Genetics Selection Evolution

RESEARCH Open Access

# Genomic prediction when some animals are not genotyped

Ole F Christensen*, Mogens S Lund

## Abstract

**Background:** The use of genomic selection in breeding programs may increase the rate of genetic improvement, reduce the generation time, and provide higher accuracy of estimated breeding values (EBVs). A number of different methods have been developed for genomic prediction of breeding values, but many of them assume that all animals have been genotyped. In practice, not all animals are genotyped, and the methods have to be adapted to this situation.

**Results:** In this paper we provide an extension of a linear mixed model method for genomic prediction to the situation with non-genotyped animals. The model specifies that a breeding value is the sum of a genomic and a polygenic genetic random effect, where genomic genetic random effects are correlated with a genomic relationship matrix constructed from markers and the polygenic genetic random effects are correlated with the usual relationship matrix. The extension of the model to non-genotyped animals is made by using the pedigree to derive an extension of the genomic relationship matrix to non-genotyped animals. As a result, in the extended model the estimated breeding values are obtained by blending the information used to compute traditional EBVs and the information used to compute purely genomic EBVs. Parameters in the model are estimated using average information REML and estimated breeding values are best linear unbiased predictions (BLUPs). The method is illustrated using a simulated data set.

**Conclusions:** The extension of the method to non-genotyped animals presented in this paper makes it possible to integrate all the genomic, pedigree and phenotype information into a one-step procedure for genomic prediction. Such a one-step procedure results in more accurate estimated breeding values and has the potential to become the standard tool for genomic prediction of breeding values in future practical evaluations in pig and cattle breeding.

## Background

Genomic selection [1] has become the new paradigm in animal breeding programs using marker-assisted selection. It may increase the rate of genetic improvement, reduce the generation time, and provide higher accuracy of estimated breeding values (EBVs). Genomic prediction of breeding values can be based on a linear mixed model using matrix computations or a non-linear mixture type of model using Markov chain Monte Carlo (McMC) procedures. In this paper we provide a natural extension of a linear mixed model to the situation with non-genotyped animals.

A marker-based relationship matrix has been used by a number of authors, in particular VanRaden in [2] and [3], but also Gianola and van Kamm [4] in a dual formulation of their model. The types of genomic relationship matrices studied here are on the form

$$G(m) = (m - p)h(m - p)^{\mathrm{T}}, \tag{1}$$

as in VanRaden [3], but other types of genomic relationship matrices are discussed in the discussion section. In VanRaden [3] it is assumed that all animals are genotyped, which is unlikely to be a common scenario. In particular, in pig breeding it is probable that only boars or other selection candidates are genotyped, and in cattle breeding, traits being recorded for millions of animals it is very unlikely that all will be genotyped. We present an

* Correspondence: OleF.Christensen@agrsci.dk
Aarhus University, Faculty of Agricultural Sciences, Dept of Genetics and Biotechnology, Blichers Allé 20, PO BOX 50, DK-8830 Tjele, Denmark

extension of matrix (1) in the situation where not all animals are genotyped. The approach presented here combines the relationship matrix (1) with a model for the markers. By marginalisation of the markers of non-genotyped animals a natural extension of (1) is obtained. The resulting extension of the genomic relationship matrix is the same as the one derived in Legarra et al. [5], but the details in the derivation are somewhat different and the derivation therefore sheds more light on this result.

To capture genetic variation not associated to the markers in a given SNP-panel, the model can also contain a polygenic genetic effect with the usual pedigree derived additive relationship matrix, as considered by [4,6] among others. The extension of the genomic relationship matrix to non-genotyped animals together with the addition of the polygenic effect provide a natural one-step procedure to blend the information from relatives and the genomic information into a combined genomically enhanced breeding value (GEBV). Genomic prediction with both a polygenic effect and with incomplete genotyping has been considered by a number of authors. Using a joint model for phenotypes and markers and using Bayesian inference, a general solution to sample missing markers in each McMC iteration has been suggested [4,7]. However, with a large number of SNP markers and many animals without genotypes such a solution seems computationally unfeasible in practice. In Gianola et al. [7] bivariate models are suggested, where the two traits are the traits of the genotyped and non-genotyped animals, respectively, and the genetic effect for a genotyped animal is the sum of a polygenic effect and a genomic effect whereas the genetic effect for a non-genotyped animal is just a polygenic effect (correlated with the polygenic effect of the genotyped animals). Since the model does not contain a genomic genetic effect for the non-genotyped animals, the phenotypic information from non-genotyped animals closely related to a given genotyped animal does not propagate properly into the estimate of the genomic genetic effect for this animal. Alternatively, the approach by Baruch and Weller [8] involves several steps, where first, expected genotypes are computed for non-genotyped animals, then marker effects are estimated (using expected genotypes for non-genotyped animals), phenotypes are adjusted by known or expected marker effects, and finally polygenic EBVs are computed from adjusted phenotypes. Although somewhat similar in idea to the approach taken here, the approach in [8] does not propagate any uncertainty from one step in the procedure to the next step, and the effects are not estimated simultaneously.

## Methods

We assume that markers are summarised into a gene content matrix, $m$ ($m_{ij}$ = -1, when the SNP $j$ of individual $i$ is 11, $m_{ij}$ = 0 for 12, and $m_{ij}$ = 1 for 22), and we use capital letters $M_{ij}$ to denote when the markers are random variables. For the genomic relationship matrix (1), the matrix $p$ is the expectation of $M$, i.e. the entries in column $j$ are $p_j = 2(\rho_j - 1/2)$ with $\rho_j$ being the allele frequency of the second allele at loci $j$, and $h$ is a diagonal matrix chosen such that $\mathrm{E}[G(M)] = A$, the usual pedigree derived additive relationship matrix. In VanRaden [3] three different genomic relationship matrices are presented, where the first two are on the form in (1), and here, we focus on the first one

$$G(m) = (m - p)(m - p)^{\mathrm{T}} / s \tag{2}$$

with $s = \sum_j 2\rho_j(1 - \rho_j)$.

The model is as follows

$$y = X\beta + Za + Zg + e, \tag{3}$$

where $y$ is phenotype, $X$ and $Z$ are incidence matrices, $\beta$ denotes fixed effects, $e$ is error, $a \sim N(0, \sigma_a^2 A)$ is the polygenic genetic effect, and $g \sim N(0, \sigma_g^2 G^*(m^{obs}))$ is the genomic genetic effect. Here $A$ is the usual pedigree derived additive relationship matrix, and $G^*(m^{obs})$ is the extension of (2) to be derived in the following section.

In the following sections, first, we derive the extension of the marker based relationship matrix, $G^*(m^{obs})$, and second, we study the variance-covariance matrix of the combined genetic effect $g + a$. Then procedures for parameter estimation using AI-REML, and breeding value estimation are presented. Finally, a simulation data set is described.

### Genomic relationship matrix with a relationship of markers

Gengler et al. [9] suggested that missing genotypes could be modelled using the usual mixed model methodology with relationship matrix $A$. We now combine that idea with the genomic relationship matrix on the form (1). For simplicity, the derivation is made for the form (2), but it is straight-forward to generalise to (1) also.

The model for the genomic genetic effect is as follows

$$g \mid M \sim N(0, \sigma_g^2 G(M)), \quad \text{with} \quad G(M) = (M - p)(M - p)^{\mathrm{T}} / s,$$

where $M$ is the gene content matrix. We assume that $\mathrm{E}[M_j] = 1p_j$, $\mathrm{Var}(M_j) = v_j A$, with $A$ the usual relationship matrix, $v_j = 2\rho_j(1 - \rho_j)$, and $s = \sum_j v_j$. The covariances of $M_j$, and $M_{j'}$ for two different loci $j \neq j'$ are on the form $\mathrm{Cov}(M_j, M_{j'}) = v_{j,j'} A$ where the $v_{j,j'}$ s are unspecified since they are cancelling in the derivations that follow.

We split $M$ into two sub-matrices containing the animals with observed genotypes and those without, respectively,

$$M = \begin{bmatrix} M^{obs} \\ M^{miss} \end{bmatrix},$$

and in the following we distinguish between small letter $m^{obs}$ (observed realisation of random variables $M^{obs}$) and capital letter $M^{miss}$ (unobserved markers are random variables). In Appendix A, the mean vector and variance-covariance matrix of the conditional distribution $[g|m^{obs}]$ (with $M^{miss}$ marginalised out) are shown to be

$$\mathrm{E}[g \mid m^{obs}] = 0, \quad \mathrm{Var}[g \mid m^{obs}] = \sigma_g^2 G^*(m^{obs}),$$

Where

$$G^*(m^{obs}) = \begin{bmatrix} G(m^{obs}) & G(m^{obs})A_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1}G(m^{obs}) & A_{21}A_{11}^{-1}G(m^{obs})A_{11}^{-1}A_{12} + A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}. \quad (4)$$

When all animals have been genotyped, $G^*(m^{obs}) = G(m^{obs})$, and when no animals have been genotyped, $G^*(m^{obs}) = A$, which makes the extension in (4) rather elegant. We assume that the distribution of $[g|m^{obs}]$ is multivariate normal, which for the non-genotyped animals is not strictly true, but an approximation.

The inverse of the genomic relationship matrix may be obtained from the inverse of $A$,

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{bmatrix} \quad (5)$$

Using some algebra, the inverse of the genomic relationship matrix becomes

$$\begin{aligned} &G^*(m^{obs})^{-1} \\ &= \begin{bmatrix} G(m^{obs})^{-1} + A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \\ -(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1} & (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{bmatrix} \\ &= \begin{bmatrix} G(m^{obs})^{-1} - A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + A^{-1}. \end{aligned} \quad (6)$$

Considering the terms in (6), because of the low dimension of $G(m^{obs})$ and $A_{11}$ a direct inversion of these matrices should be possible for practical computations, and $A^{-1}$ is a sparse matrix which can be computed directly without constructing $A$ itself and using standard techniques. To compute $A_{11}$ there might be cases where most of the $A$ matrix has to be computed, potentially causing a memory storage problem.

Alternatively, $A_{11} = ((A^{-1})^{-1})_{11}$ may be computed using the formula (5) on $A^{-1}$ and using sparse matrix computation. The formula (6) requires that $G(m^{obs})$ is invertible which may not actually be the case. In the next section this problem is automatically solved by combining the genomic genetic effect $g$ with the polygenic effect $a$.

We also note that the determinant equals

$$\det(G^*(m^{obs})) = \det(G(m^{obs}))\det(A_{22} - A_{21}A_{11}^{-1}A_{12}),$$

where $A_{22}$ - $A_{21}$ $A_{11}^{-1}A_{12}$ is easily obtained from $A^{-1}$, and the determinant can be computed using sparse matrix computation.

### The combined genetic effect

The combined genetic effect is the sum of the genomic genetic effect and the polygenic effect, $\tilde{g}$ = g + a, and using this notation the model (3) may now be written as

$$y = X\beta + Z\tilde{g} + e, \quad (7)$$

where $\tilde{g} \sim N(0, \sigma_g^2 G^*(m^{obs}) + \sigma_a^2 A)$. Introducing the notation $w = \sigma_a^2 / (\sigma_g^2 + \sigma_a^2)$ and $\sigma_{\tilde{g}}^2 = \sigma_g^2 + \sigma_a^2$, then

$$\tilde{g} \sim N(0, \sigma_{\tilde{g}}^2 \tilde{G}_w),$$

with $\tilde{G}_w$ = (1 - $w$)$G^*(m^{obs})$ + $wA$. Substituting (4) and rearranging the terms, we obtain

$$\tilde{G}_w = \begin{bmatrix} G_w & G_w A_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1}G_w & A_{21}A_{11}^{-1}G_w A_{11}^{-1}A_{12} + A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix},$$

where

$$G_w = (1 - w)G(m^{obs}) + wA_{11}.$$

The parameter $w$ is interpreted as the relative weight on the polygenic effect, and it may be estimated from data as shown in the next section or be chosen to equal a small value.

Similar to the previous section the inverse equals

$$(\tilde{G}_w)^{-1} = \begin{bmatrix} G_w^{-1} - A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + A^{-1}, \quad (8)$$

and here $G_w$ is necessarily invertible when $w > 0$ (even when $G(m^{obs})$ is singular).

### Variance component estimation

Here we consider parameter estimation using average information (AI)-REML based on the mixed model equations [10,11]

$$\begin{bmatrix} X^{\mathrm{T}}X & X^{\mathrm{T}}Z \\ Z^{\mathrm{T}}X & Z^{\mathrm{T}}Z + (\tilde{G}_w)^{-1}\lambda_{\tilde{g}}^2 \end{bmatrix} \begin{bmatrix} \beta \\ \tilde{g} \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}y \\ Z^{\mathrm{T}}y \end{bmatrix}, \quad (9)$$

where $\lambda_{\tilde{g}} = (\sigma_{\tilde{g}}^2 / \sigma_e^2)^{-1}$. We will not enter into details, but just note that the sparse structure of the left hand side matrix in (9) is the cornerstone for the fast computation of the AI-matrix used in the numerical maximisation of the REML likelihood. Considering the terms in this matrix, then $Z^{\mathrm{T}}Z$ is a sparse matrix, and from (4) we see that $\tilde{G}_w^{-1}$ has some sparse structure, although $G_w^{-1}$ is a dense matrix. Depending on the proportion of animals genotyped it may in some cases not be necessarily advantageous to compute the AI-matrix using (9), but instead an AI-REML algorithm based on the inverse phenotypic variance-covariance matrix, $(\sigma_g^2\tilde{G}_w + \sigma_e^2 I)^{-1}$, could be used, see [12]. Here, we assume that the majority of animals are not genotyped and use the sparse structure of $G^*(m^{obs})^{-1}$ for AI-REML based on the mixed model equations.

The AI-REML method based on the mixed model equations is implemented in software DMU [13] and requires input in the form of the vector of phenotypes, the nonzero entries of $\tilde{G}_w^{-1}$ and the log-determinant log (det($\tilde{G}_w$)) = log(det($G_w$)) + log(det($A_{22}$ - $A_{21}\,A_{11}^{-1}A_{12}$)). For a given $w$ the software provides estimates of $\sigma_g^2$ and $\sigma_e^2$, values of the REML log-likelihood at the maximum and (when required) BLUE solution $\hat{\beta}$ and BLUP solution $\hat{\tilde{g}}$. Here, the parameter $w$ is estimated by using a grid of values, i.e. $w$ = 0.01, 0.03, ..., 0.19, and computing the REML log-likelihood for each value. The resulting profile likelihood curve, log $\hat{L}(w)$, has a peak at the estimate $\hat{w}$, and a measure of the associated uncertainty is the interval {$w$|log $\hat{L}(w)$ > log $\hat{L}(\hat{w})$ - 3.84} where 3.84 is the 95% quantile of a $\chi^2$(1)-distribution.

### Breeding value estimation

Here we consider estimation (prediction) of breeding values. For animals included in the parameter estimation (animals with phenotypes, and some additional animals whose markers provide information about the unknown markers for non-genotyped animals with phenotypes), the GEBVs are the solution vector $\hat{\tilde{g}}$ to (9) with the parameter values being the estimated ones from the previous section. The software DMU provides these GEBVs and their precision.

For animals not included in the parameter estimation, then denoting this subset of animals by index 3 the GEBVs $\hat{\tilde{g}}_3$ are obtained by solving

$$\begin{bmatrix} X^{\mathrm{T}}X & X^{\mathrm{T}}Z_{all} \\ Z_{all}^{\mathrm{T}}X & Z_{all}^{\mathrm{T}}Z_{all} + (\tilde{G}_{all,w})^{-1}\lambda_{\tilde{g}}^2 \end{bmatrix}\begin{bmatrix} \beta \\ \tilde{g}_{all} \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}y \\ Z^{\mathrm{T}}y \end{bmatrix},$$

where $\hat{\tilde{g}}_{all}^{\mathrm{T}} = (\hat{\tilde{g}}^{\mathrm{T}}, \hat{\tilde{g}}_3^{\mathrm{T}})$, $Z_{all}$ and $\tilde{G}_{all,w}$ now contain all animals. Again software DMU provides these GEBVs and their precision.

For a scenario with a large number of genotyped animals whose marker information does not provide information for the parameter estimation, Appendix B presents a method for breeding value estimation where only part of the $\tilde{G}_{all,w}$ needs to be computed.

### A simulated data set

The simulated data set is inspired by a pig nucleus breeding program, but is formulated in a simplified form. We assume, 10 chromosomes each 160 cM long, and a panel of $p$ = 5000 equidistant SNP markers is used. It is assumed that 500 QTLs affect the phenotype, and the size of these effects is simulated from a Gamma (5.4, 0.42)-distribution. First, a base population consisting of 150 boars and 1500 sows is generated by assuming random mating for 50 generations in a population with an effective population size of 100. Then the following mating and selection scheme is followed for five generations. In each generation, 150 boars are mated with 1500 sows to produce 15000 offspring (half of them males). For the next generation, the 150 boars with the highest value of their own phenotype are selected, and 1500 sows are selected randomly. It is assumed that family records are available for all five generations, phenotypes of all boars available for all five generations (35000 records), and the selected boars in the last three generations are genotyped (450 animals). In addition, to estimate the allele frequencies required for the method, the 150 boars in the base population are genotyped (and the allele frequencies used are the estimated frequencies from these 150 boars). For prediction, it is assumed that 300 selection candidates (without phenotypes) for generation 6 are genotyped.

To evaluate the method advocated in this paper (one-step), two other methods are investigated. The first method (ped) computes traditional EBVs using the pedigree based relationship matrix (without using markers). The second method (two-step) is a two-step procedure similar to methods used in practical genomic selection [14,15] and is based on genotyped animals only using the model

$$y_{EBV} = \mu + \tilde{g} + e, \tag{10}$$

where $y_{EBV}$ is the vector of traditional EBVs, and $\tilde{g} \sim N(0, \sigma_{\tilde{g}}^2 G_w)$ with $G_w$ = 0.99$G(m^{obs})$ + 0.01$A_{11}$.

For the one-step method, the genotypes of the selection candidates provide information about the genotypes of their (non-genotyped) mothers and hence information about other non-genotyped animals further back in the pedigree. Therefore they also provide some information about the genotypes of the boars without offspring, and since these boars have phenotypes but not genotypes then the selection candidates should be included in the parameter estimation. However, to investigate how important it is to include these animals, a second analysis (one-step-2) is also performed where they are not included. Finally, to investigate the importance of obtaining the allele frequencies in the base population, the scenario where the boars in the base population have not been genotyped is also studied. The use of three different allele frequencies are compared: 1) true allele frequencies (obtained from the 150 boars in the base population), 2) estimated allele frequencies for boars used in generation 3, 3) allele frequencies estimated using the approach by Gengler et al. [9].

## Results

For the one-step method, the profile likelihood curve for $w$ is shown in Figure 1. It is seen that the data do not support a large polygenic effect, with the estimate being about zero and the 5% confidence interval being about [0; 0.06]. For computational reasons, we decided to use $\hat{w} = 0.01$.

The parameter estimates and the correlation between GEBVs and true breeding values (BVs) are shown in Table 1. For comparison, the prediction using the pedigree based relationship matrix (ped method) and the genomic prediction using (10) based on genotyped animals (two-step) are also shown. We observe that the two methods using a marker-based relationship matrix perform better than the method using the pedigree based relationship matrix, but as expected the one-step method performs the best.

Column four in Table 1 shows the result obtained when ignoring the genotypes of the 300 selection candidates in the parameter estimation (one-step-2). Even though the parameter estimates are somewhat different betweeen one-step and one-step-2, only a minor difference in the correlation between GEBVs and the true breeding values is seen. Hence, for this data set this specific computational short-cut performs well. Finally, the results from the analyses where the boars in the base population are not genotyped show that the choice of allele frequencies is very important for parameter estimation. When using the true allele frequencies, $\hat{w} \approx 0$ is obtained, whereas when using allele frequencies estimated from the observed genotypes, $\hat{w} = 1$ is obtained for both methods estimating the allele frequencies. Since $\hat{w} = 1$ corresponds to the usual animal model, no further results from this comparison are shown here. We conclude that for this data set the parameter estimation is sensitive to the allele frequencies used in the one-step method.

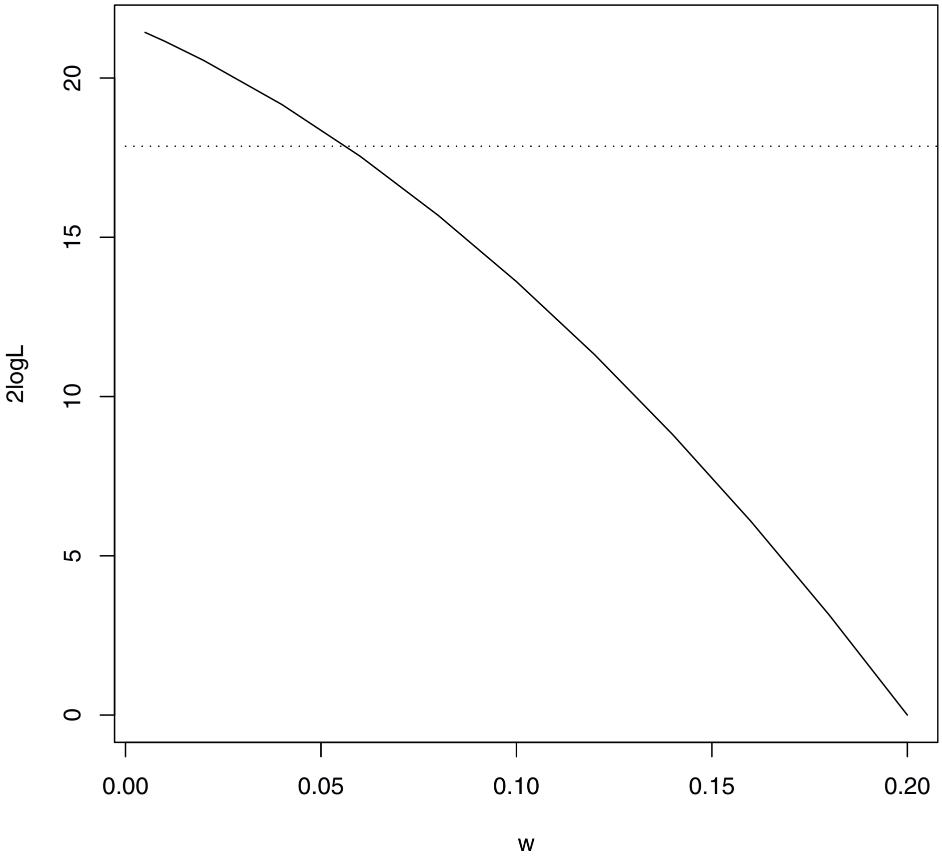

**Figure 1 The profile log-likelihood curve for *w***. The dotted line corresponds to a the 95% quantile for a $\chi^2(1)$ distribution, and provides a 5% confidence interval of [0; 0.06] for *w*.

**Table 1 Results from model with $\hat{w}$ = 0.01.**

| Method | $\hat{\sigma}_g^2$ | $\hat{\sigma}_e^2$ | Cor. true BV |
|---|---|---|---|
| one-step | 4.16 | 16.22 | 0.6598 |
| ped | 5.03 | 15.80 | 0.3537 |
| two-step | 7.56 | 0.069 | 0.5869 |
| one-step-2 | 5.98 | 15.58 | 0.6596 |

Method one-step is the method advocated in this paper, method ped uses the pedigree based relationship matrix, and method two-step is the genomic prediction method using only genotyped animals (note that parameter estimates for this method cannot be compared to parameter estimates from the other two methods). Finally, one-step-2 differs from one-step in that it ignores the markers of selection candidates in the parameter estimation. The right-most column shows the correlation between the estimated and the true breeding value (BV).

## Discussion

For genomic prediction an extension of a linear mixed model to non-genotyped animals has been derived here. The extension of the method makes it possible to integrate in an optimal way the genomic, pedigree and phenotype information into a one-step procedure for breeding value estimation. Due to the simplicity of the method, the fact that it extends the traditional breeding value estimation method in a natural way, and the possibilities of handling large populations, such a one-step procedure has the potential to become the standard tool for genomic prediction of breeding values in practical pig or cattle evaluations in the future. The practical implementation of the approach uses an existing software DMU, and therefore the approach can be easily extended to other types of models implemented by that software, in particular multivariate analysis and generalised linear mixed models.

For such a one-step procedure to become the standard tool for computing GEBVs in practical pig or cattle evaluations, some technical issues of the method need further development. First, computing times necessary for the construction and the inversion of $G(m^{obs})$ are proportional to $n_1^2 p$ and $n_1^3$, respectively. These computations seem to be the computational bottle-necks for the method, and for a very large number of genotyped animals the method may not be feasible. Further research on efficient computation of $G(m^{obs})^{-1}$ seems necessary. Second, some computational short-cuts in the method could be imagined, as illustrated in our results by the good performance of the one-step method even when the marker information from selection candidates is ignored in the parameter estimation. Investigations by extensive simulation studies may reveal the benefits of other potential short-cuts. Third, the allele frequencies in the base population are considered known, or at least easily accessible. As illustrated in the results, the parameter estimation seems to be sensitive to the choice of these allele frequencies in a scenario with selection and where the base population itself has not been genotyped. To investigate whether the problems may be related to the strong selection on phenotype for the simulation data set, this analysis was repeated for a simulation with boars selected randomly. Here more sensible parameter estimates were obtained in the sense that $\hat{w} \approx 0$ when allele frequencies were estimated from observed genotypes. For practical dairy cattle evaluations, Misztal et al. [16] investigated the use of a number of different allele frequencies and obtained the best results by using $\rho_j$ = 1/2 for all $j$ but replacing $s = 2\Sigma_j\rho_j(1 - \rho_j) = p/2$ with a another scaling $s$ which in practice was larger than $p/2$. Of course, whether that result is due to selection in this real data set is not known. Further research on the effect of selection and on how to handle appropriately the issue with allele frequencies is needed.

An assumption behind the genomic relationship matrix (2) is that all regions of the genome are equally important for the trait of interest. It is possible to instead use $G(m) \propto (m - p)h(m - p)^T$ where $h$ is a diagonal matrix with known weights $h_{jj} = b_j^2$ with $b_j$s being estimated SNP effects (estimated using for example a non-linear mixture type of model as in [1]). However, incorporating uncertainty on such estimated SNP effects into the method seems less straight-forward.

Considering other types of marker based relationship matrices, then

$$K(M)_{ii'} = \exp(-\sum_j (M_j^i - M_j^{i'})^2 / \phi), \tag{11}$$

with correlation parameter $\varphi$, corresponds to the method in [4] in it's dual formulation as a linear mixed model. For this choice of marker-based relationship matrix, the derivation of $K^*(m^{obs})$ = Var $[g|m^{obs}]$ is also possible, but as shown in Appendix C the form of the result differs from (4) in a number of ways. The implication is that using (4) and (6) with a marker based relationship matrix defined by (11) is possible, but lacks theoretical justification.

## Appendix A

Here the mean and variances of the conditional distribution $[g \mid m^{obs}]$ (with $M^{miss}$ marginalised out) are derived using formulas for conditional expectations, variances and covariances.

The mean vector

$$\mathrm{E}[g \mid m^{obs}] = \mathrm{E}[\mathrm{E}[g \mid m^{obs}, M^{miss}] \mid m^{obs}] = 0,$$

and the variance-covariance matrix

$$\begin{aligned}\mathrm{Var}[g \mid m^{obs}] &= \mathrm{E}[Var[g \mid M^{miss}, m^{obs}] \mid m^{obs}] + \mathrm{Var}[\mathrm{E}[g \mid M^{miss}, m^{obs}] \mid \boldsymbol{m}^{obs}] = \sigma_g^2 \mathrm{E}[\mathrm{G}(M^{miss}, m^{obs}) \mid m^{obs}] \\ &= \frac{\sigma_g^2}{s}\begin{bmatrix} (m^{obs} - p)(m^{obs} - p)^{\mathrm{T}} & (m^{obs} - p)(\mathrm{E}[M^{miss} \mid m^{obs}] - p)^{\mathrm{T}} \\ (\mathrm{E}[M^{miss} \mid m^{obs}] - p)(m^{obs} - p)^{\mathrm{T}} & (\mathrm{E}(M^{miss} \mid \boldsymbol{m}^{obs}] - p)(\mathrm{E}[M^{miss} \mid m^{obs}] - p)^{\mathrm{T}} + \sum_j \mathrm{Var}[M_j^{miss} \mid m^{obs}] \end{bmatrix},\end{aligned}$$

where

$$\mathrm{E}[M_j^{miss} \mid m^{obs}] = 1p_j + A_{21}A_{11}^{-1}(m_j^{obs} - 1p_j),$$

and

$$\mathrm{Var}[M_j^{miss} \mid m^{obs}] = v_j(A_{22} - A_{21}A_{11}^{-1}A_{12}),$$

with

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

and subdivision corresponding to ($M^{obs}$, $M^{miss}$). Using that $\Sigma_j\ v_j = s$, we obtain Var [$g \mid m^{obs}$] = $\sigma_g^2\, G^*(m^{obs})$ where

$$G^*(m^{obs}) = \begin{bmatrix} G(m^{obs}) & G(m^{obs})A_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1}G(m^{obs}) & A_{21}A_{11}^{-1}G(m^{obs})A_{11}^{-1}A_{12} + A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}.$$

In the calculations above it is assumed that the conditional mean $\mathrm{E}[M_j^{miss} \mid m^{obs}] = \mathrm{E}[M_j^{miss} \mid m_j^{obs}]$ and the conditional variance-covariance $\mathrm{Var}[M_j^{miss} \mid m^{obs}] = \mathrm{Var}[M_j^{miss} \mid m_j^{obs}]$, and this is correct since

$$\mathrm{E}[M^{miss} \mid m^{obs}] = p + (I \otimes A_{21}A_{11}^{-1})(m^{obs} - p)$$

$$\mathrm{Var}[M^{miss} \mid m^{obs}] = V \otimes (A_{22} - A_{21}A_{11}^{-1}A_{12})$$

when Var($M$) = $V \otimes A$.
In the main text we assume

$$g \mid m^{obs} \sim N(0, \sigma_g^2 G^*(m^{obs})),$$

where $G^*(m^{obs})$ is defined in (4). However, this is not strictly correct for a non-genotyped animal $i$ where $g_i \mid X \sim N\ (0, X)$ with $X$ here being a random variable with distribution $[\Sigma_j\ (M_{ij} - p_j)^2 | m^{obs}]$. This conditional distribution will never lead to a marginal normal distribution for $g_i$ (the only exception is when $X$ is a constant). The normal distribution of $g|m^{obs}$ is therefore only an approximation.

## Appendix B

In some scenarios the number of genotyped animals not included in the parameter estimation may be large, for example if phenotypes are expensive to obtain and therefore only observed on a small subset of the population. To reduce the computational burden of creating the whole $G_{all}^*\,(m^{obs,other})$ for all animals, a procedure is presented where only a part of this matrix needs to be computed.

For genotyped animals used in the parameter estimation, let $\hat{\tilde{g}}_1$ be the corresponding sub-vector of $\hat{\tilde{g}}$. Estimated breeding values of other genotyped animals not included in the parameter estimation (denoting this subset of animals by index 3) are obtained by

$$\hat{\tilde{g}}_3 = [\tilde{G}_{w,31} \quad \tilde{G}_{w,32}](\tilde{G}_w)^{-1}\hat{\tilde{g}},$$

Where $\tilde{G}_{w,31} = (1 - w)G_{31}^* + wA_{31}$, and $G_{31}^* = G_{all}^*(m^{obs}, m^{other})_{31}$ and $A_{31}$ = $(A_{all})_{31}$ are sub-matrices of the full (containing all animals) genomic and polygenic relationship matrix, respectively. The matrices with index 32 are similarly defined. Since $m^{other}$ does not influence $M^{miss}$ directly,

$$\begin{bmatrix} G_{31}^* & G_{32}^* \end{bmatrix} = (1/s)(m^{other} - p)\left(\begin{bmatrix} m^{obs} \\ \mathrm{E}[M^{miss} \mid m^{obs}] \end{bmatrix} - p\right)^{\mathrm{T}} = G_{31}^*\begin{bmatrix} I & A_{11}^{-1}A_{12} \end{bmatrix}.$$

Considering the polygenic effect, then the assumption that $m^{other}$ does not influence $M^{miss}$ is equivalent to $A_{32} - A_{31}\,A_{11}^{-1}\,A_{12} = 0$. Using this relation we obtain

$$\begin{bmatrix} A_{31} & A_{32} \end{bmatrix} = A_{31}\begin{bmatrix} I & A_{11}^{-1}A_{12} \end{bmatrix}.$$

Hence,

$$\begin{bmatrix} \tilde{G}_{w,31} & \tilde{G}_{w,32} \end{bmatrix} = \tilde{G}_{w,31}\begin{bmatrix} I & A_{11}^{-1}A_{12} \end{bmatrix},$$

and therefore by using (8) and (5) the following form is obtained

$$\hat{\tilde{g}}_3 = \tilde{G}_{w,31}\begin{bmatrix} I & A_{11}^{-1}A_{12} \end{bmatrix}(\tilde{G}_w)^{-1}\hat{\tilde{g}} = \tilde{G}_{w,31}\begin{bmatrix} (G_w)^{-1} & 0 \end{bmatrix}\hat{\tilde{g}} = \tilde{G}_{w,31}(G_w)^{-1}\hat{\tilde{g}}_1. \quad (12)$$

This shows that the GEBVs of such genotyped animals only depend on $\hat{\tilde{g}}_1$. It also shows that only a part of the full genomic relationship matrix for genotyped animals is necessary to compute, since $G_{w,33}$ = (1 - $w$)$G$($m^{other}$) + $wA_{33}$ does not enter into (12).

In some cases the matrix $A_{31}$ may be prohibitive to compute directly due to a large number of animals. In such a case, $\hat{\tilde{g}}_3 = (1 - w)\hat{g}_3 + w\hat{a}_3$, where $\hat{g}_3 = G_{31}^*(G_w)^{-1}\hat{\tilde{g}}_1$ is computed directly and $\hat{a}_3 = A_{31}(G_w)^{-1}\hat{\tilde{g}}_1$ may be obtained as the solution to the sparse system of equations

$$(A_{all})^{-1}\begin{bmatrix} \bar{a}_1 \\ \bar{a}_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} (G_w)^{-1}\tilde{g}_1 \\ 0 \\ 0 \end{bmatrix},$$

where $(A_{all})^{-1}$ is sparse and is computed directly, and $\bar{a}_1$ and $\bar{a}_2$ are dummy variables.

## Appendix C

Here follows the derivation of the extension of the marker-based relationship matrix

$$K(M)_{ii'} = \exp(-\sum_{j}(M_j^i - M_j^{i'})^2 / \phi),$$

to non-genotyped animals.

The extension of the genomic relationship matrix is

$$\begin{aligned} K^*(m^{obs}) &= \mathrm{Var}[g \mid m^{obs}] = \mathrm{E}[Var[g \mid M^{miss}, m^{obs}] \mid m^{obs}] + \mathrm{Var}[\mathrm{E}[g \mid M^{miss}, m^{obs}] \mid m^{obs}] \\ &= \mathrm{E}[K(M^{miss}, m^{obs}) \mid m^{obs}] + 0 = \mathrm{E}[K(M^{miss}, m^{obs}) \mid m^{obs}]. \end{aligned}$$

As written in the discussion, the form of this matrix differs from (4) in a number of ways. First, all diagonal elements $K^*(m^{obs})_{ii} = 1$, and hence $K^*(m^{obs})$ does not simplify to the $A$ matrix when no animals are genotyped. Second, the resulting matrix depends on the off-diagonal elements $v_{jj'}$ of $V$, since for non-genotyped animals $i$ and $i'$ the derivation

$$\mathrm{E}[K(m^{miss}, m^{obs}) \mid m^{obs}]_{ii'} = \prod_{j} \mathrm{E}[\exp(-M_j^i - M_j^{i'})^2 / \phi \mid m^{obs}]$$

requires that $M^1,\ldots, M^p$ are statistically independent (implying that $V$ is a diagonal matrix). Third, the conditional expectation $\mathrm{E}[\exp(-M_j^i - M_j^{i'})^2 / \phi) \mid m^{obs}]$ depends on the distributional assumptions of the model for $M$, not just first and second moments. Fourth, assuming a multivariate normal distribution of $M$, then

$$\mathrm{E}[\exp(-M_j^i - M_j^{i'})^2 / \phi) \mid m^{obs}] = \exp(-\nu^2 / (1 + \tau^2)) / \sqrt{1 + \tau^2},$$

with $\nu = \mathrm{E}[(M_j^i - M_j^{i'}) / \sqrt{\phi} \mid m^{obs}]$ and $\tau^2 = \mathrm{Var}[(M_j^i - M_j^{i'}) / \sqrt{\phi} \mid m^{obs}]$ where these expectations and variances can be computed from the conditional expectations and variances given in Appendix A. The form $\exp(-\nu^2/(1 + \tau^2))/ \sqrt{1 + \tau^2}$ with the variance $\tau^2$ occurring in two places, implies that that the elements in $K^*(m^{obs})$ cannot be expressed in matrix form as in (4) but are on a more complicated form.

### Acknowledgements

The work was part of the project "Svineavl, Genomisk selektion" funded by the Danish Ministry of Food, Agriculture and Fisheries, and Danish Pig Production. Guosheng Su is acknowledged for help in relation to the generation of the simulation study, and Per Madsen is acknowledged for his unselfish work on creating and maintaining the software DMU. A reviewer is thanked for his suggestions on how to improve the presentation.

### Authors' contributions

OFC derived and implemented the methods, created and analysed the simulation study, and wrote the paper. MSL conceived the study, took part in discussions, and provided input to the writing of the paper. Both authors have read and approved the paper.

### Competing interests

The authors declare that they have no competing interests.

Received: 28 September 2009
Accepted: 27 January 2010 Published: 27 January 2010

### References

1. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819-1829.
2. VanRaden PM: **Efficient methods to compute genomic predictions.** *Interbull Bull* 2007, **37**:111-114.
3. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91**:4414-4423.
4. Gianola D, van Kamm BCHM: **Reproducing kernel Hilbert spaces regression methods for genomic prediction of quantitative traits.** *Genetics* 2008, **178**:2289-2303.
5. Legarra A, Aguilar I, Misztal I: **A relationship matrix including full pedigree and genomic information.** *J Dairy Sci* 2009, **92**:4656-4663.
6. Calus MPL, Veerkamp RF: **Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM.** *J Anim Breed Genet* 2007, **124**:362-368.
7. Gianola D, Fernando RL, Stella A: **Genomic-assisted prediction of genetic value with semiparametric procedures.** *Genetics* 2006, **173**:1761-1776.
8. Baruch E, Weller JI: **Incorporation of genotype effects into animal model evaluations when only a small fraction of the population has been genotyped.** *Animal* 2009, **3**:16-23.
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. Gilmour AR, Thompson R, Cullis BR: **Average information REML: an efficient algorithm for parameter estimation in linear mixed models.** *Biometrics* 1995, **51**:1440-1450.
11. Johnson DL, Thompson R: **Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information.** *J Dairy Sci* 1995, **78**:449-456.
12. Lee SH, Werf van der JHJ: **An efficient variance component approach implementing an average REML suitable for combined LD and linkage mapping with a general pedigree.** *Genet Sel Evol* 1995, **38**:25-43.
13. Madsen P, Jensen J: **A users guide to DMU, version 6, release 4.7.** *Manual, Faculty of agricultural science, University of Aarhus* 2008.
14. VanRaden PM, Van Tassel CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: **Invited review: reliability of genomic predictions for North American Holstein bulls.** *J Dairy Sci* 2009, **92**:16-24.
15. Su G, Guldbrandtsen B, Gregersen VR, Lund MS: **Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population.** *J Dairy Sci* 2010.
16. Misztal I, Legarra A, Aguilar I: **Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information.** *Proceedings of the annual meeting EAAP: 24-27 August 2009; Barcelona, Spain* 2009.

doi:10.1186/1297-9686-42-2
**Cite this article as:** Christensen and Lund: **Genomic prediction when some animals
are not genotyped.** *Genetics Selection Evolution* 2010 **42**:2.