var}\left(\begin{bmatrix}\mathbf{u}_2' & \mathbf{g}'\end{bmatrix}\right)$ given genomic information, and subscripts 1 and 2 = nongenotyped and genotyped animals, respectively. Details are given in Supplementary Appendixes A1, A2, and A3.

Inverse of the numerator relationship matrix with random UPGs

Assume $\mathbf{u} \sim \text{MVN}(\mathbf{0}, \mathbf{A}\sigma_u^2)$. When some pedigree information is missing and random UPGs are considered, $\mathbf{u}$ is adjusted as $\mathbf{u}^* = \mathbf{u} + \mathbf{Qg}$, where $\mathbf{g}$ is a vector of random genetic group effects, and $\mathbf{Q}$ is a known matrix relating individuals to UPGs (Quaas, 1988). The conditional distribution of $\mathbf{u}^*$ is assumed to be $\mathbf{u}^*|\mathbf{g}, \mathbf{A}, \sigma_u^2 \sim \text{MVN}(\mathbf{Qg}, \mathbf{A}\sigma_u^2)$. Also, we assume $\mathbf{g} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}\sigma_u^2)$, where $\boldsymbol{\Sigma} = \{\Sigma_{jk}\}$ is the additive relationship between UPGs j and k. Matrix $\boldsymbol{\Sigma}$ might be computed using a recursive algorithm (VanRaden, 1992; Aguilar and Misztal, 2008) to estimate inbreeding coefficients for UPGs and additive relationships among UPGs. Another option is $\boldsymbol{\Sigma} = \mathbf{I}$ for non-inbred and unrelated UPGs.

Let $\mathbf{A}^*_\Sigma$ be the inverse of $\text{var}\begin{bmatrix}\mathbf{u}^* & \mathbf{g}'\end{bmatrix}$ computed using the rules of the inverse of a partitioned matrix (Searle, 1982):

$$\mathbf{A}^*_\Sigma = \left(\text{var}\begin{bmatrix}\mathbf{u}+\mathbf{Qg}\\ \mathbf{g}\end{bmatrix}\right)^{-1} = \begin{bmatrix}\mathbf{A}+\mathbf{Q}\boldsymbol{\Sigma}\,\mathbf{Q}' & \mathbf{Q}\boldsymbol{\Sigma}\\ \boldsymbol{\Sigma}\,\mathbf{Q}' & \boldsymbol{\Sigma}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{Q}\\ -\mathbf{Q}'\mathbf{A}^{-1} & \mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}+\boldsymbol{\Sigma}^{-1}\end{bmatrix} \quad (1)$$

Omitting σ_u^2. The matrix can also be derived from the posterior joint density $p(\mathbf{u}^*, \mathbf{g}|\mathbf{A}, \boldsymbol{\Sigma}, \sigma_u^2) = p(\mathbf{u}^*|\mathbf{g}, \mathbf{A}, \sigma_u^2)\, p(\mathbf{g}|\boldsymbol{\Sigma}, \sigma_u^2)$ as shown in Supplementary Appendix A1.

The formula (1) shows $\text{var}(\mathbf{u}^*) = \text{var}(\mathbf{u} + \mathbf{Qg}) = \mathbf{A} + \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}'$ for incomplete $\mathbf{A}$ augmented with $\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}'$, the additional relationship matrix due to UPG contributions. Quaas (1988) derived the same inverse using a transformation of the mixed model equations (MMEs) and a joint density function of the vector of phenotypes ($\mathbf{y}$), $\mathbf{u}$, and $\mathbf{g}$ with UPGs assumed to be fixed effects, equivalently, $\boldsymbol{\Sigma}^{-1} \to 0$ (Supplementary Appendix A1).

Define $\mathbf{u}' = [\mathbf{u}'_1 \;\; \mathbf{u}'_2]$, $\mathbf{A} = \begin{bmatrix}\mathbf{A}_{11}\mathbf{A}_{12}\\ \mathbf{A}_{21}\mathbf{A}_{22}\end{bmatrix}$ and $\mathbf{A}^{-1} = \begin{bmatrix}\mathbf{A}^{11}\mathbf{A}^{12}\\ \mathbf{A}^{21}\mathbf{A}^{22}\end{bmatrix}$, where subscript 1 is for nongenotyped and 2 for genotyped animals. Accordingly, $\mathbf{Q}' = [\mathbf{Q}'_1 \; \mathbf{Q}'_2]$ and

$$\mathbf{A}^*_\Sigma = \begin{bmatrix}\mathbf{A}^{11} & \mathbf{A}^{12} & -\begin{bmatrix}\mathbf{A}^{11} & \mathbf{A}^{12}\end{bmatrix}\mathbf{Q}\\ \mathbf{A}^{21} & \mathbf{A}^{22} & -\begin{bmatrix}\mathbf{A}^{21} & \mathbf{A}^{22}\end{bmatrix}\mathbf{Q}\\ -\mathbf{Q}'\begin{bmatrix}\mathbf{A}^{11}\\ \mathbf{A}^{21}\end{bmatrix} & -\mathbf{Q}'\begin{bmatrix}\mathbf{A}^{12}\\ \mathbf{A}^{22}\end{bmatrix} & \mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}+\boldsymbol{\Sigma}^{-1}\end{bmatrix} = \begin{bmatrix}\mathbf{A}^{11} & \mathbf{A}^{12} & \mathbf{A}^{13}\\ \mathbf{A}^{21} & \mathbf{A}^{22} & \mathbf{A}^{23}\\ \mathbf{A}^{31} & \mathbf{A}^{32} & \mathbf{A}^{33}_\Sigma\end{bmatrix},$$

where subscript 3 refers to UPGs.

Inverse of the unified relationship matrix

We assume the standard animal model with $\mathbf{u}|\mathbf{A}, \sigma_u^2 \sim \text{MVN}(\mathbf{0}, \mathbf{A}\sigma_u^2)$. After obtaining genotypes, the distribution of $\mathbf{u}_2$ is updated with $\mathbf{u}_2|\mathbf{G}, \sigma_u^2 \sim \text{MVN}(\mathbf{0}, \mathbf{G}\sigma_u^2)$, where $\mathbf{G}$ is the matrix of genomic relationships. Christensen et al. (2010) and Aguilar et al. (2010) showed that the inverse of the unified relationship matrix for genotyped and non-genotyped animals is

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix}0 & 0\\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\end{bmatrix}, \quad (2)$$

which is derived from the posterior density $p(\mathbf{u}_1, \mathbf{u}_2 \mid \mathbf{A}, \mathbf{G}, \sigma_u^2) = p(\mathbf{u}_1|\mathbf{u}_2, \mathbf{A}, \sigma_u^2)\, p(\mathbf{u}_2|\mathbf{G}, \sigma_u^2)$.

The Schur complement of the block $\mathbf{A}^{11}$ of the matrix $\mathbf{A}^{-1}$ is

$$\mathbf{A}_{22}^{-1} = \mathbf{A}^{22} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}. \quad (3)$$

This formula consists of a series of sparse-matrix operations, and it reduces the computing cost for $\mathbf{A}_{22}^{-1}$ times an arbitrary vector, say $\mathbf{A}^{-1}\mathbf{q}$, when solving MMEs by iterative algorithms (Masuda et al. 2017; Strandén et al. 2017).

Quaas–Pollak-transformed (QP) H-inverse

Misztal et al. (2013) suggested the QP-transformation of the MMEs (Quaas and Pollak, 1981) with the following H-inverse:

$$\mathbf{H}^*_{\mathbf{Q}\Sigma} = \mathbf{A}^*_\Sigma + \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2\\ \mathbf{0} & -\mathbf{Q}'_2(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & \mathbf{Q}'_2(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2\end{bmatrix}. \quad (4)$$

Misztal et al. (2013) assumed fixed UPG effects, and their H-inverse did not contain $\boldsymbol{\Sigma}^{-1}$. We will refer to this inverse as the "QP H-inverse" regardless of fixed or random UPGs.

There are 2 ways to derive the QP H-inverse (4) using the joint density function. The first approach assumes that the joint density $p(\mathbf{u}_1, \mathbf{u}_2|\mathbf{H}, \sigma_u^2)$ is known a priori, then incorporates $\mathbf{g} \sim \text{MVN}(0, \boldsymbol{\Sigma}\sigma_u^2)$ into the density. The posterior density is $p(\mathbf{u}_1^*, \mathbf{u}_2^*, \mathbf{g} \mid \mathbf{H}, \boldsymbol{\Sigma}, \sigma_u^2) = p(\mathbf{u}_1, \mathbf{u}_2|\mathbf{H}, \mathbf{g}, \sigma_u^2)\, p(\mathbf{g}|\boldsymbol{\Sigma}, \sigma_u^2)$. This means that genomic markers are obtained before UPGs are defined, or equivalently, that $\mathbf{H}$ is known a priori before defining $\mathbf{g}$ (Supplementary Appendix A1 and Figure 1). The second approach assumes that $p(\mathbf{u}^*, \mathbf{g}|\mathbf{A}, \boldsymbol{\Sigma}, \sigma_u^2)$ is known a priori, then incorporates $\mathbf{c}_2^*|\mathbf{K}, \sigma_u^2 \sim \text{MVN}(0, \mathbf{K}\sigma_u^2)$ into the density, where $\mathbf{c}_2^{*\prime} = \begin{bmatrix}\mathbf{u}_2^{*\prime} & \mathbf{g}'\end{bmatrix}$. The matrix $\mathbf{K}$ is the genomic relationship matrix for $\mathbf{u}_2^*$ and $\mathbf{g}$ given the genomic information. The posterior density is $p(\mathbf{u}_1^*, \mathbf{u}_2^*, \mathbf{g} \mid \mathbf{A}, \mathbf{K}, \boldsymbol{\Sigma}, \sigma_u^2) = p(\mathbf{u}_1^*|\mathbf{u}_2^*, \mathbf{g}, \mathbf{A}, \boldsymbol{\Sigma}, \sigma_u^2)\, p(\mathbf{u}_2^*, \mathbf{g}|\mathbf{K}, \sigma_u^2)$. It means that UPGs are defined before genomic markers are obtained, or equivalently, that $\mathbf{A}^*_\Sigma$ is known a priori before obtaining $\mathbf{K}$. This approach leads to (4) assuming a covariance matrix between $\mathbf{u}_2^*$ and $\mathbf{g}$ (Supplementary Appendix A2), whereas it also results in a separate H-inverse, shown below as formula (5), assuming no covariance between $\mathbf{u}_2^*$ and $\mathbf{g}$.

Altered QP H-inverse.

Masuda et al. (2018a, 2019a) suggested the following alternative QP H-inverse by eliminating $\mathbf{G}^{-1}$ in the UPG terms (i.e., eliminating terms $\mathbf{G}^{-1}\mathbf{Q}_2$ and $\mathbf{Q}'_2\mathbf{G}^{-1}\mathbf{Q}_2$) from the QP H-inverse:

$$\mathbf{H}^*_{A\Sigma} = \mathbf{A}^*_\Sigma + \begin{bmatrix}0 & 0 & 0\\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(-\mathbf{A}_{22}^{-1})\mathbf{Q}_2\\ 0 & -\mathbf{Q}'_2(-\mathbf{A}_{22}^{-1}) & \mathbf{Q}'_2(-\mathbf{A}_{22}^{-1})\mathbf{Q}_2\end{bmatrix}. \quad (5)$$

We will refer to this inverse as the "altered H-inverse" regardless of fixed or random UPGs.

The altered H-inverse was derived using the joint-density function as described as the second approach for the QP H-inverse. The matrix $\mathbf{K}$ is assumed to be block-diagonal, i.e., $\text{var}(\mathbf{u}_2^*) = \mathbf{G}\sigma_u^2$ and $\text{var}(\mathbf{g}) = \boldsymbol{\Sigma}\sigma_u^2$ but $\text{cov}(\mathbf{u}_2^*, \mathbf{g}') = 0$ (Supplementary Appendix A2).

UPG-encapsulated (EUPG) H-inverse

A new H-inverse was derived when the joint density function of $\mathbf{u}^*$ and $\mathbf{g}$ was updated in the same way as the altered H-inverse, but the genomic information updated $\mathbf{u}_2^*$, i.e., $\mathbf{u}_2^*|\mathbf{G}, \sigma_u^2 \sim \text{MVN}(0, \mathbf{G}\sigma_u^2)$, and $\mathbf{g}$ is indirectly updated through $\mathbf{u}_2^*$, as shown in Supplementary Appendix A3:

$$\mathbf{H}^{*}_{\Sigma\Sigma} = \mathbf{A}^{*}_{\Sigma} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}^{*}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad (6)$$

where

$$\mathbf{A}^{*}_{22} = \mathbf{A}^{22} - \begin{bmatrix} \mathbf{A}^{21} & \mathbf{A}^{23} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{13} \\ \mathbf{A}^{31} & \mathbf{A}^{33}_{\Sigma} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}^{12} \\ \mathbf{A}^{32} \end{bmatrix}. \quad (7)$$

The H-inverse can be derived from the posterior density $p(\mathbf{u}^{*}_{1}, \mathbf{u}^{*}_{2}, \mathbf{g} \mid \mathbf{A}, \mathbf{G}, \boldsymbol{\Sigma}, \sigma^{2}_{u}) = p(\mathbf{u}^{*}_{1}, \mathbf{g} | \mathbf{u}^{*}_{2}, \mathbf{A}, \boldsymbol{\Sigma}, \sigma^{2}_{u}) p(\mathbf{u}^{*}_{2} | \mathbf{G}, \sigma^{2}_{u})$. We will refer to this H-inverse as the "EUPG H-inverse," defined only for random UPGs. Note that $\mathbf{A}^{*}_{22}$ still consists of sparse matrices that can take advantage of existing iterative solvers.

MF H-inverse

Let $\mathbf{g}_{m}$ be the random effect of MFs, and assume that $\mathbf{g}_{m} \sim \text{MVN}(0, \boldsymbol{\Gamma}\sigma^{2}_{u})$, where $\boldsymbol{\Gamma} = \{\gamma_{jk}\}$ is the covariance matrix among MFs. Let $\mathbf{u}_{\Gamma}' = [\mathbf{u}_{\Gamma 1}' \; \mathbf{u}_{\Gamma 2}' \; \mathbf{g}_{m}']$ be a vector of additive genetic effects. The numerator relationship matrix with MFs ($\mathbf{A}_{\Gamma}$) and its inverse ($\mathbf{A}^{-1}_{\Gamma}$) are as follows:

$$\mathbf{A}_{\Gamma} = \begin{bmatrix} \mathbf{A}_{\Gamma 11} & \mathbf{A}_{\Gamma 12} & \mathbf{A}_{\Gamma 1m} \\ \mathbf{A}_{\Gamma 21} & \mathbf{A}_{\Gamma 22} & \mathbf{A}_{\Gamma 2m} \\ \mathbf{A}_{\Gamma m1} & \mathbf{A}_{\Gamma m2} & \boldsymbol{\Gamma} \end{bmatrix} \text{ and } \mathbf{A}^{-1}_{\Gamma} = \begin{bmatrix} \mathbf{A}^{11}_{\Gamma} & \mathbf{A}^{12}_{\Gamma} & \mathbf{A}^{1m}_{\Gamma} \\ \mathbf{A}^{21}_{\Gamma} & \mathbf{A}^{22}_{\Gamma} & \mathbf{A}^{2m}_{\Gamma} \\ \mathbf{A}^{m1}_{\Gamma} & \mathbf{A}^{m2}_{\Gamma} & \mathbf{A}^{mm}_{\Gamma} + \boldsymbol{\Gamma}^{-1} \end{bmatrix}$$

with subscript m for MFs. Legarra et al. (2015) showed that the H-inverse with MFs is:

$$\mathbf{H}^{-1}_{\Gamma} = \mathbf{A}^{-1}_{\Gamma} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}_{05} - \mathbf{A}^{-1}_{\Gamma 22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where G_{05} is the genomic relationship matrix with allele frequencies equal to 0.5 for all markers. We will refer to this H-inverse as the "MF H-inverse." As shown in Supplementary Appendix A3, the MF H-inverse can be derived using the same approach as the EUPG H-inverse.

Applying the rules for the computation of the inverse of partitioned matrixes to A_{Γ} and A^{-1}_{Γ} (Searle, 1982), we obtain:

$$\mathbf{A}^{-1}_{\Gamma 22} = \mathbf{A}^{22}_{\Gamma} - \begin{bmatrix} \mathbf{A}^{21}_{\Gamma} & \mathbf{A}^{2m}_{\Gamma} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{11}_{\Gamma} & \mathbf{A}^{1m}_{\Gamma} \\ \mathbf{A}^{m1}_{\Gamma} & \mathbf{A}^{mm}_{\Gamma} + \boldsymbol{\Gamma}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}^{12}_{\Gamma} \\ \mathbf{A}^{m2}_{\Gamma} \end{bmatrix}. \quad (8)$$

This formula is identical to $\mathbf{A}^{*}_{22}$ (7) except for the replacement of $\mathbf{A}^{*}_{\Sigma}$ with $\mathbf{A}^{-1}_{\Gamma}$. The above formula permits the application of MFs to a large genotyped population. Masuda et al. (2019b) applied MFs to 2.3 million genotypes using formula (8).

Simulation study

Simulated population

The simulation study aimed to demonstrate whether each H-inverse is robust in an extreme case where the pedigree has many missing parents in a purebred population. The population was simulated with software QMSim (Sargolzaei and Schenkel, 2009). The genetic architecture and the historical population were the same as in the study by Tsuruta et al. (2019). The heritability was assumed to be 0.5, and only females had phenotypes. The founder population (generation 1) consisted of 50 males and 10,000 females sampled from the historical population. Each founder female had a phenotype. Parents of founders were assumed to be unknown.

In each subsequent generation, the following steps were performed to produce the progeny. Five thousand males and 5,000 females were generated by random mating of selected animals in the previous generation. A phenotype was assigned to all the females born in the current generation. Estimated breeding values (EBVs) were calculated by a single-trait, pedigree-based animal model with the true heritability (the general mean as a fixed effect and the additive genetic effect as a random effect) using the data available at the time. The genomic information was not used for selection, and the EBVs of young males were parental average. The top 45 youngest males on the EBVs were selected as new sires, which replaced the bottom 45 sires, whereas the top 50% females on the EBVs replaced the bottom 50% of the current dams. This strategy kept 50 selected bulls and 10,000 selected females as parents to produce the next generation. The simulation was terminated after the selection in generation 13.

As a result, the selected males had 100 daughters on average in each generation. We will refer to the males as proven bulls. Litter size was fixed to 1, and some selected females did not have progeny. The females in the last (13th) generation did not have phenotypes. Genomic information was collected from all animals born in generation 8 or later. The nongenotyped bulls without progeny (generations 2 through 7) were excluded from the data. The data included 65,000 phenotypes, 100,320 pedigree animals, and 60,000 genotyped animals with 58,000 SNP markers. All genotyped females (25,000) were phenotyped.

After generating the dataset, dam identification was randomly removed from the pedigree for 25% of females in generations 2 to 7, and 50% of females in generations 8 to 13. The unknown parents were replaced with UPGs. A total of 7 UPGs were defined for unknown parents of animals. The 1st was for generations 1 and 2, the 2nd for generations 3 and 4, the 3rd for generations 5 and 6, the 4th for generations 7 and 8, the 5th for generations 9 and 10, the 6th for generations 11 and 12, and the 7th for generation 13.

Five replicates were obtained with the above configuration. Because all replicates resulted in the same conclusion, we will only show the results from the first replicate.

Basic model comparison

Single-step GBLUP models with UPG (QP, altered, and EUPG) and MF H-inverses were used to predict genomic EBVs (GEBVs) and genetic group effects. We also tested a model where UPGs were applied only to $\mathbf{A}^{-1}$ in $\mathbf{H}^{-1}$, called "Omega" H-inverse after an omega constant (ω) used in previous studies (Tsuruta et al., 2011; Misztal et al., 2013). We used $\mathbf{I}$ as $\boldsymbol{\Sigma}$ for the UPG models. We estimated $\boldsymbol{\Gamma}$ by maximum likelihood (Garcia-Baccino et al., 2017). The calculated matrix is

$$\boldsymbol{\Gamma} = \begin{bmatrix} 0.034 & 0.027 & 0.032 & 0.034 & 0.036 & 0.037 & 0.038 \\ & 0.067 & 0.039 & 0.046 & 0.046 & 0.045 & 0.045 \\ & & 0.067 & 0.067 & 0.069 & 0.070 & 0.070 \\ & & & 0.096 & 0.101 & 0.104 & 0.104 \\ & & & & 0.125 & 0.136 & 0.139 \\ & \text{Sym.} & & & & 0.168 & 0.177 \\ & & & & & & 0.199 \end{bmatrix}.$$

The pedigree best linear unbiased prediction (BLUP) model was also tested. This model included the overall mean as a fixed effect, and animal breeding value, genetic group, and residual as random effects. Variance components used in the simulation were also used for prediction.

The genomic relationship matrix was calculated with the VanRaden's first method (VanRaden, 2008). The raw genomic relationships were blended with the identity matrix as $0.99\mathbf{G} + 0.01\mathbf{I}$. Allele frequencies were calculated from marker genotypes, thus, $\mathbf{G}$ was aligned to $\mathbf{A}_{22}$ (Chen et al., 2011; Gao et al., 2012): $\beta\mathbf{G} + \alpha\mathbf{1}\mathbf{1}'$, where α and β were calculated using the following equations:

$$\begin{aligned} \beta\,\overline{\text{diag}\mathbf{G}} + \alpha &= \overline{\text{diag}\mathbf{A}_{22}} \\ \beta\,\overline{\text{of fdiag}\mathbf{G}} + \alpha &= \overline{\text{of fdiag}\mathbf{A}_{22}} \end{aligned}$$

where $\overline{\text{diag}\mathbf{X}}$ and $\overline{\text{of fdiag}\mathbf{X}}$ are the averages of diagonals and off-diagonals of matrix $\mathbf{X}$, respectively. This alignment was expected to remove the bias of GEBV for genotyped animals in a selected population (Vitezica et al., 2011). The matrix $\mathbf{A}_{22}$ was calculated using the indirect method (Colleau, 2002) with inbreeding coefficients calculated using the method by Meuwissen and Luo (1992). We refer to such inbreeding coefficients as Standard Inbreeding.

We defined the true breeding value (TBV) as the simulated breeding value based on causal genotypes, and GEBVs as the prediction of $\mathbf{u}^*$. Predicted genetic group effects and GEBVs were adjusted by the average GEBV of phenotyped animals in generations 1 and 2. Then, models were compared in terms of predictions of genetic group effects, genetic trends, and predictive abilities of young males. Predictions of genetic group effects were compared with averages of TBVs of parents replaced with UPGs in the pedigree file. Trends of GEBV were compared with TBV genetic trends. Predictive abilities were defined as correlations between GEBVs and TBVs for males in generation 12 and 13. We also computed linear regressions of TBVs on GEBVs for the same males, and utilized the slope coefficient (b_1) as an inflation indicator.

Additional model comparisons

We examined additional factors that were expected to affect genetic trends and the inflation of GEBVs. Firstly, we estimated inbreeding coefficients and additive relationships between UPGs to calculate $\boldsymbol{\Sigma}$ (Aguilar and Misztal, 2008), and used these values in $\mathbf{A}^*_\Sigma$ as suggested by VanRaden (1992). We refer to this inverse as $\tilde{\mathbf{A}}^{-1}_\Sigma$. The $\tilde{\mathbf{A}}^{-1}_{22}$ matrix was formed indirectly during the construction of matrix $\tilde{\mathbf{A}}^*_\Sigma$ using formula (3). Matrix $\mathbf{G}$ was aligned to $\mathbf{A}_{22}$ assuming Standard Inbreeding based on pedigree with missing parents. Figure 2 shows inbreeding trends across generations using inbreeding coefficients from the complete pedigree that is the original pedigree before removing the dam identifications (Complete Inbreeding), inbreeding coefficients from the pedigree with missing parents (Standard Inbreeding), inbreeding coefficients from the pedigree with UPGs estimated using VanRaden's method (Estimated Inbreeding), and inbreeding coefficients from the pedigree with MFs (Metafounders).

Secondly, we aligned $\mathbf{G}$ to $\tilde{\mathbf{A}}_{22}$, which is equal to $(\tilde{\mathbf{A}}^{-1}_{22})^{-1}$, say $\mathbf{M}$. Because the explicit calculation of $\mathbf{M}$ was expensive, we approximated the sum of all elements and the trace of $\mathbf{M}$ required to compute α and β. The sum of all elements is $s = \mathbf{1}'\mathbf{M}\mathbf{1}$, equivalent to a combination of two formulas: $\mathbf{y} = \mathbf{M}\mathbf{1}$ and $s = \mathbf{1}'\mathbf{y}$. Vector $\mathbf{y}$ is a solution of $\mathbf{M}^{-1}\mathbf{y} = \mathbf{1}$, i.e., $\tilde{\mathbf{A}}^{-1}_{22}\mathbf{y} = \mathbf{1}$, which can be easily obtained using the preconditioned conjugate gradient method (PCG). The trace of $\mathbf{M}$ was approximated with a Monte-Carlo approach (Bai et al., 1996), a variant of PCG based on the symmetric Lanczos algorithm.

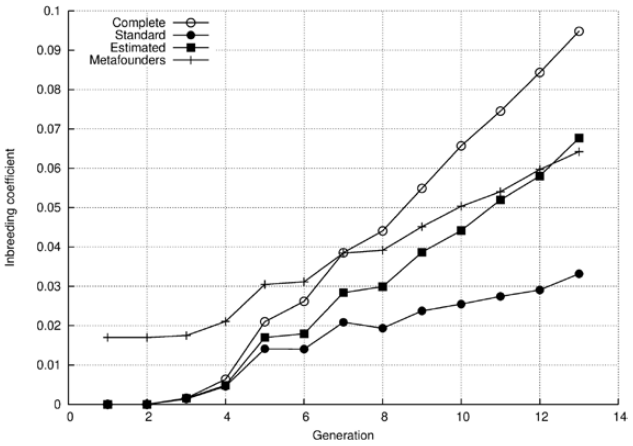

Figure 2. Inbreeding trends across generations. Inbreeding trends were calculated using inbreeding coefficients from the complete pedigree before removing parent identifications (Complete), inbreeding coefficients from the pedigree with missing parents (Standard), inbreeding coefficients from the pedigree with UPGs estimated with VanRaden's method (Estimated), and inbreeding coefficients from the pedigree with metafounders (Metafounders).

Computations

The preGSf90 program (Misztal et al., 2018; Lourenco et al., 2020) was used to calculate $\mathbf{G}$. Custom programs were developed by Fortran and Julia for other computations. We used PCG to solve MMEs. The convergence criteria was $||\mathbf{C}\mathbf{x} - \mathbf{b}||^2/||\mathbf{b}||^2 < 10^{-16}$, where $\mathbf{C}$ is the left-hand-side matrix of the MMEs, $\mathbf{b}$ is the right-hand-side vector of the MMEs, $\mathbf{x}$ is the solution vector, and $||\cdot||$ is the Euclidean norm of a vector.

Results and Discussion

Theoretical development

All the H-inverses were derived by first incorporating UPGs into the pedigree relationship matrix ($\mathbf{A}^{-1}$) and then integrating genomic information into the unified relationship matrix ($\mathbf{H}^{-1}$). The QP H-inverse can also be obtained in the opposite way by firstly integrating genomic information into $\mathbf{A}^{-1}$, and secondly incorporating UPG information into $\mathbf{H}^{-1}$.

Differences among the H-inverses depend on how genomic information contributes to the density function. Undoubtedly, genomic information updates the distribution of $\mathbf{u}^*_2$. However, there is a question about whether genomic information directly updates the distribution of $\mathbf{g}$ or not. The genomic information may add new knowledge that can change the distribution of $\mathbf{g}$, and if this is true, we should simultaneously update the distributions of $\mathbf{g}$ and $\mathbf{u}^*_2$. In such case, we should consider $\text{var}(\mathbf{c}^*_2) = \mathbf{K}$, given genomic information. Under this condition, if $\text{cov}(\mathbf{u}^*_2, \mathbf{g}') \neq 0$ we obtain the QP H-inverse, otherwise we get the altered H-inverse. When the genomic information is assumed to update the distribution of $\mathbf{u}^*_2$, meaning that $\mathbf{g}$ is updated through $\mathbf{u}^*_2$, the EUPG H-inverse is derived.

There is a clear separation in assumptions between the QP and the other H-inverses. The QP H-inverse assumes that $\mathbf{G}$ does not contain the variation due to UPG contributions, i.e., $\mathbf{Q}_2\boldsymbol{\Sigma}\,\mathbf{Q}_2'$ (Supplementary Appendix A2). Thus, $\boldsymbol{K}$ should be supplied to account for UPG variation already included in $\mathbf{A}^*_\Sigma$. The other H-inverses assume that $\mathbf{G}$ already contains missing relationships, and that $\mathbf{A}^*_\Sigma$ also accounts for them, and therefore, a function of $\mathbf{A}^{-1}_{22}$ should eliminate overlapping information. Usually, $\mathbf{G}$ is expected to describe relationships among animals

regardless of pedigree completeness (Christensen, 2012; Misztal et al., 2013; Legarra et al., 2015). Hence, the assumption of the QP H-inverse is inappropriate. The altered and the EUPG H-inverses share nearly the same assumptions; the only difference is whether the genomic information directly updates var(**g**) (altered H-inverse) or not (EUPG H-inverse).

When all animals in the pedigree are genotyped (i.e., $\mathbf{A} = \mathbf{A}_{22}$), only the genomic information remains in the H-inverse. For example, without UPGs, $\mathbf{A}_{22}^{-1}$ completely offsets $\mathbf{A}^{-1}$, and $\mathbf{H}^{-1}$ reduces to $\mathbf{G}^{-1}$, which contains the genomic information added to the joint-density function. With UPGs and $\mathbf{A} = \mathbf{A}_{22}$, the H-inverse reduces to $\mathbf{K}^{-1}$ in the QP and the altered H-inverses and to $\mathbf{G}^{-1}$ in the EUPG H-inverse. For the QP H-inverse, $\mathbf{K}$ is inappropriate as discussed above. For the altered H-inverse, $\mathbf{G}^{-1}$ contributes to prediction but $\boldsymbol{\Sigma}^{-1}$ does not, suggesting that $\boldsymbol{\Sigma}$ is redundant to be added to $\mathbf{K}$. The EUPG H-inverse is reasonable because the prediction is based on $\mathbf{G}^{-1}$ that is the complete information of genomic markers.

In this study, we did not consider a residual polygenic effect, which accounts for the incomplete linkage disequilibrium between the genetic markers and QTL or genes (Liu et al., 2016). When the residual polygenic effect is consolidated into $\mathbf{u}$ as a separate effect, the genomic relationship matrix is blended with the pedigree relationships, $\mathbf{G}_w = (1-w)\,\mathbf{G} + w\mathbf{A}_{22}$ with a weight w (Liu et al., 2016). In the UPG model, applying $\mathbf{G}_w^{-1}$ to each of H-inverses, a portion of pedigree relationships could be merged into pre-existing $\mathbf{A}_{22}^{-1}$ in the H-inverse. Future research should cover further development on this topic.

A possible question is what is the unified relationship matrix (H-matrix) for $\begin{bmatrix}\mathbf{u}^{*'} & \mathbf{g}'\end{bmatrix}$, i.e., the inverse of an H-inverse. The H-matrix is explicitly available for the QP H-inverse based on (1) or (17) in Supplementary Appendix A2, for the altered H-inverse based on (17), for the EUPG and the MF H-inverses based on (24) in Supplementary Appendix A3, and for the Omega H-inverse as shown by Martini et al. (2018). A diagonal element of the H-matrix should be required to compute the prediction-error variance of individual GEBV. There is an efficient method to compute the diagonals of $\mathbf{H}$ (Legarra et al., 2020), and this method would need to be extended to each H-matrix.

The EUPG and the MF H-inverses have the same structure, and the difference is in the modeling of the group effects. In the UPG model, a group represents a base population, and the group effect is the expectation of breeding values in the base population (Quaas, 1988). One can interpret a group as a virtual animal randomly sampled from the corresponding base population. Matrix $\boldsymbol{\Sigma}$ is the additive relationship matrix among groups. An MF is a proxy of animals in a base population so that the inbreeding coefficient of the MF is –1 for non-inbred base animals (Legarra et al., 2015). Matrix $\boldsymbol{\Gamma}$ describes the genomic relationships among groups. An MF is treated as a real parent in the pedigree relationships, whereas a UPG is considered a missing parent.

Computations

All the PCG iterations converged. The number of iterations was 273 for the Omega, 571 for the QP, 385 for the altered, 390 for the EUPG, and 357 for the MF H-inverses when the starting values were set to zero.

When (8) is used with Metafounder H-inverse, $\mathbf{A}_{\Gamma}^{11}$ can only have ancestors of genotyped animals, as in $\mathbf{A}_{22}^{-1}$ (3). The same principle can be applied to the EUPG H-inverse using $\boldsymbol{\Sigma} = \mathbf{I}$ with Standard Inbreeding or an arbitrary $\boldsymbol{\Sigma}$ with Estimated Inbreeding. The sparse matrices in (7) should be constructed by the rapid algorithm (Quaas 1988) using the subset pedigree with genotyped animals and their ancestors with UPGs. With Estimated Inbreeding, the formula (7) results in the inverse of a subset of $\tilde{\mathbf{A}}_{22}$ for genotyped animals.

Further, when (7) is used with the EUPG H-inverse, a bottleneck occurs during the computation of the Cholesky decomposition of the 2 × 2 block matrix containing $\mathbf{A}^{11}$, which is performed once prior to solving for the MMEs (Masuda et al., 2017; Strandén et al., 2017). When $\mathbf{A}^{11}$ is large (e.g., 5 million non-genotyped animals), the Cholesky decomposition can still be executed because the matrix is extremely sparse; however, the operation becomes very slow. Two solutions to this problem could be to utilize only ancestors of genotyped animals to approximate $\mathbf{A}^{11}$, or to use MFs. Finding a practical solution to a large $\mathbf{A}^{11}$ in (7) is left for future research.

Simulation study

Inbreeding coefficients

Many animals had a missing dam in the simulated pedigree. Standard Inbreeding was underestimated, and the discrepancy with Complete Inbreeding became greater in later generations as expected. Estimated Inbreeding recovered the lost inbreeding, and the inbreeding trend was between the Complete Inbreeding and Standard Inbreeding trends, as Lutaaya et al. (1999) showed. In this method, an animal's inbreeding is replaced with an estimated value even when one parent is missing, and the estimates will propagate from generation to generation rapidly.

With MFs, the animals in the earliest generations had non-zero inbreeding, but compared with Standard Inbreeding, the rate of inbreeding was similar up to generation 7 and slightly greater in generation 8 or later. One possible explanation for this phenomenon is that the inbreeding rate could reflect the increase of γ_{jk} by generation. Assigning MFs to missing parents, an animal's inbreeding coefficient is half of the additive relationship between the parents, say $a_{sd}/2$, regardless that one parent is an MF or a real animal (Legarra et al., 2015). Let us take a simple case where an animal has a known sire; the dam is MF j; the dam of the sire is MF j-1 (or j if the two dams are in the same group). The inbreeding coefficient of this animal is approximated as $\gamma_{j-1,j}/2$ (and $\gamma_{jj}/2$), and the inbreeding rate is roughly determined by a change of $a_{sd}/2$ and $\gamma_{j-1,j}/2$ (and $\gamma_{jj}/2$) from one generation to the next. Up to generation 7 associated with groups from 1 to 4 (25% of females had missing dam), based on the calculated $\boldsymbol{\Sigma}$, the change of $\gamma_{j-1,j}/2$ (and $\gamma_{jj}/2$) ranged between 0.000 and 0.017, which is close to Standard Inbreeding, and the change of $a_{sd}/2$ can be at the same level as Standard Inbreeding. After generation 7 for groups from 4 to 7 (50% of females had missing dam), the change of $\gamma_{j-1,j}/2$ (and $\gamma_{jj}/2$) was from 0.015 to 0.022, which might be slightly compensated by a lower change of $a_{sd}/2$. The resulting inbreeding-rate can be similar to but slightly greater than Standard Inbreeding.

Trends for GEBV and genetic group effects

Figure 3A shows GEBV trends for proven bulls. We show only the results in generation 6 and later because the divergence among models is more apparent than in previous generations. The models using the EUPG and MF H-inverses and the pedigree BLUP model showed genetic trends similar to the true genetic trend with slight upward and downward biases in generations 6 and 7. Models with the QP and altered H-inverses overestimated genetic trends up to generation 7 and underestimated genetic trends in generations 8 and 9. In generations 10 and 11, all models except for the model with the Omega H-inverse showed a similar trend.

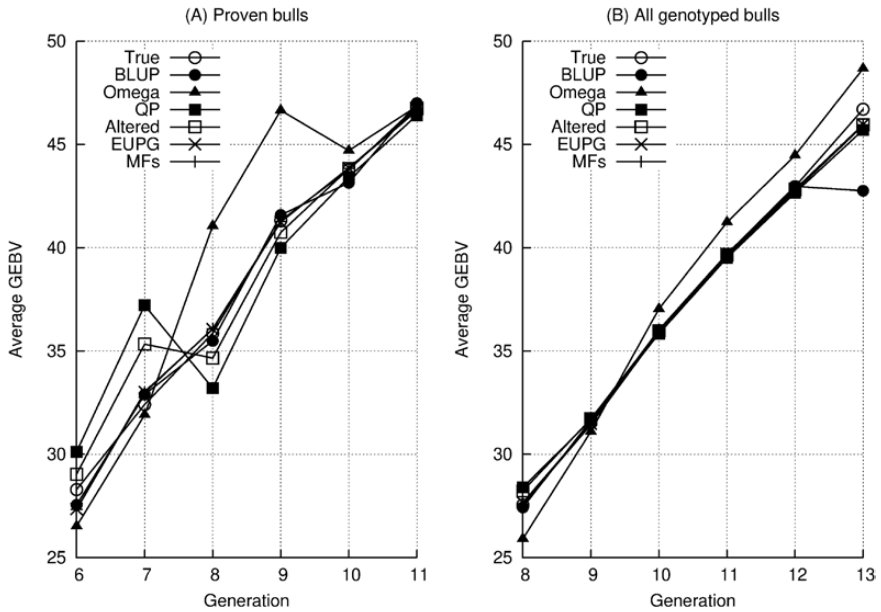

Figure 3. GEBV trends for proven bulls with phenotyped daughters (45 bulls in each generation; (A) and all genotyped bulls (5,000 bulls in each generation; (B) compared with the true genetic trends. Models include pedigree BLUP and single-step GBLUP models with Omega, QP, altered, EUPG, and MF H-inverses. The genetic trend was adjusted by the average prediction ($\mathbf{u}^*$) of phenotyped animals in generations 1 and 2 in each model.

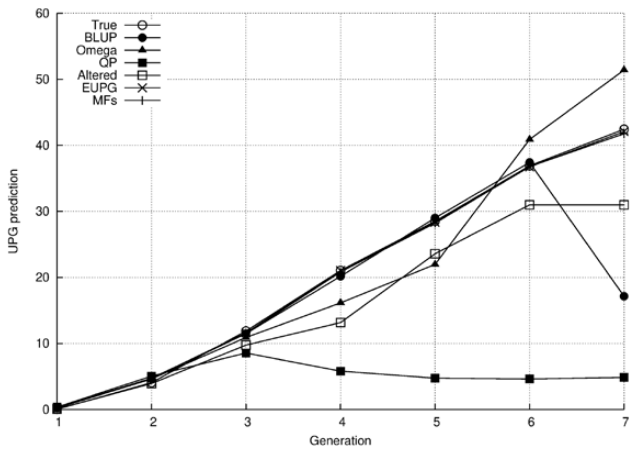

Figure 4. Trends for predicted genetic group effects and true genetic group values (True). Models include pedigree BLUP and single-step GBLUP models with Omega, QP, altered, EUPG, and MF H-inverses. The genetic trend was adjusted by the average prediction ($\mathbf{u}^*$) of phenotyped animals in generations 1 and 2 in each model.

Figure 3B shows GEBV trends for all genotyped bulls. Models with the QP and altered H-inverses overestimated genetic trends in generation 8. All models except for the model with the Omega H-inverse showed identical genetic trends up to generation 12. In generation 13, the models with the QP, altered, EUPG, and MF H-inverses slightly underestimated genetic trends relative to the true genetic trend, and the pedigree BLUP model yielded the lowest GEBVs.

Genetic trends for phenotyped animals were almost identical and close to the true genetic trend for all models, except for the model with the Omega H-inverse (results not shown).

Trends for predicted genetic group and true genetic group values are shown in Figure 4. Only the models using the EUPG and MF H-inverses gave genetic group predictions close to the true genetic group value. The pedigree BLUP model successfully predicted the group effects up to the 6th UPG but failed in the 7th UPG (g_7) because g_7 had no information for prediction, i.e., no phenotypes of progeny and relatives. In fact, the raw solution of g_7 was 0. Note that, in Figure 4, the predicted value appears to be greater than 0 because of the adjustment by the average prediction ($\mathbf{u}^*$) of phenotyped animals in generations 1 and 2. The model with the altered H-inverse showed a positive but underestimated genetic trend, whereas the model with the QP H-inverse produced a flat trend.

Results suggested that the models with the QP and altered H-inverses may prevent $\mathbf{g}$ from separating from $\mathbf{u}$. This may occur because the QP and altered H-inverses were formulated based on the assumption that genomic information directly contributed to both $\mathbf{u}_2^*$ and $\mathbf{g}$. Matrix $\mathbf{G}^{-1}$ in the QP H-inverse creates a direct link between $\mathbf{u}_2^*$ and $\mathbf{g}$, most likely inseparable. Similarly, matrix $\mathbf{A}_{22}^{-1}$ in the altered H-inverse associates $\mathbf{u}_2^*$ and $\mathbf{g}$. Further, genetic group effects are integrated into GEBVs in the QP and altered H-inverse models. Although the model with the altered H-inverse produced a GEBV trend with less bias than the model with the QP H-inverse, the bias still

remained. However, the genetic trend for $\mathbf{u}_2^*$ with the altered H-inverse should be reasonable because **G** accounts for selection (Figure 3).

With the QP and altered H-inverses, a problem arose when a nongenotyped sire had genotyped daughters with missing dams in generation 7, and many animals were genotyped in generation 8. In generation 7, daughters had GEBV (u_o), but their sire predictions were approximately the traditional EBV. When daughters had UPGs as dams and UPG effects were equal to g_d, daughter contributions to their sires (daughter deviations) are a function of $u_o - 0.5g_d$ (VanRaden and Wiggans, 1991). If UPG effects are underestimated, daughter deviations are biased upward, and therefore the EBV of nongenotyped sires will be overestimated. In generation 8, all animals were genotyped; the downward bias compensated for the overestimated trend in the previous generations. The bias continued for the next few generations because the sires of proven bulls had underestimated GEBV. Eventually, the genetic trend settled down to a certain level when many animals were genotyped.

The GEBVs from models using the EUPG and the MF H-inverses had smaller biases than those from models with the other H-inverses because $\mathbf{u}_2^*$ and **g** were related only through the traditional $\mathbf{A}_\Sigma^{-1}$. In addition, the UPG equations were absorbed into $\mathbf{A}_{22}^*$ and $\mathbf{A}^{-1}{}_{\Gamma\,22}$, and these matrices contain the same information as $\mathbf{A}_\Sigma^*$. Notably, the GEBVs were identical for these 2 models, implying that the models are closely related in a purebred population. This suggests that the stability of the genetic trends from the model with a MF H-inverse originated from the integration of UPG contributions into $\mathbf{A}_{\Gamma 22}^{-1}$ when all animals are related in the pedigree. When **Γ** is not stably calculated (Calus and Vandenplas, 2019; Kudinov et al., 2020), the model with the EUPG H-inverse can be a substitute for the model with the MF H-inverse.

Tsuruta et al. (2019) applied the ssGBLUP models with the QP and altered H-inverses to simulated data and observed both overestimation and underestimation of genetic trends for proven bulls. However, their GEBV biases were much lower than the ones observed in this study, and they did not encounter severe biases for UPG predictions. Their results suggested that confounding may not always occur. The larger number of simulated animals (2,500 sires and 25,000 dams per generation) and genotypes (200,000 in the last 4 generations) by Tsuruta et al. (2019) may have contributed to make their predictions for genetic group effects more stable than those in our study. In addition, Tsuruta et al. (2019) showed that the model with the QP H-inverse did not give a unique prediction for the last UPG because they treated UPGs as fixed effects, indicating that the QP H-inverse was not full rank. The undetermined UPG effects in the model with the QP H-inverse caused a severe underestimation of GEBVs in the last generation. Based on UPG estimates, we numerically confirmed that the QP and the altered H-inverses may not be full rank for fixed UPGs. With random UPGs, all H-inverses are full rank.

The models with the QP, altered, MF, and Omega H-inverses gave reasonable genetic trends in real populations (Masuda et al., 2018b; Tsuruta et al., 2019; Kudinov et al., 2020). In livestock populations, usually, the missing-pedigree rate is much lower than in our simulation, the number of genotyped animals is limited in the first genotyping generation, generations are overlapping, and possibly more data can be used for genomic prediction. Such conditions could lead to less extreme genetic trends for proven bulls than those observed in this simulation.

Predictions for young genotyped animals

Table 1 shows predictive abilities (correlations between TBVs and GEBVs), and inflation (regression coefficients of TBVs on GEBVs) values for young genotyped animals in generations 12 and 13. All ssGBLUP models except for the model using the Omega H-inverse showed nearly the same predictive ability and inflation values. Tsuruta et al. (2019) reported that the model with the altered H-inverse gave a better accuracy than the model with the QP H-inverse. Bradford et al. (2019a) showed that the model with the MF H-inverse yielded slightly better predictive ability, inflation, and bias values than the model with the QP H-inverse. In US Holstein, Masuda et al. (2018a) reported low predictive abilities for young bulls with a model using the QP H-inverse. Kudinov et al. (2020) and Macedo et al. (2020) reported more desirable GEBV in actual populations with a model using the MF H-inverse than with models using the QP and Omega H-inverses.

Additional model comparisons

The first additional comparison involved using Estimated Inbreeding in the numerator relationship matrices of all the UPG models. Genetic trends for proven bulls, young genotyped animals, and phenotyped females were almost identical to those obtained with Standard Inbreeding (inbreeding coefficients from pedigree with missing parents) for models with each of the H-inverses (results not shown). Further, the predictive ability and inflation did not improve. We calculated and compared matrix elements from $\mathbf{A}_{22}$ and $\tilde{\mathbf{A}}_{22}$, and the average absolute difference between these two matrices was 0.02 for diagonals, and close to 0 for off-diagonals. Some diagonal elements from $\tilde{\mathbf{A}}_{22}$ were lower than 1. The correlation between the diagonal elements of the 2 matrices was 0.95 and the correlation between the off-diagonals was 1.0. Thus, in our simulation, the 2 matrices were similar enough not to change genetic trends. Misztal et al. (2017), using US Holstein data, showed that Estimated Inbreeding removed the convergence issue in iterative solvers and reduced the inflation of GEBVs compared with Standard Inbreeding; however, accuracy remained unchanged. Further research is required to test Estimated Inbreeding in real populations.

The second additional comparison indicated that when **G** was aligned to $\tilde{\mathbf{A}}_{22}$ the resulting GEBVs were identical to those obtained when **G** was aligned to $\mathbf{A}_{22}$ for both genotyped and non-genotyped animals within the same UPG model. Additionally, we tested current allele frequencies and 0.5 allele frequencies in **G** without alignment, and the adjusted GEBVs were numerically the same as the ones from the aligned **G**. This indicated that with any alignment, GEBVs are numerically identical except for the average GEBVs for genotyped animals (μ_g), and μ_g defines the average GEBVs

Table 1. Correlations (r) between TBV and GEBV and regression coefficients of TBV on GEBV (b_1) for young genotyped animals in generations 12 and 13

	Generation 12		Generation 13	
H-inverse	r	b_1	r	b_1
Omega	0.54	0.63	0.42	0.50
QP	0.59	0.82	0.51	0.76
Altered	0.60	0.82	0.51	0.77
EUPG	0.60	0.83	0.52	0.77
MF	0.60	0.83	0.52	0.78

Omega = UPG applied only to $\mathbf{A}^{-1}$; QP = Quaas–Pollak; altered = altered QP; EUPG = encapsulated UPG; MF = metafounder.

for nongenotyped animals. The value of μ_g is determined by the average of all elements in **G** (Vitezica et al., 2011). One hypothesis is that alignment of **G** may not be needed to obtain unbiased GEBVs when the matrix is large (i.e., (nearly) singular). When massive genomic information describes most of the additive variation in the population and there are enough phenotypes, the genotyped population giving μ_g would be the "base population" like genomic BLUP. In this case, μ_g would adjust $\mathbf{u}_1$ to be comparable with $\mathbf{u}_2$ regardless of the compatibility between **G** and $\mathbf{A}_{22}$. Additional research is needed to determine specific conditions where alignment is not required.

Conclusion

The QP H-inverse is based on an inappropriate assumption that makes GEBVs inseparable from UPG effects. Although the altered H-inverse is more reasonable in theory, it can still cause confounding. These 2 H-inverses could bias genetic trends when ungenotyped sires have many genotyped daughters. The EUPG H-inverse is a new approach to yield stable GEBV and genetic group trends. The EUPG H-inverse contains $\mathbf{A}^*_{22}$, which is equivalent to $\mathbf{A}^{-1}_{\Gamma 22}$ from the MF H-inverse, and models with these 2 H inverses give essentially the same GEBVs in a purebred population. GEBV predictive abilities and inflation values for young genotyped animals were similar among ssGBLUP models with all H-inverses. Models using any of the H-inverses in this study may yield unbiased GEBV and genetic group trends. However, models using the EUPG and MF H-inverses should be preferred because of theoretical justification and their possibility to reduce biases.

Supplementary Data

Supplementary data are available at *Journal of Animal Science* online.

Conflict of interest statement

The authors declare no real or perceived conflicts of interest.

We appreciate Daniela Lourenco (University of Georgia) for useful comments on the manuscript and Paul VanRaden (USDA ARS) for discussion on unknown-parent groups. We gratefully acknowledge the editing of the manuscript by Mauricio Elzo (University of Florida). Two anonymous reviewers provided helpful comments on an earlier draft of the manuscript. This study was partially funded by Agriculture and Food Research Initiative Competitive Grant no. 2020-67015-31030 from the US Department of Agriculture's National Institute of Food and Agriculture.

Literature Cited

Aguilar, I., and I. Misztal. 2008. Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. *J. Dairy Sci.* **91**:1669–1672. doi:10.3168/jds.2007-0575

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 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

Bai, Z., G. Fahey, and G. Golub. 1996. Some large-scale matrix computation problems. *J. Comput. Appl. Math.* **74**:71–89. doi:10.1016/0377-0427(96)00018-0

Bradford, H. L., Y. Masuda, J. B. Cole, I. Misztal, and P. M. VanRaden. 2019a. Modeling pedigree accuracy and uncertain parentage in single-step genomic evaluations of simulated and US Holstein datasets. *J. Dairy Sci.* **102**:2308–2318. doi:10.3168/jds.2018-15419

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

Calus, M., and J. Vandenplas. 2019. Computation of many relationships between metafounders replacing phantom parents. Book of Abstracts of the 70th Annual Meeting of the European Federation of Animal Science. p. 596-596. doi:10.3920/978-90-8686-890-2

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci.* **89**:2673–2679. doi:10.2527/jas.2010-3555

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 M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* **42**:2. doi:10.1186/1297-9686-42-2

Colleau, J. J. 2002. An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.* **34**:409–421. doi:10.1186/1297-9686-34-4-409

Gao, H., O. F. Christensen, P. Madsen, U. S. Nielsen, Y. Zhang, M. S. Lund, and G. Su. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genet. Sel. Evol.* **44**:8. doi:10.1186/1297-9686-44-8

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. Cantet. 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

Graser, H.-U., S. P. Smith, and B. Tier. 1987. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *J. Anim. Sci.* **64**:1362–1370. doi:10.2527/jas1987.6451362x

Kennedy, B. W., L. R. Schaeffer, and D. A. Sorensen. 1988. Genetic properties of animal models. *J. Dairy Sci.* **71**:17–26. doi:10.1016/S0022-0302(88)79975-0

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

Legarra, A., I. Aguilar, and J. J. Colleau. 2020. Methods to compute genomic inbreeding for ungenotyped individuals. *J. Dairy Sci.* **103**:3363–3367. doi:10.3168/jds.2019–17750

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* **92**:4656–4663. doi:10.3168/jds.2009-2061

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

Liu, Z., M. E. Goddard, B. J. Hayes, F. Reinhardt, and R. Reents. 2016. Technical note: equivalent genomic models with a residual polygenic effect. *J. Dairy Sci.* **99**:2016–2025. doi:10.3168/jds.2015-10394

Lourenco, D. A. L., A. Legarra, S. Tsuruta, Y. Masuda, I. Aguilar, and I. Misztal. 2020. Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90. *Genes* **11**:790. doi:10.3390/genes11070790

Lutaaya, E., I. Misztal, J. K. Bertrand, and J. W. Mabry. 1999. Inbreeding in populations with incomplete pedigrees. *J. Anim. Breed. Genet.* **116**:475–480. doi:10.1046/j.1439-0388.1999.00210.x

Macedo, F. L., O. F. Christensen, J. M. Astruc, I. Aguilar, Y. Masuda, and A. Legarra. 2020. Bias and accuracy of dairy sheep

evaluations using BLUP and SSGBLUP with metafounders and unknown parent groups. *Genet. Sel. Evol.* **52**:47. doi:10.1186/s12711-020-00567-1

Martini, J. W. R., M. F. Schrauf, C. A. Garcia-Baccino, E. C. Pimentel, S. Munilla, A. Rogberg-Muñoz, R. J. Cantet, C. Reimer, N. Gao, and V. Wimmer. 2018. The effect of the H –1 scaling factors τ and ω on the structure of H in the single-step procedure. *Genet. Sel. Evol.* **50**:16. doi:10.1186/s12711-018-0386-x

Masuda, Y., I. Misztal, A. Legarra, S. Tsuruta, D. A. L. Lourenco, B. O. Fragomeni, and I. Aguilar. 2017. Avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. *J. Anim. Sci.* **95**:49–52. doi:10.2527/jas.2016.0699

Masuda, Y., I. Misztal, P. M. VanRaden, and T. J. Lawlor. 2018a. Genomic predictability of single-step GBLUP for production traits in US Holstein. *J. Dairy Sci.* **101** (Suppl. 2):182. doi:10.1016/S0022-0302(20)30814-6

Masuda, Y., S. Tsuruta, E. Nicolazzi, and I. Misztal. 2019a. Genomic prediction with missing pedigrees in single-step GBLUP for production traits in US Holstein. Book of Abstracts of the 70th Annual Meeting of the European Federation of Animal Science. p. 593–593. doi:10.3920/978-90-8686-890-2

Masuda, Y., S. Tsuruta, E. Nicolazzi, and I. Misztal. 2019b. Single-step GBLUP including more than 2 million genotypes with missing pedigrees for production traits in US Holstein. Interbull Meeting. Available from https://interbull.org/static/web/10_30_Masuda_final.pdf

Masuda, Y., P. M. VanRaden, I. Misztal, and T. J. Lawlor. 2018b. Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins. *J. Dairy Sci.* **101**:5194–5206. doi:10.3168/jds.2017-13310

Matilainen, K., I. Strandén, G. P. Aamand, and E. A. Mäntysaari. 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

Meuwissen, T. H. E., and Z. Luo. 1992. Computing inbreeding coefficients in large populations. *Genet. Sel. Evol.* **24**:305. doi:10.1186/1297-9686-24-4-305

Misztal, I., H. L. Bradford, D. A. L. Lourenco, S. Tsuruta, Y. Masuda, A. Legarra, and T. J. Lawlor. 2017. Studies on inflation of GEBV in single-step GBLUP for type. *Interbull Bull.* **51**:38–42. https://journal.interbull.org/index.php/ib/article/view/1425

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2018. Manual for blupf90 family of programs. University of Georgia. Available from http://nce.ads.uga.edu/wiki/doku.php

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

Quaas, R. L. 1988. Additive genetic model with groups and relationships. *J. Dairy Sci.* **71**:91–98. doi:10.1016/S0022-0302(88)79986-5

Quaas, R. L., and E. J. Pollak. 1981. Modified equations for sire models with groups. *J. Dairy Sci.* **64**:1868–1872. doi:10.3168/jds.S0022-0302(81)82778-6

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25**:680–681. doi:10.1093/bioinformatics/btp045

Searle, S. R. 1982. *Matrix algebra useful for statistics*. Hoboken, NJ: John Wiley & Sons.

Strandén, I., K. Matilainen, G. P. Aamand, and E. A. Mäntysaari. 2017. Solving efficiently large single-step genomic best linear unbiased prediction models. *J. Anim. Breed. Genet.* **134**:264–274. doi:10.1111/jbg.12257

Tsuruta, S., D. A. L. Lourenco, Y. Masuda, I. Misztal, and T. J. Lawlor. 2019. Controlling bias in genomic breeding values for young genotyped bulls. *J. Dairy Sci.* **102**:9956–9970. doi:10.3168/jds.2019-16789

Tsuruta, S., I. Misztal, I. Aguilar, and T. J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci.* **94**:4198–4204. doi:10.3168/jds.2011-4256

VanRaden, P. M. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. *J. Dairy Sci.* **75**:3136–3144. doi:10.3168/jds.S0022-0302(92)78077-1

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**:4414–4423. doi:10.3168/jds.2007-0980

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. *J. Dairy Sci.* **74**:2737–2746. doi:10.3168/jds.S0022-0302(91)78453-1

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

Westell, R. A., R. L. Quaas, and L. D. Van Vleck. 1988. Genetic groups in a/n animal model. *J. Dairy Sci.* **71**:1310–1318. doi:10.3168/jds.S0022-0302(88)79688-5