Genetics Selection Evolution

RESEARCH Open Access

Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation

Ole F Christensen

Abstract

Background: Single-step methods provide a coherent and conceptually simple approach to incorporate genomic information into genetic evaluations. An issue with single-step methods is compatibility between the marker-based relationship matrix for genotyped animals and the pedigree-based relationship matrix. Therefore, it is necessary to adjust the marker-based relationship matrix to the pedigree-based relationship matrix. Moreover, with data from routine evaluations, this adjustment should in principle be based on both observed marker genotypes and observed phenotypes, but until now this has been overlooked. In this paper, I propose a new method to address this issue by 1) adjusting the pedigree-based relationship matrix to be compatible with the marker-based relationship matrix instead of the reverse and 2) extending the single-step genetic evaluation using a joint likelihood of observed phenotypes and observed marker genotypes. The performance of this method is then evaluated using two simulated datasets.

Results: The method derived here is a single-step method in which the marker-based relationship matrix is constructed assuming all allele frequencies equal to 0.5 and the pedigree-based relationship matrix is constructed using the unusual assumption that animals in the base population are related and inbred with a relationship coefficient γ and an inbreeding coefficient $\gamma/2$. Taken together, this γ parameter and a parameter that scales the marker-based relationship matrix can handle the issue of compatibility between marker-based and pedigree-based relationship matrices. The full log-likelihood function used for parameter inference contains two terms. The first term is the REML-log-likelihood for the phenotypes conditional on the observed marker genotypes, whereas the second term is the log-likelihood for the observed marker genotypes. Analyses of the two simulated datasets with this new method showed that 1) the parameters involved in adjusting marker-based and pedigree-based relationship matrices can depend on both observed phenotypes and observed marker genotypes and 2) a strong association between these two parameters exists. Finally, this method performed at least as well as a method based on adjusting the marker-based relationship matrix.

Conclusions: Using the full log-likelihood and adjusting the pedigree-based relationship matrix to be compatible with the marker-based relationship matrix provides a new and interesting approach to handle the issue of compatibility between the two matrices in single-step genetic evaluation.

Introduction

Single-step methods for genetic evaluation [1-3] have recently become popular because they provide an approach to incorporate genomic information into genetic evaluations that is both coherent and conceptually simple. A single-step method extends the usual pedigree-based method by replacing the additive relationship matrix constructed from pedigree by an additive relationship matrix that combines the marker-based relationship matrix for genotyped animals with the pedigree-based relationship matrix.

An issue with a single-step method is compatibility between the marker-based relationship matrix for genotyped animals and the pedigree-based relationship matrix [4-6]. To handle this problem, it is necessary to determine

Correspondence: OleF.Christensen@agrsci.dk
Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Blichers Allé 20, P.O. BOX 50, DK-8830 Tjele, Denmark

which allele frequencies should be used in the marker-based relationship matrix and to adjust this matrix to the pedigree-based relationship matrix. In theory, one should use the allele frequencies in the founder population of the pedigree (base animals) for the marker-based and pedigree-based relationships to be compatible, but these allele frequencies are rarely available in practice since base animals are not genotyped. Chen et al. [5] and Forni et al. [7] concluded that using the observed allele frequencies improved accuracy of prediction compared to using allele frequencies equal to 0.5. Studies on how to adjust the marker-based relationship matrix to be compatible with the submatrix of the pedigree-based relationship matrix for genotyped animals have been reported [4-6]. These adjustments consisted of scaling and adding a number to all elements in the marker-based relationship matrix based on equating means of diagonal and off-diagonal elements in the two matrices, and it was demonstrated that accuracy of prediction increased and bias decreased. In relation to this problem of compatibility between marker-based and pedigree-based relationship matrices, with data from routine evaluations, selection affects the allele frequencies over time, and in principle both observed marker genotypes and observed phenotypes contain information about allele frequencies in the base population. Therefore, studies on the adjustment of the marker-based relationship matrix to the pedigree-based relationship matrix have overlooked the fact that it should in principle incorporate information on observed phenotypes.

This work explores two possibilities to solve the problem, i.e. 1) in which the pedigree-based relationship matrix is adjusted to the marker-based relationship matrix and 2) in which the single-step genetic evaluation is extended by using a joint likelihood of observed phenotypes and observed marker genotypes. This results in a single-step method in which the marker-based relationship matrix is constructed assuming all allele frequencies are equal to 0.5 and the pedigree-based relationship matrix is constructed using the unusual assumption that animals in the base population are related and inbreed with a relationship coefficient γ and an inbreeding coefficient $\gamma/2$. The log-likelihood function used for parameter inference contains two terms. The first term is the REML-log-likelihood for the phenotypes conditional on the observed marker genotypes and the second term is the log-likelihood for the observed marker genotypes. The performance of the proposed method was evaluated using two simulated datasets.

Methods

This section presents, first, the statistical model on which the single-step methods are based along the lines of Christensen and Lund [3], then the proposal on how to adjust the pedigree-based relationship matrix and finally the adjustment of the marker-based relationship matrix as previously used.

Single-step model

The marker genotypes are summarised into a marker matrix $\mathbf{m}$, where $m_{ij} = -1, 0$ or 1 if SNP j of individual i is 11, 12, or 22, respectively. In the following, capital $\mathbf{M}$ and lowercase $\mathbf{m}$ indicate whether marker genotypes are considered as random variables or as non-random variables (observed variables or integration variables), respectively.

Let us consider the simple model

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{Za} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypes, μ is the general mean, $\mathbf{1}$ is a vector of ones, $\mathbf{a}$ is the vector of breeding values, $\mathbf{Z}$ is an incidence matrix, and $\mathbf{e}$ is the vector of residual errors. The breeding value may be decomposed into $\mathbf{a} = \mathbf{g} + \mathbf{a}_r$, where $\mathbf{g}$ is the vector of genomic effects and $\mathbf{a}_r = (\mathbf{a} - \mathbf{g})$ is the vector of residual polygenic effects. The residual polygenic effects $\mathbf{a}_r \sim N(\mathbf{0}, \sigma_a^2 \omega \mathbf{A})$, where matrix $\mathbf{A}$ is the pedigree-based additive relationship matrix and $\omega \in [0;1]$ is the relative weight on the residual polygenic effect. The genomic values $[\,\mathbf{g} \mid \mathbf{M}] \sim N(\mathbf{0}, \sigma_a^2(1-\omega)\mathbf{G}(\mathbf{M}))$, with $\mathbf{G}(\mathbf{M}) = (\mathbf{M} - (2\boldsymbol{\rho} - 1)\mathbf{1}^{\mathrm{T}})(\mathbf{M} - (2\boldsymbol{\rho} - 1)\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}/s$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$ with ρ_j being the allele frequency for the jth marker and $s = s(\mathbf{v}) = \sum_j v_j$ with v_j being defined below. The model for the marker genotypes is that $\mathbf{M}$ is a multivariate Gaussian distribution with

$$\mathrm{E}[M_{ij} \mid \rho_j] = 2\rho_j - 1, \quad \mathrm{Cov}[M_{ij}, M_{i'j} \mid v_j] = v_j A_{ii'},$$

$$\mathrm{Cov}[M_{ij}, M_{i'j'} \mid v_{jj'}] = v_{jj'} A_{ii'},$$

where $v_j = v_{j,j}$ is a parameter and $v_{jj'} = 0$ for $j \neq j'$. The crude assumption that the multivariate distribution is Gaussian is crucial in the derivation, whereas the unrealistic assumption that $v_{jj'} = 0$ for $j \neq j'$ is made for simplicity. Dividing marker genotypes $\mathbf{m}$ into observed marker genotypes $\mathbf{m}^o$ and un-observed marker genotypes $\mathbf{m}^u$, the joint marginal density of observed phenotypes $\mathbf{y}$ and observed marker genotypes $\mathbf{m}^o$ is

$$\begin{aligned} f(\mathbf{y}, \mathbf{m}^o) &= \int f(\mathbf{y}, \mathbf{m}^o, \mathbf{m}^u) d\mathbf{m}^u \\ &= \int f(\mathbf{y} \mid \mathbf{m}^o, \mathbf{m}^u) f(\mathbf{m}^u \mid \mathbf{m}^o) f(\mathbf{m}^o) d\mathbf{m}^u, \end{aligned}$$

where

$$\begin{aligned} &f(\mathbf{y} \mid \mathbf{m}^o, \mathbf{m}^u) \\ &= \int\int\int f(\mathbf{y} \mid \mathbf{a}_r, \mathbf{g}, \mu) f(\mathbf{a}_r) f(\mathbf{g} \mid \mathbf{m}^o, \mathbf{m}^u) d\mathbf{a}_r d\mathbf{g} d\mu. \end{aligned}$$

By rearranging terms and using that $\int f(\mathbf{g} \mid \mathbf{m}^o, \mathbf{m}^u) f(\mathbf{m}^u \mid \mathbf{m}^o) d\mathbf{m}^u = f(\mathbf{g} \mid \mathbf{m}^o)$, this becomes

$$\begin{aligned} & f(\mathbf{y}, \mathbf{m}^o) \\ & = \int \int \int f(\mathbf{y} \mid \mathbf{a}_r, \mathbf{g}, \mu) f(\mathbf{a}_r) f(\mathbf{g} \mid \mathbf{m}^o) d\mathbf{a}_r d\mathbf{g} d\mu \\ & \quad \times f(\mathbf{m}^o) \\ & = f(\mathbf{y} \mid \mathbf{m}^o) f(\mathbf{m}^o), \end{aligned}$$

where $f(\mathbf{y} \mid \mathbf{m}^o)$ is defined implicitly. The first term $f(\mathbf{y} \mid \mathbf{m}^o)$ is the density for the phenotypes given the observed marker genotypes, whereas the second term $f(\mathbf{m}^o)$ is the multivariate Gaussian density for the observed marker genotypes.

Single-step methods are based on the density $f(\mathbf{y} \mid \mathbf{m}^o)$, which may be written (see [3]) as

$$f(\mathbf{y} \mid \mathbf{m}^o) = \int \int f(\mathbf{y} \mid \mathbf{a}, \mu) f(\mathbf{a} \mid \mathbf{m}^o) d\mathbf{a} d\mu,$$

where the vector $\mathbf{a}$ has a mean zero and a variance-covariance matrix $\mathbf{H}_\omega$ with inverse

$$\mathbf{H}_\omega^{-1} = \begin{bmatrix} \mathbf{G}_\omega^{-1} - \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \mathbf{A}^{-1}, \tag{1}$$

with $\mathbf{G}_\omega = (1 - \omega)\mathbf{G}(\mathbf{m}^o) + \omega \mathbf{A}_{11}$. In the above formula, the sub-division is made according to genotyped and non-genotyped animals. The matrices $\mathbf{G}(\mathbf{m}^o) = (\mathbf{m}^o - (2\boldsymbol{\rho} - 1)\mathbf{1}^{\mathrm{T}})(\mathbf{m}^o - (2\boldsymbol{\rho} - 1)\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}/s$ and $\mathbf{A}_{11}$ are the marker-based relationship matrix for the genotyped animals and the submatrix of the pedigree-based relationship matrix corresponding to genotyped animals, respectively. The sparse structure of the matrix in equation (1) is the corner-stone for efficient computing using a single-step method. Assuming that $[\mathbf{a} \mid \mathbf{m}^o]$ is Gaussian distributed, mixed model equations can be solved for BLUP predictions and AI-REML [8] provides REML parameter estimates.

Issues raised by the single-step method are 1) how are the allele frequencies $\boldsymbol{\rho}$ that are used in $\mathbf{G}(\mathbf{m}^o)$ obtained and 2) how is the required compatibility between $\mathbf{G}(\mathbf{m}^o)$ and $\mathbf{A}_{11}$ that is evident in equation (1) reached. To investigate these issues, the joint density of observed phenotypes and marker genotypes, $f(\mathbf{y}, \mathbf{m}^o) = f(\mathbf{y} \mid \mathbf{m}^o) f(\mathbf{m}^o)$, is taken as the starting point. From this, the full marginal log-likelihood becomes

$$\begin{aligned} & \ell(\sigma_a^2, \sigma_e^2, \omega, \boldsymbol{\rho}, \mathbf{v}) \\ & = \ell_{\mathbf{y}|\mathbf{m}^o}(\sigma_a^2, \sigma_e^2, \omega, \boldsymbol{\rho}, s(\mathbf{v})) + \ell_{mark}(\boldsymbol{\rho}, \mathbf{v}), \end{aligned} \tag{2}$$

where $\ell_{\mathbf{y}|\mathbf{m}^o}(\sigma_a^2, \sigma_e^2, \omega, \boldsymbol{\rho}, s(\mathbf{v}))$ for fixed ω, $\boldsymbol{\rho}$ and $s(\mathbf{v}) = \sum_j v_j$ is the single-step REML-log-likelihood for the phenotypes conditional on the observed marker genotypes, and

$$\begin{aligned} \ell_{mark}(\boldsymbol{\rho}, \mathbf{v}) = {} & \text{const} - (n_1/2) \sum_j \log(v_j) \\ & - \sum_{ii'j} \frac{(m_{ij}^o - (2\rho_j - 1))(\mathbf{A}_{11}^{-1})_{ii'}(m_{i'j}^o - (2\rho_j - 1))}{2v_j}, \end{aligned}$$

with n_1 denoting the number of genotyped animals, is the log-likelihood of the observed marker genotypes. Thus, the parameter $\boldsymbol{\rho}$ enters into both terms of the full log-likelihood in equation (2), and this implies that estimation of allele frequencies $\boldsymbol{\rho}$ should be in principle based on this log-likelihood. In particular, if selection has been performed the phenotypes will contain information about the allele frequencies which may cause bias if ignored. However, estimating $\boldsymbol{\rho}$ and $s = s(\mathbf{v})$ by maximising $\ell(\sigma_a^2, \omega, \sigma_e^2, \boldsymbol{\rho}, \mathbf{v})$ for all parameters jointly is not feasible in practice since $\boldsymbol{\rho}$ is a very high-dimensional parameter. Instead, in general the observed marker genotypes are used to estimate $\boldsymbol{\rho}$ and then they are plugged into $\ell_{\mathbf{y}|\mathbf{m}^o}$, and this log-likelihood is used to estimate the remaining parameters. Estimation of $\boldsymbol{\rho}$ based on observed marker genotypes may consist of simply using the observed allele frequencies [5,7], or may be done by maximising $\ell_{mark}(\boldsymbol{\rho}, \mathbf{v})$ as a function of $\boldsymbol{\rho}$ (this is essentially the method of Gengler et al. [9], although, for computational reasons, that method adds a small residual error to the distribution of $\mathbf{M}$). The high-dimensional parameter $\mathbf{v}$ also enters into both terms of the full log-likelihood, although it only enters into $\ell_{\mathbf{y}|\mathbf{m}^o}$ via the scaling parameter $s = s(\mathbf{v})$, and therefore estimation of $\mathbf{v}$ should also be in principle based on maximising the full log-likelihood. Usually, estimating $s = \sum_j v_j$ is based on the observed marker genotypes, either using $v_j = 2\hat{\rho}_j(1 - \hat{\rho}_j)$, implying $s = \sum_j 2\hat{\rho}_j(1 - \hat{\rho}_j)$ as in [2,3], or using s such that the average diagonal of $\mathbf{G}$ equals the average diagonal of $\mathbf{A}_{11}$ as in [7]. Furthermore, it has been demonstrated that the accuracy of prediction is improved and bias is reduced by adjusting the marker-based relationship matrix as $\mathbf{G}(\mathbf{m}^o)\beta + \alpha$ where α and β are based on the elements in $\mathbf{G}(\mathbf{m}^o)$ and $\mathbf{A}_{11}$; further details are given below. It should be noted that such an adjustment is based on observed marker genotypes only, and lacks a theoretical justification within the framework considered here. To summarise, the allele frequencies in $\mathbf{G}(\mathbf{m}^o)$ and the adjustments necessary for the compatibility of $\mathbf{G}(\mathbf{m}^o)$ and $\mathbf{A}_{11}$ should be in principle derived based on both observed marker genotypes and observed phenotypes.

Furthermore, when computing BLUP breeding values $\hat{\mathbf{a}}$ with plugged-in parameter estimates, the uncertainty in these parameters estimates is ignored, which is an important issue in the case of the high-dimensional parameter $\boldsymbol{\rho}$. Alternatively, uncertainty in parameter estimates may be incorporated into the predictions by using a Bayesian

Table 1 Example pedigree

id	sire	dam
1	0	0
2	0	0
3	1	2
4	1	3
5	1	2
6	5	4

approach. Demonstration that a Bayesian approach with prior distributions on $\boldsymbol{\rho}$ and $\mathbf{v}$ results in an method in which the pedigree-based relationship matrix is adjusted to a marker-based relationship matrix, is presented below.

Adjusting the pedigree-based relationship matrix

The derivation of the proposed adjustment of the pedigree-based relationship matrix is based on assigning priors on the high-dimensional parameters $\boldsymbol{\rho}$ and $\mathbf{v}$ in the previously described single-step model, and then considers the first and second order moments of the marginal distribution of $\mathbf{M}$ (integrating $\boldsymbol{\rho}$ and $\mathbf{v}$). Appendix A shows that the resulting marker distribution satisfies

$$\mathrm{E}[M_{ij}] = 0,$$

$$\mathrm{Cov}[M_{ij}, M_{i'j'}] = \tilde{v}_{jj'}\tilde{A}(\gamma)_{ii'},$$

with $\tilde{v}_j = \tilde{v}_{jj} = \tilde{s}/p$, $j = 1, \ldots, p$, $\tilde{s}$ being a parameter, and $\tilde{v}_{jj'} = 0$ when $j \neq j'$. Matrix $\tilde{\mathbf{A}}(\gamma)$ is an additive relationship matrix that satisfies the usual recursions but with the peculiar feature that base animals in the pedigree are related and inbred. Within the population of base animals, the relationship coefficient is γ and the inbreeding coefficient is $\gamma/2$. Table 1 contains data for a small pedigree used to derive matrix $\tilde{\mathbf{A}}$ as shown in Table 2. The mean and variance-covariance structure of $\mathbf{M}$ shown above is of the form in [3], with scaling parameter $\tilde{s} = \sum_j \tilde{v}_j$, and hence the breeding values $\mathbf{g}$ have a combined relationship matrix $\mathbf{H}_{\gamma,\tilde{s},\omega}$, where the inverse is

$$\mathbf{H}^{-1}_{\gamma,\tilde{s},\omega} = \begin{bmatrix} \mathbf{G}^{-1}_{\gamma,\tilde{s},\omega} - \tilde{\mathbf{A}}(\gamma)^{-1}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \tilde{\mathbf{A}}(\gamma)^{-1}, \qquad (3)$$

with $\mathbf{G}_{\gamma,\tilde{s},\omega} = (1-\omega)\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}}/\tilde{s} + \omega\tilde{\mathbf{A}}(\gamma)_{11}$. This construction is a single-step method for which the individuals in the base population are related and inbred, and the marker-based relationship matrix, $\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}}/\tilde{s}$, has allele frequencies equal to 0.5.

Based on the derivation above (see also Appendix A), the full marginal log-likelihood function for parameter inference becomes

$$\begin{aligned} &\tilde{\ell}(\sigma^2_a, \sigma^2_e, \omega, \gamma, \tilde{s}) \\ &= \tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}(\sigma^2_a, \sigma^2_e, \omega, \gamma, \tilde{s}) + \tilde{\ell}_{mark}(\gamma, \tilde{s}), \end{aligned} \qquad (4)$$

where $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}(\sigma^2_a, \sigma^2_e, \omega, \gamma, \tilde{s})$ for fixed ω, γ and $\tilde{s}$ is the single-step REML-log-likelihood for the phenotypes conditional on the observed marker genotypes, with marker-based relationship matrix $\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}}/\tilde{s}$ and pedigree-based relationship matrix $\tilde{\mathbf{A}}(\gamma)$, and

$$\begin{aligned} &\tilde{\ell}_{mark}(\gamma, \tilde{s}) = \text{const} - (pn_1/2)\log(\tilde{s}) \\ &- (p/2)\log(\det(\tilde{\mathbf{A}}(\gamma)_{11})) \\ &- \frac{p}{2\tilde{s}}\sum_{ii'}(\tilde{\mathbf{A}}(\gamma)^{-1}_{11})_{ii'}(\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}})_{ii'} \end{aligned} \qquad (5)$$

is the log-likelihood of the observed marker genotypes. Using the log-likelihood in equation (4) instead of the log-likelihood in equation (2) makes the estimation of parameters feasible in practice by numeric maximization methods since the high-dimensional parameters $\boldsymbol{\rho}$ and $\mathbf{v}$ are replaced by two parameters, a parameter γ that determines the relationship and inbreeding of individuals in the base population and a parameter $\tilde{s}$ that scales the marker-based relationship matrix.

The computations require algorithms that compute $\mathbf{A}(\gamma)^{-1}$ and $\mathbf{A}_{11}(\gamma)$ efficiently. Computations of inbreeding coefficients with related base animals have been considered by [10,11] in the context of incomplete pedigrees. In Appendix B, algorithms are presented that extend the approaches of Quaas [12] for computing $\mathbf{A}^{-1}$ and of Colleau [13] for computing $\mathbf{A}_{11}$ to the case where base animals are related and inbred.

Maximisation of the full log-likelihood in equation (4) is done by first specifying a discrete three-dimensional grid of values for parameters $\omega, \gamma, \tilde{s}$ and then, for each value

Table 2 $\tilde{\mathrm{A}}(\gamma)$ for the pedigree in Table 1

id	1	2	3	4	5	6
1	$1+\gamma/2$					
2	γ	$1+\gamma/2$				
3	$1/2+3\gamma/4$	$1/2+3\gamma/4$	$1+\gamma/2$			
4	$3/4+5\gamma/8$	$1/4+7\gamma/8$	$3/4+5\gamma/8$	$5/4+3\gamma/8$		
5	$1/2+3\gamma/4$	$1/2+3\gamma/4$	$1/2+3\gamma/4$	$1/2+3\gamma/4$	$1+\gamma/2$	
6	$5/4+3\gamma/8$	$3/8+13\gamma/32$	$5/8+11\gamma/16$	$7/8+9\gamma/16$	$3/4+3\gamma/16$	$5/4+3\gamma/8$

of $(\omega, \gamma, \tilde{s})$, computing the maximum values of the log-likelihood $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}$ and the log-likelihood of observed marker genotypes $\tilde{\ell}_{mark}$. This provides a three-dimensional profile log-likelihood $\hat{\tilde{\ell}}(\omega, \gamma, \tilde{s})$, which can then be assessed to find the maximum.

For faster computing, an alternative to using the full log-likelihood is to determine parameters γ and $\tilde{s}$ based on observed marker genotypes only, i.e. by maximising $\tilde{\ell}_{mark}(\gamma, \tilde{s})$, and to estimate the remaining parameter based on $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}(\sigma_a^2, \sigma_e^2, \omega, \gamma, \tilde{s})$ for a grid of values for ω, with estimates of γ and $\tilde{s}$ plugged in. Setting the derivative of $\tilde{\ell}_{mark}(\gamma, \tilde{s})$ with respect to $\tilde{s}$ equal to zero gives

$$0 = -\frac{pn_1}{2\tilde{s}} + \frac{p}{2\tilde{s}^2}\sum_{ii'}(\tilde{\mathbf{A}}(\gamma)_{11}^{-1})_{ii'}(\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}})_{ii'},$$

which has the solution

$$\hat{\tilde{s}}(\gamma) = \sum_{ii'}(\tilde{\mathbf{A}}(\gamma)_{11}^{-1})_{ii'}(\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}})_{ii'}/n_1.$$

Substituting $\hat{\tilde{s}}(\gamma)$ into equation (5), we obtain

$$\hat{\tilde{\ell}}_{mark}(\gamma) = \text{const} - (pn_1/2)\log(\hat{\tilde{s}}(\gamma)) - (p/2)\log(\det(\tilde{\mathbf{A}}(\gamma)_{11})),$$

which has to be maximised numerically to estimate γ.

Adjusting the marker-based relationship matrix (G-adjust)

Alternatively, an adjustment of the form $\mathbf{G}_a = \mathbf{G}\beta + \alpha$ is used, where $\mathbf{G}$ is the marker-based relationship matrix with allele frequencies $\hat{\boldsymbol{p}}$ equal to the observed ones and scaling parameter $\hat{s} = \sum_j 2\hat{p}_j(1 - \hat{p}_j)$, and parameters α and β are determined by fitting $\mathbf{G}_a$ to $\mathbf{A}_{11}$. This adjustment was used by VanRaden [14], Christensen et al. [6] and Gao et al. [15], with the first paper suggesting that α and β should be estimated by least square estimation, i.e. by minimizing the sum of squares of $\mathbf{G}\beta + \alpha - \mathbf{A}_{11}$ and the other two papers suggesting that they should be to estimated by equating means of diagonal elements and all elements in the two matrices. Here, the later is applied and α and β are estimated by solving the two equations

$$\bar{G}\beta + \alpha = \bar{A}_{11} \text{ and } dG\beta + \alpha = dA_{11}, \tag{6}$$

for α and β, where $\bar{G}$ and $\bar{A}_{11}$ denote means of all elements in the two matrices, and dG and dA denote means of diagonal elements in the two matrices.

Simulated example 1

This example is deliberately simple and very extreme, and constructed for the purpose of showing that parameter estimates of γ and $\tilde{s}$ can depend on the observed phenotypes.

The base population consists of two individuals, one sire and one dam with 15 bi-allelic markers, and their genotypes are simulated assuming independence between markers and with equal allele frequencies. Furthermore, it is assumed that the markers are independently inherited, are all QTL with allele substitution effect equal to 1, and heritability of the phenotype is equal to 1.

The two base individuals produce 100 offspring (generation 1) that all have observed phenotypes. The two individuals in generation 1 with the largest own phenotype value are selected as parents and produce 100 offspring (generation 2) that are all genotyped.

Two different approaches to estimate parameters are compared. In the first approach, all parameters are estimated using the full log-likelihood in equation (4). In the second approach, γ and $\tilde{s}$ are estimated based only on the log-likelihood of the observed marker genotypes in equation (5), and the remaining parameters are estimated based on the REML-log-likelihood $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}(\sigma_a^2, \sigma_e^2, \omega, \gamma, \tilde{s})$, with estimates of γ and $\tilde{s}$ plugged-in.

Simulated example 2

This example is inspired by a pig nucleus breeding scheme and consists of five generations in which all animals have recorded phenotypes. In each generation, 150 boars are mated to 1 500 sows to produce 15 000 offspring (50/50 males/females). For the next generation, boars with a high own phenotype value are chosen and sows are selected at random. The last three generations of selected boars are genotyped and a sixth generation of 300 candidate boars are also genotyped. The breeding value is the sum of 500 independent QTL effects simulated from a Gamma$(5.4, 0.42)$ distribution, and the heritability of the phenotype is 0.22. This dataset is described in more detail in Christensen and Lund [3].

Three different approaches are compared. In the first approach, all parameters are estimated using the full log-likelihood in equation (4). In the second approach, parameters γ and $\tilde{s}$ are estimated based only on the log-likelihood of the observed marker genotypes in equation (5), and the remaining parameters are estimated based on the REML-log-likelihood $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}(\sigma_a^2, \sigma_e^2, \omega, \gamma, \tilde{s})$, with estimates of γ and $\tilde{s}$ plugged in. Finally, the G-adjust approach is used for which α and β are estimated using equation (6) and the remaining parameters are estimated by $\ell_{\mathbf{y}|\mathbf{m}^o}$. For all three approaches, the correlation between predicted breeding value $\hat{\mathbf{a}}$ and true breeding value $\mathbf{a}$ for the candidate boars is reported as well as the estimated regression coefficient (reg) for the regression of $\mathbf{a}$ on $\hat{\mathbf{a}}$, where deviation from one indicates bias.

Results

Simulated example 1

Table 3 shows that both γ and $\tilde{s}$ were smaller when estimated with the full log-likelihood than with the log-likelihood of the observed marker genotypes.

Table 3 Parameter estimates obtained with simulated dataset 1

	$\tilde{\ell}_{\mathbf{y}\vert\mathbf{m}^o} + \tilde{\ell}_{mark}$	$\tilde{\ell}_{mark}$ & $\tilde{\ell}_{\mathbf{y}\vert\mathbf{m}^o}$
$\hat{\gamma}$	0.524	0.542
$\hat{\tilde{s}}/(p/2)$	1.068	1.081
$\hat{\omega}$	0.005	0.005
$\hat{\sigma}_a$	1.186	1.355
$\hat{\sigma}_e$	2.370	2.312

Parameters were estimated either, jointly using the full log-likelihood $\tilde{\ell} = \tilde{\ell}_{\mathbf{y}|\mathbf{m}^o} + \tilde{\ell}_{mark}$, or by first estimating γ and $\tilde{s}$ using $\tilde{\ell}_{mark}$ and then the other parameters using $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}$.

Simulated example 2

Table 4 shows that γ and $\tilde{s}$ were slightly smaller when estimated with the full log-likelihood than with the log-likelihood of the observed marker genotypes. Parameter ω was about 0.375 whether estimated with the full log-likelihood or with the log-likelihood $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}$, with parameter estimates of γ and $\tilde{s}$ from $\tilde{\ell}_{mark}$ plugged-in. When presented in three dimensions, the profile log-likelihood showed a very weak association between $(\gamma, \tilde{s})$ and ω. A contour plot of the profile log-likelihood surface for γ and $\tilde{s}$ is shown in Figure 1 and the profile log-likelihood function for ω is shown in Figure 2.

Estimated parameters obtained with the G-adjust approach are also shown in Table 4, where it should be noted that $\hat{\omega} = 0.6$.

In terms of prediction performance, the correlation between predicted and true breeding value and the regression coefficient were 0.536 and 1.17, respectively, in both cases when the pedigree-based matrix was adjusted, but 0.493 and 1.34, respectively, when the G-adjust approach was used.

Table 4 Parameter estimates and prediction performance obtained with simulated dataset 2

	$\tilde{\ell}_{\mathbf{y}\vert\mathbf{m}^o} + \tilde{\ell}_{mark}$	$\tilde{\ell}_{mark}$ & $\tilde{\ell}_{\mathbf{y}\vert\mathbf{m}^o}$	*G*-adjust
$\hat{\gamma}$	0.4605	0.4615	
$\hat{\tilde{s}}/(p/2)$	0.9996	1.0003	
α			0.0134
β			1.0074
$\hat{\omega}$	0.375	0.375	0.60
$\hat{\sigma}_a$	6.507	6.512	5.084
$\hat{\sigma}_e$	15.816	15.816	15.797
Cor$(\hat{\mathbf{a}}, \mathbf{a})$	0.536	0.536	0.493
reg	1.17	1.17	1.34

Parameters were estimated either, jointly using the full log-likelihood $\tilde{\ell} = \tilde{\ell}_{\mathbf{y}|\mathbf{m}^o} + \tilde{\ell}_{mark}$, or by first estimating γ and $\tilde{s}$ using $\tilde{\ell}_{mark}$ and then the other parameters using $\tilde{\ell}_{\mathbf{y}|\mathbf{m}^o}$; the last column contains parameter values estimated with the single-step method with an adjusted marker-based relationship matrix, in which first α and β were estimated using equation (6) and then the remaining parameters were estimated using $\ell_{\mathbf{y}|\mathbf{m}^o}$; the last two rows show the correlation between predicted and true breeding values, Cor$(\hat{\mathbf{a}}, \mathbf{a})$, for candidate boars, which is a measure of prediction performance, and the regression coefficient (reg) for the regression of $\mathbf{a}$ on $\hat{\mathbf{a}}$.

Discussion

This paper provides a coherent approach to handle the issue of compatibility between pedigree-based and marker-based relationship matrices in single-step genetic evaluation. The approach is computationally fast and feasible for large datasets, as is the case for previously developed single-step methods. Parameter γ can adjust the pedigree-based relationship matrix to the marker-based relationship matrix that is scaled by parameter $\tilde{s}$, and these two parameters should in principle be estimated using the full log-likelihood for observed phenotypes and observed marker genotypes. This is computationally feasible by computing the full log-likelihood values for parameters γ, $\tilde{s}$ and ω in a three-dimensional grid. However, in practice this can be computationally burdensome and an appealing alternative is to estimate γ and $\tilde{s}$ based on the observed marker genotypes only. Analysis of simulated datasets, shows that the estimates of parameters γ and $\tilde{s}$ depend on the observed phenotypes as well as on the observed marker genotypes, although this dependence is not too large. A conjecture is that in a scenario with a small number of genotyped animals it is more important to use the full log-likelihood than in a scenario with a large number of genotyped animals. Based on these two simulated datasets only, no general conclusion can be made and further studies are needed to determine in which scenarios it would be safe to base inference on a two-step procedure in which γ and $\tilde{s}$ are estimated based on the observed marker genotypes only, and the remaining parameters are then estimated by the log-likelihood of the phenotypes conditional on the observed marker genotypes.

Using the approach developed in this paper may also provide insight on performance of other approaches for making marker-based and pedigree-based relationship matrices compatible in single-step genetic evaluation. Examples of such approaches are the adjustments of the marker-based relationship matrix reported in [4-6,14,15] and also investigated here, and the approach by Meuwissen et al. [16]. In [16], first, an average of position specific identical by descent matrices based on linkage information [17] (hereafter denoted $\mathbf{G}_{FG}$) were used instead of the combined relationship matrix (1), and second, the final method consisted of replacing the pedigree-based relationship matrix $\mathbf{A}$ by the matrix $\mathbf{G}_{FG}$ in matrix (1) and then adjusting the marker-based relationship matrix to the $\mathbf{G}_{FG}$ matrix. Based on simple simulated datasets, without selection, the method resulted in high accuracy and low bias. The computation of $\mathbf{G}_{FG}$ is computationally burdensome and therefore this approach was not considered here. It should be noted that for these other approaches, the combined relationship matrix is constructed based

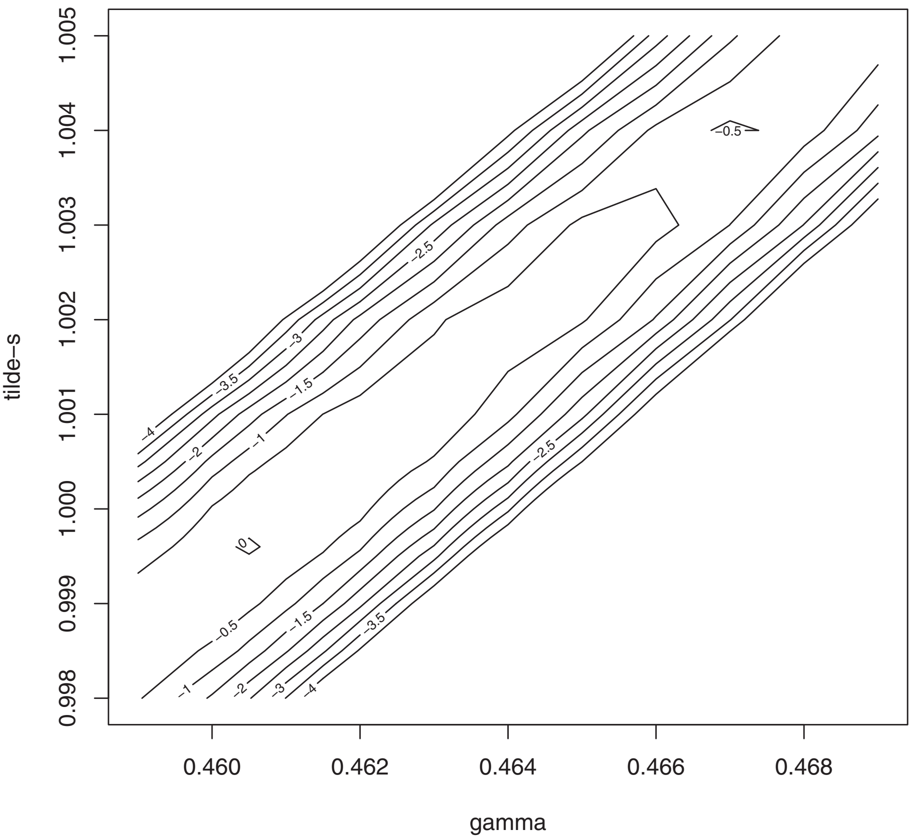

Figure 1 Profile log-likelihood for γ and $\tilde{s}$. A contour-plot of the profile log-likelihood for parameters γ and $\tilde{s}$ based on the full log-likelihood in equation (5); the plot is constructed with values in a discrete grid (explaining the roughness of the plot) that have been standardised such that the maximum value is zero.

on observed marker genotypes only, and the inference on remaining parameters is based on the log-likelihood conditional on the observed marker genotypes. The approach in this paper may provide guidelines on when it is safe to base inference on such procedures in which, first the observed marker genotypes are used to construct a combined relationship matrix, and second the remaining inference is based on the log-likelihood for the phenotypes conditional on observed marker genotypes.

Evaluation of accuracy and bias of prediction with the simulated dataset 2 shows that the approaches based on adjusting the pedigree-based relationship matrix, where $\mathrm{Cor}(\hat{\mathbf{a}}, \mathbf{a}) = 0.536$ and reg $= 1.17$, seem to perform better than the G-adjust approach, where $\mathrm{Cor}(\hat{\mathbf{a}}, \mathbf{a}) = 0.493$ and reg $= 1.34$. However, $\hat{\omega}$ is also larger in the second case (0.60) than in the first case (0.375). If ω is set at 0.375 in the second case then $\mathrm{Cor}(\hat{\mathbf{a}}, \mathbf{a}) = 0.534$, which is close to the value in the first case, but reg is somewhat larger, i.e. 1.30. If ω is set at 0.1, then $\mathrm{Cor}(\hat{\mathbf{a}}, \mathbf{a}) = 0.545$ and reg $= 1.00$ in the first case, and $\mathrm{Cor}(\hat{\mathbf{a}}, \mathbf{a}) = 0.550$ and reg $= 1.12$ in the second case, and if ω is set at 0.01, then $\mathrm{Cor}(\hat{\mathbf{a}}, \mathbf{a}) = 0.548$ and reg $= 1.05$ in the second case. Thus, the first two approaches seem to perform better in terms of estimating a more proper ω and have somewhat less bias, while accuracy is similar for all three approaches.

This paper suggests that the pedigree-based relationship matrix should be adjusted instead of the marker-based relationship matrix. On the one hand, from a practical point of view this would be rather inconvenient since standard software used for REML estimation and BLUP would have to be modified, but on the other hand, it is conceptually simpler and may be easier to extend. For example, when the population consists of a mixture of breeds, it may be simpler to extend the approach in this paper and specify a parametric structure on the relationships of the animals in the base population and estimate those parameters, instead of developing an appropriate way of adjusting the marker-based relationship matrix of the genotyped animals across breeds. An additional issue is the interpretation of the genetic variance parameter σ_a^2. The parameter estimates in Table 1 are $\hat{\sigma}_a^2 = 6.507$ when $\hat{\gamma} = 0.4605$, $\hat{\sigma}_a^2 = 6.512$ when $\hat{\gamma} = 0.4615$, and $\hat{\sigma}_a^2 = 5.084$ for G-adjust, where base animals are unrelated, i.e. $\gamma = 0$. This may seem counter-intuitive since the variance of breeding

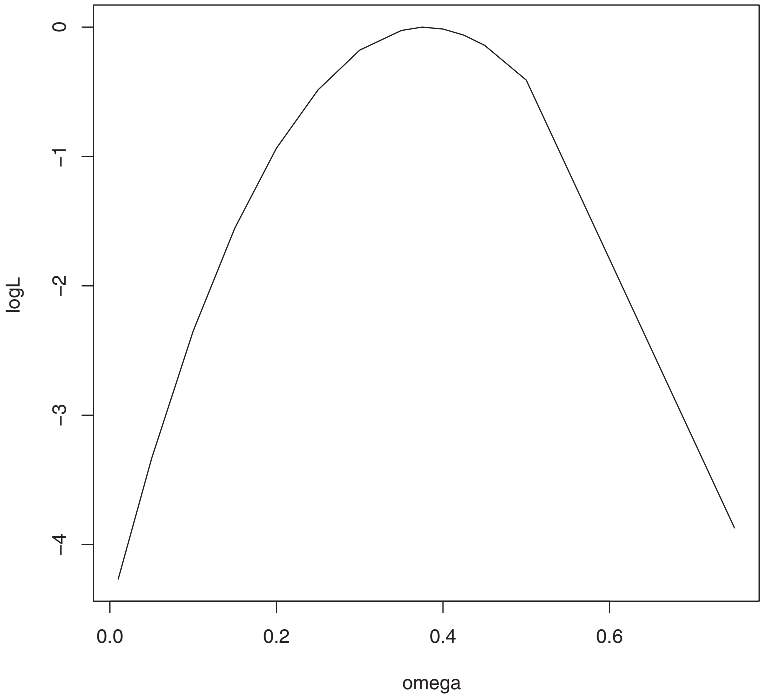

Figure 2 Profile log-likelihood for ω. The profile log-likelihood function for parameter ω based on the full log-likelihood in equation (5); the plot is constructed with values in a discrete grid that have been standardised such that the maximum value is zero.

values for a base animal is $\sigma_a^2(1 + \gamma/2)$. However, when studying the inverse relationship matrix $\tilde{\mathbf{A}}(\gamma)^{-1}$, then the averages of diagonal elements are 3.807, 3.809 and 2.93 for the three cases, respectively, and since 3.807/6.507, 3.809/6.512 and 2.93/5.084 are roughly the same, the parameter estimates actually make good sense, although the interpretation of σ_a^2 is unclear. Finally, it should be noted that parameter γ could influence to some extent accuracy and bias of prediction, and neither $\gamma = 0$ nor γ estimated to make marker-based and pedigree-based relationships compatible, would be optimal for that purpose. Adjusting the pedigree-based relationship instead of the marked-based relationship matrix to make the two matrices compatible in a single-step method is an interesting alternative.

For the simulated dataset 1, the estimate of the relative polygenic weight ω was about 0, as was expected since in the simulation the markers were assumed to capture all the genetic variation. However, for the simulated dataset 2, $\hat{\omega}$ was large when adjusting the pedigree-based relationship matrix, i.e. 0.375, and even larger, i.e. 0.60, when the G-adjust approach was used. For the second dataset, the prediction biases were reduced when ω was set at 0.1. This poses the question whether it is actually possible to estimate ω at a reasonable value from data, or whether ω should be determined manually to control the bias. It should be noted that confidence intervals of this parameter are large in these examples and further studies are needed.

Conclusions

Using the full log-likelihood and adjusting the pedigree-based relationship matrix to be compatible with the marker-based relationship matrix provides a new and interesting approach to handle the issue of compatibility between the two matrices in single-step genetic evaluation.

Appendix A

The derivation of the proposed adjustment of the pedigree-based relationship matrix is based on assigning priors on the high-dimensional parameters $\boldsymbol{\rho}$ and $\mathbf{v}$ in the previously described single-step model. This derivation is presented here.

The model for the marker genotypes, $\mathbf{M}$, given the allele frequencies $\boldsymbol{\rho}$ and variances $\mathbf{v}$, is as in [3],

$$\mathrm{E}[M_{ij} \mid \rho_j] = 2\rho_j - 1, \;\; \mathrm{Cov}[M_{ij}, M_{i'j} \mid v_j] = v_j A_{ii'},$$

$$\mathrm{Cov}[M_{ij}, M_{i'j'} \mid v_{jj'}] = v_{jj'} A_{ii'},$$

where $v_j = v_{j,j}$ is a parameter and $v_{jj'} = 0$ for $j \neq j'$. Prior distributions for $\boldsymbol{\rho}$ and $\mathbf{v}$ are not specified explicitly, only the necessary assumptions are stated. Assuming that the alleles are randomly labeled 1 and 2, then a priori the expected allele frequency is $\mathrm{E}[\rho_j] = 0.5$. Additional assumptions are that $\eta_1 = \mathrm{Var}[\rho_j]$ and $\eta_2 = \mathrm{E}[v_j]$ do not depend on j, and that $\mathrm{Cov}[\rho_j, \rho_{j'}] = 0$ for $j \neq j'$.

With these prior distributions, a marginal distribution of $\mathbf{M}$ may be obtained by integrating $\boldsymbol{\rho}$ and $\mathbf{v}$. Using well-known formulas for conditional expectations and covariances, the mean and covariances become

$$\mathrm{E}[M_{ij}] = \mathrm{E}[\mathrm{E}[M_{ij} \mid \rho_j]] = \mathrm{E}[2\rho_j - 1] = 0,$$

$$\begin{aligned}
&\mathrm{Cov}[M_{ij}, M_{i'j}] \\
&= \mathrm{E}[\mathrm{Cov}[M_{ij}, M_{i'j} \mid \rho_j, v_j]] \\
&\quad + \mathrm{Cov}[\mathrm{E}[M_{ij} \mid \rho_j, v_j], \mathrm{E}[M_{i'j} \mid \rho_j, v_j]] \\
&= \mathrm{E}[v_j A_{ii'}] + \mathrm{Var}(2\rho_j - 1) = \eta_2 A_{ii'} + 4\eta_1,
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Cov}[M_{ij}, M_{i'j'}] \\
&= \mathrm{E}[\mathrm{Cov}[M_{ij}, M_{i'j'} \mid \boldsymbol{\rho}, \mathbf{v}]] \\
&\quad + \mathrm{Cov}[\mathrm{E}[M_{ij} \mid \boldsymbol{\rho}, \mathbf{v}], \mathrm{E}[M_{i'j'} \mid \boldsymbol{\rho}, \mathbf{v}]] \\
&= \mathrm{E}[0] + 4\mathrm{Cov}[\rho_j, \rho_{j'}] = 0 \quad \text{for } j \neq j'.
\end{aligned}$$

Defining $\tilde{v}_j = 2\eta_1 + \eta_2$, and $\gamma = 4\eta_1/(2\eta_1 + \eta_2)$, we obtain

$$\mathrm{Cov}[M_{ij}, M_{i'j}] = \tilde{A}(\gamma)_{ii'} \tilde{v}_j,$$

with $\tilde{A}(\gamma)_{ii'} = (1 + \gamma/2)A_{ii'} + \gamma$. It is not difficult to see that $\tilde{\mathbf{A}}(\gamma)$ satisfies the usual recursions for an additive relationship matrix. Requiring that $0 \leq \gamma < 1$ is equivalent to making a further assumption that $2\eta_1 < \eta_2$ (this is for example satisfied when $\rho_j \sim U]0; 1[$ and $v_j = 2\rho_j(1 - \rho_j)$, where $\eta_1 = 1/12$ and $\eta_2 = 1/3$). The first and second order moments of $\mathbf{M}$ are therefore of the form considered in [3], with pedigree-based relationship matrix $\tilde{\mathbf{A}}(\gamma)$, allele frequencies $\tilde{p}_j = 0.5$, scaling parameter $\tilde{s} = \sum_j \tilde{v}_j$ with $\tilde{v}_j$ not depending on j, and marker-based relationship matrix $\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}}/\tilde{s}$.

Note that the random labelling of alleles as 1 or 2 is not important because the resulting expression does not depend on how alleles are labelled.

Using $\tilde{v}_j = \tilde{s}/p$, the log-likelihood for the observed marker genotypes becomes

$$\begin{aligned}
&\tilde{\ell}_{mark}(\gamma, \tilde{s}) = \mathrm{const} - (pn_1/2)\log(\tilde{s}) \\
&- (p/2)\log(\det(\tilde{\mathbf{A}}(\gamma)_{11})) \\
&- (1/2)\sum_{ii'}((\tilde{s}\tilde{\mathbf{A}}_{11}(\gamma)/p)^{-1})_{ii'}(\mathbf{m}^o(\mathbf{m}^o)^{\mathrm{T}})_{ii'}.
\end{aligned}$$

This log-likelihood may also be viewed as the log-likelihood for $\mathbf{M}^o(\mathbf{M}^o)^{\mathrm{T}}$ being Wishart distributed $W_{n_1}(\tilde{s}\tilde{\mathbf{A}}_{11}(\gamma)/p, p)$.

Appendix B

The aim here is to present algorithms for computing $(\tilde{\mathbf{A}}(\gamma))^{-1}$ and $\tilde{\mathbf{A}}(\gamma)_{11}$ by recursions. Computations of inbreeding coefficients with related based animals were considered by [10,11]. For simplicity, let's assume that, either both parents are known or both parents are unknown. This is not an important restriction in practice, since the case with one unknown parent can be handled by assigning an artificial animal-id to this unknown parent and by letting both its parents be unknown.

Let us partition matrix $\tilde{\mathbf{A}}$ (skipping the dependence on γ in the notation from now and onwards) according to whether the animals are base animals (both parents unknown) or not

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}^{bas} & \tilde{\mathbf{A}}^{bas,nonbas} \\ \tilde{\mathbf{A}}^{nonbas,bas} & \tilde{\mathbf{A}}^{nonbas} \end{bmatrix}.$$

Similar to the usual $\mathbf{A}$ matrix (see [18]) a decomposition exists

$$\tilde{\mathbf{A}} = \mathbf{T} \begin{bmatrix} \tilde{\mathbf{A}}^{bas} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}} \end{bmatrix} \mathbf{T}^{\mathrm{T}},$$

where $\mathbf{T}$ is a lower triangular matrix with entries $T_{ii} = 1$, $T_{ik} = (T_{f(i)k} + T_{m(i)k})/2$ when i has known parents $f(i)$ and $m(i)$ and $i > k$, and $T_{ik} = 0$ otherwise ($k > i$ or i has unknown parents), and $\tilde{\mathbf{D}}$ is a diagonal matrix containing the variance of the Mendelian sampling part for matrix $\tilde{\mathbf{A}}^{nonbas}$, i.e. $\tilde{D}_{ii} = 1 - (\tilde{A}_{f(i)f(i)} + \tilde{A}_{m(i)m(i)})/4$. Comparing this decomposition to the decomposition for the usual $\mathbf{A}$ matrix, then matrix $\tilde{\mathbf{A}}^{bas}$ has replaced an identity matrix and $\tilde{\mathbf{D}}$ has replaced a diagonal matrix containing the Mendelian sampling terms related to $\mathbf{A}$ for the animals with known parents.

The inverse matrix becomes

$$\tilde{\mathbf{A}}^{-1} = (\mathbf{T}^{-1})^{\mathrm{T}} \begin{bmatrix} (\tilde{\mathbf{A}}^{bas})^{-1} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}^{-1} \end{bmatrix} \mathbf{T}^{-1}.$$

Here $\mathbf{T}^{-1}$ is a lower triangular matrix with ones on the diagonal and the only non-zero elements being $-1/2$ for offspring-parent entries, and matrix $(\tilde{\mathbf{A}}^{bas})^{-1}$ has diagonal elements equal to $(1 + (n_{bas} - 3/2)\gamma)/((1 - \gamma/2)(1 + (n_{bas} - 1/2)\gamma))$ and off-diagonal elements equal to $-\gamma/((1 - \gamma/2)(1 + (n_{bas} - 1/2)\gamma))$, where n_{bas} is the dimension of $\tilde{\mathbf{A}}^{bas}$.

A procedure to obtain the diagonal of $\tilde{\mathbf{A}}$ and the inverse $(\tilde{\mathbf{A}})^{-1}$ can be constructed similar to the algorithm by Quaas [12], which utilises the form $\tilde{\mathbf{A}} = \mathbf{L}\mathbf{L}^T$ where

$$\mathbf{L} = \mathbf{T} \begin{bmatrix} \sqrt{\tilde{\mathbf{A}}^{bas}} & \mathbf{0} \\ \mathbf{0} & \sqrt{\tilde{\mathbf{D}}} \end{bmatrix}.$$

Matrix $\sqrt{\tilde{\mathbf{A}}_{bas}}$ is a lower triangular matrix such that $\tilde{\mathbf{A}}_{bas} = \sqrt{\tilde{\mathbf{A}}_{bas}} \left(\sqrt{\tilde{\mathbf{A}}_{bas}}\right)^{\mathrm{T}}$. First, matrix $\sqrt{\tilde{\mathbf{A}}_{bas}}$ is computed and $\tilde{A}_{ii}$ is set to $1 + \gamma/2$ for the base animals. Second, the remaining part of the algorithm is the same as in Quaas [12], where recursively, rows in L are computed using $\tilde{D}_{ii} = 1 - (\tilde{A}_{f(i)f(i)} + \tilde{A}_{m(i)m(i)})/4$ and $\tilde{A}_{ii} = \sum_{k=1}^{i} L_{ik}^2$. Obtaining $(\tilde{\mathbf{A}})^{-1}$ at the same time is done by first setting elements in $(\tilde{\mathbf{A}})^{-1}$ for base animals equal to the elements in $(\tilde{\mathbf{A}}^{bas})^{-1}$, and then adding elements to $(\tilde{\mathbf{A}})^{-1}$ in the usual way (assuming a half-stored format, add $1/\tilde{D}_{ii}$ to the (i, i) element, add $-1/(2\tilde{D}_{ii})$ to the $(i, f(i))$ and $(i, m(i))$ elements, and add $1/(4\tilde{D}_{ii})$ to the $(f(i), f(i))$, $(m(i), m(i))$ and $(m(i), f(i))$ elements).

Matrix $\tilde{\mathbf{A}}_{11}$ can be obtained by a modification of the Colleau algorithm [13,19,20] by using an algorithm to compute $\tilde{\mathbf{A}}\mathbf{x}$, where $\mathbf{x}$ is a vector. The product of matrix $\tilde{\mathbf{A}}$ and vector $\mathbf{x}$ is

$$\tilde{\mathbf{A}}\mathbf{x} = \mathbf{T}\begin{bmatrix}\tilde{\mathbf{A}}^{bas} & \mathbf{0}\\ \mathbf{0} & \tilde{\mathbf{D}}\end{bmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{x}. \tag{7}$$

Therefore, an algorithm (with notation similar to that in Appendix A in Misztal et al. [19]) consists of first computing $\mathbf{r} = \mathbf{T}^{\mathrm{T}}\mathbf{x}$ by solving the sparse system $(\mathbf{T}^{-1})^{\mathrm{T}}\mathbf{r} = \mathbf{x}$ for $\mathbf{r}$, then computing

$$\mathbf{t} = \begin{bmatrix}\tilde{\mathbf{A}}^{bas} & \mathbf{0}\\ \mathbf{0} & \tilde{\mathbf{D}}\end{bmatrix}\mathbf{r} = \begin{bmatrix}(1-\gamma/2)\mathbf{r}^{bas} + \gamma\mathbf{1}^{\mathrm{T}}\mathbf{r}^{bas}\mathbf{1}\\ \tilde{\mathbf{D}}\mathbf{r}^{nonbas}\end{bmatrix},$$

and finally computing $\tilde{\mathbf{A}}\mathbf{x} = \mathbf{T}^{\mathrm{T}}\mathbf{t}$ by solving the sparse system $(\mathbf{T}^{-1})^{\mathrm{T}}(\tilde{\mathbf{A}}\mathbf{x}) = \mathbf{r}$ for $\tilde{\mathbf{A}}\mathbf{x}$.

Competing interests
The author declares that he has no competing interests.

Authors' contributions
The author has read and approved the final manuscript.

Acknowledgements
The work was performed in a project funded through the Green Development and Demonstration Programme (grant no. 3405-11-0279) by the Danish Ministry of Food, Agriculture and Fisheries, the Pig Research Centre and Aarhus University. The author thanks two anonymous reviewers, an associate editor and editor Jack Dekkers for their comments, and in particular editor Hélène Hayes for her careful and competent language revisions.

Received: 3 September 2012 Accepted: 20 November 2012
Published: 3 December 2012

References

1. Legarra A, Aguilar I, Misztal I: **A relationship matrix including full pedigree and genomic information.** *J Dairy Sci* 2009, **92:**4656–4663.
2. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ: **Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluations of Holstein final score.** *J Dairy Sci* 2010, **93:**743–752.
3. Christensen OF, Lund MS: **Genomic prediction when some animals are not genotyped.** *Genet Sel Evol* 2010, **42:**2.
4. Vitezica ZG, Aguilar I, Misztal I, Legarra A: **Bias in genomic predictions for populations under selection.** *Genet Res* 2011, **93:**357–366.
5. Chen CY, Misztal I, Aguilar I, Legarra A, Muir WM: **Effect of different genomic relationship matrices on accuracy and scale.** *J Anim Sci* 2011, **89:**2673–2679.
6. Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G: **Single-step methods for genomic evaluation in pigs.** *Animal* 2012, **6:**1565–1571.
7. Forni S, Aguilar I, Misztal I: **Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information.** *Genet Sel Evol* 2011, **43:**1.
8. Gilmour AR, Thompson R, Cullis BR: **Average information REML: an efficient algorithm for parameter estimation in linear mixed models.** *Biometrics* 1995, **51:**1440–1450.
9. Gengler N, Mayeres P, Szydlowski M: **A simple method to approximate gene content in large pedigree populations: application to the myostation gene in dual-purpose Belgian Blue cattle.** *Animal* 2007, **1:**21–28.
10. VanRaden PM: **Accounting for inbreeding and crossbreeding in genetic evaluation of large populations.** *J Dairy Sci* 1992, **75:**3136–3144.
11. Colleau JJ, Sargolzaei M: **MIM: an indirect method to assess inbreeding and coancestry in large incomplete pedigrees of selected dairy cattle.** *J Anim Breed Genet* 2011, **128:**163–173.
12. Quaas RL: **Computing the diagonal elements and inverse of a large numerator relationship matrix.** *Biometrics* 1976, **32:**949–953.
13. Colleau JJ: **An indirect approach to the extensive calculation of relationship coefficients.** *Genet Sel Evol* 2002, **34:**409–421.
14. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91:**4414–4423.
15. Gao H, Christensen OF, Madsen P, Nielsen US, Zhang Y, Lund MS, Su G: **Comparison on genomic predictions using GBLUP models and two single-step blending methods with different relationship matrices in the Nordic Holstein population.** *Genet Sel Evol* 2012, **44:**8.
16. Meuwissen THE, Luan T, Woolliams JA: **The unified approach to the use of genomic and pedigree information in genomic evaluations revisited.** *J Anim Breed Genet* 2011, **128:**429–439.
17. Fernando RL, Grossman M: **Marker assisted selection using best linear unbiased prediction.** *Genet Sel Evol* 1989, **21:**467–477.
18. Mrode RA: *Linear models for the prediction of animal breeding values.* Wallingford: CABI Publishing; 2005.
19. Misztal I, Legarra A, Aguilar I: **Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information.** *J Dairy Sci* 2009, **92:**4648–4655.
20. Aguilar I, Misztal I, Legarra A, Tsuruta S: **Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation.** *J Anim Breed Genet* 2011, **128:**422–428.

doi:10.1186/1297-9686-44-37
Cite this article as: Christensen: Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genetics Selection Evolution* 2012 **44**:37.