J. Dairy Sci. 92:4656–4663
doi:10.3168/jds.2009-2061

A relationship matrix including full pedigree and genomic information

A. Legarra,*[1] I. Aguilar,†‡ and I. Misztal†
*INRA, UR631 SAGA, BP 52627, 32326 Castanet-Tolosan, France
†Department of Animal and Dairy Science, University of Georgia, Athens 30602
‡Instituto Nacional de Investigación Agropecuaria, Las Brujas, Uruguay

ABSTRACT

Dense molecular markers are being used in genetic evaluation for parts of the population. This requires a two-step procedure where pseudo-data (for instance, daughter yield deviations) are computed from full records and pedigree data and later used for genomic evaluation. This results in bias and loss of information. One way to incorporate the genomic information into a full genetic evaluation is by modifying the numerator relationship matrix. A naive proposal is to substitute the relationships of genotyped animals with the genomic relationship matrix. However, this results in incoherencies because the genomic relationship matrix includes information on relationships among ancestors and descendants. In other words, using the pedigree-derived covariance between genotyped and ungenotyped individuals, with the pretense that genomic information does not exist, leads to inconsistencies. It is proposed to condition the genetic value of ungenotyped animals on the genetic value of genotyped animals via the selection index (e.g., pedigree information), and then use the genomic relationship matrix for the latter. This results in a joint distribution of genotyped and ungenotyped genetic values, with a pedigree-genomic relationship matrix **H**. In this matrix, genomic information is transmitted to the covariances among all ungenotyped individuals. The matrix is (semi)positive definite by construction, which is not the case for the naive approach. Numerical examples and alternative expressions are discussed. Matrix **H** is suitable for iteration on data algorithms that multiply a vector times a matrix, such as preconditioned conjugated gradients.
Key words: genetic evaluation, genomic selection, relationship matrix, mixed model

INTRODUCTION

Availability of dense molecular markers of type SNP has lead to the recent introduction of the so-called genome-wide or genomic selection evaluation models. Most such models are based on variants of simultaneous genome-wide association analysis, in which marker or haplotype effects (**a**) are estimated. Differences among methods are mostly on the a priori distribution of **a** (e.g., Meuwissen et al., 2001; Gianola et al., 2006).

Although these methods are very promising for animal breeding, genotyping is not feasible for an entire population because of its high cost or logistical constraints (i.e., culled, slaughtered, or foreign animals). This is of importance, for example, for foreign bulls for which no genotyping is possible. Animals that are genotyped include prospective and old males, and possibly prospective mothers of future candidates (e.g., embryo transfer dams).

As not all animals can be genotyped, a 2- or 3-step procedure has to be followed; first, a regular genetic evaluation is run; then, corrected phenotypes or pseudo-data are used in the second step, where the marker-assisted selection model is effectively applied (Guillaume et al., 2008; VanRaden et al., 2009). These phenotypes are daughter yield deviations (**DYD**) and yield deviations (**YD**) for dairy cattle.

After computation of pseudo-data, genomic or marker-assisted predictions can be obtained by either simultaneously fitting polygenic and QTL effects (Guillaume et al., 2008), or by computing the genomic prediction and combining it with estimated breeding values from the animal model (VanRaden et al., 2009). Genomic predictions can be obtained either by estimating **a** effects caused by markers or by using mixed model equations with a genomic relationship matrix **G** (VanRaden, 2008). This assumes that a priori marker effects are normally distributed with a common variance. Although the assumption is arguable, positing a more complicated prior distribution resulted in little gain in practice (VanRaden et al., 2009). On the other hand, the genomic relationship matrix is simple to interpret and handle.

Received January 26, 2009.
Accepted April 28, 2009.
[1]Corresponding author: andres.legarra@toulouse.inra.fr

Advantages of the multistage system include no change to the regular evaluations and simple steps for predicting genomic values for young genotyped animals. Disadvantages include weighting parameters, such as variance components (Guillaume et al., 2008) or selection index coefficients (VanRaden et al., 2009), and loss of information. Furthermore, the extension to multiple traits is not obvious and tracing back anomalies in a two-step procedure might become very complicated.

As for the loss of information, several problems exist in the use of DYD and YD. These problems are weights (caused by different amount of information in the original data set), bias (caused by selection, for example), accuracy (for animals in small herds), and collinearity (for example, the YD of two cows in the same herd). As for the bias, if genomic selection is used, the expectation of Mendelian sampling in selected animals is not zero (Party and Ducrocq, 2009).

These problems may offset the benefit of marker-assisted selection, particularly for cows (Neuner et al., 2008, 2009). Also, in other species (sheep, swine, beef cattle) or traits (e.g., maternal traits, calving ease) DYD are more difficult to compute or even to define, or they might be poorly estimated—for example, if the contemporary groups are small.

One simplification of the current strategy would be to perform a joint evaluation using all phenotypic, pedigree, and genomic information. A possibility is to impute markers in ungenotyped animals via marker and pedigree information (i.e., linkage analysis) and estimate marker effects once imputation is done. However, in order to get a "best" predictor (in the sense of Henderson (1984), i.e., the conditional expectation), the incertitude in marker imputation, which is very high for most ungenotyped individuals, has to be accounted for via integration over the posterior distribution of marker imputations and marker effects. This can be achieved for example by peeling or Markov chain Monte Carlo (Abraham et al., 2007). However, this is unfeasible for a data set of even medium size when there are many loci or when many markers are missing, and particularly in the presence of loops, which are common in livestock pedigrees.

Another possibility is to use the same methodology as in the current evaluation (i.e., Henderson's mixed model equations) except that the relationship matrix **A** needs to be modified to include the genomic information. The purpose of this study is to provide such a relationship matrix, based on transmissions from genotyped animals to their offspring, or selection indexes from genotyped to ungenotyped animals. This will blend complementary information from recorded pedigree and molecular markers. Computational methods for such a modified numerator relationship matrix, even if complex, can be found in Misztal et al. (2009).

METHODS

Covariance Matrix of Breeding Values Including Genomic Information

Let $\mathbf{u}$ be a vector of genetic effects. Under a polygenic infinitesimal model of inheritance, $\mathrm{Var}(\mathbf{u}) = \mathbf{A}\sigma_u^2$, where $\mathbf{A}$ is the numerator relationship matrix based on pedigree. Consider three types of animals in $\mathbf{u}$: 1) ungenotyped ancestors with breeding values $\mathbf{u}_1$; 2) genotyped animals, with breeding values $\mathbf{u}_2$ (no ancestor is genotyped and phantom parents can be generated if necessary); and 3) ungenotyped animals with breeding values $\mathbf{u}_3$, which might descend from either one of the three types of animals. A particular case is one in which ungenotyped animals are ancestors and progeny of genotyped animals—for example, a bull dam daughter of another bull. They are arbitrarily put in group 1. Then $\mathbf{A}$ can be partitioned as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \\ \mathbf{A}_{33} & \mathbf{A}_{32} & \mathbf{A}_{33} \end{bmatrix}.$$

Let $\mathbf{u}_2 = \mathbf{Za}$, $\mathbf{Z}$ being an incidence matrix and $\mathbf{a}$ the effects of markers. Matrix $\mathbf{Z}$ is centered by allele frequencies (VanRaden, 2008). Then $\mathrm{Var}(\mathbf{u}_2) = \mathbf{ZZ}'\sigma_a^2 = \frac{\mathbf{ZZ}'}{k}\sigma_u^2 = \mathbf{G}\sigma_u^2$, where k is twice the sum of heterozygosities of the markers (VanRaden, 2008).

In some implementations, matrix $\mathbf{G}$ can be seen as an "improved" matrix of relationships (Amin et al., 2007). Villanueva et al. (2005) and Visscher et al. (2006) propose to use a realized matrix of transmissions from parents to offspring in the data, averaging across all positions in the genome; this proposal is impractical in a general manner as genotypes are needed over entire families. VanRaden (2008) discussed how the expectation of $\mathbf{G}$ above is $\mathbf{A}$, the regular numerator relationship matrix, and that $\mathbf{G}$ represents observed, rather than average, relationships. Therefore, it accounts for Mendelian samplings (i.e., it can distinguish full-sibs) and unknown or far relationships. The gain by using $\mathbf{G}$ has been shown (González-Recio et al., 2008; Legarra et al., 2008; VanRaden et al., 2009). In principle, the additive variance using $\mathbf{G}$ is identical to that using $\mathbf{A}$ (Habier et al., 2007).

There is no need for $\mathbf{G}$ to have a particular genetic interpretation in terms of relationships. For example, using the reproducing kernel Hilbert spaces equations of González-Recio et al. (2008), $\mathrm{Var}(\mathbf{u}_2) = \mathbf{K}\sigma^2_\alpha$, where $\mathbf{K}$ is a matrix with "distances" among individuals. Matrix $\mathbf{K}$ can be scaled to $\mathbf{K}^*$ so that $\mathrm{Var}(\mathbf{u}_2) = \mathbf{K}^*\sigma^2_u$ by equating the expectation for the sum of squares of $\mathbf{u}_2$ in the data following polygenic and the reproducing kernel Hilbert spaces models. The expected sum of squares is, respectively (polygenic vs. reproducible kernel Hilbert spaces): $E(\mathbf{u}_2'\mathbf{u}_2) = tr(\mathbf{A}_{22})\sigma^2_u$ and $E(\mathbf{u}_2'\mathbf{u}_2) = tr(\mathbf{K})\sigma^2_\alpha = tr(\mathbf{K}^*)\sigma^2_u$, where tr is the trace operator. In absence of inbreeding, $tr(\mathbf{A}_{22}) = 1$ and thus $\mathbf{K}^* = \mathbf{K}/tr(\mathbf{K})$. Note that matrix $\mathbf{K}$ is also centered, and pseudo-inbreeding coefficients can be extracted from the diagonal of $\mathbf{K}^*$. Of course, by using a reproducing kernel Hilbert spaces model any "genealogical" intuition is lost.

In the following, and for simplicity of notation, it will be assumed that $\sigma^2_u = 1$.

Plug-in G. A simple way to use $\mathbf{G}$ is to plug it into $\mathbf{A}$; this results in the following modified $\mathbf{A}$:

$$\mathbf{A}_\mathrm{g} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{G} & \mathbf{A}_{23} \\ \mathbf{A}_{33} & \mathbf{A}_{32} & \mathbf{A}_{33} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}-\mathbf{A}_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad [1]$$

where $\mathbf{A}_{22}$ has been simply replaced by $\mathbf{G}$. A proposal by Gianola and De los Campos (2008) to come up with predictions of ungenotyped animals from predictions of genotyped ones is to use $\mathbf{A}_{12}\mathbf{G}^{-1}\hat{\mathbf{u}}_2$. Their proposal reduces thus to a selection index by making the assumption that the covariance among individuals is described by $\mathbf{A}_\mathrm{g}$.

Matrix $\mathbf{A}_\mathrm{g}$ is simple to use but not properly constructed. The use of $\mathbf{G}$ potentially modifies covariances in ancestors and descendants of genotyped animals. For example, assume two full-sibs in the genotyped animals whose genomic relationship is 0.6. By using $\mathbf{A}_\mathrm{g}$, it is assumed that average relationship among their daughters is 0.25, whereas in fact it is 0.3.

It can be verified by small numerical examples that $\mathbf{A}_\mathrm{g}$ is indefinite (i.e., some eigenvalues are negative and some positive); the reason is that it is not based in an underlying linear model leading to a matrix crossproduct of the type **T'T** plus a diagonal matrix, like the numerator relationship matrix (Quaas, 1988) or the marker-assisted BLUP (Fernando and Grossman, 1989). Therefore, the statistical background is ill-defined (Searle, 1971; Harville, 1976). A correct statistical inference can only be made if the covariance matrix is positive or semi-positive definite. Matrix $\mathbf{A}_\mathrm{g}$ might lead to correct inferences if the matrix is reasonable and numerical errors are not big. This ought to be checked by simulations.

Modification for Progeny. For this different approach, consider the descendants of genotyped animals. Following the decomposition of $\mathbf{A}$ (i.e., Quaas, 1988), let $\mathbf{P}$ be a matrix containing expected transmissions from ancestors to offspring, that is, with values of 0.5 in the son-dam and son-sire cells. Then $\mathbf{u} = \mathbf{P}\mathbf{u} + \boldsymbol{\varphi}$, where $\boldsymbol{\varphi}$ is a vector of Mendelian samplings and founder effects (Quaas, 1988). The variance of $\boldsymbol{\varphi}$ is indicated as $\mathbf{D}$.

In this particular group of animals:

$$\mathbf{u}_3 = \begin{bmatrix} \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix} + \boldsymbol{\varphi}_3.$$

Then

$$\mathbf{u}_3 = \mathbf{T}_{33}(\mathbf{P}_{32}\mathbf{u}_2 + \mathbf{P}_{31}\mathbf{u}_1 + \boldsymbol{\varphi}_3),$$

which can be seen as a regression equation, and where $\mathbf{T}_{33} = (\mathbf{I} - \mathbf{P}_{33})^{-1}$ (Quaas, 1988).

Then

$$\mathrm{Var}(\mathbf{u}_3) = \mathbf{T}_{33}(\mathbf{P}_{32}\mathbf{G}\mathbf{P}_{32}' + \mathbf{P}_{31}\mathbf{A}_{11}\mathbf{P}_{31}' + \mathbf{P}_{32}\mathbf{A}_{21}\mathbf{P}_{31}' + \mathbf{P}_{31}\mathbf{A}_{12}\mathbf{P}_{32}' + \mathbf{D}_3)\mathbf{T}_{33}'$$

$$\mathrm{Cov}(\mathbf{u}_3, \mathbf{u}_1) = \mathbf{T}_{33}(\mathbf{P}_{31}\mathbf{A}_{11} + \mathbf{P}_{32}\mathbf{A}_{21}) = \mathbf{A}_{31}$$

$$\mathrm{Cov}(\mathbf{u}_3, \mathbf{u}_2) = \mathbf{T}_{33}(\mathbf{P}_{32}\mathbf{G} + \mathbf{P}_{31}\mathbf{A}_{12})$$

Then the covariance matrix becomes:

$$\mathbf{A}_p = \begin{bmatrix} \mathbf{A}_{11} & & \textit{symm} \\ \mathbf{A}_{21} & \mathbf{G} & \\ \mathbf{A}_{31} & \mathbf{T}_{33}\mathbf{P}_{32}\mathbf{G} + \mathbf{P}_{31}\mathbf{A}_{12} & \mathbf{T}_{33}(\mathbf{P}_{31}\mathbf{A}_{11}\mathbf{P}_{31}' + \mathbf{P}_{32}\mathbf{G}\mathbf{P}_{32}' + \mathbf{D}_3)\mathbf{T}_{33}' \end{bmatrix}. \quad [2]$$

Variance caused by Mendelian sampling in $\mathbf{D}_3$ is related to inbreeding in the founders $(\mathrm{Var}(\varphi) = (1/2 - (F_s + F_d)/4)\sigma^2_u)$, where F_s and F_d are inbreeding coefficients of sire and dam; this can be extracted from the diagonal of $\mathbf{G}$ if needed. Otherwise, $\mathbf{D}_3$ is the same as in classical methods. Assuming that $\mathbf{D}_3$ is equivalent in both cases (i.e., parents are not inbred), $\mathbf{A}_\mathrm{p}$ can be formed by appropriately modifying $\mathbf{A}$:

$$\mathbf{A}_p = \mathbf{A} + \begin{bmatrix} 0 & 0 & symm \\ 0 & \mathbf{G} - \mathbf{A}_{22} & \\ 0 & \mathbf{T}_{33}\mathbf{P}_{32}\left(\mathbf{G} - \mathbf{A}_{22}\right) & \mathbf{T}_{33}\mathbf{P}_{31}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{P}'_{31}\mathbf{T}'_{33} \end{bmatrix}.$$

Again, matrix $\mathbf{A}_{\mathrm{p}}$ might not be fully coherent (and indeed might be indefinite) because matrix $\mathbf{G}$ also includes information about the ancestors of genotyped animals. For example, two genotyped animals, say A and B, that have no relationship in the numerator relationship matrix $\mathbf{A}$ might show some relationship in $\mathbf{G}$, because of a common, unrecorded, ancestor. Thus, a relationship can be posited between the ancestors of A and B. Matrix $\mathbf{A}_{\mathrm{p}}$ would work if all founders were genotyped (e.g., in a nucleus scheme); in this case, the system is fully coherent. For practical purposes, $\mathbf{A}_{\mathrm{p}}$ might be reasonable because most information for sire evaluation is contained in the progeny, and not in the ancestors.

Modification for the Whole Pedigree. There is no distinction between ancestors or progeny of genotyped animals in this method; animals in 1 are ungenotyped, whereas animals in 2 are genotyped.

Then $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ with inverse $\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}$.

Based on selection index theory and properties of the normal distribution, conditionally on pedigree (Sorensen and Gianola, 2002, p. 254; Gelman et al., 2004, p. 86), the distribution of breeding values of ungenotyped animals, conditioned on breeding values of genotyped animals, is:

$$p\left(\mathbf{u}_1 \middle| \mathbf{u}_2\right) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right) \qquad [3]$$

(which is the best predictor if we assume normality), or,

$$\mathbf{u}_1 = E\left(\mathbf{u}_1 \middle| \mathbf{u}_2\right) + \varepsilon = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2 + \varepsilon,$$
$$\mathrm{Var}\left(\varepsilon\right) = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} = \left(\mathbf{A}^{11}\right)^{-1}.$$

This can be seen just as a regression equation. Now substitute $\mathbf{u}_2 = \mathbf{Za}$. Then

$$\mathbf{u}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{Za} + \varepsilon$$

so that

$$\mathrm{Var}\left(\mathbf{u}_1\right) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}.$$

This can be reduced to

$$\mathrm{Var}\left(\mathbf{u}_1\right) = \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$$

$$\mathrm{Var}\left(\mathbf{u}_2\right) = \mathbf{ZZ}'/k = \mathbf{G} \text{ and}$$

$$\mathrm{Cov}\left(\mathbf{u}_1, \mathbf{u}_2\right) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}.$$

Note that $\mathbf{A}_{12}\mathbf{A}_{22}^{-1} = -\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}$. This might be convenient for computation as $\mathbf{A}^{11}$ and $\mathbf{A}^{12}$ are sparse and simpler to create, following Henderson's rules, than $\mathbf{A}_{12}$ and $\mathbf{A}_{22}$.

Let us now call $\mathbf{H}$ the covariance matrix of breeding values including genomic information. This is:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}. \qquad [4]$$

Matrix $\mathbf{H}$ is identical to $\mathbf{A}_{\mathrm{p}}$ if all founders are genotyped, because in that case $\mathbf{A}_{12} = \mathbf{T}_1\mathbf{P}_{12}\mathbf{A}_{22}$. By construction, this matrix is semipositive or positive definite, which implies that the statistical background is sound (e.g., Harville, 1976). It is possible to come up with rules for inverting $\mathbf{H}$, in the lines of Wang et al. (1995). However, $\mathbf{H}^{-1}$ might be difficult to invert because full positive definiteness of $\mathbf{G}$ is not guaranteed and therefore their inverse (which is needed to get $\mathbf{H}^{-1}$) might not exist, or might be very ill-conditioned. Positive-definitiveness of $\mathbf{H}$ is not necessary for prediction (Harville, 1976; Henderson, 1984). Two alternative expressions for $\mathbf{H}$ that might be computationally convenient are:

$$\mathbf{H} = \begin{bmatrix} \left(\mathbf{A}^{11}\right)^{-1} + \left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}\mathbf{G}\mathbf{A}^{21}\left(\mathbf{A}^{11}\right)^{-1} & -\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}\mathbf{G} \\ -\mathbf{G}\mathbf{A}^{21}\left(\mathbf{A}^{11}\right)^{-1} & \mathbf{G} \end{bmatrix} \qquad [5]$$

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G} - \mathbf{A}_{22}\right) \\ \left(\mathbf{G} - \mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}$$
$$= \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & 0 \\ 0 & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix}\left[\mathbf{G} - \mathbf{A}_{22}\right]\begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & 0 \\ 0 & \mathbf{I} \end{bmatrix}. \qquad [6]$$

Computational Suggestions. An outline of some ideas for solving mixed model equations for big data sets will be shown here including matrix $\mathbf{H}$ (similar algorithms can be conceived for $\mathbf{A}_p$ and $\mathbf{A}_g$), whereas the companion paper by Misztal et al. (2009) gives more details and examples. Henderson (1984, 1985) gave expressions for the computation of the mixed model equations without use of the inverse of the relationship matrix. These expressions are valid for singular matrices (Harville, 1976), which might be the case for $\mathbf{G}$ as it was in our experience (unpublished). For the random effects the equation is: $\left[\mathbf{HZ}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{I}\right]\hat{\mathbf{u}} = \mathbf{W}\hat{\mathbf{u}} = \mathbf{HZ}'\mathbf{R}^{-1}\mathbf{y}$.

This equation can be solved, in methods such as preconditioned conjugated gradients, by repeatedly multiplying matrix $\mathbf{W}$ times the current guess of $\mathbf{u}$. This requires computing the product $\mathbf{Hq}$, where $\mathbf{q}$ is a vector. This is feasible using [6]. Whereas $\mathbf{G}$ is created explicitly, only $\mathbf{A}^{-1}$ can be created efficiently; $\mathbf{A}_{22}$ can be created from pedigree by computing single elements of the $\mathbf{A}$ matrix using recursive (Aguilar and Misztal, 2008) or indirect (Colleau, 2002) algorithms. For large data files, matrix $\mathbf{G}$ can be computed in parallel or even using iteration on data on genotype files. It will be assumed that $\mathbf{A}_{22}$ and $\mathbf{G}$ can be computed and stored in core. First, $\mathbf{Aq}$ can be computed by Colleau's (2002) indirect algorithm by reading twice the pedigree file without explicitly creating $\mathbf{A}$. This algorithm works by reading a pedigree twice. The other part is a product of the form $\mathbf{NQRSVq}$. This product can be computed as $\mathbf{N}(\mathbf{Q}(\mathbf{R}(\mathbf{S}(\mathbf{V}(\mathbf{q})))))$. The only difficult parts are the computations of $\mathbf{s} = \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{t}_1$, where $\mathbf{t}_1$ is a vector of size equal to the number of ungenotyped animals, and its symmetric product of the form $\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$. The product $\mathbf{p} = \mathbf{A}_{21}\mathbf{t}_1$ can be found as follows. Let be the product $\mathbf{A}\begin{bmatrix}\mathbf{t}_1\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{t}_1\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{11}\mathbf{t}_1\\ \mathbf{A}_{21}\mathbf{t}_1\end{bmatrix} = \begin{bmatrix}\mathbf{z}\\ \mathbf{y}\end{bmatrix}$, whose result $\mathbf{y}$ is needed. Now let $\mathbf{A}^*$ be the ordered relationship matrix (parents before offspring), and $\mathbf{x}$ a vector containing the reordered elements in $\mathbf{t}_1$ and zero otherwise (i.e., the values in $\mathbf{x}$ corresponding to animals in $\mathbf{A}_{22}$ are zero). Then, the product $\mathbf{A}^*\mathbf{x}$ can be computed by solving the system of equations $\mathbf{A}^{*-1}\mathbf{y}^* = \mathbf{x}$ by Colleau's algorithm and rearranging $\mathbf{y}^*$ into $\mathbf{z}$ and $\mathbf{y}$.

The product $\mathbf{s} = \mathbf{A}_{22}^{-1}\mathbf{p}$ can be computed directly if $\mathbf{A}_{22}^{-1}$ has been previously computed; or done by solving $\mathbf{A}_{22}\mathbf{s} = \mathbf{p}$ if it has not. Both operations have quadratic cost on the number of genotyped animals, say n. Even if $\mathbf{A}_{22}$ cannot be stored, solving $\mathbf{A}_{22}\mathbf{s} = \mathbf{p}$ can in principle be done by an iterative solver and repeated use of the Colleau's algorithm to compute the successive products $\mathbf{A}_{22}\mathbf{s}$. The opposite (multiplication followed by indirect algorithm) strategy can be applied in computing the product with $\mathbf{N}$. Product by $\mathbf{S}$ will involve n^2 operations. If $\mathbf{G}$ is smaller than $\mathbf{Z}$, products can be computed as $\mathbf{s} = \mathbf{Gp} = \mathbf{Z}(\mathbf{Z}'\mathbf{p})/k$ at a cost of $3nm$ (m being the number of markers). Overall, one iteration of the full algorithm involves reading the pedigree file 6 times, plus a number of operations being several times n^2 or $3nm$. For example, for 10 million animals in pedigree and $n = 10{,}000$ genotyped individuals, computing time per iteration will be roughly proportional to n^2. Thus, solving the mixed model equations may be feasible even for large pedigrees. More detailed explanations on the algorithms and preliminary studies of their performances can be found in the companion paper by Misztal et al. (2009).

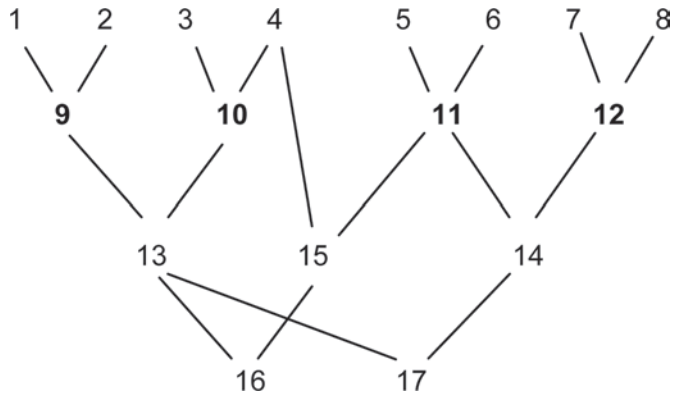

Figure 1. Example pedigree. Genotyped animals are in bold.

Example

Consider the pedigree in Figure 1. Animals 1 to 8 are unrelated founders, whereas animals 9 to 12 are genotyped. As an example, let $\mathbf{G}$ be a matrix with 1 on the diagonal and 0.7 otherwise (i.e., all animals are related although their founders are supposedly unrelated). The regular numerator relationship matrix $\mathbf{A}$ is in Table 1; only a slight modification is needed to get $\mathbf{A}_g$ (not shown). The modified $\mathbf{A}_p$, for progeny, is in Table 2, and the pedigree modified $\mathbf{H}$ is in Table 3. Even for this small example, $\mathbf{A}_p$ is indefinite, whereas $\mathbf{H}$ is positive definite.

It can be seen that in the latter, the relationships among genotyped individuals are projected backward and forward. The backward projection implies, for example, that parents of 9 and 10 are related, and 1 and 2 are not. In fact other possibilities exist (for example, that 2 and 3 were related but not 1 and 4), but the selection index gives a parsimonious solution. This is not the case in $\mathbf{A}_p$, where there is no backward projection. The nonexistence of this backward projection makes

Table 1. Numerator relationship matrix **A** for the pedigree in Figure 1[1]

1.00								0.50				0.25			0.13	0.13
	1.00							0.50				0.25			0.13	0.13
		1.00							0.50			0.25			0.13	0.13
			1.00						0.50			0.25		0.50	0.38	0.13
				1.00						0.50			0.25	0.25	0.13	0.13
					1.00					0.50			0.25	0.25	0.13	0.13
						1.00					0.50		0.25			0.13
							1.00				0.50		0.25			0.13
0.50	0.50							**1.00**				0.50			0.25	0.25
		0.50	0.50						**1.00**			0.50		0.25	0.38	0.25
				0.50	0.50					**1.00**			0.50	0.50	0.25	0.25
						0.50	0.50				**1.00**		0.50			0.25
0.25	0.25	0.25	0.25					0.50	0.50			1.00		0.13	0.56	0.50
				0.25	0.25	0.25	0.25			0.50	0.50		1.00	0.25	0.13	0.50
			0.50	0.25	0.25				0.25	0.50		0.13	0.25	1.00	0.56	0.19
0.13	0.13	0.13	0.38	0.13	0.13			0.25	0.38	0.25		0.56	0.13	0.56	1.06	0.34
0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.25	0.25	0.25	0.25	0.50	0.50	0.19	0.34	1.00

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold. Matrix $\mathbf{A}_g$ is obtained by setting the out-of-diagonal coefficients of genotyped animals to 0.7.

$\mathbf{A}_p$ for 1 to 12 indefinite, as the covariance structure it defines is ill-posed.

Also, in comparison to **A**, it can be seen that inbreeding coefficients appear in descendants of genotyped animals as these are related.

DISCUSSION

The system in [6] might also be expressed as if the overall genetic value was the sum of 2 different genetic values: the one in the infinitesimal model plus a difference whose covariance matrix is $\mathbf{G} - \mathbf{A}_{22}$. In the naive approach, this difference is not correctly accounted for in the relatives. If $\mathbf{G} = \mathbf{A}_{22}$ (which will not happen in practice), matrices **A** and **H** are identical as expected. Further, this shows that genetic variance in the population is the same on average (i.e., there is no artificial inflation). These are of course desirable properties.

The proposed matrix **H** is based on selection index principles or, equivalently, in assumptions of **A** being multivariate normal. Conditioning on breeding values of genotyped animals in [3] allowed us to develop a full multivariate distribution **H**. Thus, matrix **H** has been constructed from the joint density $p(\mathbf{u}_1,\mathbf{u}_2) = p(\mathbf{u}_1|\mathbf{u}_2)\, p(\mathbf{u}_2)$, where $p(\mathbf{u}_2)$ is obtained from genomic data. This distribution includes desirable aspects well known in genetic evaluation: the fact that sons inherit half their parents (as in the descendants of genotyped animals) and the notion of selection index (which is included in BLUP). So, these aspects are indeed used in the 2-step evaluation.

It is hard to envision other possibilities as it is not simple to come up with an underlying model and set up a probability distribution. For example, the "intuitive" expression $\hat{\mathbf{u}}_2 \mid \hat{\mathbf{u}}_1 = \mathbf{A}_{12}\mathbf{G}^{-1}\hat{\mathbf{u}}_1$ follows the logic of a se-

Table 2. Modified relationship matrix $\mathbf{A}_p$ including genomic information for genotyped animals and their progeny for the pedigree in Figure 1[1]

1.00								0.50				0.25			0.13	0.13
	1.00							0.50				0.25			0.13	0.13
		1.00							0.50			0.25			0.13	0.13
			1.00						0.50			0.25		0.50	0.38	0.13
				1.00						0.50			0.25	0.25	0.13	0.13
					1.00					0.50			0.25	0.25	0.13	0.13
						1.00					0.50		0.25			0.13
							1.00				0.50		0.25			0.13
0.50	0.50							**1.00**	**0.70**	**0.70**	**0.70**	0.85	0.70	0.35	0.60	0.78
		0.50	0.50					**0.70**	**1.00**	**0.70**	**0.70**	0.85	0.70	0.60	0.73	0.78
				0.50	0.50			**0.70**	**0.70**	**1.00**	**0.70**	0.70	0.85	0.50	0.60	0.78
						0.50	0.50	**0.70**	**0.70**	**0.70**	**1.00**	0.70	0.85	0.35	0.53	0.78
0.25	0.25	0.25	0.25					0.85	0.85	0.70	0.70	1.35	0.70	0.48	0.91	1.03
				0.25	0.25	0.25	0.25	0.70	0.70	0.85	0.85	0.70	1.35	0.43	0.56	1.03
			0.50	0.25	0.25			0.35	0.60	0.50	0.35	0.48	0.43	1.00	0.74	0.45
0.13	0.13	0.13	0.38	0.13	0.13			0.60	0.73	0.60	0.53	0.91	0.56	0.74	1.33	0.74
0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.78	0.78	0.78	0.78	1.03	1.03	0.45	0.74	1.53

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold.

Table 3. Modified relationship matrix **H** including genomic information for genotyped animals and all relatives for the pedigree in Figure 1[1]

1.00		0.18	0.18	0.18	0.18	0.18	0.18	0.50	0.35	0.35	0.35	0.43	0.35	0.26	0.34	0.39
	1.00	0.18	0.18	0.18	0.18	0.18	0.18	0.50	0.35	0.35	0.35	0.43	0.35	0.26	0.34	0.39
0.18	0.18	1.00		0.18	0.18	0.18	0.18	0.35	0.50	0.35	0.35	0.43	0.35	0.18	0.30	0.39
0.18	0.18		1.00	0.18	0.18	0.18	0.18	0.35	0.50	0.35	0.35	0.43	0.35	0.68	0.55	0.39
0.18	0.18	0.18	0.18	1.00		0.18	0.18	0.35	0.35	0.50	0.35	0.35	0.43	0.34	0.34	0.39
0.18	0.18	0.18	0.18		1.00	0.18	0.18	0.35	0.35	0.50	0.35	0.35	0.43	0.34	0.34	0.39
0.18	0.18	0.18	0.18	0.18	0.18	1.00		0.35	0.35	0.35	0.50	0.35	0.43	0.26	0.31	0.39
0.18	0.18	0.18	0.18	0.18	0.18		1.00	0.35	0.35	0.35	0.50	0.35	0.43	0.26	0.31	0.39
0.50	0.50	0.35	0.35	0.35	0.35	0.35	0.35	**1.00**	**0.70**	**0.70**	**0.70**	0.85	0.70	0.53	0.69	0.78
0.35	0.35	0.50	0.50	0.35	0.35	0.35	0.35	**0.70**	**1.00**	**0.70**	**0.70**	0.85	0.70	0.60	0.73	0.78
0.35	0.35	0.35	0.35	0.50	0.50	0.35	0.35	**0.70**	**0.70**	**1.00**	**0.70**	0.70	0.85	0.68	0.69	0.78
0.35	0.35	0.35	0.35	0.35	0.35	0.50	0.50	**0.70**	**0.70**	**0.70**	**1.00**	0.70	0.85	0.53	0.61	0.78
0.43	0.43	0.43	0.43	0.35	0.35	0.35	0.35	0.85	0.85	0.70	0.70	1.35	0.70	0.56	0.96	1.03
0.35	0.35	0.35	0.35	0.43	0.43	0.43	0.43	0.70	0.70	0.85	0.85	0.70	1.35	0.60	0.65	1.03
0.26	0.26	0.18	0.68	0.34	0.34	0.26	0.26	0.53	0.60	0.68	0.53	0.56	0.60	1.18	0.87	0.58
0.34	0.34	0.30	0.55	0.34	0.34	0.31	0.31	0.69	0.73	0.69	0.61	0.96	0.65	0.87	1.41	0.80
0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.78	0.78	0.78	0.78	1.03	1.03	0.58	0.80	1.53

[1]Cells with 0 are empty to show the pattern. Coefficients for genotyped animals are in bold.

lection index (or a multivariate normal distribution), but the covariances of $\mathbf{u}_1$ and $\mathbf{u}_2$ do not account for **G** as they should. It is not coherent to use **G** to derive $\text{Var}(\mathbf{u}_2)$ and not to derive $\text{Cov}(\mathbf{u}_1,\mathbf{u}_2)$. These covariances can be derived for descendants using the transmission vectors **P** and **T** as shown above, including **G** in the expression; however, it is more difficult to come up with a similar expression for ancestors. The selection index used as a conditional distribution overcomes this problem and accounts for **G** to generate the covariance of $\mathbf{u}_1$ and $\mathbf{u}_2$. This resulted in a parsimonious inclusion of all information (full pedigree and genomic relationships).

All of these assumptions are actually applied in the 2- or 3-step procedure for genomic selection mentioned previously, but as we discussed, information is lost by doing the steps procedure. A full relationship matrix would allow a joint evaluation and all the information would be accounted for automatically. We have also sketched how computations could be feasible in practice. Some aspects, like computation of reliabilities, deserve further research.

ACKNOWLEDGMENTS

Discussions with P. VanRaden (USDA, Beltsville, MD), C. Robert-Granié (INRA), and S. Neuner (Bavarian State Research Center for Agriculture) are gratefully acknowledged. Thanks to D. Gianola and G. De los Campos (University of Wisconsin-Madison) for sharing the unpublished article with us. Also acknowledged is the encouragement to pursue this study by T. Lawlor (Holstein Association) and the financial support by the Holstein Association (I. Misztal and I. Aguilar) and to the EADGENE network of excellence and ANR project AMASGEN (Legarra). A visit of A. Legarra to the University of Georgia was financed by Maison de Relations Internationales (INRA) and the Holstein Association of America. Two reviewers made very constructive comments.

REFERENCES

Abraham, K. J., L. R. Totir, and R. L. Fernando. 2007. Improved techniques for sampling complex pedigrees with the Gibbs sampler. Genet. Sel. Evol. 39:27–38.

Aguilar, I., and I. Misztal. 2008. Recursive algorithm for inbreeding coefficients assuming non-zero inbreeding of unknown parents. J. Dairy Sci. 91:1669–1672.

Amin, N., C. M. van Duijn, and Y. S. Aulchenko. 2007. A genomic background based method for association analysis in related individuals. PLoS One 2:e1274.

Colleau, J. J. 2002. An indirect approach to the extensive calculation of relationship coefficients. Genet. Sel. Evol. 34:409–421.

Fernando, R. L., and M. Grossman. 1989. Marker assisted prediction using best linear unbiased prediction. Genet. Sel. Evol. 21:467–477.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton, FL.

Gianola, D., and G. De los Campos. 2008. Inferring genetic values for quantitative traits non-parametrically. Genet. Res. 90:525–540.

Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776.

González-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa, and S. Avendaño. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178:2305–2313.

Guillaume, F., S. Fritz, D. Boichard, and T. Druet. 2008. Short communication: correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. J. Dairy Sci. 91:2520–2522.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Harville, D. 1976. Extension of the Gauss-Markov theorem to include the estimation of random effects. Ann. Stat. 4:384–395.

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph, Canada.

Henderson, C. R. 1985. Best linear unbiased prediction using relationship matrices derived from selected base populations. J. Dairy Sci. 68:443–448.

Legarra, A., C. Robert-Granié, E. Manfredi, and J. M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611–618.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. J. Dairy Sci. 92:4648–4655.

Neuner, S., C. Edel, R. Emmerling, G. Thaller, and K.-U. Götz. 2009. Precision of genetic parameters and breeding values estimated in marker assisted BLUP genetic evaluation. Genet. Sel. Evol. 41:26.

Neuner, S., R. Emmerling, G. Thaller, and K.-U. Götz. 2008. Strategies for estimating genetic parameters in marker-assisted best linear unbiased predictor models in dairy cattle. J. Dairy Sci. 91:4344–4354.

Party, C., and V. Ducrocq. 2009. Bias due to genomic selection. Interbull Bull. 39. Uppsala, Sweden.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71:1338–1345.

Searle, S. R. Linear Models. 1971. John Wiley, New York, NY.

Sorensen, D. A., and D. Gianola. 2002. Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. Springer-Verlag, New York, NY.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

Villanueva, B., R. Pong-Wong, J. Fernández, and M. A. Toro. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J. Anim. Sci. 83:1747–1752.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu, B. K. Cornes, G. W. Montgomery, and N. G. Martin. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2:e41.

Wang, T., R. L. Fernando, S. Vanderbeek, M. Grossman, and J. A. M. Van Arendonk. 1995. Covariance between relatives for a marked quantitative trait locus. Genet. Sel. Evol. 27:251–274.