Genet. Sel. Evol. 35 (2003) 469–487
© INRA, EDP Sciences, 2003
DOI: 10.1051/gse:2003035

Original article

Bayesian inference on genetic merit under uncertain paternity

Fernando F. CARDOSO*, Robert J. TEMPELMAN

Department of Animal Science, Michigan State University,
East Lansing, MI 48824, USA

(Received 3 October 2002; accepted 3 April 2003)

Abstract – A hierarchical animal model was developed for inference on genetic merit of livestock with uncertain paternity. Fully conditional posterior distributions for fixed and genetic effects, variance components, sire assignments and their probabilities are derived to facilitate a Bayesian inference strategy using MCMC methods. We compared this model to a model based on the Henderson average numerator relationship (ANRM) in a simulation study with 10 replicated datasets generated for each of two traits. Trait 1 had a medium heritability (h^2) for each of direct and maternal genetic effects whereas Trait 2 had a high h^2 attributable only to direct effects. The average posterior probabilities inferred on the true sire were between 1 and 10% larger than the corresponding priors (the inverse of the number of candidate sires in a mating pasture) for Trait 1 and between 4 and 13% larger than the corresponding priors for Trait 2. The predicted additive and maternal genetic effects were very similar using both models; however, model choice criteria (Pseudo Bayes Factor and Deviance Information Criterion) decisively favored the proposed hierarchical model over the ANRM model.

uncertain paternity / multiple-sire / genetic merit / Bayesian inference / reduced animal model

1. INTRODUCTION

Multiple-sire mating is common on large pastoral beef cattle operations in Argentina, Australia, Brazil and parts of the United States, for example. Here, groups of cows are exposed to several males within the same breeding season. Consequently, pedigrees in these herds are uncertain, adversely affecting accuracy of genetic evaluations and selection intensities.

A number of statistical models have been proposed for genetic evaluation of animals with uncertain paternity. One simple solution appears to be genetic grouping [19], whereby "phantom parent" groups are assigned to animals within the same mating pasture. In genetic grouping, phantom parent groups are typically defined to be a contemporary cluster of unknown parents in order

* Correspondence and reprints
E-mail: cardosof@msu.edu

to minimize bias on breeding value predictions due to genetic trend [2,13]. The use of genetic grouping for multiple-sire mating, however, is equivalent to assuming an infinite number of non-inbred, unrelated candidate sires within each group, each candidate having the same probability of being the correct sire [12,17] of the animal with an uncertain paternity. However, only the candidate sires actually used within a group or pasture should be considered.

This requirement is more aptly handled with the average numerator relationship matrix (ANRM) proposed by Henderson [11]. The ANRM is based on knowledge of true probabilities of each candidate male being the correct sire. The ANRM helps specify the correct genetic variance-covariance matrix when these probabilities are presumed known [12], thereby facilitating best linear unbiased predictions (BLUP) of genetic merit. A simple and rapid algorithm to compute the inverse of the ANRM is available [6] and the advantage in selection response of using ANRM *versus* genetic grouping, when candidate sires are recorded, has been demonstrated by simulation studies [12,17]. Equal prior probabilities might be assumed for each sire; however information from blood typing, genetic markers, mating behavior, fertility, breeding period and gestation length could also be used to make these probabilities more distinctive [7,11].

A novel empirical Bayes procedure to infer upon uncertain paternity was proposed by Foulley *et al.* [7,8]. Their sire model implementation combines data and prior information to determine the posterior probabilities of sire assignments for each animal with uncertain paternity. With the advent of Markov chain Monte Carlo (MCMC) techniques in animal breeding [18], it is now possible to extend their method to an animal model and allow a more formal assessment of statistical uncertainty on genetic merit and of probabilities of sire assignments.

The objectives of this study were to: (1) develop a hierarchical animal model and Bayesian MCMC inference strategy for the prediction of genetic merit on animals having an uncertain paternity; (2) use this model to estimate posterior probabilities of paternity, by combining prior and data information; and (3) compare the performance of the proposed model with a model based on the use of the Henderson ANRM having equal prior probability assignments for all candidate sires.

2. THE BAYES HIERARCHICAL MODEL

2.1. The reduced animal model with maternal effects

Consider an $n \times 1$ data vector $\mathbf{y} = \left\{y_{ij}\right\}$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, q$. Here i identifies the record and j the animal associated with the ith record. We allow for the possibility of any animal j having no record; nevertheless, a genetic evaluation on that same animal may be desired if it is related to other animals having data. In the reduced animal model (RAM) of Quaas and Pollak [14], $\mathbf{y}$ is partitioned into two major subsets:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_p \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \mathbf{X}_p \\ \mathbf{X}_t \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_{1p} \\ \mathbf{Z}_{1t}\mathbf{P}_t \end{bmatrix} \mathbf{a}_p + \begin{bmatrix} \mathbf{Z}_{2p} \\ \mathbf{Z}_{2t} \end{bmatrix} \mathbf{m}_p + \begin{bmatrix} \mathbf{e}_p \\ \mathbf{Z}_{1t}\boldsymbol{\gamma}_t + \mathbf{e}_t \end{bmatrix}. \tag{1}$$

The first $n_p \times 1$ subset $\mathbf{y}_p$ of $\mathbf{y}$ is observed on q_p animals that are identified as *parents* or ancestors of other animals having data. In equation (1), $\mathbf{y}_p$ is a linear function of a $p \times 1$ vector of "fixed" effects $\boldsymbol{\beta}$, a $q_p \times 1$ vector of additive direct genetic effects $\mathbf{a}_p$, and a $q_p \times 1$ vector of additive maternal genetic effects $\mathbf{m}_p$. Here, $\mathbf{a}_p$ and $\mathbf{m}_p$ correspond to effects on the q_p parents. The design matrices connecting $\mathbf{y}_p$ to $\boldsymbol{\beta}$, $\mathbf{a}_p$ and $\mathbf{m}_p$ are $\mathbf{X}_p$, $\mathbf{Z}_{1p}$, and $\mathbf{Z}_{2p}$, respectively. The remaining $n_t \times 1$ data subset $\mathbf{y}_t$ is recorded on *terminal* or non-parent animals who are not parents of any other animals with data. As with $\mathbf{y}_p$, $\mathbf{y}_t$ is modeled similarly as a function of $\boldsymbol{\beta}$, and $\mathbf{m}_p$ except that t rather than p is used as the subscript index for the respective design matrices in equation (1). Furthermore, $\mathbf{y}_t$ is modeled as a linear function (through $\mathbf{Z}_{1t}\mathbf{P}_t$) of $\mathbf{a}_p$. Here $\mathbf{Z}_{1t}$ is a $n_t \times q_t$ design matrix and $\mathbf{P}_t$ is a $q_t \times q_p$ matrix connecting the genetic effects of q_t non-parent animals to that of their parents. That is, in $\mathbf{P}_t$, row j, indexed $j = q_p + 1, q_p + 2, \ldots, q$ connects the genetic effect of non-parent animal j to that of its sire s_j^* and dam d_j^* such that the j, s_j^* and j, d_j^* elements of $\mathbf{P}_t$ for identified parents of animal j are equal to 0.5. The "residual" vector is composed of error terms $\mathbf{e}_p$ and $\mathbf{e}_t$, respectively of parent and terminal animals, and additionally, for terminal animals, of additive Mendelian genetic sampling terms in the vector $\boldsymbol{\gamma}_t$, which is connected to $\mathbf{y}_t$ through $\mathbf{Z}_{1t}$.

We assume that the variance covariance matrix of the RAM residual vector is:

$$\mathbf{R} = \text{var} \begin{bmatrix} \mathbf{e}_p \\ \mathbf{Z}_{1t}\boldsymbol{\gamma}_t + \mathbf{e}_t \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{q_p}\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{1t}\boldsymbol{\Omega}_{tt}\mathbf{Z}'_{1t}\sigma_a^2 + \mathbf{I}_{q_t}\sigma_e^2 \end{bmatrix}, \tag{2}$$

where $\boldsymbol{\Omega}_{tt} = \text{diag}\left\{\omega_j\right\}_{j=q_p+1}^{q}$ is a $q_t \times q_t$ diagonal matrix, with the jth element corresponding to the proportion of the additive genetic variance (σ_a^2) on animal j that is due to Mendelian sampling [13]; and σ_e^2 is the residual variance.

The structural prior specifications on the genetic effects are defined accordingly to include only parent terms, *i.e.*

$$p\left(\begin{bmatrix} \mathbf{a}_p \\ \mathbf{m}_p \end{bmatrix} | \mathbf{G}\right) = N\left(\mathbf{0}, \mathbf{G} \otimes \mathbf{A}_{pp}\right), \tag{3}$$

where $\mathbf{G} = \begin{bmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{bmatrix}$ is the genetic variance-covariance matrix for direct and maternal genetic effects with σ_m^2 being the maternal genetic variance, and σ_{am}

the covariance between direct and maternal genetic effects. Furthermore, $\mathbf{A}_{pp}$ is the numerator relationship matrix amongst all q_p parent animals and $\otimes$ is the Kronecker or direct product. For conjugate convenience, a joint bounded uniform or normal prior $p(\boldsymbol{\beta})$ may be specified for $\boldsymbol{\beta}$, an inverted Wishart prior density $p(\mathbf{G})$ specified for $\mathbf{G}$ and an inverted gamma density $p(\sigma_e^2)$ specified for σ_e^2.

In addition we have that

$$\begin{bmatrix} \mathbf{a}_t \\ \mathbf{m}_t \end{bmatrix} = \begin{bmatrix} \mathbf{I}_2 \otimes \mathbf{P}_t \end{bmatrix} \begin{bmatrix} \mathbf{a}_p \\ \mathbf{m}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{\gamma}_t \\ \boldsymbol{\delta}_t \end{bmatrix}, \tag{4}$$

where $\mathbf{a}_t$ and $\mathbf{m}_t$ are respectively, the $q_t \times 1$ vectors of additive and maternal genetic effects associated with terminal animals. Furthermore, $\boldsymbol{\gamma}_t$ and $\boldsymbol{\delta}_t$ are each $q_t \times 1$ vectors of additive and maternal Mendelian genetic sampling terms, respectively, also associated with terminal animals and such that

$$\begin{bmatrix} \boldsymbol{\gamma}_t \\ \boldsymbol{\delta}_t \end{bmatrix} | \mathbf{G} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{G} \otimes \boldsymbol{\Omega}_{tt} \right). \tag{5}$$

2.2. Modeling uncertain paternity

In populations undergoing multiple-sire mating, a number of males are possible candidate sires for each of several animals. This translates into uncertainty on various elements of $\mathbf{P}_t$ for non-parent animals and on various elements of $\mathbf{A}_{pp}$ for parent animals.

We first considered uncertain paternity on the q_t non-parent animals indexed $j = q_p+1, q_p+2, \ldots, q$ and associated with n_t records in $\mathbf{y}_t$. Let $\mathbf{Z}_1 = \begin{bmatrix} \mathbf{Z}_{1p} \\ \mathbf{Z}_{1t} P_t \end{bmatrix}$. Then if non-parent j has uncertain paternity, this uncertainty translates into the j, s_j^* element of $\mathbf{P}_t$ being unknown or, equivalently, the s_j^* element of $\mathbf{z}'_{1ij}$ being unknown, where $\mathbf{z}'_{1ij}$ is the row of $\mathbf{Z}_1$ matching with the address of y_{ij} in $\mathbf{y}$. Suppose, that for animal j, there are v_j possible candidate sires with identifications listed in $\mathbf{s}_j = \left\{ s_j^{(1)}, s_j^{(2)}, \ldots, s_j^{(v_j)} \right\}$. The distribution of y_{ij}, conditional on a given sire assignment $s_j^* = s_j^{(k)}$, $1 \leq k \leq v_j$, on animal j and all other parameters is given by:

$$\begin{aligned} y_{ij} | \boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, s_j^* = s_j^{(k)}, \sigma_a^2, \sigma_e^2 \\ \sim N \left(\mathbf{x}'_{ij} \boldsymbol{\beta} + 0.5 a_{s_j^{(k)}} + 0.5 a_{d_j^*} + \mathbf{z}'_{2ij} \mathbf{m}_p, \sigma_e^2 + \omega_j^{(k)} \sigma_a^2 \right), \\ i = n_p + 1, n_p + 2, \ldots, n; j = q_p + 1, q_p + 2, \ldots, q. \end{aligned} \tag{6a}$$

Here $\mathbf{x}'_{ij}$, and $\mathbf{z}'_{2ij}$ are, respectively, the rows of $\mathbf{X}$ and $\mathbf{Z}_2$ matching the address of y_{ij} in $\mathbf{y}$. When animal j has certain paternity, $v_j = 1$ such that then s_j^* is not random. Note that the conditioning on known d_j^* (dam identification) is implied for all animals throughout this paper whereas the conditioning on $s_j^* = s_j^{(k)}$ is explicitly provided given that s_j^* may be uncertain. This uncertainty is further reflected in equation (6a) by the term $\omega_j^{(k)} = \omega_j\Big|_{s_j^*=s_j^{(k)}}$ indicating that fraction $\omega_j^{(k)}$ of genetic variance attributable to Mendelian sampling for animal j is a function of its inbreeding coefficient and hence of the sire assignment $s_j^* = s_j^{(k)}$.

Now consider the possibility that at least one of the parent animals, indexed from 1 to q_p, has uncertain paternity such that elements of $\mathbf{A}_{pp}$ are also uncertain. The sampling distribution of y_{ij}, on parent animal j, $j = 1, 2, \ldots, q_p$, is not conditioned on uncertainty on sires, that is,

$$y_{ij}|\boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \sigma_e^2 \sim N\left(\mathbf{x}'_{ij}\boldsymbol{\beta} + a_j + \mathbf{z}'_{2ij}\mathbf{m}_p, \sigma_e^2\right), \\ i = 1, 2, \ldots, n_p; j = 1, 2, \ldots, q_p. \quad (6b)$$

Uncertain paternity on parent animals is modeled with the second stage structural prior on $\mathbf{a}_p$ and $\mathbf{m}_p$ in equation (3). A useful decomposition of $\mathbf{A}_{pp}^{-1}$ as shown by Henderson [10] and Quaas [13] is $\mathbf{A}_{pp}^{-1} = \mathbf{T}_p\boldsymbol{\Omega}_{pp}^{-1}\mathbf{T}'_p$, where $\mathbf{T}_p$ is a $q_p \times q_p$ lower triangular matrix and $\boldsymbol{\Omega}_{pp} = \text{diag}\left\{\omega_j\right\}_{j=1}^{q_p}$ is a $q_p \times q_p$ diagonal matrix analogous to $\boldsymbol{\Omega}_{tt}$, but with elements corresponding to the fraction of σ_a^2 due to Mendelian sampling on each parent animal j. All of the diagonal elements of $\mathbf{T}_p$ are equal to 1 with at most two other elements per row, say j, s_j^* and j, d_j^*, being equal to -0.5, if the corresponding parents s_j^* and d_j^* of animal j are identified, for $j = 1, 2, \ldots, q_p$. Consequently, $\left|\mathbf{A}_{pp}^{-1}\right| = \left|\mathbf{T}_p\right|\left|\boldsymbol{\Omega}_{pp}^{-1}\right|\left|\mathbf{T}'_p\right| = \left|\boldsymbol{\Omega}_{pp}^{-1}\right|$ since $\left|\mathbf{T}_p\right| = 1$. Given this result, the joint prior density of $\mathbf{a}_p$ and $\mathbf{m}_p$ conditioned on $\mathbf{A}_{pp}$, can be written as,

$$p\left(\begin{matrix}\mathbf{a}_p \\ \mathbf{m}_p\end{matrix}\middle| \mathbf{G}, \mathbf{A}_{pp}\right) \propto |\mathbf{G}|^{-\frac{p}{2}} \left|\boldsymbol{\Omega}_{pp}^{-1}\right| \exp(-0.5 \\ \times \left(\mathbf{a}'_p\mathbf{T}_p\boldsymbol{\Omega}_{pp}^{-1}\mathbf{T}'_p\mathbf{a}_p g^{11} + 2\mathbf{a}'_p\mathbf{T}_p\boldsymbol{\Omega}_{pp}^{-1}\mathbf{T}'_p\mathbf{m}_p g^{12} + \mathbf{m}'_p\mathbf{T}_p\boldsymbol{\Omega}_{pp}^{-1}\mathbf{T}'_p\mathbf{m}_p g^{22}\right)), \quad (7)$$

where g^{ij} is the (i, j)th element of $\mathbf{G}^{-1}$ for $i, j = 1, 2$.

Let $\mathbf{t}'_j$ denote the jth row of $\mathbf{T}_p$. Then it can be readily shown that the additive and maternal Mendelian sampling terms are respectively $\gamma_j = \mathbf{t}'_j\mathbf{a}_p = a_j - 0.5a_{s_j^*} - 0.5a_{d_j^*}$ and $\delta_j = \mathbf{t}'_j\mathbf{m}_p = m_j - 0.5m_{s_j^*} - 0.5m_{d_j^*}$ for $j = 1, \ldots, q_p$. If there are no known candidates for s_j^* and d_j^* then the corresponding parental contributions of $a_{s_j^*}$ and $a_{d_j^*}$ to γ_j and $m_{s_j^*}$ and $m_{d_j^*}$ to δ_j are equal to 0, as would be true for each of the base population animals $j = 1, 2, \ldots, q_b \leq q_p$. Let $\mathbf{s}_p^* = \left\{s_j^*\right\}_{j=1}^{q_p}$ denote the vector of random sire assignments on parent animals and $\mathbf{s}_p^{(\mathbf{k})} = \left\{s_j^{(k)}\right\}_{j=1}^{q_p}$ be a particular realization of $\mathbf{s}_p^*$ from the set

$\mathbf{S}_p = \left\{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \ldots, \mathbf{s}_{q_p}\right\}$ such that the jth element of $\mathbf{s}_p^{(\mathbf{k})}$ is one of the v_j elements chosen from $\mathbf{s}_j = \left\{s_j^{(1)}, s_j^{(2)}, \ldots, s_j^{(v_j)}\right\}$ for $j = 1, 2, \ldots, q_p$. Note that for the q_b base animals, $\mathbf{s}_j$ is an empty set. We can then rewrite (7), explicitly conditioning on sire assignments as follows:

$$p\left(\left.\begin{matrix}\mathbf{a}_p\\ \mathbf{m}_p\end{matrix}\right| \mathbf{G}, \mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})}\right) \propto |\mathbf{G}|^{-\frac{q_p}{2}} \prod_{j=1}^{q_p} \left(\left(\omega_j^{(k)}\right)^{-1}\right.$$
$$\left.\times \exp\left(-0.5\left(\omega_j^{(k)}\right)^{-1}\left(\left(\gamma_j^{(k)}\right)^2 g^{11} + \left(\delta_j^{(k)}\right)^2 g^{22} + 2\gamma_j^{(k)}\delta_j^{(k)} g^{12}\right)\right)\right), \qquad (8)$$

where $\delta_j^{(k)} = \delta_j\big|_{s_j^* = s_j^{(k)}}$ and $\gamma_j^{(k)} = \gamma_j\big|_{s_j^* = s_j^{(k)}}$, indicating the natural dependence of Mendelian sampling terms on the sire assignment $s_j^* = s_j^{(k)}$. Since there is no need to infer upon uncertain paternity for the q_b base animals, $\omega_j^{(k)} = 1$ for $j = 1, 2, \ldots, q_b$ with $\left\{s_j^{(k)}\right\}_{j=1}^{q_b}$ being an empty subset of $\mathbf{s}_p^{(\mathbf{k})}$.

The third stage of the model specifies the prior probability for each of v_j males being the correct sire of an animal j. As we do similarly for parents, we let $\mathbf{s}_t^* = \left\{s_j^*\right\}_{j=q_p+1}^{q}$ denote the vector of random sire assignments on the non-parent animals and $\mathbf{s}_t^{(\mathbf{k})} = \left\{s_j^{(k)}\right\}_{j=q_p+1}^{q}$ denote a particular realization of $\mathbf{s}_t^*$ from the set $\mathbf{S}_t = \left\{\mathbf{s}_{q_p+1}, \mathbf{s}_{q_p+2}, \ldots, \mathbf{s}_q\right\}$. For all q animals, we then write $\mathbf{s}^{(\mathbf{k})} = \begin{bmatrix}\mathbf{s}_p^{(\mathbf{k})}\\ \mathbf{s}_t^{(\mathbf{k})}\end{bmatrix} = \left\{s_j^{(k)}\right\}_{j=1}^{q}$ as being a realization of $\mathbf{s}^* = \begin{bmatrix}\mathbf{s}_p^*\\ \mathbf{s}_t^*\end{bmatrix} = \left\{s_j^*\right\}_{j=1}^{q}$ from the set $\mathbf{S} = \left\{\mathbf{S}_p, \mathbf{S}_t\right\}$. The probability that $s_j^{(k)}$ is the sire of animal j is defined as $\pi_j^{(k)} = \text{Prob}\left(s_j^* = s_j^{(k)}\right)$ for $k = 1, 2, \ldots, v_j$ such that $\sum_{k=1}^{v_j} \pi_j^{(k)} = 1$. For animals with certain paternity, there is only one candidate $s_j^* \equiv s_j^{(1)}$ such that $\pi_j^{(1)} = 1$ and hence is constant. For each of the q_b base animals, $\pi_j^{(k)}$ is not specified since there are no candidate sires. The set of probabilities $\boldsymbol{\pi}_j = \left\{\pi_j^{(1)}, \pi_j^{(2)}, \ldots, \pi_j^{(v_j)}\right\}$ for each one of v_j candidate sires for non-base animal j ($j = q_b + 1, q_b + 2, \ldots, q$) may be conceptually elicited using external information (*e.g.* genetic markers). The entire set of probabilities $\boldsymbol{\pi} = \left\{\boldsymbol{\pi}_j\right\}_{q_b+1}^{q}$ is rarely known with absolute certainty, and so we might regard them as random quantities from a Dirichlet distribution:

$$p\left(\boldsymbol{\pi}_j|\boldsymbol{\alpha}_j\right) \propto \prod_{k=1}^{v_j} \left(\pi_j^{(k)}\right)^{\alpha_j^{(k)}} \qquad (9)$$

where $\boldsymbol{\alpha}_j = \left\{\alpha_j^{(k)}\right\}_{k=1}^{v_j}$, $\alpha_j^{(k)} > 0$ for $k = 1, 2, \ldots, v_j$ and $\pi_j^{(v_j)} = 1 - \sum_{k=1}^{v_j-1} \pi_j^{(k)}$ is constrained accordingly. Specification of the set of hyper-parameters $\boldsymbol{\alpha} = \left\{\boldsymbol{\alpha}_j\right\}_{j=q_b+1}^{q}$ might be based on the assessed reliability of the source of external information on the prior probability of each sire assignment.

We use (6a), (6b), and (8) as key expressions to determine the joint posterior density of all unknown parameters:

$$
\begin{aligned}
&p\left(\boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}, \boldsymbol{\pi}, \mathbf{G}, \sigma_e^2 | \mathbf{y}\right) \\
&\propto \prod_{i=1}^{n_p} p\left(y_{ij} | \boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \sigma_e^2\right) \prod_{i=n_p+1}^{n} p\left(y_{ij} | \boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}_t^* = \mathbf{s}_t^{(\mathbf{k})}, \sigma_a^2, \sigma_e^2\right) \\
&\times p\left(\mathbf{a}_p, \mathbf{m}_p | \mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})}, \mathbf{G}\right) p\left(\boldsymbol{\beta}\right) \mathrm{Prob}\left(\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})} | \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi} | \boldsymbol{\alpha}\right) p\left(\mathbf{G}\right) p\left(\sigma_e^2\right). \quad (10)
\end{aligned}
$$

Here,

$$
\begin{aligned}
\mathrm{Prob}\left(\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})} | \boldsymbol{\pi}\right) &= \prod_{j=q_b+1}^{q} \mathrm{Prob}\left(s_j^* = s_j^{(k)} | \boldsymbol{\pi}_j\right) \\
&= \prod_{j=q_b+1}^{q} \prod_{k=1}^{v_j} \left(\pi_j^{(k)}\right)^{I_j^{(k)}},
\end{aligned}
$$

where $I_j^{(k)} = 1$ if $s_j^* = s_j^{(k)}$ and $I_j^{(k)} = 0$ otherwise. Furthermore, $p\left(\boldsymbol{\pi} | \boldsymbol{\alpha}\right) = \prod_{j=q_b+1}^{q} p\left(\boldsymbol{\pi}_j | \boldsymbol{\alpha}_j\right) = \prod_{j=q_b+1}^{q} \prod_{k=1}^{v_j} \left(\pi_j^{(k)}\right)^{\alpha_j^{(k)}}$. The fully conditional distributions (FCD) of all unknown parameters/quantities or blocks thereof in (10) necessary to conduct MCMC inference with some details on the sampling strategy itself are derived in the Appendix of this paper. A good exposition on MCMC implementations in hierarchical animal breeding models analogous to that presented in this paper is provided by Wang *et al.* [19].

3. SIMULATION STUDY

A simulation study was carried out to compare two models for the prediction of genetic merit allowing for uncertain paternity on some animals. The first model is the hierarchical model proposed in this paper (section 2), which infers upon this uncertainty using phenotypic data; the other model is based on the use of the Henderson average numerator relationship.

Ten datasets were generated for each of two different types of traits. Trait 1 had medium direct heritability ($h_a^2 = 0.3$), medium maternal heritability ($h_m^2 = 0.2$) and a slightly negative direct-maternal correlation ($r_{am} = -0.2$) as, for example, would characterize weaning weight. Trait 2 had a high direct heritability ($h_a^2 = 0.5$), but null h_m^2 as would characterize post-weaning gain. The residual variance (σ_e^2) was 60 and 50, respectively for Traits 1 and 2.

Each population included 80 sires, 400 dams (480 parents) and 2000 non-parent animals, all of which descended from 20 base sires and 100 base dams. From these base animals, five generations were created. Fifteen males and 75 females were randomly selected from each generation to be parents of the next generation. Furthermore, five sires and 25 dams from the previous generation's breedstock were retained, such that a total of 20 sires and 100 dams were used as the breeding group for each generation. That is, the population was structured to have overlapping generations. The probability of any offspring being assigned to an uncertain paternity situation was 0.3. If an animal had uncertain paternity, it was randomly assigned to one of six possible multiple-sire groups in each of the five generations. These groups had six different sizes: $v_j = 2, 3, 4, 6, 8$ or 10 candidate sires. Once the group was chosen, one of the males in the group was selected to be the true sire with either equal ($1/v_j$) or unequal probability relative to the rest of the candidate sires (the actual probabilities used to assign progeny to sires in each group can be obtained from the corresponding author upon request). The latter scenario was intended to represent the dominant male situation, common in beef cattle [5]. The five sires selected from the previous generation's breedstock had only certain progeny. An additional ten sires were used in group matings but also had certain progeny, whereas the remaining five sires had only uncertain progeny. One group of three sires in each population was formed with sires having only uncertain progeny with the purpose of comparing the performance of the two models in the case where sires have only their own record and pedigree as the only source of information for their genetic evaluation, other than uncertain progeny. All other mating groups had at least one sire that was known to be sires of other animals. We deliberately intended to mimic the situation observed in some ranches under genetic evaluation in Brazil. These ranches select their own young bulls to serve their herd by natural service (NS) and also collect semen from their own top bulls to be used in artificial insemination (AI). Moreover, they import external genetics especially through AI. In this scenario, the sires can be categorized in three different ways: (1) sires having only known progeny (*i.e.* imported AI bulls); (2) sires having both known and uncertain progeny assignments, such as top herd bulls that are used by AI or known NS mating but also by uncertain NS in multiple sire pastures during the breeding season and (3) sires having only uncertain progeny assignments.

Only one record was generated per animal. For both traits, the overall mean was equal to 100 and a fixed effects factor with three levels, having values 25, −25 and 0, was randomly assigned to generate the individual records.

The ten replicates for each of the two traits were analyzed using three different models:

(1) HIER: A hierarchical mixed effects model fully accounting for uncertainty on sire assignments as proposed in section 2.
(2) ANRM: A linear mixed effects model based on the average numerator relationship matrix [11]. Equal and fixed probabilities were assigned to each candidate sire of animals pertaining to uncertain paternity.
(3) TRUE: A linear mixed effects model based on the true sire assignments, as if there was no uncertainty on assignments. This model was included to serve as a positive control for the other two models.

For all three models, a MCMC sampling chain of $G = 20\,000$ cycles was run after a burn-in period of 4000 cycles. In order to concentrate our attention on the relative performance of the models for breeding value prediction, variance components were considered to be known. Flat bounded priors were placed on each fixed effect. Naïve equal prior probabilities, *i.e.* inverse of the number of candidate sires within each group, were specified on each sire assignment to an animal. By setting $\alpha_j^{(k)} = \dfrac{1}{v_j}$ for $k = 1, 2, \ldots, v_j$, we have that $\sum_{k=1}^{v_j} \alpha_j^{(k)} = 1$, and the same weight is statistically given to prior and data information in the sampling of sire assignments for the jth animal in the set of animals with uncertain paternity.

The parameters used to compare the methods studied were the mean squared error of prediction (MSEP), the mean bias of prediction (MBIAS) and rank (Spearman) correlations between estimated and true genetic values. The MSEP for each model was estimated as $\sum_{h=1}^{10} \sum_{j=1}^{q} \left(\hat{u}_{hj} - u_{hj}\right)^2 /q\big/10$, where 10 denotes the number of replicates, q is the total number of parent or non-parent animals with uncertain paternity per replicate, $\hat{u}_{hj}$ is the estimated genetic additive or maternal effect for animal j in replicate h and u_{hj} is the true genetic additive or maternal effect for animal j in replicate h. MBIAS was similarly estimated as $\sum_{h=1}^{10} \sum_{j=1}^{q} \left(\hat{u}_{hj} - u_{hj}\right) /q\big/10$.

Variables describing uncertain paternity, specifically, s_j^* and $\pi_j^{(k)}$, were analyzed separately for parents and non-parents, since parents were considered to have greater amounts of information on their genetic merit compared to non-parents. Sires had on average 23.6 progeny, while dams averaged 5.9 progeny. Within each group size category, the animals with certain paternity and with uncertain paternity were considered separately. Pairwise comparisons based on genetic merits estimated under the three different models were performed using a t-test.

We also considered two model choice criteria: the *Pseudo Bayes Factor* (PBF) [9] and the *Deviance Information Criterion* (DIC) [16]. For comparing, say, models M_1 and M_2, the corresponding PBF was determined to be:

$$PBF_{1,2} = \prod_{i=1}^{n} \frac{p\left(y_{ij}|\mathbf{y}_{(-ij)}, M_1\right)}{p\left(y_{ij}|\mathbf{y}_{(-ij)}, M_2\right)},$$

where $p\left(y_{ij}|\mathbf{y}_{(-ij)}, M_r\right)$ is the conditional predictive ordinate (CPO) for observation y_{ij}, intended to be a cross-validation density, which suggests what values of y_{ij} are likely when Model M_r is fit to all other observations $\mathbf{y}_{(-ij)}$ except y_{ij}. An MCMC approximation for the CPO of Model M_r with parameters θ is obtained by a harmonic mean of the G MCMC cycles:

$$p\left(y_{ij}|\mathbf{y}_{(-ij)}, M_r\right) \approx \frac{1}{\frac{1}{G}\sum_{l=1}^{G} p^{-1}\left(y_{ij}|\boldsymbol{\theta}^{(l)}, M_r\right)}.$$

The DIC is composed of a measure of global fit, posterior mean of the deviance, and a penalization for complexity of the model. The deviance for Model M_r using the null standardization from Spiegelhalter *et al.* [16] can be estimated by $\bar{D}_r = \frac{1}{G}\sum_{l=1}^{G} -2\log p\left(\mathbf{y}|\boldsymbol{\theta}^{(l)}, M_r\right)$. The "complexity" of Model M_r is determined as the effective number of parameters given by $p_{D(r)} = \bar{D}_r - D_r(\bar{\boldsymbol{\theta}})$ where $D_r(\bar{\boldsymbol{\theta}}) = -2\log p(\mathbf{y}|\bar{\boldsymbol{\theta}}, M_r)$ with $\bar{\boldsymbol{\theta}}$ being the posterior mean of $\boldsymbol{\theta}$. That is, $p_{D(r)}$ represents the difference between the posterior mean of the deviance and the deviance based on the posterior mean of the parameters under Model M_r. The DIC for Model M_r is then determined as:

$$DIC_r = \bar{D}_r + p_{D(r)}.$$

Smaller values of DIC are indicative of a better-fitting model.

4. RESULTS

Since it was unclear to us whether the indicator variable s_j^* or parameter $\pi_j^{(k)}$ should be used for inferring uncertainty with respect to the assignment of sire k to animal j, we considered both variables. Inference on the probabilities of the true sires for animals with uncertain paternity in the HIER model was based on determining the frequency of the MCMC samples of s_j^* that were equal to the true sire, designated as Prob $\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$, and by determining E $\left(\pi_j^{(true)}|\mathbf{y}\right)$ the posterior mean of $\pi_j^{(true)}$, the probability parameter identified with $s_j^{(true)}$, the true sire of j. These summaries are presented separately for parent and non-parent animals with uncertain paternity in Table I for both Traits 1 and 2.

Table I. Posterior means of probabilities of sires being true sires $\left(\mathrm{E}\left(\pi_j^{(true)}|\mathbf{y}\right)\right)$ and probability of sire assignments being equal to true sires $\left(\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)\right)$ averaged across sires and replicates for Traits 1 and 2 by multiple-sire group size and parents *versus* non-parent animals.

		Multiple-sire group size					
Parameter	Animal Category	2	3	4	6	8	10
Trait 1							
$\left(\mathrm{E}\left(\pi_j^{(true)}|\mathbf{y}\right)\right)$	Parents	0.513	0.341	0.259	0.175	0.126[a]	0.105
$\left(\mathrm{E}\left(\pi_j^{(true)}|\mathbf{y}\right)\right)$	Non-parents	0.509	0.339	0.259	0.172	0.130	0.103
$\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$	Parents	0.525	0.349	0.269	0.183	0.127	0.110
$\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$	Non-parents	0.517	0.345	0.268	0.178	0.134	0.105
Trait 2							
$\left(\mathrm{E}\left(\pi_j^{(true)}|\mathbf{y}\right)\right)$	Parents	0.510	0.343	0.265	0.177	0.132	0.105
$\left(\mathrm{E}\left(\pi_j^{(true)}|\mathbf{y}\right)\right)$	Non-parents	0.520	0.346	0.270	0.179	0.134	0.106
$\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$	Parents	0.521	0.352	0.280	0.188	0.138	0.111
$\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$	Non-parents	0.540	0.360	0.289	0.191	0.143	0.111

[a] posterior probability is not statistically different from the prior of its group size at $\alpha = 0.05$.

The average posterior probabilities attributed to the true sire (*i.e.* based on $\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$) were between 1 and 10% larger than the respective priors ($1/v_j$ for a respective mating group of size v_j) for Trait 1 and between 4 and 13% larger than the priors for Trait 2. Inference on uncertain paternity using $\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$ had a slightly better general performance than an inference based on $\mathrm{E}\left(\pi_j^{(true)}|\mathbf{y}\right)$. The larger differences between the average posterior and prior probabilities in Trait 2 may be a result of the higher heritability. These differences were generally statistically significant ($P < 0.05$), based on one-sample *t*-tests.

The consistently higher probability attributed to $s_j^{(true)}$ by HIER indicates that this model tends to infer towards the correct sire; however, the small magnitude of these differences suggests that phenotypes may not be sufficiently informative to precisely infer upon paternity assignments under these two trait scenarios. The average $\mathrm{Prob}\left(s_j^* = s_j^{(true)}|\mathbf{y}\right)$ for mating groups of size

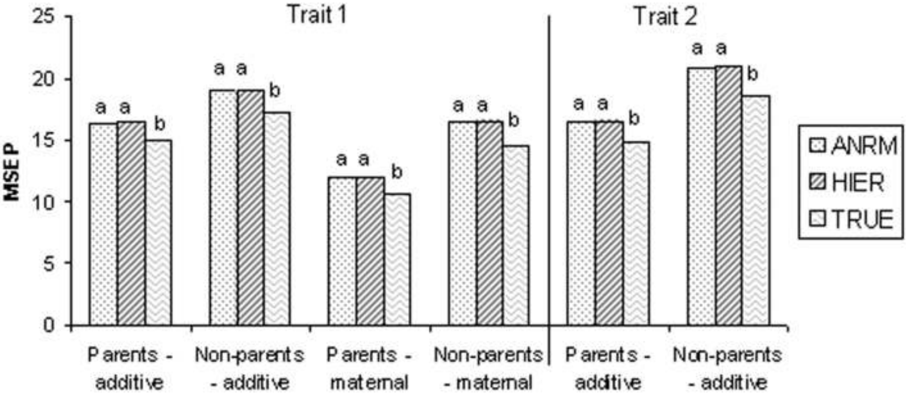

Figure 1. Mean squared error of prediction (MSEP) of posterior means of additive and maternal genetic effects of parent and non-parent animals with uncertain paternity for Traits 1 and 2 under three models: (1) HIER based on proposed hierarchical model; (2) ANRM based on the Henderson average numerator relationship matrix; and (3) TRUE based on knowledge of the true sire as a positive control. Within each group, bars sharing the same letter are not statistically different at $\alpha = 0.05$.

$v_j = 3$ and formed with sires with exclusively uncertain progeny were 0.348 for Trait 1 and 0.360 for Trait 2. These probabilities were consistent with those determined for other groups of size $v_j = 3$ but including sires that had also certain progeny. That is, the HIER model performed similarly in terms of probabilities of assignments to sires whether or not sires have both certain and uncertain progeny or only uncertain progeny as the source of information.

In terms of MBIAS, none of the three models were significantly different from each other under all situations analyzed, and the results are not presented here. The mean squared error of prediction (MSEP) and rank correlation on additive and maternal genetic effects of parents and non-parents, with uncertain paternity for Trait 1 (medium h_a^2 – additive and maternal effects) are presented in Figures 1 and 2, respectively. As expected, the MSEP was always smaller and rank correlation higher for TRUE compared to ANRM and HIER, showing that the use of multiple-sire matings adversely affects the accuracy of genetic evaluations [17]. Posterior means of additive and maternal genetic effects were very similar for HIER and ANRM with no significant difference in MSEP and rank correlations on these posterior means between these models. There was, however, a tendency of having a smaller MSEP and a higher rank correlation under HIER for animals with uncertain paternity. There does not seem to be enough information, at least in this simulated scenario, to discriminate between ANRM and HIER for MSEP and rank correlation of genetic evaluations using only phenotypic records. This result may be due to the small differences between prior and posterior probabilities of sire assignments under HIER.

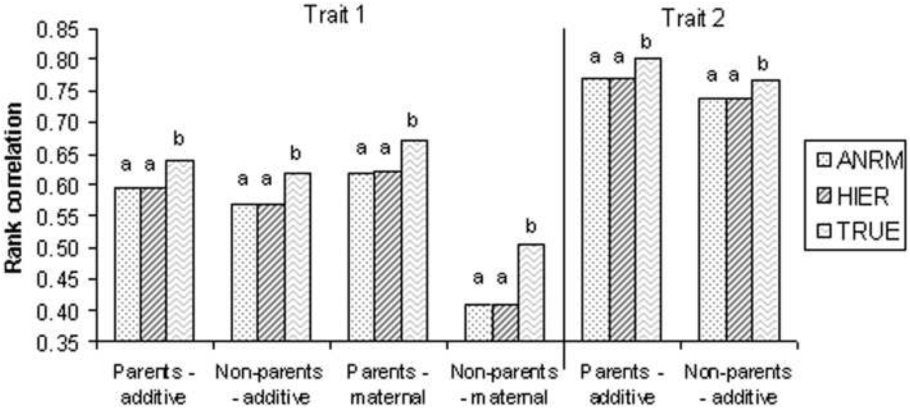

Figure 2. Rank correlation of additive and maternal genetic effects of parent and non-parent animals with uncertain paternity for Traits 1 and 2 under three models: (1) HIER based on proposed hierarchical model; (2) ANRM based on the Henderson average numerator relationship matrix; and (3) TRUE based on knowledge of the true sire as a positive control. Within each group, bars sharing the same letter are not statistically different at $\alpha = 0.05$.

For Trait 2, the MSEP and rank correlation were also not statistically different between ANRM and HIER across the ten simulated datasets (Figs. 1 and 2). Here, the differences in terms of rank correlation among models were somewhat smaller relative to Trait 1. This result may be due to the higher h^2, and therefore the decreased importance of pedigree information, *i.e.* sire assignments, relative to phenotypes for the prediction of genetic effects.

We applied two model choice criteria, the PBF and DIC as previously described, to compare the statistical fit of the two models, ANRM and HIER. The PBF for all replicates were always favorable to HIER compared to ANRM, with magnitudes ranging from 2.1×10^2 to 2.4×10^7 for Trait 1, and from 6.3×10^7 to 2.6×10^{24} for Trait 2. The calculated DIC were also always in favor of HIER compared to ANRM ranging from differences of 9 to 41 for Trait 1 and from 33 to 115 for Trait 2. These results appear to be decisively in favor of the HIER model since Spiegelhalter *et al.* [16] has suggested a DIC difference of 7 to be an important difference in the model fit. For Trait 1, the average DIC over the ten replicates was 17 843 for HIER ($\bar{D}_{HIER} = 17\,135$ and $p_{D(HIER)} = 709$) and 17 866 for ANRM ($\bar{D}_{ANRM} = 17\,164$ and $p_{D(ANRM)} = 702$); and for Trait 2 we obtained an average DIC of 17 553 for HIER ($\bar{D}_{HIER} = 16\,605$ and $p_{D(HIER)} = 949$) and of 17 630 for ANRM ($\bar{D}_{ANRM} = 16\,704$ and $p_{D(ANRM)} = 926$). The primary reason for a smaller DIC for HIER compared to ANRM was the smaller mean deviance ($\bar{D}_r$) of HIER. The difference in terms of $\bar{D}_r$ was large enough to compensate for the penalty of a larger effective number of parameters ($p_{D(r)}$) applied to HIER. These two model choice criteria (PBF and DIC) clearly indicate that the HIER model provides a better statistical fit than the ANRM model to the simulated data involving animals with uncertain paternity.

5. DISCUSSION

We proposed in this study a fully Bayesian approach for prediction of genetic merit of animals having uncertain paternity. Similar to the empirical Bayes sire model method of Foulley *et al.* [7], our procedure combines data and prior information to determine posterior probabilities of sire assignments. Nevertheless, our method represents an important extension since it uses more recently developed MCMC tools to provide small sample inference based on the animal model, the most common model for current genetic evaluations. Our method can be readily extended to multiple-trait or other quantitative genetic (*e.g.* random regression) models without great conceptual difficulty. It could also be easily generalized to the case of uncertain dams; however, this is not a typical scenario in livestock breeding.

The results obtained from our simulation study indicate that a model accounting for uncertainty on sire assignments provides a better fit to data characterized by uncertain paternity relative to a model based on the use of the average numerator relationship matrix [11]. The relative performance between the two models might be expected to increase with h^2 since the power of discriminating between candidate sires should intuitively increase. We previously have shown that when $h^2 = 0.10$, there was no significant difference between prior and posterior probabilities of sire assignments [3]. However, the lower the h^2, the greater the importance of data on uncertain progeny in the prediction of a sire's genetic merit [17]. The difference between the two models, nevertheless, does not then necessarily increase with higher heritabilities as the importance of pedigree information relative to phenotypic information decreases with respect to the prediction of genetic merit. Our work then suggests that the largest differences in performance between the two models exist for traits with medium h^2. Nonetheless, due to similarity in terms of rank correlation, and especially in the absence of prior information from *e.g.* genetic markers, the ANRM model may be preferable for genetic evaluation of large populations given the potential savings in computational time.

In the presence of prior information on sire assignments, the hierarchical model presented in this study represents an important alternative for genetic prediction. That is, in addition to the incorporation of prior probabilities on sire assignments, as also possible with ANRM, the HIER model allows for the integration of the uncertainty about these prior probabilities in the prediction of genetic merit. Genetic markers, for example, represent an important objective source of prior information. Moreover, the HIER model represents a general framework which could be extended to model the quality of genetic marker information contributing to sire assignment [15].

The use of multiple-sire mating is common in large beef cattle populations raised on pastoral conditions. Currently, about 25–30% of the calves evaluated by the beef cattle improvement programs in Brazil derive from multiple-sire mating. Multiple sire matings are used to improve pregnancy rates, since the average size of breeding groups, as a function of paddock size, is too large to be sired by a single bull. Other examples of uncertain parentage include the use of AI followed by NS, accidental or unplanned breedings, and AI with pooled semen as is common in swine production. Multiple-sire matings are also commonly found in some sheep production systems.

The impact of modeling uncertain paternity, either through ANRM or HIER, is expected to be particularly important for large herds. These herds provide sizable gene pools for selection, thereby offering great potential for genetic improvement programs; however, the exclusive use of single matings is costly and generally impractical on these operations due to their size and labor commitments. Genetic evaluation systems that model uncertain paternity will aid genetic improvement of economically important traits in large populations raised in pastoral conditions and undergoing multiple-sire mating.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from the National Beef Cattle Evaluation Consortium (USA) and from the Michigan Agricultural Experiment Station (Project MICL 01822). F.F. Cardoso further acknowledges financial support from CAPES – Brasilia/Brazil. We are grateful to Ignacy Misztal for making available Sparsem90, Fspak90 and Gibbsf90.

REFERENCES

[1] Bink M., Quaas R.L., van Arendonk J.A.M., Bayesian estimation of dispersion parameters with a reduced animal model including polygenic and QTL effects, Genet. Sel. Evol. 30 (1998) 103–125.

[2] Cantet R.J.C., Gianola D., Misztal I., Fernando R.L., Estimates of dispersion parameters and of genetic and environmental trends for weaning weight in Angus cattle using a maternal animal-model with genetic grouping, Livest. Prod. Sci. 34 (1993) 203–212.

[3] Cardoso F.F., Tempelman R.J., Bayesian inference on uncertain paternity for prediction of genetic merit, J. Anim. Sci. 79 Suppl. 1 (2001) 111.

[4] Chib S., Carlin B.P., On MCMC sampling in hierarchical longitudinal models, Stat. Comput. 9 (1999) 17–26.

[5] DeNise S., Using parentage analysis in commercial beef operations, in: Proceedings of the Beef Improvement Federation, 31st Annual Research Symposium and Annual Meeting, June 1999, Roanoke, Virginia, pp. 183–190.

[6] Famula T.R., Simple and rapid inversion of additive relationship matrices incorporating parental uncertainty, J. Anim. Sci. 70 (1992) 1045–1048.

[7] Foulley J.L., Gianola D., Planchenault D., Sire evaluation with uncertain paternity, Génét. Sél. Évol. 19 (1987) 83–102.

[8] Foulley J.L., Thompson R., Gianola D., On sire evaluation with uncertain paternity, Genet. Sel. Evol. 22 (1990) 373–376.

[9] Gelfand A.E., Model determination using sampling-based methods, in: Gilks W.R., Richardson S., Spiegelhalter D.J. (Eds.), Marcov chain Monte Carlo in practice, 1st edn., Chapman & Hall, London, 1996, pp. 145–161.

[10] Henderson C.R., A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values, Biometrics 32 (1976) 69–83.

[11] Henderson C.R., Use of an average numerator relationship matrix for multiple-sire joining, J. Anim. Sci. 66 (1988) 1614–1621.

[12] Perez-Enciso M., Fernando R.L., Genetic evaluation with uncertain parentage – a comparison of methods, Theor. Appl. Genet. 84 (1992) 173–179.

[13] Quaas R.L., Additive genetic model with groups and relationships, J. Dairy Sci. 71 (1988) 1338–1345.

[14] Quaas R.L., Pollak E.J., Mixed model methodology for farm and ranch beef cattle testing programs, J. Anim. Sci. 51 (1980) 1277–1287.

[15] Rosa G.J.M., Yandell B.S., Gianola D., A Bayesian approach for constructing genetic maps when markers are miscoded, Genet. Sel. Evol. 34 (2002) 353–369.

[16] Spiegelhalter D.J., Best N.G., Carlin B.P., van derLinde A., Bayesian measures of model complexity and fit, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2002) 583–616.

[17] Sullivan P.G., Alternatives for genetic evaluation with uncertain parentage, Can. J. Anim. Sci. 75 (1995) 31–36.

[18] Wang C.S., Rutledge J.J., Gianola D., Bayesian-analysis of mixed linear-models *via* Gibbs sampling with an application to litter size in Iberian pigs, Genet. Sel. Evol. 26 (1994) 91–115.

[19] Westell R.A., Quaas R.L., Van Vleck L.D., Genetic groups in an animal-model, J. Dairy Sci. 71 (1988) 1310–1318.

APPENDIX

Specification of fully conditional distributions

Let $\boldsymbol{\theta} = \left[\boldsymbol{\beta}', \mathbf{a}'_p, \mathbf{m}'_p\right]'$; $\mathbf{W}_t^{(\mathbf{k})} = \left[\mathbf{X}\ \mathbf{Z}_1|_{\mathbf{s}_t^*=\mathbf{s}_t^{(\mathbf{k})}}\ \mathbf{Z}_2\right]$, with $\mathbf{Z}_1|_{\mathbf{s}_t^*=\mathbf{s}_t^{(\mathbf{k})}}$ indicating the dependency of this design matrix on sire assignments $\mathbf{s}_t^* = \mathbf{s}_t^{(\mathbf{k})}$ for non-parent animals; and $\left(\boldsymbol{\Sigma}_p^{(\mathbf{k})}\right)^- = \begin{bmatrix} \mathbf{V}_{\boldsymbol{\beta\beta}}^{-1} & \mathbf{0}_{p\times 2q_p} \\ \mathbf{0}_{2q_p\times p} & \mathbf{G}^{-1} \otimes \mathbf{A}_{pp}^{-1}|_{\mathbf{s}_p^*=\mathbf{s}_p^{(\mathbf{k})}} \end{bmatrix}$, with $\mathbf{A}_{pp}^{-1}|_{\mathbf{s}_p^*=\mathbf{s}_p^{(\mathbf{k})}}$ indicating the dependence of parental relationships on sire assignments $\mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})}$ for parent animals, and $\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}$ being a $p \times p$ diagonal matrix consistent with a $N\left(\boldsymbol{\beta}_o, \mathbf{V}_{\boldsymbol{\beta\beta}}\right)$ prior assignment on $\boldsymbol{\beta}$. If $\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1} = \mathbf{0}_{p\times p}$, then $p(\boldsymbol{\beta}) \propto 1$. We, however, adopted a proper bounded uniform prior on $\boldsymbol{\beta}$, which is equivalent to specifying $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} = \mathbf{0}_{p\times p}$ but with values of $\boldsymbol{\beta}$ constrained to be within the specified bounds. Then, it can be readily shown using results from Wang *et al.* [18] that the FCD of θ is multivariate normal, that is,

$$\boldsymbol{\theta}|\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}, \mathbf{G}, \sigma_e^2, \mathbf{y} \sim N(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{C}^{(\mathbf{k})}) \tag{A.1}$$

where

$$\hat{\boldsymbol{\theta}}^{(\mathbf{k})} = \mathbf{C}^{(\mathbf{k})}\left(\mathbf{W}_t^{(\mathbf{k})'}\left(\mathbf{R}_t^{(\mathbf{k})}\right)^{-1}\mathbf{y} + \begin{bmatrix}\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\boldsymbol{\beta}_\mathbf{o}\\ \mathbf{0}_{2q_px1}\end{bmatrix}\right)$$

for $\mathbf{C}^{(\mathbf{k})} = \left(\mathbf{W}_t^{(\mathbf{k})'}\left(\mathbf{R}_t^{(\mathbf{k})}\right)^{-1}\mathbf{W}_t^{(\mathbf{k})} + \left(\boldsymbol{\Sigma}_p^{(\mathbf{k})}\right)^{-}\right)^{-1}$, with $\mathbf{R}_t^{(\mathbf{k})} = \mathbf{R}|_{\mathbf{s}_t^*=\mathbf{s}_t^{(\mathbf{k})}}$ indicating the dependency of $\mathbf{R}$ on sire assignments $\mathbf{s}_t^* = \mathbf{s}_t^{(\mathbf{k})}$ on non-parents.

The FCD of sire assignments in $\mathbf{s}^*$ are considered separately for parents and non-parent animals. For parent animals, the FCD of the sire assignment on animal j is:

$$\text{Prob}\left(s_j^* = s_j^{(k)}|\boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}_{-j}^* = \mathbf{s}_{-j}^{(\mathbf{k})}, \boldsymbol{\pi}, \mathbf{G}, \sigma_e^2, \mathbf{y}\right)$$
$$= \frac{\pi_j^{(k)}\left(\omega_j^{(k)}\right)^{-1}\exp\left(-0.5\left(\omega_j^{(k)}\right)^{-1}\left(\left(\gamma_j^{(k)}\right)^2 g^{11} + \left(\delta_j^{(k)}\right)^2 g^{22} + 2\left(\gamma_j^{(k)}\right)\left(\delta_j^{(k)}\right)g^{12}\right)\right)}{\sum_{k=1}^{v_j}\pi_j^{(k)}\left(\omega_j^{(k)}\right)^{-1}\exp\left(-0.5\left(\omega_j^{(k)}\right)^{-1}\left(\left(\gamma_j^{(k)}\right)^2 g^{11} + \left(\delta_j^{(k)}\right)^2 g^{22} + 2\left(\gamma_j^{(k)}\right)\left(\delta_j^{(k)}\right)g^{12}\right)\right)},$$
$$j = q_b + 1, q_b + 2, \ldots, q, \tag{A.2}$$

where $\mathbf{s}_{-j}^* = \mathbf{s}_{-j}^{(\mathbf{k})}$ is used to denote the conditioning on sire assignments for all animals other than j. For *non-parent* animals, the FCD of the sire assignment on animal j is:

$$\text{Prob}\left(s_j^* = s_j^{(k)}|\boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}_{-j}^* = \mathbf{s}_{-j}^{(\mathbf{k})}, \boldsymbol{\pi}, \mathbf{G}, \sigma_e^2, \mathbf{y}\right)$$
$$= \frac{\pi_j^{(k)}\left(\sigma_e^2 + \omega_j^{(k)}\sigma_a^2\right)^{-1/2}\exp\left(-0.5\left(\sigma_e^2 + \omega_j^{(k)}\sigma_a^2\right)^{-1}\left(e_{ij}^{(k)}\right)^2\right)}{\sum_{k=1}^{v_j}\pi_j^{(k)}\left(\sigma_e^2 + \omega_j^{(k)}\sigma_a^2\right)^{-1/2}\exp\left(-0.5\left(\sigma_e^2 + \omega_j^{(k)}\sigma_a^2\right)^{-1}\left(e_{ij}^{(k)}\right)^2\right)}, \tag{A.3}$$

where $e_{ij}^{(k)} = y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta} - 0.5a_{s_j^{(k)}} - 0.5a_{d_j^*} - \mathbf{z}'_{2ij}\mathbf{m}_p$ and $j = q_p + 1, q_p + 2, \ldots, q$. Therefore, MCMC inference on sire assignments require random draws from generalized Bernoulli (*i.e.* single trial multinomial) distributions.

The FCD's for the probabilities of sire assignments are given by:

$$p\left(\boldsymbol{\pi}_j|\boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}, \mathbf{G}, \sigma_e^2, \mathbf{y}\right) \propto \prod_{k=1}^{v_j}\left(\pi_j^{(k)}\right)^{\alpha_j^{(k)} + I_j^{(k)} - 1}, \tag{A.4}$$

which corresponds to a series of Dirichlet distributions for $j = q_b + 1, q_b + 2, \ldots, q$.

The FCD's of each of σ_e^2 and $\mathbf{G}$ using the RAM specification do not have recognizable forms. Bink *et al.* [1] suggested univariate Metropolis-Hastings sampling updates for various functions of variance components in their RAM-based specification. We alternatively base our MCMC algorithm on the method of composition using specifically Algorithm 2 of Chib and Carlin [4] except that their data distribution is fully marginalized over the random effects whereas the RAM specification in (1) is only marginalized over the non-parent genetic effects. The joint posterior density of all parameters in a full animal model can be obtained from the reduced animal model as follows:

$$\begin{aligned} &p\left(\boldsymbol{\beta}, \mathbf{a}, \mathbf{m}|\mathbf{s}^* = \mathbf{s^{(k)}}, \mathbf{G}, \sigma_e^2, \mathbf{y}\right) \\ &\quad = p\left(\boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p|\mathbf{s}^* = \mathbf{s^{(k)}}, \mathbf{G}, \sigma_e^2, \mathbf{y}\right) p\left(\boldsymbol{\gamma}_t, \boldsymbol{\delta}_t|\mathbf{a}_p, \mathbf{m}_p, \mathbf{G}, \mathbf{s}^* = \mathbf{s^{(k)}}\right). \end{aligned} \quad \text{(A.5)}$$

That is, a random draw from (A.5) is equivalent to a random draw from (A.1) followed by a random draw from $p\left(\boldsymbol{\gamma}_t, \boldsymbol{\delta}_t|\mathbf{a}_p, \mathbf{m}_p, \mathbf{G}, \mathbf{s}^* = \mathbf{s^{(k)}}\right)$ that can be readily derived as a sequence of univariate draws from the additive $\gamma_j^{(k)}$ and maternal $\delta_j^{(k)}$ Mendelian sampling terms. Specifically, this involves sampling first from

$$\begin{aligned} &\gamma_j^{(k)}|\boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}^* = \mathbf{s^{(k)}}, \mathbf{G}, \sigma_e^2, \mathbf{y} \\ &\qquad \sim NID\left(\left(\frac{1}{\sigma_e^2} + \frac{\left(\omega_j^{(k)}\right)^{-1}}{\sigma_a^2}\right)^{-1} \frac{e_{ij}^{(k)}}{\sigma_e^2}, \left(\frac{1}{\sigma_e^2} + \frac{\left(\omega_j^{(k)}\right)^{-1}}{\sigma_a^2}\right)^{-1}\right) \\ &\qquad\qquad j = q_p + 1, q_p + 2, \ldots, q, \end{aligned} \quad \text{(A.6)}$$

followed by

$$\begin{aligned} &\delta_j^{(k)}|\boldsymbol{\gamma}_t, \boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}^* = \mathbf{s^{(k)}}, \mathbf{G}, \sigma_e^2, \mathbf{y} \sim NID\left(-\frac{g^{12}}{g^{22}}\gamma_j^{(k)}, \frac{\omega_j^{(k)}}{g^{22}}\right) \\ &\qquad\qquad j = q_p + 1, q_p + 2, \ldots, q. \end{aligned} \quad \text{(A.7)}$$

Let $p(\mathbf{G})$ be a conjugate inverted Wishart prior density with parameters v_g and $\mathbf{G}_\text{o}$ such that $\mathrm{E}\left(\mathbf{G}|v_g, \mathbf{G}_\text{o}\right) = \frac{1}{v_g - 3}\mathbf{G}_\text{o}^{-1}$. The FCD of $\mathbf{G}$ given the augmentation of the RAM joint posterior density in (8) with γ_t and δ_t is:

$$\begin{aligned} &p\left(\mathbf{G}|\boldsymbol{\beta}, \mathbf{a}, \mathbf{m}, \mathbf{s}^* = \mathbf{s^{(k)}}, \sigma_e^2, \mathbf{y}\right) \\ &\qquad \propto |\mathbf{G}|^{-\frac{q+v_g+3}{2}} \exp\left(-0.5 \text{ trace } \left(\mathbf{G^{-1}}\left(\mathbf{S_G} + \mathbf{G_o^{-1}}\right)\right)\right), \end{aligned} \quad \text{(A.8)}$$

where

$$\mathbf{S_G}=\begin{bmatrix}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} & \mathbf{a}'\mathbf{A}^{-1}\mathbf{m}\\ \mathbf{m}'\mathbf{A}^{-1}\mathbf{a} & \mathbf{m}'\mathbf{A}^{-1}\mathbf{m}\end{bmatrix}.$$

These components of $\mathbf{S_G}$ can be readily computed without explicitly determining $\mathbf{a}_t$ and or $\mathbf{m}_t$. For example, using results from Quaas [13] and those in this paper, $\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}=\mathbf{a}'_p\mathbf{A}_{pp}^{-1}\mathbf{a}_p+\boldsymbol{\gamma}'_t\boldsymbol{\Omega}_{tt}^{-1}\boldsymbol{\gamma}_t$, where $\boldsymbol{\gamma}'_t\boldsymbol{\Omega}_{tt}^{-1}\boldsymbol{\gamma}_t=\sum_{j=q_p+1}^{q}\frac{\gamma_j^2}{\omega_j}$.

Finally, let $p\left(\sigma_e^2\right)$ be an inverted-gamma density with parameters α_e and β_e. Then the FCD of σ_e^2 is also inverted-gamma and given by:

$$p\left(\sigma_e^2|\boldsymbol{\beta},\mathbf{a},\mathbf{m},\mathbf{s}=\mathbf{s}^{(k)},\mathbf{G},\mathbf{y}\right)\propto\left(\sigma_e^2\right)^{-(n/2+\alpha_e-1)}\exp\left(-\frac{1}{\sigma_\mathbf{e}^2}\left(\frac{\mathbf{e}'\mathbf{e}}{2}+\beta_e\right)\right). \tag{A.9}$$

The first n_p elements of $\mathbf{e}$ are $\mathbf{e}_p=\left\{e_{ij}\right\}_{j=1}^{q_p}$ which are residuals due to records on the parents. The last n_t elements of $\mathbf{e}$ are $\mathbf{e}_t^{(\mathbf{k})}=\left\{e_{ij}^{(k)}-\gamma_j^{(k)}\right\}_{j=q_p+1}^{q}$ with $\mathbf{e}_t^{(\mathbf{k})}=\mathbf{e}_t|_{\mathbf{s}_t^*=\mathbf{s}_t^{(\mathbf{k})}}$ indicating the dependence of $\mathbf{e}_t$ on sire assignments $\mathbf{s}_t^*=\mathbf{s}_t^{(\mathbf{k})}$ on non-parent animals.

The MCMC sampling scheme can thus be summarized as follows:

(1) Draw samples of $\boldsymbol{\beta}$, $\mathbf{a}_p$, and $\mathbf{m}_p$ from (A.1) using the proposition from the appendix of Wang *et al.* [18].
(2) Draw samples of γ_t and δ_t from (A.6) and (A.7).
(3) Compute $\mathbf{S_G}$ using the samples of $\mathbf{a}_p$, $\mathbf{m}_p$, γ_t, and δ_t in order to sample $\mathbf{G}$ from a scaled inverted Wishart distribution (A.8).
(4) Determine $\mathbf{e}_t^{(\mathbf{k})}$ and combine with $\mathbf{e}_p$ to sample σ_e^2 from an inverted-gamma distribution (A.9).
(5) For each animal j with uncertain paternity, independently draw a sire s_j^* using as the probability of assignment either (A.2) if the animal is a parent or (A.3) if the animal is a non-parent.
(6) For each animal j with uncertain paternity, independently draw $\boldsymbol{\pi}_j$ from a Dirichlet distribution (A.4).