Genetic evaluation including phenotypic, full pedigree and genomic data

by

Ignacio Aguilar

(Under the direction of Ignacy Misztal)

Abstract

Genomic evaluations could be obtained using a unified methodology that combines phenotypic, pedigree and genomic information. A national single-step approach (**SSP**) for a comprehensive information (phenotype, pedigree and genotype markers) genetic evaluation was developed for final score of US Holsteins. Data included final scores recorded from 1955 to 2009 for 6,232,548 Holsteins cows. BovineSNP50 genotypes from the Cooperative Dairy DNA Repository were available for 6,508 bulls. Analyses used a repeatability animal model as is currently used for the national US evaluation. Analyses included pedigree and genomic-based relationships matrices. Full data sets and a subset of records (final scores up to 2004) were used to estimate the increase in accuracy due to genomic information. Also, comparisons include a multiple-step approach for genomic selection. The SSP genetic evaluation with the pedigree relationship matrix augmented with genomic information provided genomic predictions with accuracy and bias comparable to multiple-step procedures.

The implementation of such SSP requires the inverse of a join relationship matrix based on pedigree and genomic relationships. A second study investigated efficient computing options for creating relationship matrices based on genomic markers and pedigree information as well as their inverses. A matrix of incidence of SNP marker information was simulated for a

panel of 40K SNPs. The number of genotyped animals varied from 1,000 to 30,000. Efficient methods to create the matrices used in the unified approach are presented. Optimizations can be obtained either by modifications of the existing code or by the use of automatic optimizations provided by open source or third-party libraries.

The third study evaluated the feasibility and accuracy of multiple trait evaluation for conception rate (CR) defined as outcomes of all inseminations in US Holsteins using all available phenotypic, pedigree and genomic information. Genetic evaluations used a national data set and a multiple trait model. The evaluations were obtained by regular BLUP or by the SSP, using genomic information. The R^2 obtained with the SSP were almost doubled compared to BLUP. Computing with SSP took 33% more time than with BLUP. A multiple trait evaluation of CR using the genomic information is possible and advantageous.

INDEX WORDS: BLUP, genomic selection, SNP, genetic evaluation, computing methods, relationship matrix

Genetic evaluation including phenotypic, full pedigree and genomic data

by

Ignacio Aguilar

Ing. Agr. Universidad de la Republica, Uruguay, 1995

M.S., The University of Georgia, 2008

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2010

Genetic evaluation including phenotypic, full pedigree and genomic data

by

Ignacio Aguilar

Approved:

Major Professor: Ignacy Misztal

Committee: Romdhane Rekaya
J. Keith Bertrand
Shogo Tsuruta
George R. Wiggans

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2010

Acknowledgments

I would like to express my sincere gratitude to Dr. Ignacy Misztal for his guidance, supervision and economic support. It was a pleasure and a privilege to work with him during my academic program.

I would also like to thank the other members of my committee: Drs. J. Keith Bertrand, Shogo Tsuruta, George Wiggans and Romdhane Rekaya for their advice and assistance.

Special recognition to the Instituto Nacional de Investigacion Agropecuaria (INIA) Uruguay, for allowing me to do these studies and for their financial support.

I want to acknowledge the friendship and assistance from Drs. Juan Pablo Sanchez, Ching-Yi Chen, Luiz O. Silva and Selma Forni and from the current and past graduate students from the Animal Breeding and Genetics group. I appreciate the help of Jamie Williams and Regina Simeone in editing, as well as of Mike Kelly with the informatics support.

The assistance and discussions with Dr. Andres Legarra are appreciated. Thanks to Dr. Paul VanRaden for sharing his programs and to Dr. Rodrigo Villaroel for giving me helpful tips for making this manuscript in LaTeX.

I would like to thank my friends and colleagues in Uruguay, especially Olga Ravagnolo, Gabriel Ciappesoni, Mario Lema and Diego Gimeno for their support.

To my parents, my sister, brother and family, for their constant support.

Finally I want to express my special thanks to my wife Andrea, for her unconditional encouragement, patience and support and to my son Pedro; his happiness and love makes everything better.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Traditional genetic evaluations use phenotypic and pedigree information to predict breeding values of selection candidates for economically important traits. In recent years, the availability of high-density markers of type single-nucleotide polymorphisms (**SNP**) and the cost-effective genotyping has led to the so called genome-wide or genomic selection methods. Genomic selection can be defined as a form of marked-assisted selection, where dense genetic markers, covering the whole genome are in linkage disequilibrium with quantitative trait loci. Results from simulation studies and more recently using field data, have shown that substantial increases in accuracy could be obtained compared to a regular genetic evaluation for animals with no records (i.e. young bulls in dairy). This can improve the genetic gain by reducing the generation interval.

Genomic breeding values usually are obtained by estimating the effect of each of the genetic makers, and then summing their effect over all the markers. The SNP marker effects can be estimated with different assumptions regarding the prior distribution of such effects. Assuming a normal prior distribution with constant variance for each marker effect results in a method known as GBLUP. Genomic breeding values can also be estimated with a simple model that includes a genomic relationship matrix derived from genotypes and variances of the SNP marker effects. Such a matrix includes information on the Mendelian sampling deviations.

In general, not all animals in a population are genotyped, and multiple breeding values can be computed. These include: regular estimated breeding values from phenotypic and pedigree data, and genomic breeding values for the genotyped animals. Genomic evaluations

are currently calculated with a multiple-step procedure in dairy cattle. A typical evaluation requires 1) traditional evaluation with an animal model, 2) extraction of pseudo-observations such as deregressed evaluations or daughter deviations, 3) estimation of genomic effects for genotyped animals, and possibly 4) combining the genomic index with traditional parent averages and EBV.

The best approach for genomic evaluations could be by a unified approach, where all available information (pedigree, phenotypic data and genomic markers) is considered simultaneously. This could eliminate a number of assumptions and parameters, and possibly deliver more accurate genomic evaluations than multiple-step procedures. Therefore, the objective of these studies are 1) to use a single-step procedure including phenotypic, pedigree and marker information, for genomic evaluation in a national evaluation setting and to compare its performance to a multiple-step genomic evaluation procedure, 2) to study efficient computing options to create relationship matrices based on genomic markers and pedigree information as well as their inverses, 3) to study the feasibility and accuracy of multiple trait evaluation for a lowly heritable trait such as the outcome of artificial insemination.

Chapter 2

Review of Literature

Traditional genetic evaluations use phenotypic and pedigree information to predict breeding values of selection candidates for traits of economical importance. In a review, Hill (2008) showed that the genetic improvement is successful in several domestic species.

In general, genes with known polymorphisms that affect quantitative traits do not add to selection based on estimated breeding values (**EBV**) from the pedigree and phenotypic information (Goddard, 2009). The author presented four reasons supporting his conclusions based on several studies. First, the traditional selection based on EBV is effective. Second, there are many genes that affect a trait so the variance explained by each gene is small. Third, given that traits are controlled by many genes, the estimates of their effects are small and therefore it is hard to have accurate estimates. Finally, few genes are known to be responsible for large variation in important traits.

In recent years, the availability of high-density markers of type single-nucleotide polymorphisms (**SNP**) as well as the cost-effective genotyping led to genome-wide or genomic selection methods (Meuwissen et al., 2001). Genomic selection can be defined as a form of marked-assisted selection, where dense genetic markers that cover the whole genome are in linkage disequilibrium with quantitative trait loci (**QTL**) (Meuwissen et al., 2001).

In humans, results from the International HapMap Consortium identified over 3.1 million SNPs (Frazer et al., 2007); assays of 500,000 SNPs have currently been used in genome-wide association (**GWAs**) studies (e.g. Weedon et al. (2008)). In cattle, an assay interrogating approximately 57,000 SNP loci was developed (Van Tassell et al., 2008; Matukumalli et al., 2009) and is commercially available with the Illumina BovineSNP50 BeadChip (Illumina

Inc., San Diego, CA). A subset of SNPs from this chip were selected and are being used in a national genomic evaluation of dairy cattle in North America (Wiggans et al., 2009).

Results from simulation studies that considered the SNP information (Meuwissen et al., 2001; VanRaden, 2008; Solberg et al., 2008; Habier et al., 2007) show that a substantial increase in accuracy can be obtained compared to a regular genetic evaluation where no information was available (i.e. young bulls in dairy). Genomic selection increases the realized genetic gain, reduces the generation interval, and reduces the cost of testing bulls by approximately 90% (Schaeffer, 2006). Konig et al. (2009) found benefits from genomic breeding programs due to the substantial reduction in the generation interval, to increasing the accuracy of estimated breeding values, and to increasing the selection intensity of cow sires.

Several studies using real data were carried out to asses the accuracy of genomic selection in animal and plant species. These studies involved mice (Legarra et al., 2008; de los Campos et al., 2009), chickens (Gonzalez-Recio et al., 2009), wheat (de los Campos et al., 2009) and several other plant species (Lorenzana and Bernardo, 2009). Studies in dairy cattle genomic selection included several populations: in North America (VanRaden et al., 2009b), Australia (Hayes et al., 2009c,a), Canada (Van Doormaal et al., 2009), New Zealand (Harris et al., 2008), Norway (Luan et al., 2009) and Denmark (Su et al., 2010).

Genomic breeding values are usually obtained by estimating the effect of each of the genetic markers and then summing their effect over all markers (Meuwissen et al., 2001). The SNP marker effects can be estimated with different assumptions regarding the prior distribution of such effects (Meuwissen et al., 2001). Meuwissen et al. (2001) defined two Bayesian methods using different types of prior distribution for the marker variance. The first method (**BayesA**) uses an inverted chi-square distribution for the marker variance, and the second method (**BayesB**) uses a prior that has a high density of zeros, allowing some markers to have a null effect. Gianola et al. (2009) discussed theoretical statistical concepts of the methods presented by Meuwissen et al. (2001) and suggested an alternative methodology. VanRaden (2008) also presented non-linear models to estimate marker effects,

which are analogous to BayesA and BayesB of Meuwissen et al. (2001). Different authors proposed alternative methods to estimate marker effects: semi-parametric methods (Gianola et al., 2006; Gianola and De los Campos, 2008), Bayesian Lasso (de los Campos et al., 2009) and variable selection methodology (Verbyla et al., 2009).

Assuming a normal prior distribution with constant variance for marker effects results in the method known as GBLUP (Meuwissen et al., 2001; Habier et al., 2007; VanRaden, 2008). Genomic breeding values also can be estimated with a simple model that includes a genomic relationship matrix derived from genotypes and variances of the SNP marker effects (Nejati-Javaremi et al., 1997; VanRaden, 2008; Habier et al., 2007). Using a realized relationship matrix, genomic selection will exploit the Mendelian sampling deviations from the average relationship matrix based on pedigree information (Goddard, 2009). Both methods are equivalent except for numerical properties (VanRaden, 2007).

Nejati-Javaremi et al. (1997) proposed to use total allelic relationships as an alternative to pedigree-based relationships. They defined the total allelic identity between two individuals as:

$$TA_{xy} = \frac{\sum_{l=1}^{L} TA_l}{L} = \frac{\sum_{l=1}^{L} (\frac{\sum_{i=1}^{2}\sum_{j=1}^{2} I_{lij}}{2})}{L}$$

where I_{lij} is the identity at locus l for the i^{th} allele of the first individual (x) with the j^{th} allele of the second individual (y) taking the value 1 if both alleles are the same and 0 if not; and L is the number of markers. Using two simulated populations, one population with random mating and another under selection, they found that breeding values estimated using relationships based on the marker information were more accurate and resulted in higher response to selection. The authors concluded that accounting for the identity by state and for deviations from the average relationships based on pedigree increased accuracy of the estimated breeding values.

Genomic relationships ($\mathbf{G}$) could also be created as follows (VanRaden, 2008):

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{k},$$

where $\mathbf{Z}$ is an incidence matrix for SNP markers and k is a scaling parameter.

Elements of $\mathbf{Z}$ are:

$$z_{ij} = \begin{cases} 0 - 2p_j & \text{if homozygous 11} \\ 1 - 2p_j & \text{if heterozygous 12 or 21,} \\ 2 - 2p_j & \text{if homozygous 22} \end{cases}$$

for animal i and SNP j with allele frequency p_j.

Different allele frequencies could be used: estimates from the current or the base population, or constant frequency (i.e. $p_j = 0.5$) for each marker (VanRaden, 2008). Pedigree-based relationships are created with respect to the base population which assume no inbreeding or selection, so the allele frequency p should be the estimates of the base population (VanRaden, 2008). Gengler et al. (2007) presented a method to estimate allele frequencies in the base population by linear regression of gene content.

The scaling parameter k usually is defined as follows (VanRaden, 2008; Habier et al., 2007):

$$k = 2\sum p_j(1 - p_j),$$

which assumes a priori independence of SNP effects (Gianola et al., 2009).

Gianola et al. (2009) proposed another scaling parameter that accounts for the random ascertainment of SNP and their frequencies, which results in:

$$k = \left[(p_0 - q_0)^2 + 2\left(\frac{\sum p_j(1 - p_j)}{n}\right)\left(\frac{\alpha + \beta + 2}{\alpha + \beta}\right)\right] n,$$

where $p_0 = \alpha/(\alpha + \beta)$ is the expected allele frequency, $q_0 = (1 - p_0)$; α and β are parameters of the beta distribution fitting the base allelic frequency, and n is the number of SNP.

Matrices $\mathbf{G}$ are sometimes singular or close to singularity (VanRaden, 2008). In order to facilitate inversion, VanRaden (2008) proposed to use weighted $\mathbf{G}^*$ as:

$$\mathbf{G}^* = w\mathbf{G} + (1 - w)\mathbf{A}$$

where **A** is the pedigree-based relationship matrix, and w is a weight parameter between 0 and 1. VanRaden (2008) suggested to use a value of $w = 0.95$. Van Doormaal et al. (2009) in the Canadian implementation of genomic selection, suggested to use a $w = 0.80$, in order to have a polygenic effect of 20% instead of 5%.

Hayes et al. (2009b) used a similar approach as in Nejati-Javaremi et al. (1997) to study the increase in accuracy by using the realized relationship matrix. Hayes and Goddard (2008) found that estimated heritability was close to the simulated one with marker-based relationships compared to regular relationships based on pedigree information. As the number of markers used to create the relationship matrix increased, the estimate of heritability approached the true one.

Villanueva et al. (2005) studied benefits of marker-assisted selection for a genetic model based on a large number of additive loci of small effect. Comparisons were between a regular BLUP and a BLUP that used an identical-by-descent (**IBD**) matrix that combined the pedigree and marker information. Extra gains in response were observed by increasing accuracy from the marker information in genetic relationships.

In a simulation study, Habier et al. (2007) showed that markers can capture the genetic relationship between genotyped animals and thus affect the accuracy of estimated genomic breeding values.

Goddard (2009) derived expressions for the accuracy of genomic selection. He showed that the accuracy of genomic breeding values depends on the LD between the marker and the QTL and on the accuracy with which the markers effects are estimated. Different accuracies could be obtained assuming different prior distribution of QTL effects. Assumption of a normal distribution with a constant variance for all marker effects (GBLUP), results in a robust method regardless of the true distribution of QTLs. The accuracy of genomic selection could reach 100% with sufficient data (Goddard, 2009).

When a trait is under the control of many QTL with small effect, Daetwyler (2009, Chap 2) found that GBLUP results in a more accurate estimation of breeding values com-

pared with methods which assume a prior distribution of marker effects where a large number of markers have zero effects (e.g. BayesB Meuwissen et al. (2001)).

Luan et al. (2009) assessed the accuracy of genomic selection in Norwegian Red Cattle. They observed that GBLUP results in higher accuracies compared with BayesB or with a mixture model approach for several production and health traits. Their results indicated a strong relationship between the accuracy and heritability of the trait. Lower accuracy and greater bias was obtained for traits with low heritability.

Experiences with actual data from dairy cattle (Hayes et al., 2009c; VanRaden et al., 2009a) indicated that using a large number of markers with equal variance for all markers is appropriate for most traits. Limiting the number of SNP markers to only those with large effects resulted in reduced accuracy (Cole et al., 2009). However, little (if any) loss of accuracy occurred for most dairy cattle traits by assuming equal rather than different variance for each SNP marker (Cole et al., 2009; VanRaden et al., 2009a). Further, assuming equal variance allows the use of the same genomic relationship matrix for all traits.

Results from human genome-wide association studies shows that a small fraction of the total variance is explained by genetic variants (Maher, 2008). Studies done in human height, which has estimates of heritability around 80-90%, found that the genetic variants only explain about 5% of the total variance (Weedon et al., 2008). Goldstein (2009) estimated that the number of SNP that are required to explain 80% of the variation for human height was about 93,000. In a review study about mapping genes for complex traits in domestic animals, Goddard and Hayes (2009) arrived at similar conclusions.

Genomic relationships are beneficial in GWA studies (Kang et al., 2010, 2008; Amin et al., 2007). These authors proposed to use linear mixed models with relationships between individuals by using the kinship matrix estimated through genomic markers. Using the genetic relatedness between individual avoids spurious associations (Kang et al., 2010, 2008; Amin et al., 2007).

When variances are not equal, e.g., as in BayesA or BayesB of Meuwissen et al. (2001), or Bayes-Lasso (de los Campos et al., 2009), an equivalent G can be constructed by scaling contributions from different markers. Weighted genomic relationship matrices were also used in human studies (Amin et al., 2007; Leutenegger et al., 2003) and by VanRaden (2008). In such cases weights were defined by the expected variance:

$$\mathbf{G} = \mathbf{ZDZ}'$$

where $\mathbf{D}$ is a diagonal matrix with elements: $D_{ii} = \frac{1}{m[2p_i(1-p_i)]}$ and m is the number of genotyped individuals.

Single-step genomic evaluation

In general, not all the animals in a population are genotyped, and evaluations by different methods can provide different breeding values. Regular estimated breeding values from phenotypic and pedigree data are available for all animals, and genomic breeding values can be obtained for genotyped animals.

Genomic evaluations in dairy cattle are currently calculated with a multiple-step procedure (Hayes et al., 2009c; VanRaden, 2008). A typical evaluation requires 1) traditional evaluation with an animal model, 2) extraction of pseudo-observations such as deregressed evaluations or daughter deviations, 3) estimation of genomic effects for genotyped animals usually using simple sire models, and possibly 4) combining the genomic index with traditional parent averages (PA) and EBV (Hayes et al., 2009c; VanRaden et al., 2009b). Those steps are dependent on many parameters and assumptions. For example, there are several options for estimating genomic effects (Meuwissen et al., 2001; Gianola et al., 2006; VanRaden, 2008; de los Campos et al., 2009).

Advantages of the multistage procedure include no change to the regular evaluation and simple steps for predicting genomic values for young genotyped animals. Disadvantages are requirements for parameters in steps 3) and 4) such as prior variances and weights, as well as the loss of accuracy and biases due to selection. While the model in 1) uses the information

on all animals and can be multi-trait, the model in 3) is equivalent to a single-trait sire model for a highly selected set of sires. Incorrect parameters in 3) and 4) can result in unexpected changes for high reliability bulls. Neuner et al. (2008) claimed that problems associated with the multi-step procedure reduce its benefits, especially for cows.

Current experiences with genomic evaluations from the multiple-step procedure seem mixed. Genomic evaluations are more accurate than PA and approach the accuracy of evaluations for progeny-tested bulls, but they also seem inflated (VanRaden et al., 2009b). Use of regression coefficients to measure bias was described by Reverter et al. (1994) and forms the basis of Method R estimation of variance components. Values of the regression coefficient different from 1 indicate bias with overestimation (underestimation) for values lower (greater) than 1.

Inflation of genetic evaluations by genomic information causes top young bulls to have an unfair advantage over older progeny-tested bulls. Some of the problems with genomic evaluations may be caused by incorrect parameters and strong assumptions used in multiple-step procedures. However, effects of those parameters and assumptions are difficult to verify, particularly in the presence of selection.

A more serious problem in the multi-step method is when pseudo-observations are poorly defined or of poor quality (e.g., for animals with small progeny numbers), which is often the case for monogastric species and for beef cattle. Then, genomic predictions may be poor.

Misztal et al. (2009) (see Appendix A) proposed a single-step evaluation in which the pedigree-based relationship matrix is augmented by contributions from the genomic relationship matrix. They proposed that the numerator relationship matrix ($\mathbf{A}$) can be modified to a matrix ($\mathbf{H}$) that includes both pedigree-based relationships and differences between pedigree-based and genomic-based relationships ($\mathbf{A}_\Delta$) : $\mathbf{H} = \mathbf{A} + \mathbf{A}_\Delta$. where $\mathbf{A}_\Delta$ is a matrix that can be stored explicitly and accounts for deviation due to genomic information.

They also suggested a computing procedure (Bi-CGSTAB; van der Vorst (1992)) based on a non-symmetric system of mixed model equations that was suitable for millions of animals.

The PCG algorithm (Barrett, 1994), which is implemented in current software for solving large scale genetic evaluation (Tsuruta et al., 2001), is only applicable to symmetric systems of equations.

In their examples, they used

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix},$$

where subscripts 1 and 2 represent ungenotyped and genotyped animals, respectively, and $\mathbf{G}$ is a genomic relationship matrix. In tests, such $\mathbf{H}$ did not work because off-diagonals of $\mathbf{H}$ were not functions of $\mathbf{G}$.

Legarra et al. (2009) (see Appendix B) derived a joint relationship matrix based on pedigree and genomic relationships. They suggested deriving the joint density of $\mathbf{u}_1$ and $\mathbf{u}_2$ as $p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_1|\mathbf{u}_2)p(\mathbf{u}_2)$. The conditional distribution $p(\mathbf{u}_1|\mathbf{u}_2)$ is based on pedigree through the selection index or multivariate normal properties; $p(\mathbf{u}_2)$ is based only on genomic information, possibly from genomic relationships. The covariance of the joint distribution of $\mathbf{u}_1$ and $\mathbf{u}_2$ is thus $\mathbf{H}$:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} (\mathbf{G} - \mathbf{A}_{22}) \begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & 0 \\ 0 & \mathbf{I} \end{bmatrix},$$

which could be implemented in tests by using computing algorithms such as in Misztal et al. (2009) with only a few more computations per round of iteration than for traditional evaluations. Christensen and Lund (2010) using other derivation arrived to the same expression as Legarra et al. (2009). Even though the matrix was complex, computations were feasible even for large data sets.

Computing methods for genomic matrix

VanRaden (2008) presented methods to create genomic relationship matrices. The kernel of such methods involves the multiplication of the matrix of marker incidence (with dimension

number of genotyped animals by number of SNP markers) by its transpose. Matrix operations, and in particular matrix multiplication has been studied in computer science field and results in the widely used linear algebra kernels called Basic Linear Algebra Subroutines (**BLAS**; http://www.netlib.org/blas) (Dongarra et al., 1988, 1990).

Memory hierarchy can be partitioned into two basic types: main and cache. While the cache memory usually has a capacity of 256 Kbyte to 16 Mbyte, the main memory in current computer ranges from 1 Gbyte to 128 Gbyte. Processors have fast access to the cache memory but much slower to the main memory. Accessing large amount of memory, where blocks of memory are not contiguous and the capacity of the cache memory is exceeded, increases the computing time due to slow traffic to the main memory.

An optimized version of BLAS subroutines was developed by Whaley and Dongarra (1998) and an open source of theses libraries are available in the Automatically Tuned Linear Algebra Software (**ATLAS** http://math-atlas.sourceforge.net). These optimized libraries take into account features of a specific processor (memory speed and cache size) in several subroutines. Third-party libraries like the Intel Math Kernel Library (**MKL**) also implements an optimized version of BLAS.

The use of parallel processing is now simplified by hardware, as many computer chips contain two to eight processors, and by specific software tools for parallelization. Automatic parallelization by OpenMP (http://www.openmp.org) with MKL libraries requires only appropriate flag option during the compilation. Parallel multiplication can be facilitated by the open-source optimized subroutine (i.e. ATLAS) combined with OpenMP directives (Bentz and Kendall, 2005).

Computing methods for numerator relationship matrix

The degree of inbreeding of an individual and the relationship between two individuals (Wright, 1922) are represented by the numerator relationship matrix (**A**). Wright (1922)

developed a path coefficient method to calculate such coefficients. Henderson (1976) presented a recursive method that is computationally suitable compared to Wright's path formulas. Colleau (2002) presented an indirect method to calculate relationship coefficients. His method is based on a decomposition of the numerator relationship matrix, and coefficients for selected individuals are computed with sequential reading of the pedigree file (see Appendix A). Relationship coefficients for pairs of individuals can also be obtained using methods described by Aguilar and Misztal (2008).

Fertility traits

Worldwide declines of fertility in Holsteins creates a need for accurate evaluation of fertility traits. Fertility in dairy can be evaluated on a number of traits (Jamrozik et al., 2005; Gonzalez-Recio et al., 2005). Typical fertility traits such as Non-Return Rate and Days Open have their advantages and disadvantages (Huang et al., 2007). A desirable trait is conception rate (**CR**) defined as an outcome of individual service (Averill et al., 2004; Huang et al., 2007). Treating each service separately allows for adjusting specific effects influencing each service. Kuhn et al. (2008); Kuhn and Hutchison (2008) presented methodology to analyze individual service records for evaluations of CR in dairy cattle in the US. International evaluations for fertility traits of dairy bulls is based on five groups of traits, with CR considered in several groups (Jorjani, 2007).

Estimates of heritability for CR are low (Tsuruta et al., 2009; Gonzalez-Recio et al., 2006, 2005), and accuracies of estimated breeding values are low. Such accuracies can be improved by using all available services in each parity. Furthermore, it can be boosted by utilizing the genomic information (Veerkamp and Beerda, 2007).

References

Aguilar, I. and Misztal, I. (2008). Technical note: Recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. *J. Dairy Sci.*, 91(4):1669–1672.

Amin, N., van Duijn, C. M., and Aulchenko, Y. S. (2007). A genomic background based method for association analysis in related individuals. *PLoS ONE*, 2(12):e1274.

Averill, T. A., Rekaya, R., and Weigel, K. (2004). Genetic analysis of male and female fertility using longitudinal binary data. *J. Dairy Sci.*, 87(11):3947–3952.

Barrett, R. (1994). *Templates for the solution of linear systems: building blocks for iterative methods*. SIAM, Philadelphia.

Bentz, J. L. and Kendall, R. A. (2005). Parallelization of general matrix multiply routines using OpenMP. In *Shared Memory Parallel Programming with Open MP*, pages 1–11. Springer Berlin / Heidelberg.

Christensen, O. and Lund, M. (2010). Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.*, 42(1):2.

Cole, J. B., VanRaden, P. M., O'Connell, J. R., Van Tassell, C. P., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Wiggans, G. R. (2009). Distribution and location of genetic effects for dairy traits. *J. Dairy Sci.*, 92(6):2931–2946.

Colleau, J. J. (2002). An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.*, 34(4):409–421.

Daetwyler, H. D. (2009). *Genome-wide evaluation of populations*. PhD thesis, Wageningen University, Wageningen, NL.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375–385.

Dongarra, J. J., Croz, J. D., Hammarling, S., and Duff, I. S. (1990). A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Softw.*, 16(1):1–17.

Dongarra, J. J., Croz, J. D., Hammarling, S., and Hanson, R. J. (1988). An extended set of fortran basic linear algebra subprograms. *ACM Trans. Math. Softw.*, 14(1):1–17.

Frazer, K. A., Ballinger, D. G., and Cox, D. R. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

Gengler, N., Mayeres, P., and Szydlowski, M. (2007). A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal*, 1(01):21–28.

Gianola, D. and De los Campos, G. (2008). Inferring genetic values for quantitative traits non-parametrically. *Genet. Res.*, 90(06):525–540.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347–363.

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173(3):1761–1776.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136:245–257.

Goddard, M. E. and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.*, 10(6):381–391.

Goldstein, D. B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.*, page NEJMp0806284.

Gonzalez-Recio, O., Alenda, R., Chang, Y. M., Weigel, K. A., and Gianola, D. (2006). Selection for female fertility using censored fertility traits and investigation of the relationship with milk production. *J. Dairy Sci.*, 89(11):4438–4444.

Gonzalez-Recio, O., Chang, Y. M., Gianola, D., and Weigel, K. A. (2005). Number of inseminations to conception in Holstein cows using censored records and time-dependent covariates. *J. Dairy Sci.*, 88(10):3655–3662.

Gonzalez-Recio, O., Gianola, D., Rosa, G., Weigel, K., and Kranis, A. (2009). Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genet. Sel. Evol.*, 41(1):3.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397.

Harris, B., Johnson, D., and Spelman, R. (2008). Genomic selection in New Zealand and the implications for national genetic evaluation. *Proc. 36th ICAR Biennial Session, Niagara Falls, NY. Interbull Bull., Uppsala, Sweden*, pages 325–330.

Hayes, B., Bowman, P., Chamberlain, A., Verbyla, K., and Goddard, M. (2009a). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.*, 41(1):51.

Hayes, B., Visscher, P., and Goddard, M. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.*, 91(01):47–60.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009c). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.*, 92(2):433–443.

Hayes, B. J. and Goddard, M. E. (2008). Technical note: Prediction of breeding values using marker-derived relationship matrices. *J. Anim Sci.*, 86(9):2089–2092.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32(1):69–83.

Hill, W. G. (2008). Estimation, effectiveness and opportunities of long term genetic improvement in animals and maize. *Lohmann Information*, 43(1):3–20.

Huang, C., Misztal, I., Tsuruta, S., and Lawlor, T. J. (2007). Methodology of evaluation for female fertility. *Interbull Bull.*, 37:156–160.

Jamrozik, J., Fatehi, J., Kistemaker, G. J., and Schaeffer, L. R. (2005). Estimates of genetic parameters for Canadian Holstein female reproduction traits. *J. Dairy Sci.*, 88(6):2199–2208.

Jorjani, H. (2007). International genetic evaluation of female fertility traits in five major breeds. *Interbull Bull.*, 37:144–147.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, 42(4):348–354.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Konig, S., Simianer, H., and Willam, A. (2009). Economic evaluation of genomic breeding programs. *J. Dairy Sci.*, 92(1):382–391.

Kuhn, M. T. and Hutchison, J. L. (2008). Prediction of dairy bull fertility from field data: Use of multiple services and identification and utilization of factors affecting bull fertility. *J. Dairy Sci.*, 91(6):2481–2492.

Kuhn, M. T., Hutchison, J. L., and Norman, H. D. (2008). Modeling nuisance variables for prediction of service sire fertility. *J. Dairy Sci.*, 91(7):2823–2835.

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.*, 92(9):4656–4663.

Legarra, A., Robert-Granie, C., Manfredi, E., and Elsen, J.-M. (2008). Performance of genomic selection in mice. *Genetics*, 180(1):611–618.

Leutenegger, A.-L., Prum, B., Gnin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.*, 73(3):516–523.

Lorenzana, R. and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor. Appl. Genet.*, 120(1):151–161.

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. E. (2009). The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. *Genetics*, 183(3):1119–1126.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. S., and Van Tassell, C. P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE*, 4(4):e5350.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.*, 92(9):4648–4655.

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim Sci.*, 75(7):1738–1745.

Neuner, S., Emmerling, R., Thaller, G., and Gotz, K.-U. (2008). Strategies for estimating genetic parameters in marker-assisted best linear unbiased predictor models in dairy cattle. *J. Dairy Sci.*, 91(11):4344–4354.

Reverter, A., Golden, B. L., Bourdon, R. M., and Brinks, J. S. (1994). Technical note: detection of bias in genetic predictions. *J. Anim Sci.*, 72(1):34–37.

Schaeffer, L. (2006). Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.*, 123(4):218–223.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *J. Anim Sci.*, 86(10):2447–2454.

Su, G., Guldbrandtsen, B., Gregersen, V. R., and Lund, M. S. (2010). Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *J. Dairy Sci.*, 93(3):1175–1183.

Tsuruta, S., Misztal, I., Huang, C., and Lawlor, T. J. (2009). Bivariate analysis of conception rates and test-day milk yields in Holsteins using a threshold-linear model with random regressions. *J. Dairy Sci.*, 92(6):2922–2930.

Tsuruta, S., Misztal, I., and Stranden, I. (2001). Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim Sci.*, 79(5):1166–1172.

van der Vorst, H. (1992). Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644.

Van Doormaal, J., Kistemaker, G., Sullivan, P. G., Sargolzaei, M., and Schenkel, F. S. (2009). Canadian implementation of genomic evaluations. *Interbull Bull.*, 40:214–218.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C., and Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods*, 5(3):247–252.

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull Bull.*, 37:33–36.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91(11):4414–4423.

VanRaden, P. M., Tooker, M. E., and Cole, J. B. (2009a). Can you believe those genomic evaluations for young bulls? *J. Dairy Sci.*, 92(Suppl. 1).

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. (2009b). Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, 92(1):16–24.

Veerkamp, R. F. and Beerda, B. (2007). Genetics and genomics to improve fertility in high producing dairy cows. *Theriogenology*, 68:S266–S273.

Verbyla, K. L., Hayes, B. J., Bowman, P. J., and Goddard, M. E. (2009). Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res.*, 91(05):307–311.

Villanueva, B., Pong-Wong, R., Fernandez, J., and Toro, M. A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim Sci.*, 83(8):1747–1752.

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R. B., Stevens, S., Hall, A. S., Samani, N. J., Shields, B., Prokopenko, I., Farrall, M., Dominiczak, A., Johnson, T., Bergmann, S., Beckmann, J. S., Vollenweider, P., and Waterworth, D. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.*, 40(5):575–583.

Whaley, R. C. and Dongarra, J. J. (1998). Automatically tuned linear algebra software. IEEE Computer Society.

Wiggans, G. R., Sonstegard, T. S., VanRaden, P. M., Matukumalli, L. K., Schnabel, R. D., Taylor, J. F., Schenkel, F. S., and Van Tassell, C. P. (2009). Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *J. Dairy Sci.*, 92(7):3431–3436.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.*, 56(645):330–338.

Chapter 3

Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score[1]

[1]I. Aguilar, I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. *Online Journal of Dairy Science*. 93 (2) : 743–752. Reprinted here with permission of publisher.

Abstract

The first national single-step, full-information (phenotype, pedigree, and marker genotype) genetic evaluation was developed for final score of US Holsteins. Data included final scores recorded from 1955 to 2009 for 6,232,548 Holsteins cows. BovineSNP50 genotypes from the Cooperative Dairy DNA Repository were available for 6,508 bulls. Three analyses used a repeatability animal model as currently used for the national US evaluation. The first 2 analyses used final scores recorded up to 2004. The first analysis used only a pedigree-based relationship matrix. The second analysis used a relationship matrix based on both pedigree and genomic information (single-step approach). The third analysis used the complete data set and only the pedigree-based relationship matrix. The fourth analysis used predictions from the first analysis (final scores up to 2004 and only a pedigree-based relationship matrix) and prediction using a genomic based matrix to obtain genetic evaluation (multiple-step approach). Different allele frequencies were tested in construction of the genomic relationship matrix. Coefficients of determination between predictions of young bulls from parent average, single-step, and multiple-step approaches and their 2009 daughter deviations were 0.24, 0.37 to 0.41, and 0.40, respectively. The highest coefficient of determination for a single-step approach was observed when using a genomic relationship matrix with assumed allele frequencies of 0.5. Coefficients for regression of 2009 daughter deviations on parent-average, single-step, and multiple-step predictions were 0.76, 0.68 to 0.79, and 0.86, respectively, which indicated some inflation of predictions. The single-step regression coefficient could be increased up to 0.92 by scaling differences between the genomic and pedigree-based relationship matrices with little loss in accuracy of prediction. One complete evaluation took about 2h of computing time and 2.7 gigabytes of memory. Computing times for single-step analyses were slightly longer (2%) than for pedigree-based analysis. A national single-step genetic evaluation with the pedigree relationship matrix augmented with genomic information provided genomic predictions with accuracy and bias comparable to multiple-step procedures

and could account for any population or data structure. Advantages of single-step evaluations should increase in the future when animals are pre-selected on genotypes.

(*Key words*: BLUP, genomic prediction, SNP, genetic evaluation)

Introduction

Genomic evaluations are currently calculated with a multiple-step procedure (Hayes et al., 2009; VanRaden, 2008) . A typical evaluation requires 1) traditional evaluation with an animal model, 2) extraction of pseudo-observations such as deregressed evaluations or daughter deviations (DD), 3) estimation of genomic effects for genotyped animals usually using simple sire models, and possibly 4) combining the genomic index with traditional parent averages (PA) and EBV (Hayes et al., 2009; VanRaden et al., 2009b). Those steps are dependent on many parameters and assumptions. For example, estimation of genomic effects has several options (Meuwissen et al., 2001; Gianola et al., 2006; VanRaden, 2008; de los Campos et al., 2009). The SNP marker effects can be estimated with different assumptions regarding the prior distribution of such effects. Genomic effects also can be estimated with a simple model that includes a genomic relationship matrix derived from genotypes and variances of the SNP marker effects (Nejati-Javaremi et al., 1997). Both methods are equivalent except for numerical properties (VanRaden, 2007).

Initially, genomic evaluation was tested with simulated data and a variety of assumptions (VanRaden, 2008). Experiences with actual data from dairy cattle (Hayes et al., 2009; VanRaden et al., 2009a) indicated that using a large number of markers with equal variance for all markers is appropriate for most traits. Limiting the number of SNP markers to only those with large effects resulted in reduced accuracy (Cole et al., 2009). However, little (if any) loss of accuracy occurred for most dairy cattle traits by assuming equal rather than different variance for each SNP marker (Cole et al., 2009; VanRaden et al., 2009a). Further, assuming equal variance allows the use of the same genomic relationship matrix for all traits.

Current experiences with genomic evaluations from the multiple-step procedure seem mixed. Genomic evaluations are more accurate than PA and approach the accuracy of evaluations for progeny-tested bulls, but they also seem inflated (VanRaden et al., 2009b). Although their inflation is lower than that of current PA, the potentially great utilization of top genomically evaluated young sires increases the importance of high accuracy and minimum bias. Inflation of genetic evaluations by genomic information causes top young bulls to have an unfair advantage over older progeny-tested bulls. Some of the problems with genomic evaluations may be caused by incorrect parameters and strong assumptions used in multiple-step procedures. However, effects of those parameters and assumptions are extremely difficult to verify, particularly in the presence of selection. An alternative explanation for the mixed results is that observed regressions and estimated reliabilities are biased downward by selective genotyping. A more serious problem is when pseudo-observations are poorly defined or of poor quality (e.g., for animals with small progeny numbers), which is often the case for monogastric species and for beef cattle.

Misztal et al. (2009) proposed a single-step evaluation in which the pedigree-based relationship matrix is augmented by contributions from the genomic relationship matrix. They also suggested a computing procedure based on a nonsymmetric system of mixed model equations that was suitable for millions of animals. Legarra et al. (2009) derived a joint relationship matrix based on pedigree and genomic relationships. Even though the matrix was expense and complex to create, computations were feasible even for large data sets.

The single-step procedure provides a unified framework, eliminates a number of assumptions and parameters, and provides the opportunity to calculate more accurate genomic evaluations than with multiple-step procedures. The objective of this study was to utilize a single-step procedure for genomic evaluation in a national evaluation setting and compare its performance to a multiple-step procedure.

Materials and Methods

Data

Data were US Holstein information for final score used for May 2009 official evaluations (Sires Summaries, 2009). A total of 10,466,066 records were available for 6,232,548 cows. Pedigrees were available for 9,100,106 animals. Genotypes for 6,508 bulls were generated using the Illumina BovineSNP50 BeadChip and DNA from semen contributed by US and Canadian AI organizations to the Cooperative Dairy DNA Repository; genotypes were provided by the Animal Improvement Programs Laboratory, ARS, USDA (Beltsville, MD).

Relationship Matrix with Pedigree and Genomic Information

Misztal et al. (2009) suggested that a numerator relationship matrix ($\mathbf{A}$) can be modified to a matrix ($\mathbf{H}$) that includes both pedigree-based relationships and differences between pedigree-based and genomic-based relationships ($\mathbf{A}_\Delta$) : $\mathbf{H} = \mathbf{A} + \mathbf{A}_\Delta$. In their examples, they used

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix},$$

where subscripts 1 and 2 represent ungenotyped and genotyped animals, respectively, and $\mathbf{G}$ is a genomic relationship matrix. In tests, such $\mathbf{H}$ did not work because off-diagonals of $\mathbf{H}$ were not functions of $\mathbf{G}$. Assume, for example, that no animal in $\mathbf{G}$ has records; then, according to $\mathbf{H}$, the predicted breeding value for genotyped animals ($\mathbf{u}_2$) would be $\mathbf{u}_2|\mathbf{u}_1 = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{u}_1$, where $\mathbf{u}_1$ is the predicted breeding value for ungenotyped animals, and $\mathbf{G}$ would have no role whatsoever.

Legarra et al. (2009) suggested deriving the joint density of $\mathbf{u}_1$ and $\mathbf{u}_2$ as $p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_1|\mathbf{u}_2)p(\mathbf{u}_2)$. The conditional distribution $p(\mathbf{u}_1|\mathbf{u}_2)$ is based on pedigree through the selection index or multivariate normal properties; $p(\mathbf{u}_2)$ is based only on genomic information, possibly from genomic relationships. The covariance of the joint distribution of $\mathbf{u}_1$ and $\mathbf{u}_2$ is

thus **H**:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} (\mathbf{G} - \mathbf{A}_{22}) \begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & 0 \\ 0 & \mathbf{I} \end{bmatrix},$$

which could be implemented in tests by using computing algorithms such as in Misztal et al. (2009) with only a few more computations per round of iteration than for traditional evaluations. Convergence was readily obtained for medium-sized data sets (up to 1 million); however, for larger data sets, convergence was strongly dependent on the type of **G** used.

An inverse of **H** that allows for drastically simpler computations (see Appendix C) is

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

where $\mathbf{A}_{22}^{-1}$ is the inverse of a pedigree-based relationship matrix for genotyped animals only. This expression has also been independently derived by Christensen and Lund (2009). However, the new formula introduces a small problem: G is usually singular and, therefore, is not invertible without additional steps.

Models and Analyses

A repeatability animal model was used for analysis as is currently done for US national evaluation of Holstein conformation traits (Sires Summaries, 2009). The first 2 analyses used final scores through 2004 only. The first analysis (**Ped$_{04}$**) used only the pedigree-based relationship matrix; the second analysis (**PedGen1$_{04}$**) used relationships based on both pedigree and genomic information in a single-step approach. The third analysis (**Ped$_{09}$**) used the complete data set and only the pedigree-based relationship matrix. The fourth analysis (**PedGenM$_{04}$**) used predictions from Ped_{04} and a multiple-step approach to obtain genomic predictions (**GP**) as described by VanRaden et al. (2009b). Options in the last analysis were genomic relationship matrix and base allele frequencies. Both $PedGen1_{04}$ and $PedGenM_{04}$ assumed equal variances per SNP marker effect.

"Raw" genomic relationships ($\mathbf{G}_b$) were created as

$$\mathbf{G}_b = \frac{\mathbf{Z}\mathbf{Z}'}{k},$$

where $\mathbf{Z}$ is an incidence matrix for SNP effects with elements

$$z_{ij} = \begin{cases} 0 - 2p_j & \text{if homozygous 11} \\ 1 - 2p_j & \text{if heterozygous 12 or 21,} \\ 2 - 2p_j & \text{if homozygous 22} \end{cases}$$

for animal i and SNP j with allele frequency p_j. Several allele frequencies were used to center the matrix: 0.5, base population estimated by linear regression of gene content (Gengler et al., 2007), and current population. The scaling parameter k was defined as

$$k = 2\sum p_j(1 - p_j)$$

(VanRaden, 2008), which assumes a priori independence of SNP effects (Gianola et al., 2009). Another scaling parameter has been proposed by Gianola et al. (2009) with

$$k = \left[(p_0 - q_0)^2 + 2\left(\frac{\sum p_j(1-p_j)}{n}\right)\left(\frac{\alpha+\beta+2}{\alpha+\beta}\right)\right] n,$$

where $p_0 = \alpha/(\alpha+\beta)$ is the expected allele frequency, $q_0 = (1-p_0)$; α and β are parameters of the beta distribution fitting the base allelic frequency, and n is the number of SNP. That modification accounts for random ascertainment of SNP and their frequencies.

Matrices $\mathbf{G}_b$ were sometimes singular or close to singularity. In order to facilitate inversion, final analyses used a weighted $\mathbf{G}$ as proposed by VanRaden (2008): $\mathbf{G} = 0.95\mathbf{G}_b + 0.05\mathbf{A}_{22}$. The weights were not critical, and replacing them with 0.98 and 0.02 caused negligible differences.

Because GP could be scaled incorrectly, a series of analyses used $\mathbf{H}^{-1}$:

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

where λ scales differences between genomic and pedigree-based information. More exactly (see Appendix D), λ sets the value of $\mathbf{G}$ in $\mathbf{H}$ to a new value ($\mathbf{G}^*$):

$$\mathbf{G}^* = \left[\lambda \mathbf{G}^{-1} + (1-\lambda)\mathbf{A}_{22}^{-1})\right]^{-1},$$

thus blending genomic and pedigree information. For $\lambda = 1$, $\mathbf{G}^* = \mathbf{G}$; for $\lambda = 0$, $\mathbf{G}^* = \mathbf{A}_{22}$ and $\mathbf{H} = \mathbf{A}$. In fact, this corresponds to the following prior for genotyped animals:

$$p\left(u_2|\mathbf{G}, \mathbf{A}_{22}, \lambda\right) = p\left(u_2|\mathbf{G}, \lambda\right) p\left(u_2|\mathbf{A}_{22}, \lambda\right) = N\left(0, \mathbf{G}/\lambda\right) N\left(0, \mathbf{A}_{22}/\left(1-\lambda\right)\right)$$

Comparisons were based on the regressions

$$DD = \mu + \delta EBV_{04} + e$$

and

$$EBV_{09} = \mu + \delta EBV_{04} + e$$

where DD were deregressed evaluations (VanRaden et al., 2009b) from genotyped bulls without daughter records in 2004 but with daughter records in 2009 that were computed with complete final score data but without genomic information; EBV_{09} are breeding values based on final scores up to 2009 but without genomic information; μ is a mean; δ is a regression coefficient; EBV_{04} are breeding values based on final scores up to 2004; and e is residual error. Breeding values were calculated for 2 sets of genotyped bulls: 1) 2,575 young bulls with no daughter records in 2004 but with daughter records in 2009 and 2) 3,933 evaluated bulls with daughter records in 2004. The most accurate method for prediction for young bulls would have μ close to 0, δ close to 1, and R^2 as high as possible.

Both DD and EBV_{09} regressions were examined to allow more detailed comparison. Although DD computed through deregressed evaluations allow partial removal of the effect of PA, the removal is contingent on the accuracy of approximate reliabilities. Also, the goal of GP is not to predict DD but to predict future breeding values.

Software

Initial software for the construction of G and the multiple-step evaluation was provided by P. M. VanRaden (Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD). Additional software for creating G was contributed by B. J. Hayes (Biosciences Research Division, Department of Primary Industries Victoria, Bundoora, Australia). Software refinement included rearrangements of code in Fortran 95 for efficient matrix multiplication, matrix inversion, and parallelization. Computation of $\mathbf{A_{22}}$ followed the formulas of Misztal et al. (2009), which used the algorithm of Colleau (2002). Genetic evaluation was performed by modified BLUP90IOD (Tsuruta et al., 2001; Misztal et al., 2002), which uses iteration on data with the preconditioned conjugate gradient algorithm.

Results and Discussion

Precomputation of $\mathbf{G}$ and $\mathbf{A_{22}}$ took 650 s and 45 s, respectively, on an Opteron 64-bit processor with a clock speed of 3.02 GHz and a cache size of 1Mbyte, using one processor; their inversion took approximately 150 s. Time per 1 preconditioned conjugate gradient round for $PedGen1_{04}$ was 13 s, which was 2% greater than 1 round for Ped_{04}. Convergence rates (not shown) for $PedGen1_{04}$ and Ped_{04} were almost identical. A complete analysis with $PedGen1_{04}$ took approximately 2 h. Memory requirement for precomputation of $\mathbf{G}$ was 2.7 gigabytes.

Table 3.1 shows R^2 and δ for regression of 2009 DD and corresponding EBV_{09} on various 2004 predictions for young bulls. For PA, R^2 was 24% with δ of 0.76. The δ showed that PA overestimated the genetic evaluation with progeny included by 27%. For the multiple-step approach, R^2 increased to 40% and δ to 0.86. The increase in R^2 of 16% compared with PA R^2 was slightly higher than the increase of 13% reported by VanRaden et al. (2009b). VanRaden et al. (2009a) reported a regression coefficient of 0.74. Differences from the results of VanRaden et al. (2009a,b) were due partly to slightly different data (theirs included

Canadian evaluations but fewer genotypes and US records) and methodology details (e.g., different computation of approximate reliabilities).

For the single-step approaches (Table 3.1), R^2 for DD varied between 37 and 41%, and δ varied between 0.68 and 0.79 depending on $\mathbf{G}$. The highest single-step increase in R^2 over prediction from PA was 1% higher than the multiple-step increase, which indicated that single-step breeding values were slightly more accurate than those by the multiple-step as implemented here. The best δ was 0.07 lower than the multiple-step δ, which indicated greater inflation of prediction for young bulls. The highest single-step R^2 and δ (least inflation) were for $\mathbf{G}$ based on equal allele frequencies with extra benefits from modifications by Gianola et al. (2009). For simplification, subsequent comparisons used the equal allele frequency $\mathbf{G}$ but without the modifications.

R^2 obtained using $\mathbf{G}$ matrix with equal allele frequency was greater compared with a $\mathbf{G}$ matrix created using base allele frequency. This was in the opposite direction compared with a similar study (VanRaden et al., 2008). In addition, the latter study reported correlations of 0.6 between genomic and pedigree based inbreeding coefficients, whereas a correlation of 0.2 using base allele frequencies was found in the current study. Further analyses need to be done to address such differences.

Results for EBV_{09} (Table 3.1) generally were similar to those for DD but with a slight advantage for the multiple-step approach. The δ indicated much greater inflation than for DD. Inflation on the EBV_{09} scale is important for producers because their comparisons are based on EBV and not on DD. It is debatable whether the results with EBV_{09} are valid in this case, because they contain information from PA. On the other hand, DD computed using approximated reliabilities may contain an extra noise.

Parent average was, in general, similar for runs with and without $\mathbf{G}$. Thus, inflation higher than that in PA could be caused by too much indirect weight on genomic relationships. Inflation could be lowered by weighting $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ by λ (see Appendix B). Table 3.2 shows R^2 and δ for DD and EBV_{09} with such a weighting. As λ decreased from 1.0 to 0.5, R^2

gradually decreased for DD but had an interim maximum for EBV_{09}. At $\lambda = 0.7$, EBV_{09} R^2 increased to 51%, which was 1% better than for the multiple-step approach (Table 3.1); δ also was higher than for the multiple-step approach by 0.01. The δ can be increased to 0.92 with only a slight decrease in R^2. Because the primary interest of breeders is to identify animals with the highest genetic merit, a moderate reduction in bias (i.e., higher δ) would be preferred to a small increase in overall accuracy (R^2).

Accuracy of the single-step approach was dependent on the choice of $\mathbf{G}$ and the weighting placed on the difference between $\mathbf{G}$ and $\mathbf{A}$. With the proper choice, accuracy of the single-step approach was superior to the multiple-step approach. One reason why the choice of $\mathbf{G}$ is critical is that genomic and pedigree relationship matrices should be compatible both in scale and in structure. The importance of structure can be seen from the decomposition of the genomic breeding value in Appendix B. The weight of PA relative to genomic information depends on λ and even more on diagonals of $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$. In general, the diagonal of $\mathbf{G}^{-1}$ depends on the genomic relationships and measures the amount of information provided to individual i by other animals.

The primary influence of the weighting factor (λ) appears to be related to the proportion of the additive variance explained by the genomic information (Appendix B). Snelling et al. (2009) found that different numbers of SNP genotypes used for the construction of $\mathbf{G}$ resulted in different decomposition of the additive variance between the genomic and polygenic effects. Genomic information from the best genotyped bulls would add relationship information for a number of animals and most likely result in higher additive variance. The Canadian official genomic evaluation system for Holsteins (Van Doormaal et al., 2009) assumes that only 80% of the additive variance is explained by the SNP information. Other factors behind the weighting factor may be related to final score as a trait in US Holsteins. For example, heritability based on records of grade animals is lower than with records on registered animals (Koduru, 2006). Other issues are preferential treatment of bull dams and the nature of final score, for which the definition changes over time (Tsuruta et al., 2005). Future studies

with more traits and species will clarify the influence of the weighting factor as well as alternative weighting factors. While our decomposition between the genomic and polygenic effects involved inverses of the respective matrices, it can also be done on the direct scale, by assuming that only part of the genetic variance is explained by the genomic information (Christensen and Lund, 2009).

What **G** should be is still undetermined. As implemented for this study, **G** was constructed so that linear effects were assumed for SNP genotypes while also collecting information about realized relationships (VanRaden, 2008). Other alternatives exist. For example, matrix **K** in Gonzalez-Recio et al. (2008) included a similarity index across genotypes. Probabilities for identity by descent can also be used and averaged across loci (Villanueva et al., 2005).

Use of regression coefficients to measure bias was described by Reverter et al. (1994) and forms the basis of Method $\Re$ estimation of variance components. However, the use of δ to calibrate GP might be problematic. First, it relies on the same set of equations being used for old and recent evaluations, which was not true for this study; the "old" evaluation ($PedGen1_{04}$) used H, whereas the "recent" evaluation (Ped_{09}) used A. Second, as seen by experience from Method $\Re$, the estimated regression coefficient has large error and might be biased, especially by selection (Schenkel and Schaeffer, 2000; Cantet et al., 2000). On the other hand, little bias and very efficient computations were reported by Druet et al. (2001), who traced the bias to the use of fixed effects estimated from subsets of the data.

For this study, **G** was constructed with equal variances assumed for SNP marker effects. When variances are not equal, e.g., as in Bayes-A or Bayes-B (Meuwissen et al., 2001), an equivalent **G** can be constructed by scaling contributions from different markers. Such construction requires precomputing those variances based on genotyped individuals and pseudo-data.

The generalization of the single-step approach to multiple traits is obvious when **G** is identical for each trait. However, separate **G** matrices for each trait may require single-trait

analyses. For several traits, the benefits and simplicity of multiple-trait analysis using the same **G** may overcome the loss of accuracy from using less than the optimal **G** for each trait.

The single-step approach to evaluation as described in this study is easy to implement just by modifying the relationship matrix for current evaluations. Aside from simplification of genomic evaluation, the procedure is expected to improve evaluations for all ungenotyped animals. Updated PA and PTA for descendant of genotyped animals are possible using multiple-step methods with additional calculations (see http://aipl.arsusda.gov/reference/changes/eval0901.html). Advantages of single-step evaluations should increase in the future when animals have been pre-selected on genotypes. Traditional evaluations expect that Mendelian sampling averages 0, but in the future only animals with positive Mendelian sampling may receive phenotypes.

To demostrate the utility of genomic evaluation, it is necessary to validate them, particulary for young animals. In contrast, in BLUP based on pedigree information, such a validation is rarely performed and is implicitly replaced by variance component estimation, although some validation is performed indirectly for analyses used by Interbull MACE evaluations (Interbull, 2001). With some assumptions, it is possible that the parameters of a single-step procedure are regular variance components plus weighting factors, either such as proposed in this study or different. In such a case, the validation steps can be replaced by parameter estimation, greatly simplifying the use of the genomic information. Ways to estimate values of weighting factors by REML, MCMC, or other methods remain to be investigated.

Conclusions

Full genomic and pedigree evaluations by the single-step approach were as good as the multiple-step approach in terms of accuracy and bias. Generalization for complex data structures or more complicated models are straight forward. Additional computational cost was small relative to pedigree evaluation. The highest accuracy was obtained with a scaled genomic relationship matrix created under the assumption of equal allele frequencies. The

main advantages of the single-step approach are its simplicity and automatic weights for the various sources of information for the overall breeding value. Moreover, advantages of single-step evaluations should increase in the future when animals are pre-selected on genotypes.

Acknowledgments

The authors thank P. M. VanRaden and G. R. Wiggans (Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD), J. R. OConnell (University of Maryland School of Medicine, Baltimore), C. P. Van Tassell (Bovine Functional Genomics Laboratory, ARS, USDA, Beltsville, MD), and L. Varona (Universidad de Zaragoza, Spain) as well as Holstein Association USA Inc. (Brattleboro, VT) and the Cooperative Dairy DNA Repository (Beltsville, MD) for providing genotypic data. Financing from Agence National de la Recherche project AMASGEN (Jouy en Josas, France) is acknowledged. Editing assistance was provided by Suzanne Hubbard. Helpful comments and suggestions from the two reviewers are acknowledged.

References

Cantet, R. J., Birchmeier, A. N., Santos-Cristal, M. G., and de Avila, V. S. (2000). Comparison of restricted maximum likelihood and method R for estimating heritability and predicting breeding value under selection. *J. Anim Sci.*, 78(10):2554–2560.

Christensen, O. F. and Lund, M. S. (2009). Genomic relationship matrix when some animals are not genotyped. In *60th Annual Meeting of the European Association for Animal Production.*, Barcelona, Spain.

Cole, J. B., VanRaden, P. M., O'Connell, J. R., Van Tassell, C. P., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Wiggans, G. R. (2009). Distribution and location of genetic effects for dairy traits. *J. Dairy Sci.*, 92(6):2931–2946.

Colleau, J. J. (2002). An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.*, 34(4):409–421.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375–385.

Druet, T., Misztal, I., Duangjinda, M., Reverter, A., and Gengler, N. (2001). Estimation of genetic covariances with method R. *J. Anim Sci.*, 79(3):605–615.

Gengler, N., Mayeres, P., and Szydlowski, M. (2007). A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal*, 1(01):21–28.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347–363.

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173(3):1761–1776.

Gonzalez-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M., and Avendano, S. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. *Genetics*, 178(4):2305–2313.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.*, 92(2):433–443.

Interbull (2001). Interbull guidelines for national and international genetic evaluation systems in dairy cattle with focus on production traits. In *Intebull Bull 28. http://www-interbull.slu.se/bulletins/bulletin28/Interbull Guidelines-2001.pdf. Accessed November 9, 2009.* Interbull Bull 28, Interbull Bull.

Koduru, V. K. R. (2006). *Changes in genetic evaluations from 1st to 2nd crop for final score in Holsteins*. M.S. Thesis, University of Georgia, Athens, GA.

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.*, 92(9):4656–4663.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.*, 92(9):4648–4655.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. (2002). BLUPF90 and related programs (BGF90). In *7th World Congress on Genetics Applied to Livestock Production*, Montpellier, France.

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim Sci.*, 75(7):1738–1745.

Reverter, A., Golden, B. L., Bourdon, R. M., and Brinks, J. S. (1994). Technical note: detection of bias in genetic predictions. *J. Anim Sci.*, 72(1):34–37.

Schenkel, F. and Schaeffer, L. (2000). Effects of nonrandom parental selection on estimation of variance components. *J. Anim. Breed. Genet.*, 117(4):225–239.

Sires Summaries, . (2009). *Holstein Association USA*. Holstein Association USA Inc. Brattleboro, VT., Brattleboro, VT, USA.

Snelling, W., Kuehn, L., Thallman, R., Keele, J., and Bennett, G. (2009). Genomic heritability of beef cattle growth. *J. Anim Sci.*, 87 (Suppl 1).

Tsuruta, S., Misztal, I., and Lawlor, T. J. (2005). Changing definition of productive life in US Holsteins: Effect on genetic correlations. *J. Dairy Sci.*, 88(3):1156–1165.

Tsuruta, S., Misztal, I., and Stranden, I. (2001). Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim Sci.*, 79(5):1166–1172.

Van Doormaal, J., Kistemaker, G., Sullivan, P. G., Sargolzaei, M., and Schenkel, F. S. (2009). Canadian implementation of genomic evaluations. *Interbull Bull.*, 40:214–218.

VanRaden, P., Tooker, M., and Gengler, N. (2008). Effects of allele frequency estimation on genomic predictions and inbreeding coefficients. *J. Dairy Sci.*, 91(E-Suppl. 1):506 (Abstr.).

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull Bull.*, 37:33–36.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91(11):4414–4423.

VanRaden, P. M., Tooker, M. E., and Cole, J. B. (2009a). Can you believe those genomic evaluations for young bulls? *J. Dairy Sci.*, 92(Suppl. 1).

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. (2009b). Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, 92(1):16–24.

VanRaden, P. M. and Wiggans, G. R. (1991). Derivation, calculation, and use of national animal model information. *J. Dairy Sci.*, 74(8):2737–2746.

Villanueva, B., Pong-Wong, R., Fernandez, J., and Toro, M. A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim Sci.*, 83(8):1747–1752.

Table 3.1: Coefficients of determination (R^2) and coefficients (δ) for regression of 2009 daughter deviations (DD) or corresponding estimated breeding values (EBV_{09}) for bulls progeny tested from 2005 through 2009 on 2004 predictions obtained by different algorithms.

	DD		EBV_{09}	
Prediction method	$R^2, \%$	δ	$R^2, \%$	δ
Parent average	24	0.76	36	0.79
Multiple-step	40	0.86	50	0.82
Single-step[a]				
G5	41	0.76	49	0.7
GB	38	0.68	45	0.63
GC	37	0.71	45	0.66
GG – G5	41	0.79	50	0.73
GG – GB	38	0.77	46	0.71
GG – GC	39	0.79	46	0.73

[a] Assumed allele frequency of 0.5 (G5), base population (GB), current population (GC), or calculated as in [30] of Gianola et al. (2009) (GG).

Table 3.2: Coefficients of determination (R^2) and coefficients (δ) for regression of 2009 daughter deviations (DD) or corresponding breeding values (EBV_{09}) for bulls progeny tested from 2005 through 2009 on 2004 predictions from a single-step approach using an allele frequency of 0.5 and different relative variances for the genomic matrix (λ).

	DD		EBV_{09}	
λ	$R^2, \%$	δ	$R^2, \%$	δ
1.0	41	0.76	49	0.70
0.9	41	0.81	50	0.76
0.8	41	0.84	51	0.79
0.7	40	0.88	51	0.83
0.6	40	0.90	50	0.85
0.5	39	0.92	50	0.88
0.3	35	0.91	47	0.89

Chapter 4

Efficient computations of genomic relationship matrix and other matrices used in the single-step evaluation[1]

[1] I. Aguilar, I. Misztal, A. Legarra, and S. Tsuruta. *To be submitted to Journal of Animal Breeding and Genetics.*

Abstract

Genomic evaluations could be calculated using a unified procedure that combines phenotypic, pedigree and genomic information. Implementation of such a procedure requires the inverse of the relationship matrix based on pedigree and genomic relationships. The objective of this study was to investigate efficient computing options to create relationship matrices based on genomic markers and pedigree information as well as their inverses. A matrix of incidence of SNP maker information was simulated for a panel of 40K SNPs. Number of genotyped animals varied from 1,000 to 30,000. Kernel of the computation for the genomic relationship matrix requires a matrix multiplication of the incidence matrix. Methods included a simple "do" loop, two optimized versions of the loop and specific matrix multiplication subroutine (DGEMM). Inversion methods were by a generalized inverse algorithm and by LAPACK subroutines. Useful matrices to implement a unified approach can be computed efficiently. Optimizations can be either by modifications of existing code or by the use of efficient automatic optimizations provided by open source or third-party libraries.

(*Key words*: relationship matrix, genomic selection, computing methods)

Introduction

Genomic evaluations in dairy cattle are currently performed using multiple step procedures (Hayes et al., 2009; VanRaden et al., 2009). A typical evaluation requires 1) traditional evaluation with an animal model, 2) extraction of pseudo-observations such as deregressed evaluations or daughter deviations (DD), 3) estimation of genomic effects for genotyped animals usually using simple sire models, and possibly 4) combining the genomic index with traditional parent averages (PA) and breeding values (Hayes et al., 2009; VanRaden et al., 2009). Genomic effects also can be estimated with a simple model that includes a genomic relationship matrix derived from genotypes and variances of the SNP marker effects (Nejati-Javaremi et al., 1997; VanRaden, 2007). Recently, Misztal et al. (2009) proposed

a single-step evaluation in which the pedigree-based relationship matrix is augmented by contributions from the genomic relationship matrix. Legarra et al. (2009) derived a joint relationship matrix based on pedigree and genomic relationships; and Aguilar et al. (2010) described the inverse for such a matrix. A similar matrix and its inverse were independently derived by Christensen and Lund (2010). VanRaden (2008) discussed methods to create genomic relationship matrices. The kernel of such methods involves multiplication of the matrix of marker incidences (with dimension number of genotyped animals by number of SNP marker) by its transpose. Matrix operations, are implemented efficiently in many packages; most of them use linear algebra kernels called Basic Linear Algebra Subroutines (BLAS; http://www.netlib.org/blas) (Dongarra et al., 1988, 1990). An optimized version of BLAS subroutines was developed by Whaley and Dongarra (1998) and an open source of these libraries is available at Automatically Tuned Linear Algebra Software (ATLAS http://math-atlas.sourceforge.net). These libraries account for features of a specific processor (memory speed and cache size).

Modifications of current software for genetic evaluations and variance component estimations to implement the unified approach for genomic evaluation (Aguilar et al., 2010) require inverses of the genomic relationship matrix and the regular relationship matrix for genotyped animals. The objectives of this research were to present efficient computing options to create these relationship matrices and their inverses.

Materials and Methods

The inverse of the relationship matrix based on both pedigree and genomic information (Aguilar et al., 2010) is:

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

where $\mathbf{A}^{-1}$ is the inverse of the numerator relationship matrix, $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix and $\mathbf{A}_{22}^{-1}$ is the inverse of the relationship matrix based on pedigree information corresponding to the genotyped animals. Modifications of current software for genetic evaluation (Tsuruta et al., 2001; Misztal et al., 2002) to use $\mathbf{H}^{-1}$ require the inclusion of $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$.

The genomic relationship matrix ($\mathbf{G}$) was created by simulation. A matrix of incidences of SNP maker information ($\mathbf{Z}$) was simulated for a panel of 40K SNPs, with values corresponding to gene content of the second allele (0, 1 and 2). Number of genotyped animals varied from 1,000 to 30,000. Pedigree-based relationship matrix ($\mathbf{A}_{22}$) was constructed using a pedigree dataset of 9,100,106 US Holsteins provided by Holstein USA Inc. (Brattleboro, VT).

Following VanRaden (2008) the genomic relationships ($\mathbf{G}$) were created as

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{k},$$

where $\mathbf{Z}$ is an incidence matrix for SNP effects with elements

$$z_{ij} = \begin{cases} 0 - 2p_j & \text{if homozygous 11} \\ 1 - 2p_j & \text{if heterozygous 12 or 21,} \\ 2 - 2p_j & \text{if homozygous 22} \end{cases}$$

for animal i and SNP j with allele frequency $p_j = 0.5$ for all SNP markers. The scaling parameter k was defined as

$$k = 2 \sum p_j(1 - p_j)$$

Memory hierarchy

Memory hierarchy can be partitioned into two types: main and cache. While the cache memory has a capacity of 256 Kbyte to 16 Mbyte, the main memory can support 128 Gbytes. One processor has fast access to the cache memory but much slower access to the main memory. Accessing large amount of memory, where block of memory are not contiguous

and exceeding the capacity of the cache memory, will increase the computing time due to slow access to the main memory.

Methods

Computations of $\mathbf{ZZ}'/k$ were performed in Fortran 95 by several methods (Figure 4.1). The first method (**ORIG**) was a simple three"do" loops, centering the matrix $\mathbf{Z}$ through indirect memory access, and performing the scaling by k after each element of G_{ij} was computed. The second method (**OPTM**) was a modification to optimize the indirect memory access. The matrix $\mathbf{Z}$ was centered once at the beginning of the process, outside of the main loop. In the third method (**OPTML**), loops were reorganized and the scaling operation was performed outside of the main loop. Having separate operations for matrix multiplication and scaling allows using general subroutines to compute $\mathbf{ZZ}'$. Also, matrix multiplications of the form $\mathbf{ZZ}'$ were computed by the original BLAS subroutine DGEMM, and by their optimized versions as in ATLAS or in the Intel Math Kernel Library (**MKL**)

Two methods were used to create the pedigree relationship matrix for genotyped animals ($\mathbf{A_{22}}$). The first method was the tabular method and the second was based formulas presented in Misztal et al. (2009), which use the algorithm of Colleau (2002).

Matrix inversion was performed by a converted Fortran 95 code of a generalized inverse algorithm from the BLUPF90 package (Misztal et al., 2002) and by the LU factorization using the DGETRF/DGETRI subroutines from LAPACK (Anderson et al., 1990). Such subroutines are available either in ATLAS or in MKL libraries.

Computations

All programs were run on an Opteron 64-bit processor with a clock speed of 3.02 GHz and a cache size of 1 Mbyte. Some programs were also run on a Xeon 64-bit processor with a clock speed of 3.5 GHz and a cache size of 6 Mbyte. Initial software for the construction of $\mathbf{G}$ and for the tabular method was provided by P. M. VanRaden (Animal Improvement Programs

Laboratory, ARS, USDA, Beltsville, MD). Programs were written in Fortran 95 and compiled by Intel Fortran Compiler with option -O3. Automatic uses of parallelization using OpenMP (http://www.openmp.org) were obtained using MKL libraries with appropriate flag option during compilation.

Results and discussion

Timing using the alternative loop codes are presented in Table 4.1. The original method was 10 times slower on the Opteron system because of lower cache memory and different memory system. Large improvement was achieved with the OPTM on the Opteron but not on the Xeon system. An alternative explanation is that the Intel compiler was efficient in optimizing codes on Xeon but not on Opteron systems. Almost 4 times speedup was obtained on both computers with the code optimized for memory and loop rearrangement.

Different algorithms to perform the matrix multiplication were tested with the Opteron system. Figure 4.2 shows the computing time for the modified code for memory access and loop ordering, the BLAS subroutine for matrix multiplication (DGEMM), and its optimized version as in ATLAS libraries. Performance of the original DGEMM deteriorated with increase of the number of individuals. The optimized ATLAS-DGEMM subroutine was the fastest. The performance of the optimized code for memory and loops ordering shows a trend similar to ATLAS-DGEMM, but is slightly slower .

Multiplication of large matrices requires optimization to fully utilize the cache memory. This operation requires fine tuning for specific system architectures. Simple modifications to optimize indirect memory access were successful in reducing the run time. Also, a simple rearrangement of the codes, splitting task and avoiding the use of static variables within the loops allows the compiler to do an automatic optimization (e.g. vectorization of “do” loops operations). Speed-ups were from 4 to 15 times, depending on the processor. However, code generation obtained by ATLAS-DGEMM run faster with no additional programming.

The use of parallel processing is now simplified by hardware, as many computer chips contain two to eight processors, and by specific software tools for parallelization. Optimized implementation of DGEMM in MKL allows parallel processing. Figure 4.3 shows the results of the optimized implementation of DGEMM in the MKL using up to 4 processors. The speedup with 3 processors and 5000 genotyped animals was 2.93, which was close to an ideal one. Matrix multiplication can be parallelized by the open source optimized DGEMM subroutine (i.e. ATLAS) combined with OpenMP directives (Bentz and Kendall, 2005). Actual time to create the genomic relationship matrix for 50K SNPs and 30,000 animals with DGEMM as implemented in MKL and using 3 processors was 79 min. The operation required 30 Gbytes of memory.

Creating the relationship matrix based on pedigree information for 6,500 genotyped animals using the tabular method required 311 s and 12.1 Gbyte of memory. The same computation with the Colleau method (2002) required 45 s and 322 Mbytes. The tabular method requires storage for a dense matrix for all genotyped animals and their ancestors (approximately 57,000 individuals for 6,500 genotyped animals) while the Colleau method needs only a few vectors with dimension equal to the number of genotyped animals. Memory requirements for the tabular method can be reduced by splitting the pedigree file in several groups, but at the cost of additional computations (VanRaden, personal communication, 2009).

Computing time of inversion for different number of genotyped animals are in Figure 4.4. For the largest matrix (30,000 animals) the inversion took approximately 13 h with the generalized inversion method, but only 3.4 h using the optimized version of LAPACK.

Further reduction in computing time could be obtained by parallel processing using OpenMP directives. Speedups for up to 4 processors are plotted in the Figure 4.5 using the optimized version of LAPACK as implemented in MKL. With four processors was used for the inversion of the largest matrix took approximately 1 h.

Conclusion

We presented methods for efficient construction of matrices required for the implementation of the single-step genomic evaluation. Optimizations were attained by modifications of existing codes, using efficient automatic optimization provided as open source or by commercial libraries. With all the optimizations building the genomic relationship matrix for 30,000 animals with 40K SNPs each took about 1 h, and similar time was necessary to obtain the inverse.

Acknowledgements

The authors thank P.M. VanRaden from Animal Improvement Programs Laboratory, USDA (Beltsville, MD), for providing the original software. This study was partially funded by the Holstein Association USA Inc. and by AFRI grants 2009-65205-05665 and 2010-65205-20366 from the USDA NIFA Animal Genome Program.

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.*, 93(2):743–752.

Anderson, E., Bai, Z., Dongarra, J., Greenbaum, A., McKenney, A., Croz, J. D., Hammerling, S., Demmel, J., Bischof, C., and Sorensen, D. (1990). LAPACK: a portable linear algebra library for high-performance computers. IEEE Computer Society Press.

Bentz, J. L. and Kendall, R. A. (2005). Parallelization of general matrix multiply routines using OpenMP. In *Shared Memory Parallel Programming with Open MP*, pages 1–11. Springer Berlin / Heidelberg.

Christensen, O. and Lund, M. (2010). Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.*, 42(1):2.

Colleau, J. J. (2002). An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.*, 34(4):409–421.

Dongarra, J. J., Croz, J. D., Hammarling, S., and Duff, I. S. (1990). A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Softw.*, 16(1):1–17.

Dongarra, J. J., Croz, J. D., Hammarling, S., and Hanson, R. J. (1988). An extended set of fortran basic linear algebra subprograms. *ACM Trans. Math. Softw.*, 14(1):1–17.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.*, 92(2):433–443.

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.*, 92(9):4656–4663.

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.*, 92(9):4648–4655.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. (2002). BLUPF90 and related programs (BGF90). In *7th World Congress on Genetics Applied to Livestock Production*, Montpellier, France.

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim Sci.*, 75(7):1738–1745.

Tsuruta, S., Misztal, I., and Stranden, I. (2001). Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim Sci.*, 79(5):1166–1172.

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Interbull Bull.*, 37:33–36.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91(11):4414–4423.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, 92(1):16–24.

Whaley, R. C. and Dongarra, J. J. (1998). Automatically tuned linear algebra software. IEEE Computer Society.

Table 4.1: Computing time (m) for alternatives codes for construction of the G matrix[a] on different machines.

		Algorithms		
Processor	Cache	Original	Memory optimized	Memory & loop optimized
Xeon 3.5 GHz	6 Mbyte	24	26	7
Opteron 3.02 GHz	1 Mbyte	265	59	17

[a] using 6,500 animals and 40K SNP markers.

Denote M as the matrix incidence of SNP marker with dimension n by p, with n = number of animals,and p = number of markers; Z is center matrix with dimension 3 by p with values corresponding to gene content-$2p_i$

Original

```
do j=1,n
   do i=j,n
      S=0
      do k=1,p
        S=S+Z(M(i,k),k)
             *Z(M(j,k),k)
      end do
      G(i,j)=0.5*S/
              sqrt(d(i)*d(j))
      G(j,i)=G(i,j)
    end do
end do
```

Optimize indirect memory access

```
do k=1,p
   X(:,k)=Z(M(:,k),k)
end do
do j=1,n
   do i=j,n
      S=0
      do k=1,p
         S=S+X(i,k)
            *X(j,k)
      end do
      G(i,j)=0.5*S/
               sqrt(d(i)*d(j))
      G(j,i)=G(i,j)
   end do
end do
```

Optimize memory and loops

```
do k=1,p
   X(:,k)=Z(M(:,k),k)
end do
do i=1,n
   do j=i,n
      do k=1,p
         G(i,j)=G(i,j)
                +X(i,k)*X(j,k)
      end do
   end do
end do
do i=1,n
   do j=1,n
      G(i,j)=0.5*G(i,j)/
              sqrt(d(i)*d(j)
   end do
end do
```

Figure 4.1: Alternative codes to create genomic relationships.

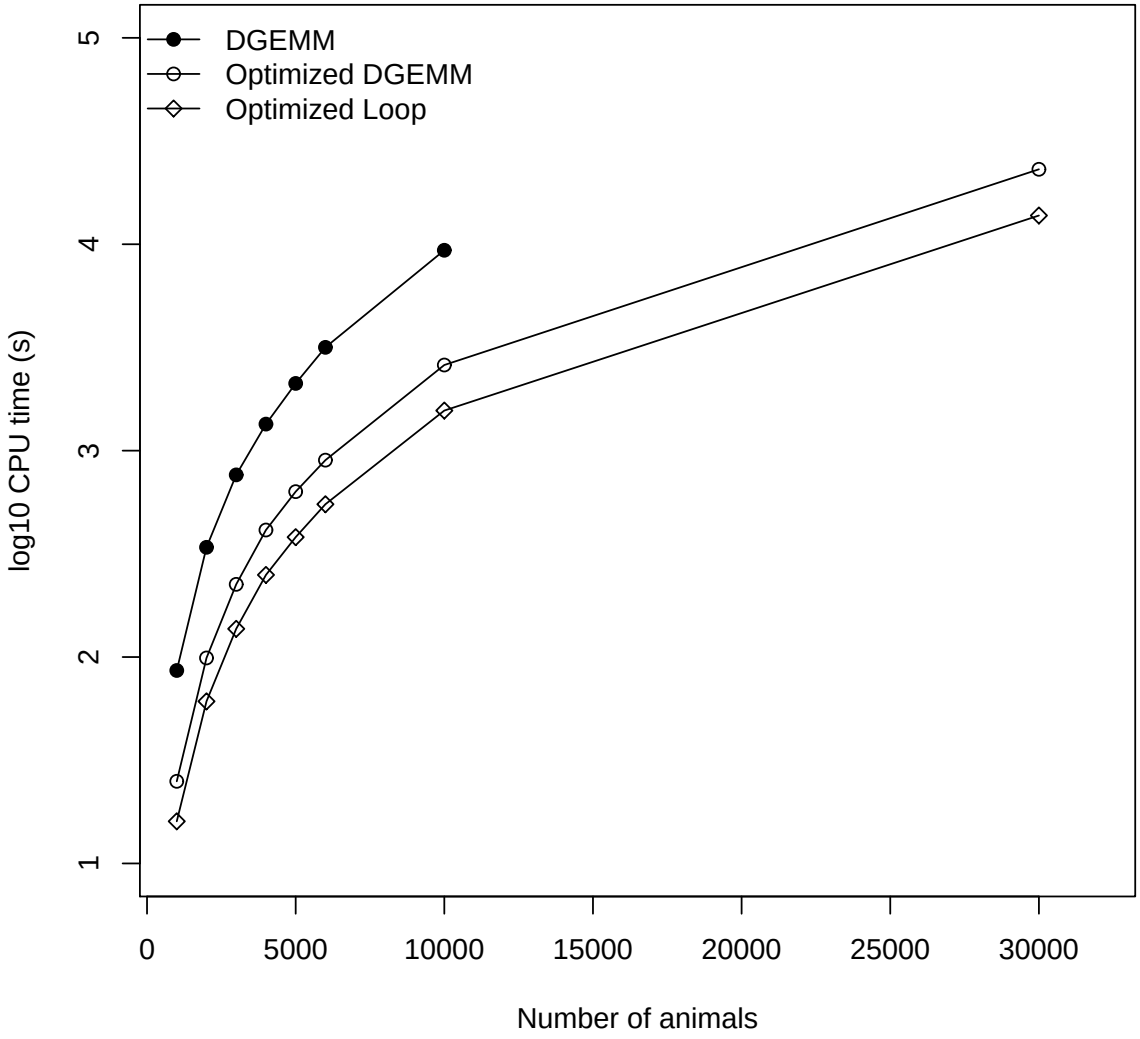

Figure 4.2: Computing time using different matrix multiplications algorithms.

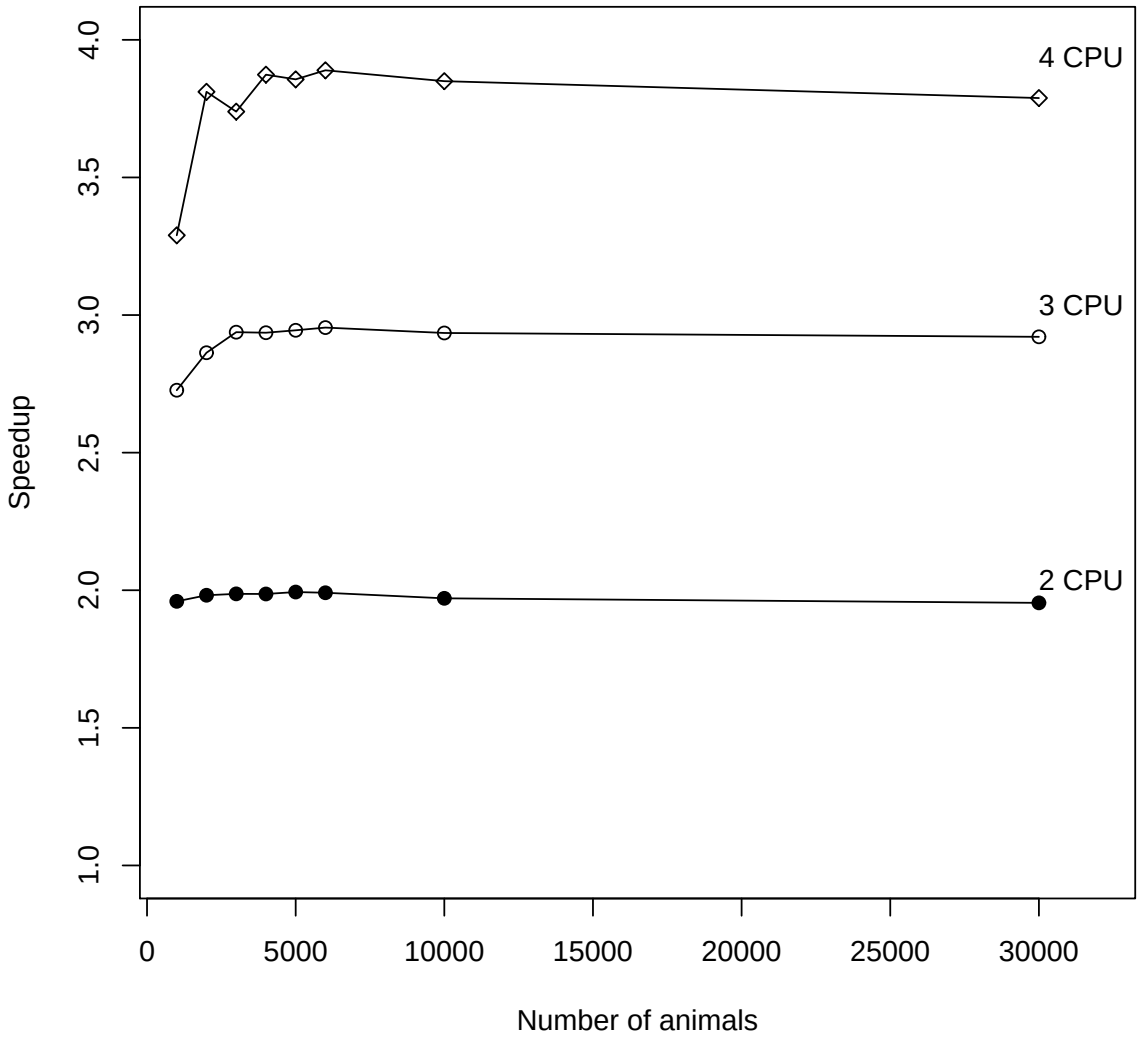

Figure 4.3: Speedup for optimized DGEMM for multiple processors using OpenMP.

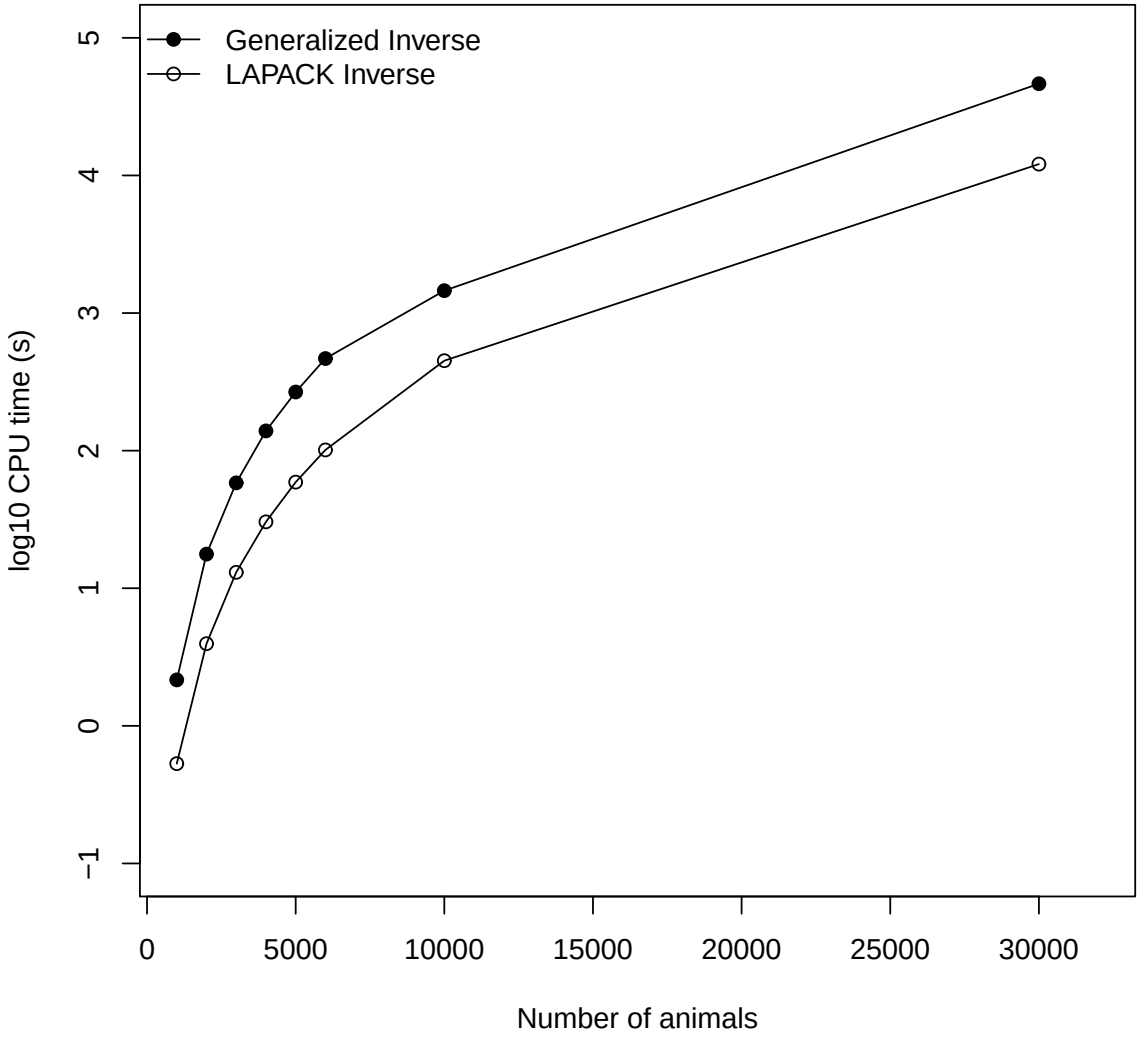

Figure 4.4: Computing time using different methods of matrix inversion.

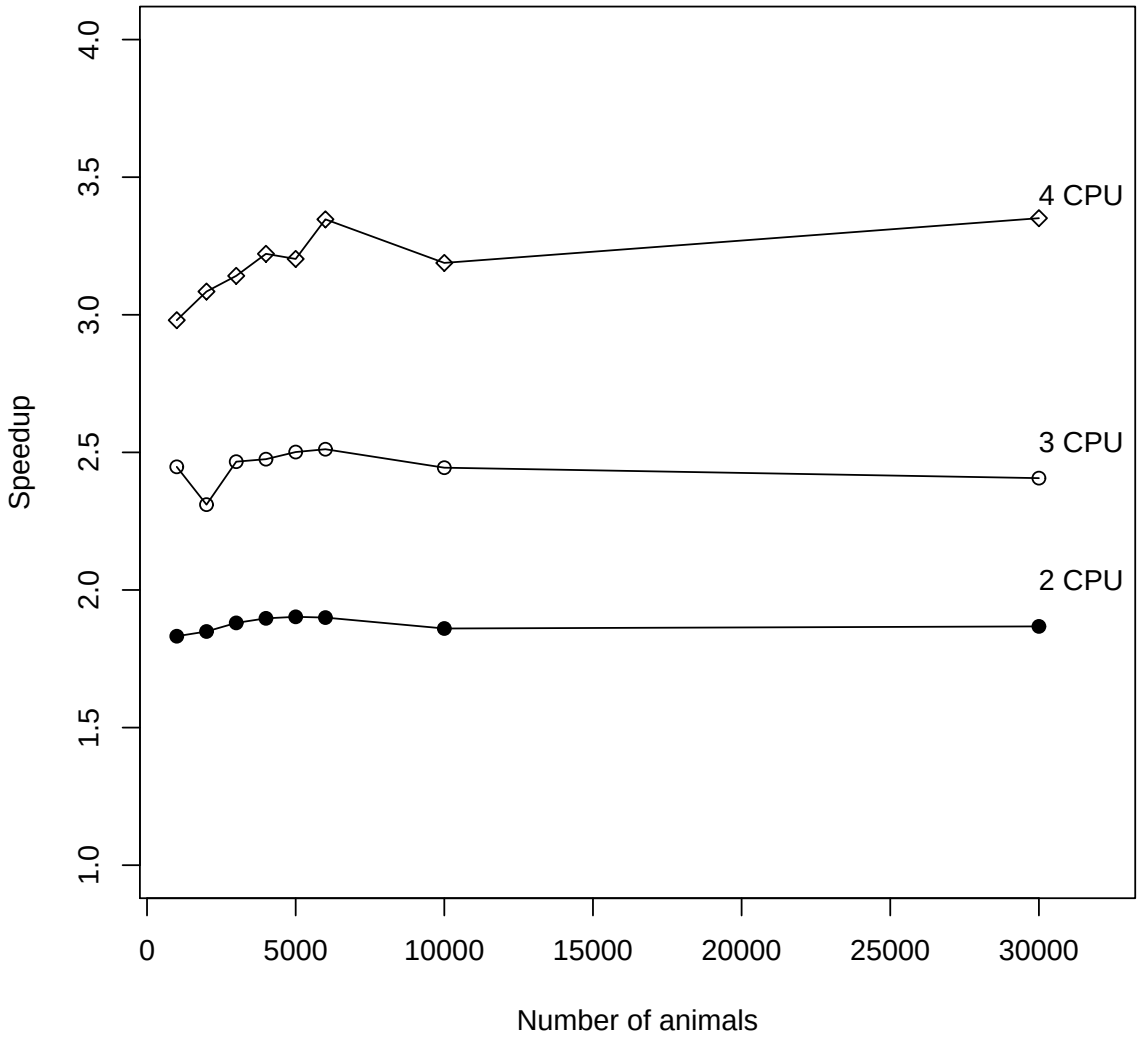

Figure 4.5: Speedup for matrix inversion using optimized LAPACK for multiple processors using OpenMP.

Chapter 5

Multiple trait genomic evaluation of conception rate in Holsteins[1]

[1]I. Aguilar, I. Misztal, and S. Tsuruta. *To be Submitted to Journal of Dairy Science.*

Abstract

This study evaluated the feasibility and accuracy of multiple trait evaluation for conception rate (**CR**) defined as outcomes of all available inseminations in US Holsteins using all available phenotypic, pedigree and genomic information.

Genetic parameters of CR in the first three parities were estimated with data from New York State. Heritability estimates were around 2% and genetic correlations > 0.73. Genetic evaluations used the national data set and a multiple trait model. The evaluations were performed by regular BLUP or by a single-step approach (**SSP**), which utilized the genomic information. R^2 obtained with the single-step approach were almost the double of those achieved with BLUP. Computing the single-step approach took 33% more time than with BLUP. A multiple trait evaluation of conception rate using the genomic information is possible and advantageous.

(*Key words*: BLUP, genomic selection, fertility, genetic evaluation)

Introduction

Worldwide declines of fertility in Holsteins creates a need for accurate evaluation of fertility traits. Fertility in dairy cattle can be evaluated based on a number of traits (Jamrozik et al., 2005; Gonzalez-Recio et al., 2005). Typical fertility traits such as Non-Return Rate and Days Open have their advantages and disadvantages (Huang et al., 2007). A desirable trait is conception rate (**CR**) which is defined as an outcome of individual services (Averill et al., 2004; Huang et al., 2007). Treating each service separately allows for adjusting specific effects that influence each service.

Because of low heritability of CR (Tsuruta et al., 2009; Gonzalez-Recio et al., 2006, 2005), accuracies of bull estimated breeding values (**EBV**) for CR are usually low. Such accuracy can be improved using all available services in each parity. Furthermore, it can be boosted by utilizing the genomic information (Veerkamp and Beerda, 2007). Even though records from

later parities come too late for young bulls, that information is useful for better prediction in the genomic analysis.

The simplest but most efficient way to utilize the genomic information is a single-step approach (Aguilar et al., 2010). In this approach, genomic information is used to improve relationship information and no changes to the model are required. The goals of this study were to estimate genetic parameters of CR in the first three parities, run a national evaluation with and without the genomic information, and estimate gains of accuracy from the genomic information.

Material and methods

Data

Holstein service records for the first three parities were obtained from the Animal Improvement Programs Laboratory, ARS, USDA (Beltsville, MD). Records from calvings registered between 2003 and 2009 were used. Information prior to 2002 was scarce. Data editing followed criteria presented by Kuhn et al. (2008). Only AI services were used and DIM at inseminations were required to be between 30 and 365 d. Success of insemination was determined through all reproductive events (heat detection, natural service, AI and pregnancy diagnostic), as well as the presence of the next calving. Service sires were restricted to Holsteins. Variance components were estimated using records from New York State. All available insemination records were used in the national genetic evaluation. A summary of both data sets is in Table 5.1.

Model

Conception rates in the first three parities were considered as correlated traits. Parameter estimation and prediction of EBVs were obtained using a multiple-trait linear model. In some analysis single trait linear models were used for the first parity service records. Fixed effects in the model included contemporary group defined by herd-year of calving, month of

service, age at calving, and days to service after calving. Random effects were service sire (s), additive genetic (a), permanent environmental (p) and residual (e) effects. The (co)variance structure was :

$$\text{var}\begin{bmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{s}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G}\otimes\mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{P}\otimes\mathbf{I} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{S}\otimes\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}\otimes\mathbf{I}\end{bmatrix}$$

where **A** is the numerator relationship matrix; **G**, **P** and **S** are 3 x 3 (co)variance matrices for additive genetic, permanent and service sire respectively; **I** are identity matrices and **R** is a 3 x 3 diagonal matrix of residual variances.

Variance components

Parameters were estimated the GIBBS2F90 program (Misztal et al., 2002) via a Bayesian approach using Gibbs sampling. Genomic data was not included for variance component estimation. Of a total of 100,000 samples, the first 10,000 were discarded as a burn-in, and every 10^{th} sample was retained to calculate posterior means and standard deviations.

Genetic evaluation

Genetic evaluations were computed using a modified BLUP90IOD (Tsuruta et al., 2001; Aguilar et al., 2010). Approximate accuracies were calculated using ACCF90 (Misztal et al., 2002). Deregressed evaluations (**DD**) were obtained from EBVs and approximate accuracies (VanRaden et al., 2009). A subset of records up to 2005 was used to assess the accuracy of prediction of breeding values. Two sets of EBV were obtained. The first set used a genetic evaluation with the regular numerator relationship matrix (**PA**). The second set used a modified relationship matrix that accounts for genomic relationships and predicts EBVs with a single-step approach (**SSP**) (Misztal et al., 2009). In SSP the numerator relationship

matrix ($\mathbf{A}$) was replaced by the $\mathbf{H}$ matrix with the following inverse (Aguilar et al., 2010; Christensen and Lund, 2010):

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

where $\mathbf{H}$ is a modified relationship matrix incorporating genomic information as described by Legarra et al. (2009), $\mathbf{G}$ is a genomic relationship matrix (VanRaden, 2008) and $\mathbf{A}_{22}$ is the pedigree-based relationship matrix for genotyped animals.

The genomic relationship matrix ($\mathbf{G}$) were created as:

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{k},$$

where $\mathbf{Z}$ is an incidence matrix for SNP effects with elements

$$z_{ij} = \begin{cases} 0 - 2p_j & \text{if homozygous 11} \\ 1 - 2p_j & \text{if heterozygous 12 or 21,} \\ 2 - 2p_j & \text{if homozygous 22} \end{cases}$$

for animal i and SNP j with allele frequency $p_j = 0.5$ for all SNPs markers. The scaling parameter k was defined as

$$k = 2 \sum p_j(1 - p_j)$$

Predictions from the two methods were compared by the regressions:

$$DD = \mu + \delta EBV_{05} + e$$

where DD were deregressed evaluations from 154 genotyped bulls without daughter records in 2005 but with daughter records in 2009 that were computed with complete data but without genomic information; μ is a mean; δ is a regression coefficient; EBV_{05} are breeding values based on insemination records up to 2005; and e is residual error. EBV_{05} were either PA (PA_{05}) or based on SSP ($SSP - EBV_{05}$).

Results and discussion

Table 5.2 contains estimates of genetic parameters in the first three parities. Heritability estimates for each parity were close to 2%, and genetic correlations among parities were high (0.7-0.9).

Results of the genetic evaluations for bulls with no daughters in 2005 and at least 50 daughters in 2009 are presented in Table 5.3. The R^2 obtained with PA in 2005 were low, probably because only 3 years of data were available and the heritability of CR is low. Few service records were available prior to 2003. Luan et al. (2009) assessed the accuracy of genomic selection in Norwegian Red Cattle and observed a strong relationship between accuracy and heritability of the trait; traits with low heritability had EBV with lower accuracy and greater bias. The R^2 obtained with EBV in 2005 and the genomic information were higher than the regular EBV based on pedigree relationship. Thus, the genomic information doubled R^2. Coefficients of regression (δ) were lower in PA compared with SSP, indicating little bias in predictions with genomic predictions.

To investigate increases in accuracy due to the use of three parities, the analyses were repeated for first parity service records only. Results for CR in first parity comparing SSP and PA using single trait or multiple trait models are in Table 5.4. Use of first parity records only result in lower R^2 and δ. Adding a genomic-based relationship matrix or using multiple parity with pedigree-based relationship, increased the accuracy of estimated breeding values for CR in the first parity by 3 times. Moreover, an additional increment in accuracy was obtained using all available information (genomic markers and multiple parities). A simple repeatability model with service records from several parities could also be applied. This topic remains to be addressed in further studies.

Eight years of data from 2003 to 2009 were sufficient to generate accurate predictions for many bulls but was insufficient to test the accuracy via the methodology used in other studies (VanRaden et al., 2009). A better assessment would be based on accuracies of EBV, however, such method with the SSP and genomic information requires further research. One

source of information about the increased accuracy using genomic information is obtained by analyzing the diagonals of the inverse of the pedigree-based ($\mathbf{A}$) and genomic based ($\mathbf{G}$) relationship matrix. While diagonal elements of $\mathbf{A}^{-1}$ for bulls are close to $2 + n/2$, where n is the number of daughters, the corresponding elements of $\mathbf{G}^{-1}$ are $2 + x/2$ where $x > n$. In that case, $x - n$ may be regarded as the additional number of daughters equivalent from the genomic information.

Initial computing with BLUP took 1.5 hrs. Computing with the added genomic information via SSP increased the time to 2 hrs. Computing with multiple-trait models and the genomic information is therefore realistic.

Conclusion

Multiple trait genetic evaluation for conception rate using outcomes of all available inseminations is technically possible. Large improvement in accuracy is possible using the genomic information and computation with the single-step approach is straightforward. More accurate assessment of such an improvement would require either records over a longer period of time or a different methodology for comparisons.

Acknowledgments

This study was partially funded by the Holstein Association USA Inc. and by AFRI grants 2009-65205-05665 and 2010-65205-20366 from the USDA NIFA Animal Genome Program. The authors thank G. R. Wiggans from AIPL for providing phenotypic and pedigree data and the Cooperative Dairy DNA Repository (Beltsville, MD) for providing genotypic data.

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.*, 93(2):743–752.

Averill, T. A., Rekaya, R., and Weigel, K. (2004). Genetic analysis of male and female fertility using longitudinal binary data. *J. Dairy Sci.*, 87(11):3947–3952.

Christensen, O. and Lund, M. (2010). Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.*, 42(1):2.

Gonzalez-Recio, O., Alenda, R., Chang, Y. M., Weigel, K. A., and Gianola, D. (2006). Selection for female fertility using censored fertility traits and investigation of the relationship with milk production. *J. Dairy Sci.*, 89(11):4438–4444.

Gonzalez-Recio, O., Chang, Y. M., Gianola, D., and Weigel, K. A. (2005). Number of inseminations to conception in Holstein cows using censored records and time-dependent covariates. *J. Dairy Sci.*, 88(10):3655–3662.

Huang, C., Misztal, I., Tsuruta, S., and Lawlor, T. J. (2007). Methodology of evaluation for female fertility. *Interbull Bull.*, 37:156–160.

Jamrozik, J., Fatehi, J., Kistemaker, G. J., and Schaeffer, L. R. (2005). Estimates of genetic parameters for Canadian Holstein female reproduction traits. *J. Dairy Sci.*, 88(6):2199–2208.

Kuhn, M. T., Hutchison, J. L., and Norman, H. D. (2008). Modeling nuisance variables for prediction of service sire fertility. *J. Dairy Sci.*, 91(7):2823–2835.

Legarra, A., Aguilar, I., and Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.*, 92(9):4656–4663.

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. E. (2009). The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. *Genetics*, 183(3):1119–1126.

Misztal, I., Legarra, A., and Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.*, 92(9):4648–4655.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. (2002). BLUPF90 and related programs (BGF90). In *7th World Congress on Genetics Applied to Livestock Production*, Montpellier, France.

Tsuruta, S., Misztal, I., Huang, C., and Lawlor, T. J. (2009). Bivariate analysis of conception rates and test-day milk yields in Holsteins using a threshold-linear model with random regressions. *J. Dairy Sci.*, 92(6):2922–2930.

Tsuruta, S., Misztal, I., and Stranden, I. (2001). Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim Sci.*, 79(5):1166–1172.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91(11):4414–4423.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, 92(1):16–24.

Veerkamp, R. F. and Beerda, B. (2007). Genetics and genomics to improve fertility in high producing dairy cows. *Theriogenology*, 68:S266–S273.

Table 5.1: Descriptive summary of national and New York data by parity.

	National			New York State		
	1[a]	2	3	1	2	3
Insemination records	3,025,115	2,033,086	945,870	165,159	116,494	55,038
Herd-Year	14,581	14,322	12,203	862	855	745
Cows	1,186,451	790,354	380,776	67,083	46,248	22,634
Conception Rate (%)	33.0	31.0	30.7	35.4	32.8	33.2
Pedigree Animals		2,489,119			132,623	

[a] Number of parity.

Table 5.2: Estimates of posterior mean and standard deviations for genetic parameters for conception rate in the first three parities[a].

	CR 1[b]	CR 2	CR 3
CR 1	0.018 ± 0.002	0.877 ± 0.045	0.732 ±0.047
CR 2	0.288 ± 0.083	0.022 ± 0.002	0.808 ±0.103
CR 3	0.162 ± 0.084	0.326 ± 0.07	0.016 ±0.005

[a] Heritability estimates ± SD on the diagonal, genetic and permanent correlations above and below the diagonal, respectively.

[b] CR 1, CR 2 and CR 3, conception rate in first, second and third parity respectively.

Table 5.3: Coefficients of determination (R^2) and coefficients of regression (δ) of daughter deviation on estimated breeding values using single-step approach ($SSP - EBV_{05}$) or parent average (PA_{05}).

	$SSP - EBV_{05}$		PA_{05}	
Traits	R^2	δ	R^2	δ
CR 1[a]	0.15	0.84	0.07	0.72
CR 2	0.13	0.81	0.06	0.66
CR 3	0.10	0.96	0.05	0.82

[a] CR 1, CR 2 and CR 3, conception rate in first, second and third parity respectively.

Table 5.4: Coefficients of determination (R^2) and coefficients of regression (δ) of daughter deviation on estimated breeding values using single-step approach ($SSP - EBV_{05}$) or parent average (PA_{05}) for first parity conception rate using single trait or multiple trait analysis.

	$SSP - EBV_{05}$		PA_{05}	
Model	R^2	δ	R^2	δ
Single Trait	0.07	0.86	0.02	0.57
Multiple Trait	0.15	0.84	0.07	0.72

Chapter 6

Conclusions

Genomic evaluation by the single-step approach is as accurate as the multiple-step approach. Generalization for complex data structures or more complicated models is straightforward. Additional computational costs are small relatively to a regular BLUP evaluation. The highest accuracy was obtained with a scaled genomic relationship matrix created under the assumption of equal allele frequencies. Main advantages of the single-step approach are simplicity and automatic weights for the various sources of information. Advantages of single-step evaluations should increase in the future when animals are pre-selected on genotypes.

Efficient methods to create the matrices used in the single-step genetic evaluations are presented. Optimizations were performed by modifications of the existing code, by automatic optimization in open source software or using commercial libraries. With all the optimizations, the construction of the genomic relationship matrix for 30,000 animals with 40K SNPs took about 1 h, and similar time was necessary to obtain the inverse.

Multiple-trait genetic evaluation for conception rate using outcomes of all available inseminations is technically possible. Large improvement in accuracy is possible using genomic information and computation with the single-step approach is straightforward. More accurate assessment of such an improvement would require either records over a longer period of time or a different methodology for assessment of accuracy.

Appendix A

Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information[1]

[1]I. Misztal, A. Legarra, and I. Aguilar. *Online Journal of Dairy Science*. 92 (9) : 4648–4655. Reprinted here with permission of publisher.

J. Dairy Sci. 92:4648–4655
doi:10.3168/jds.2009-2064

Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information

I. Misztal,*[1] A. Legarra,† and I. Aguilar*‡
*Department of Animal and Dairy Science, University of Georgia, Athens 30602
†Institut National de la Recherche Agronomique (INRA), UR631 SAGA, BP 52627, 32326 Castanet-Tolosan, France
‡Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay

ABSTRACT

Currently, genomic evaluations use multiple-step procedures, which are prone to biases and errors. A single-step procedure may be applicable when genomic predictions can be obtained by modifying the numerator relationship matrix $\mathbf{A}$ to $\mathbf{H} = \mathbf{A} + \mathbf{A}_\Delta$, where $\mathbf{A}_\Delta$ includes deviations from expected relationships. However, the traditional mixed model equations require $\mathbf{H}^{-1}$, which is usually difficult to obtain for large pedigrees. The computations with $\mathbf{H}$ are feasible when the mixed model equations are expressed in an alternate form that also applies for singular $\mathbf{H}$ and when those equations are solved by the conjugate gradient techniques. Then the only computations involving $\mathbf{H}$ are in the form of $\mathbf{Aq}$ or $\mathbf{A}_\Delta\mathbf{q}$, where $\mathbf{q}$ is a vector. The alternative equations have a nonsymmetric left-hand side. Computing $\mathbf{A}_\Delta\mathbf{q}$ is inexpensive when the number of nonzeros in $\mathbf{A}_\Delta$ is small, and the product $\mathbf{Aq}$ can be calculated efficiently in linear time using an indirect algorithm. Generalizations to more complicated models are proposed. The data included 10.2 million final scores on 6.2 million Holsteins and were analyzed by a repeatability model. Comparisons involved the regular and the alternative equations. The model for the second case included simulated $\mathbf{A}_\Delta$. Solutions were obtained by the preconditioned conjugate gradient algorithm, which works only with symmetric matrices, and by the bi-conjugate gradient stabilized algorithm, which also works with nonsymmetric matrices. The convergence rate associated with the nonsymmetric solvers was slightly better than that with the symmetric solver for the original equations, although the time per round was twice as much for the nonsymmetric solvers. The convergence rate associated with the alternative equations ranged from 2 times lower without $\mathbf{A}_\Delta$ to 3 times lower for the largest simulated $\mathbf{A}_\Delta$. When the information attributable to genomics can be expressed as modifications to the numerator relationship matrix, the proposed methodology may allow the upgrading of an existing evaluation to incorporate the genomic information.

Key words: best linear unbiased predictor, genomic selection, single nucleotide polymorphism, genetic evaluation

INTRODUCTION

Availability of dense molecular markers of type SNP led to the recent introduction of the genome-wide or genomic selection evaluation models. Those models are most often based on the simultaneous estimation of SNP marker effects $\mathbf{a}$. Differences among methods are mostly on the a priori distribution of marker effects $\mathbf{a}$ (Meuwissen et al., 2001; Gianola et al., 2006). Efficient procedures exist for the computation of $\mathbf{a}$, even for large data sets (Legarra and Misztal, 2008).

The genomic evaluation is currently implemented as a multistep procedure. For example, an implementation for US dairy cattle (VanRaden, 2008; VanRaden et al., 2009) requires 3 steps: a) regular evaluation by the animal model, b) estimation of genomic effects for a relatively small number of genotyped animals, and c) estimation of genomic breeding values by a selection index. The elements in the index include a parent average or PTA from step a), genomic solutions from step b), and a parent average or PTA computed based on genotyped ancestors. Weights in the index are functions of heritability and accuracy. The marker-assisted selection program in France simultaneously fits QTL and polygenic effects, with weights depending on associated variance components (Guillaume et al., 2008).

Advantages of the multistage procedure include no change to the regular evaluations and simple steps for predicting genomic values for young genotyped animals. Disadvantages are requirements for parameters in steps b) and c) such as prior variances and weights, and loss of accuracy and biases attributable to selection. Whereas the model in a) uses the information on all animals and can be multitrait, the model in b) is equivalent to a single-trait sire model for a highly selected set of sires. Incorrect parameters in b) and

Received January 26, 2009.
Accepted April 29, 2009.
[1]Corresponding author: ignacy@uga.edu

c) can result in unexpected changes for high-reliability bulls. Neuner et al. (2008) claimed that problems associated with the multistep procedure reduce its benefits, especially for cows.

VanRaden (2008) investigated 2 options for step b): "nonlinear," based on estimating effects attributable to SNP markers with a prior mixture distribution for those effects, and "linear," based on prior normal distribution for SNP markers. The latter is equivalent to using mixed model equations with a genomic relationship matrix. For most dairy traits, predictions based on the estimation of marker effects with nonlinear predictions were practically equivalent to linear predictions and thus to predictions with BLUP using a genomic relationship matrix (Cole et al., 2009; VanRaden et al., 2009). Therefore, using the genomic relationship matrix results in little or no loss of accuracy.

One way to simplify the multistep procedure is by incorporating the genomic information into step a), resulting in a single-step procedure. This could be accomplished by modifying the numerator relationship matrix $\mathbf{A}$ in that evaluation to include the genomic information. Such modifications are presented and discussed by Legarra et al. (2009) in a companion paper.

Assume that such a modification is known and that it involves relatively few elements of $\mathbf{A}$. The mixed model equations require $\mathbf{A}^{-1}$, which is very easy to create for large populations because of its sparsity and its special structure (Henderson, 1976). However, obtaining the inverse of the modified matrix is likely to be impossible in general for large populations. This is not only because the cost of inversion is high, but also because $\mathbf{A}$ is dense and thus too large to store for large pedigrees. Thus, an approach using a modified $\mathbf{A}$ is of little value unless a feasible computing approach is available. The purpose of this study is to develop an efficient computing strategy to obtain solutions to mixed model equations in which the numerator relationship matrix is modified by a known matrix accounting for the genomic information.

MATERIALS AND METHODS

Assume regular mixed model equations as used in a traditional genetic evaluation, for simplicity with only a single random effect:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of records, $\mathbf{b}$ is a vector of fixed effects, and $\mathbf{u}$ is a vector of animal effects. Under a polygenic infinitesimal model of inheritance, $\mathrm{var}(\mathbf{u}) = \mathbf{A}\sigma_\alpha^2$, where $\mathbf{A}$ is the numerator relationship matrix based on pedigree. Furthermore, $\mathrm{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$, and $\mathbf{X}$ and $\mathbf{Z}$ are appropriate incidence matrices.

Assume that the numerator relationship can be modified to account for genomic information:

$$\mathbf{H} = \mathbf{A} + \mathbf{A}_\Delta,$$

where $\mathbf{A}_\Delta$ is a matrix that can be stored explicitly, and $\mathbf{H}$ is the new modified matrix. In the simplest case, a genomic relationship matrix $\mathbf{G}$ replaces the numerator relationship matrix for the genotyped animals. Let indices 1 and 2 refer to ungenotyped and genotyped animals, respectively. Then

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

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix},$$

and

$$\mathbf{A}_\Delta = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}.$$

Legarra et al. (2009) proposed several $\mathbf{H}$ based on the partition of animals into several groups, including ungenotyped and genotyped animals. Although their different $\mathbf{H}$ are more complex than in the simple case, most quantities can be computed efficiently without any steps involving large matrix multiplications. Therefore, for simplicity of presentations, the following computing formulas assume the simple case above.

Solving Algorithm

Assume that $\mathbf{G}$ and $\mathbf{A}_{22}$ are available. Temporarily assume that $\mathbf{H}$ is positive definite. The regular mixed model equations are

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \alpha\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

or

$$\mathbf{LHS}\ \mathbf{w} = \mathbf{RHS}$$

using the usual notation, where **LHS** and **RHS** are the left- and right-hand side, and $\mathbf{w} = \begin{bmatrix} \hat{\mathbf{b}}' & \hat{\mathbf{u}}' \end{bmatrix}'$.

Assume that the system of equations is solved using an algorithm that does not require the elements of **LHS** explicitly but only its product by a vector, say **LHS q**, as in the preconditioned conjugate gradient (**PCG**) iteration on data (Tsuruta et al., 2001). Then

$$\mathbf{LHS\ q} = \begin{bmatrix} \mathbf{X'Xq_1 + X'Zq_2} \\ \mathbf{Z'Xq_1 + Z'Zq_2 + \alpha H^{-1} q_2} \end{bmatrix} = \begin{bmatrix} \mathbf{c_1} \\ \mathbf{c_2 + c_3} \end{bmatrix},$$

with

$$\mathbf{q} = \begin{bmatrix} \mathbf{q_1} \\ \mathbf{q_2} \end{bmatrix};\ \mathbf{c_2 = Z'Xq_1 + Z'Zq_2};\ \mathbf{c_3 = \alpha H^{-1} q_2};\ \mathbf{RHS} = \begin{bmatrix} \mathbf{r_1} \\ \mathbf{r_2} \end{bmatrix}.$$

However, $\mathbf{H}^{-1}$ can be computed only for small populations; furthermore, **H** might be singular or close to singularity. Henderson (1984, 1985) and Harville (1976) described an unsymmetric set of mixed model equations in which only **H**, not necessarily of full rank, is required:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{HZ'X} & \mathbf{HZ'Z} + \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{HZ'y} \end{bmatrix}$$

or

$$\mathbf{LHS_M w = RHS_M}.$$

For that set,

$$\mathbf{LHS_M\ q} = \begin{bmatrix} \mathbf{X'Xq_1 + X'Zq_2} \\ \mathbf{HZ'Xq_1 + HZ'Zq_2 + \alpha q_2} \end{bmatrix} = \begin{bmatrix} \mathbf{c_1} \\ \mathbf{Hc_2 + \alpha q_2} \end{bmatrix} = \begin{bmatrix} \mathbf{c_1} \\ \mathbf{Ac_2 + A_\Delta c_2 + \alpha q_2} \end{bmatrix}$$

with

$$\mathbf{RHS_M} = \begin{bmatrix} \mathbf{r_1} \\ \mathbf{Hr_2} \end{bmatrix} = \begin{bmatrix} \mathbf{r_1} \\ \mathbf{Ar_2 + A_\Delta r_2} \end{bmatrix},$$

The new formulas do not include $\mathbf{H}^{-1}$ but include $\mathbf{A_\Delta c_2}$, $\mathbf{Ac_2}$, $\mathbf{A_\Delta r_2}$, and $\mathbf{Ar_2}$. For the simplistic **H**, the first term can be computed directly at a low cost. The second term can also be computed inexpensively following the algorithm by Colleau (2002; see Appendix A), which uses only the pedigree information and is completed in the amount of time proportional to the number of animals. The same algorithm also can be used to compute $\mathbf{Ar_2}$. Selected elements of **A** can be computed recursively, for example, by using the algorithm by Aguilar and Misztal (2008).

More Complicated Models

Assume a multiple trait model, possibly with effects such as random regression or maternal. The regular mixed model equations for such models can be presented as

$$\begin{bmatrix} \dots & \dots \\ \dots & \dots + \mathbf{G_{0a}^{-1}} \otimes \mathbf{H^{-1}} \end{bmatrix} \begin{bmatrix} \dots \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \dots \\ \mathbf{Z'R^{-1}y} \end{bmatrix},$$

where parts not listed (. . .) are due to effects other than $\hat{\mathbf{u}}$. By expanding the unsymmetric model by Henderson (1984) to multiple traits, the quantities needed for the iterations become

$$\mathbf{LHS_M q} = \begin{bmatrix} \mathbf{c_1} \\ \mathbf{G_0 \otimes Ac_2 + G_0 \otimes A_\Delta c_2 + Iq_2} \end{bmatrix}$$

with

$$\mathbf{RHS_M} = \begin{bmatrix} \mathbf{r_1} \\ \mathbf{(G_0 \otimes A + G_0 \otimes A_\Delta) r_2} \end{bmatrix},$$

where quantities $\mathbf{c}_1$ and $\mathbf{r}_1$ are now associated with all effects other than the additive.

Nonsymmetric Solvers

The presented system of equations is nonsymmetric and the matrix **H** may be semipositive definite. The PCG algorithm (Barrett et al., 1994) is applicable only to symmetric systems of equations. Therefore, it is important to find a suitable conjugate-gradient type algorithm and ensure that it would converge even with a poorly conditioned **H**. Barrett et al. (1994) and Van der Vorst (2003) reviewed and presented several algorithms for solving the linear systems of equations. Based on their studies, the standard algorithm for solving sparse systems with nonsymmetric LHS is bi-conjugate gradient stabilized (**Bi-CGSTAB**; Van der Vorst, 1992; see Appendix B). This algorithm requires 2 **LHS** times a vector products per round as opposed to just one with PCG. When that product uses the majority of the computing time, Bi-CGSTAB is about twice as expensive as PCG per round of iteration.

Choice of Preconditioners

In initial tests (results not reported), Bi-CGSTAB converged very quickly with the unsymmetric equations for small models, but not for large ones. This was traced to large off-diagonal elements of the unsymmetric equations. The standard way in conjugate-gradient types

of algorithms to improve convergence is by choice of a preconditioner $\mathbf{M}$, which approximates **LHS** but is easily invertible (Van der Vorst, 2003). Then the system of equations solved is equivalent to

$$\mathbf{M}^{-1}\mathbf{LHS}\ \mathbf{w} = \mathbf{M}^{-1}\mathbf{RHS},$$

which has better numerical properties than the original system. The preconditioner is never used explicitly, but only in multiplications with a vector.

Assuming a diagonal preconditioner,

$$\mathbf{M}^{-1} = \mathrm{diag}\left(\mathbf{LHS}\right)^{-1} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix},$$

the **LHS** for regular equations after preconditioning is

$$\mathbf{M}^{-1}\ \mathbf{LHS} = \begin{bmatrix} \mathbf{D}_1\mathbf{X}'\mathbf{X} & \mathbf{D}_1\mathbf{X}'\mathbf{Z} \\ \mathbf{D}_2\mathbf{Z}'\mathbf{X} & \left[\mathbf{D}_2\left(\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{A}^{-1}\right)\right] \end{bmatrix}.$$

The symmetry can be partially restored with a modified preconditioner:

$$\mathbf{M}_{\mathrm{M}}^{-1} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2\mathbf{A}^{-1} \end{bmatrix}.$$

Then

$$\begin{aligned} &\mathbf{M}_{\mathrm{M}}^{-1}\ \mathbf{LHS}_{\mathrm{M}} \\ &= \begin{bmatrix} \mathbf{D}_1\mathbf{X}'\mathbf{X} & \mathbf{D}_1\mathbf{X}'\mathbf{Z} \\ \mathbf{D}_2\left[\left(\mathbf{I} + \mathbf{A}^{-1}\mathbf{A}_{\Delta}\right)\mathbf{Z}'\mathbf{X}\right] & \mathbf{D}_2\left[\left(\mathbf{I} + \mathbf{A}^{-1}\mathbf{A}_{\Delta}\right)\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{A}^{-1}\right] \end{bmatrix} \\ &= \mathbf{M}^{-1}\ \mathbf{LHS} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{D}_2\mathbf{A}^{-1}\mathbf{A}_{\Delta}\mathbf{Z}'\mathbf{X} & \mathbf{D}_2\mathbf{A}^{-1}\mathbf{A}_{\Delta}\mathbf{Z}'\mathbf{Z} \end{bmatrix}. \end{aligned}$$

When the genomic information is missing ($\mathbf{A}_{\Delta} = 0$), the preconditioned left-hand side of the unsymmetric system of equations is the same as with the preconditioned regular equations. With the genomic information, the off-diagonal elements are likely to be small for small $\mathbf{A}_{\Delta}$. The cost of the extra preconditioning is low because the product $\mathbf{D}_2\mathbf{A}^{-1}\mathbf{q}$, where $\mathbf{q}$ is a vector, can be done sequentially as $\mathbf{D}_2(\mathbf{A}^{-1}\mathbf{q})$.

Data

The data set included 10.5 million final scores on 6.2 million Holsteins as used for the recent genetic evaluation by the Holstein Association. Analyses were by a repeatability animal model. Two sets of mixed model equations were considered: regular and unsymmetric. For the second set, the genomic information was simulated for 5,000 randomly chosen animals as random numbers from the uniform distribution from 0 to b, where b was set to 0.0, 0.01, 0.03, and 0.05. For b = 0 there was no adjustment ($\mathbf{A}_{\Delta} = 0$). Only positive adjustments were included to avoid some elements of $\mathbf{H}$ being negative. Solving algorithms were PCG (for the regular equations only) and Bi-CGSTAB. The first algorithm used a diagonal preconditioner. The second algorithm used the modified preconditioner because no convergence was achieved with the diagonal preconditioner. In all cases, the stopping criterion was set at 10^{-12}. Computing was by the regular and modified program BLUP90IOD (Tsuruta et al., 2001) and was carried out on an Opteron system running at 3 GHz.

RESULTS AND DISCUSSION

The purpose of testing with the simulated genomic changes was to evaluate the computing feasibility of the method, and especially the robustness of the computing methodology. The results presented for the unsymmetric equations are only with the modified preconditioner. The Bi-CGSTAB diverged with the regular preconditioner and large data sets although it converged with small data sets. This is because products of $\mathbf{A}$ were very large for rows corresponding to popular bulls as all elements of $\mathbf{A}$ are positive; those products with $\mathbf{A}^{-1}$ are small because of cancellations; a contribution to a parent by a progeny in $\mathbf{A}^{-1}$ is proportional to [... 1.0 ... −0.5 ... −0.5 ...], which sums to 0.

Table 1 shows the number of rounds and computing time with PCG and Bi-CGSTAB for the regular and unsymmetric equations and with varying magnitudes of simulated changes. For the regular equations, Bi-CGSTAB was slightly faster but took twice the computing time (26 vs. 13 s). Figure 1 shows the convergence pattern for the regular equations. Whereas the pattern for PCG shows small fluctuations, the pattern for Bi-CGSTAB has more abrupt changes. Some differences in the number of rounds to convergence may be due to differences in the convergence criteria. However, the differences in solutions were very small (correlations >0.99999).

For the unsymmetric equations with no simulated changes, the number of rounds approximately doubled and the computing time increased by 30% (from 26 to 34 s). Adding small simulated changes (b = 0.01) increased the computing time per round by 10% (from 34 to 37 s) and slightly deteriorated convergence. The number of rounds increased by about 30% when changes were increased to b = 0.03 and again by 10% when

Table 1. The number of rounds (computing time per round in seconds) for different computing algorithms and different magnitudes of modification to the numerator relationship matrix

	Equation				
		Unsymmetric[2]			
Solving algorithm[1]	Regular	b = 0	b = 0.01	b = 0.03	b = 0.05
PCG	189 (13.1)	—	—	—	—
Bi-CGSTAB	166 (26.0)	318 (34.0)	369 (37.3)	477 (37.1)	520 (37.4)

[1]PCG = preconditioned conjugate gradient; Bi-CGSTAB = bi-conjugate gradient stabilized.
[2]Changes in relationships simulated from uniform (0, b) distribution for 5,000 randomly selected animals.

changes were increased by b = 0.05. Figure 2 shows the convergence pattern for the modified equations and b = 0.0 and 0.03. Much larger fluctuations than with the regular equations were observed, which may have been due to a more complex preconditioner. For a multiple-trait random regression model, Aguilar et al. (2008) observed much larger fluctuations in the convergence pattern with a block-diagonal preconditioner as compared with a diagonal one.

Additional computations will be necessary in practical applications of the method with the real genomic relationship matrix. For simple **H**, additional steps include the multiplications of $\mathbf{G}\text{-}\mathbf{A}_{22}$ and **A** by a vector. The last one can be done efficiently using the algorithm of Colleau (2002) in linear time (see Appendix A). The cost of this algorithm is equal to scanning the pedigree file twice and is small, especially with pedigrees in memory. Legarra et al. (2009) presented formulas for

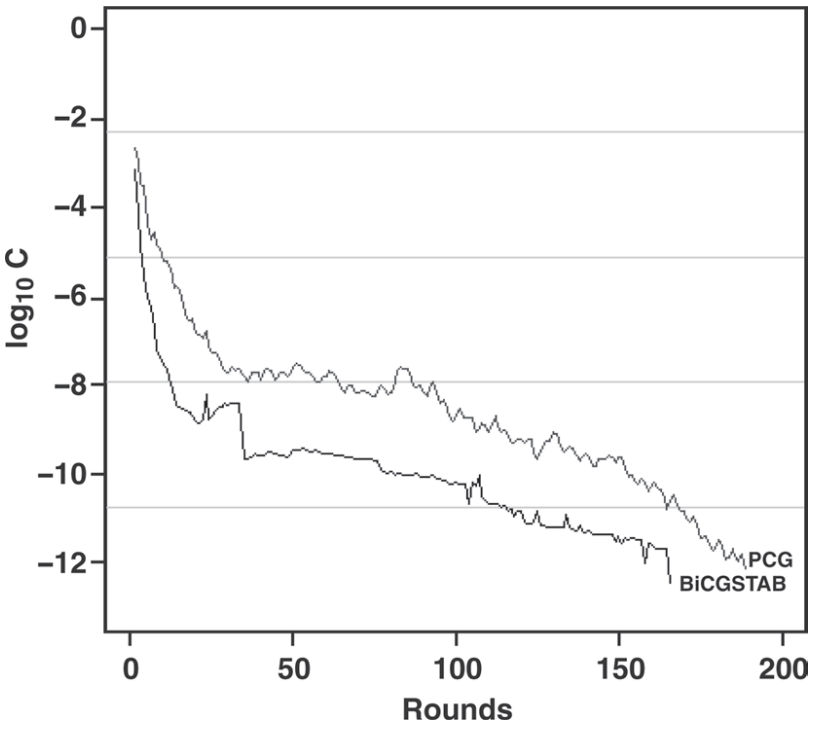

Figure 1. Convergence rate for the preconditioned conjugate gradient (PCG) and bi-conjugate gradient stabilized (Bi-CGSTAB) algorithms with the symmetric system of equations.

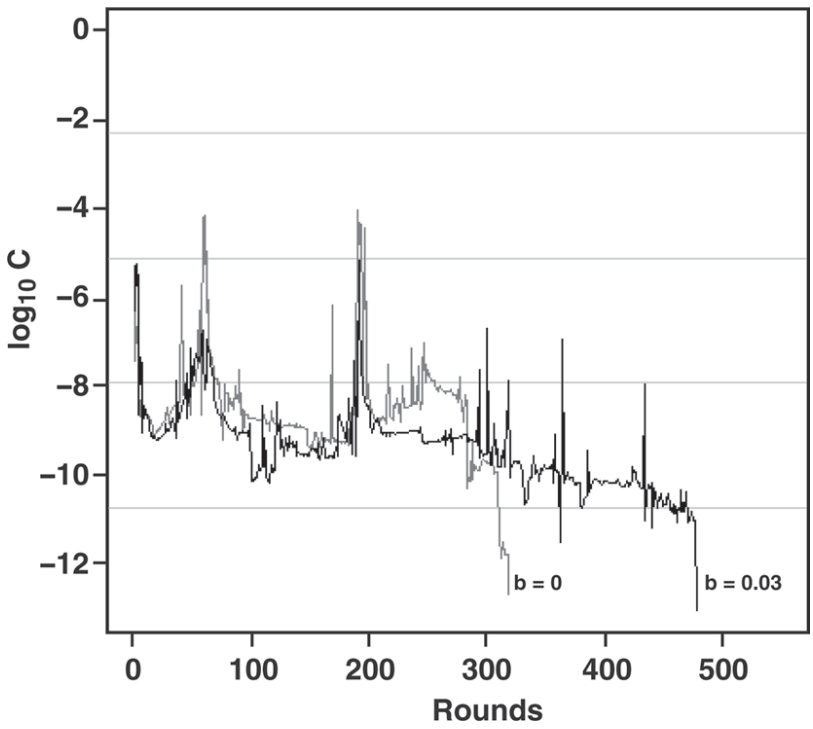

Figure 2. Convergence rate for the bi-conjugate gradient stabilized (Bi-CGSTAB) algorithm with the unsymmetric system of equations and either no (b = 0) or a middle level (b = 0.03) of simulated changes attributable to the genomic relationship matrix.

more realistic **H** and also computing details for a product of that **H** by a vector. With such a product, the only components that cannot be computed in linear but rather in quadratic time (matrix-vector multiplication) are those corresponding to **G** and possibly those due to $\mathbf{A}_{22}$. If $\mathbf{A}_{22}$ needs to be available explicitly, it can be computed by the method of Aguilar and Misztal (2008). When applied to 17 million Holsteins, that method calculated about 80,000 inbreeding coefficients/s. Assuming that computing one relationship costs no more than computing one inbreeding coefficient, on average, the computation of $\mathbf{A}_{22}$ for 20,000 animals would take 40 min. Alternatively, $\mathbf{A}_{22}$ can be computed by the repeated applications of the algorithm of Colleau (2002), in which the vector to multiply by would contain one 1.0 and zeros elsewhere.

When the number of genotyped animals is very high, say >50,000, storage and computations with matrix **G** and possibly $\mathbf{A}_{22}$ can be quite involving. A few choices may be applicable. First, some computations may easily be done in parallel. Current computers routinely include 4 processors (cores) per processor module, and computers with 4 modules are readily available. Second, some elements in $\mathbf{A}_{\Delta}$ may be very small or unimportant and could be neglected. Neglecting small elements in the computation of sparse inverse for the purpose of calculating accuracies reduced the computations by 50 times while retaining high precision (Thompson et al., 1994). Finally, genotypes of some animals may be unimportant and do not have to be included.

In summary, we have demonstrated that mixed model equations with small modifications to the numerator relationship matrix can be solved efficiently by conjugate-gradient type algorithms. Only a few modifications may be required for existing programs using the PCG algorithm.

ACKNOWLEDGMENTS

Discussions with Rohan Fernando (Department of Animal Science, Iowa State University, Ames), Jeff O'Connell (University of Maryland School of Medicine,

Baltimore), Paul VanRaden (Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD), Curt Van Tassel (Bovine Functional Genomics Laboratory, ARS, USDA, Beltsville, MD), and Bruce Tier (Animal Genetics and Breeding Unit, University of New England, Armidale, Australia) are gratefully acknowledged. Also acknowledged are the encouragement to pursue this study by Tom Lawlor and the financial support by the Holstein Association, Brattleboro, Vermont (IM, IA), the EADGENE network of excellence, Agence National de la Recherche project AMASGEN, and Maison de Relations Internationales (INRA, France). We also appreciate the very diligent work by the 2 anonymous reviewers.

REFERENCES

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

Aguilar, I., S. Tsuruta, and I. Misztal. 2008. Computing options for multiple trait test day random regression models with account of heat tolerance and national datasets. J. Dairy Sci. 91(Suppl. 1):9. (Abstr.)

Barrett, R., M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. 1994. Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods. SIAM, Philadelphia, PA.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92:2931–2946.

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

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

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.

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

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

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. Univ. Guelph, Guelph, Ontario, 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., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663.

Legarra, A., and I. Misztal. 2008. Technical note: Computing strategies in genome-wide selection. J. Dairy Sci. 91:360–366.

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.

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.

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

Thompson, R., N. R. Wray, and R. E. Crump. 1994. Calculation of prediction error variances using sparse matrix methods. J. Anim. Breed. Genet. 111:102–109.

Tsuruta, S., I. Misztal, and I. Stranden. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. J. Anim. Sci. 79:1166–1172.

Van der Vorst, H. 1992. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Statist. Comput. 13:631–644.

Van der Vorst, H. 2003. Iterative Krylov Methods for Large Linear Systems. Cambridge University Press, Cambridge, UK

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.

APPENDIX A

Below we show how to create a product $\mathbf{Aq}$, where $\mathbf{A}$ is the numerator relationship matrix and $\mathbf{q}$ is a vector. The recurrence equation for the additive effect is

$$\mathbf{a} = \mathbf{Pa} + \boldsymbol{\phi},$$

where $\mathbf{a}$ is a vector of animals ordered from oldest to youngest, $\boldsymbol{\phi}$ is a diagonal matrix of Mendelian samplings, and $\mathbf{P}$ is a matrix relating animals to their parents; this matrix has at most 2 elements per row, both equal to 0.5. Following Quaas (1988),

$$\text{Var}(\mathbf{a}) = \mathbf{A} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{D}(\mathbf{I} - \mathbf{P})^{-1'},$$

where $\mathbf{D} = \text{var}(\boldsymbol{\phi})$. Colleau (2002) showed that the product of $\mathbf{A}$ by a vector, for example,

$$\mathbf{v} = \mathbf{Aq} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{D}(\mathbf{I} - \mathbf{P})^{-1\prime}\mathbf{q} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{D}\,[(\mathbf{I} - \mathbf{P})^{-1\prime}\mathbf{q}],$$

can be solved in linear time. In particular, quantities $\mathbf{r} = (\mathbf{I} - \mathbf{P})^{-1\prime}\mathbf{q}$ and $\mathbf{v} = (\mathbf{I} - \mathbf{P})^{-1}\mathbf{Dr}$ can be obtained by solving $(\mathbf{I} - \mathbf{P})'\mathbf{r} = \mathbf{q}$ and $(\mathbf{I} - \mathbf{P})\mathbf{v} = \mathbf{Dr}$, each one in a single sweep because $(\mathbf{I} - \mathbf{P})$ is triangular. The scalar formulas are

$$r_i = r_i + q_i;\ r_{si} = r_{si} + r_i/2;\ r_{di} = r_{di} + r_i/2;\ i = n, ..., 1$$

$$v_i = d_i r_i + (v_{si} + v_{di})/2,\ i = 1, ..., n,$$

where s_i and d_i are positions of the sire and dam of animal i, respectively.

The Colleau (2002) algorithm can be used to compute products of sections of matrices. For instance, the products below show how to compute $\mathbf{A}_{12}\mathbf{q}$, $\mathbf{A}_{22}\mathbf{q}$, $\mathbf{A}_{21}\mathbf{q}$, or $\mathbf{A}_{22}\mathbf{q}$:

$$\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{q}\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{11}\mathbf{q}\\ \mathbf{A}_{21}\mathbf{q}\end{bmatrix},\ \begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{q}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{12}\mathbf{q}\\ \mathbf{A}_{22}\mathbf{q}\end{bmatrix}.$$

APPENDIX B

The pseudo-program below implements the Bi-CG-STAB (Van der Vorst, 1992) for a system of equations $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{M}$ being a preconditioner. The major expenses in the algorithm are products of $\mathbf{A}$ by a vector, possibly followed by products of $\mathbf{M}^{-1}$, but only if $\mathbf{M}$ is of complex structure.

Compute $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{Ax}^{(0)}$ for some initial guess $\mathbf{x}^{(0)}$

Choose $\tilde{\mathbf{r}}$ (for example, $\tilde{\mathbf{r}} = \mathbf{r}^{(0)}$)

for i = 1, 2, …

$\rho_{i-1} = \tilde{\mathbf{r}}'\mathbf{r}^{(i-1)}$

if ρ_{i-1} = 0 method fails

if i = 1

$\mathbf{p}^{(i)} = \mathbf{r}^{(i-1)}$

else

$\beta_{i-1} = \left(\frac{\rho_{i-1}}{\rho_{i-2}}\right)\left(\frac{\alpha_{i-1}}{\omega_{i-1}}\right)$

$\mathbf{p}^{(i)} = \mathbf{r}^{(i-1)} + \beta_{i-1}\left(\mathbf{p}^{(i-1)} - \omega_{i-1}\boldsymbol{\upsilon}^{(i-1)}\right)$

endif

solve $\boldsymbol{M}^{-1}\hat{\mathbf{p}} = \mathbf{p}^{(i)}$

$\boldsymbol{\upsilon}^{(i)} = \mathbf{A}\hat{\mathbf{p}}$

$\alpha_i = \frac{\rho_{i-1}}{\tilde{\mathbf{r}}'\boldsymbol{\upsilon}^{(i)}}$

$\mathbf{g} = \mathbf{r}^{(i-1)} - \alpha_i\boldsymbol{\upsilon}^{(i)}$

check norm of g; if small enough: set

$\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + \alpha_i\hat{\mathbf{p}}$ and stop

solve $\mathbf{M}\hat{\mathbf{g}} = \mathbf{g}$

$\mathbf{t} = \mathbf{A}\hat{\mathbf{g}}$

$\omega_i = \frac{\mathbf{t}'\mathbf{g}}{\mathbf{t}'\mathbf{t}}$

$\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} + \alpha_i\hat{\mathbf{p}} + \omega_i\hat{\mathbf{g}}$

$\mathbf{r}^{(i)} = \mathbf{g} - \omega_i\mathbf{t}$

check for convergence; continue if necessary

for continuation it is necessary that $\omega_i \neq 0$

end

Appendix B

A Relationship Matrix Including Full Pedigree and Genomic Information[1]

[1]A. Legarra, I. Aguilar and I. Misztal. *Online Journal of Dairy Science*. 92 (9) : 4656–4663. Reprinted here with permission of publisher.

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 $\mathbf{A}$:

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

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 $\mathbf{A}_g$.

Matrix $\mathbf{A}_g$ 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 $\mathbf{A}_g$, 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 $\mathbf{A}_g$ 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 $\mathbf{T'T}$ 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 $\mathbf{A}_g$ 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 $\mathbf{u} = \mathbf{Pu} + \boldsymbol{\varphi}$, 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), $\mathbf{A}_p$ 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 $\mathbf{A}_p$ 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 $\mathbf{A}_p$ would work if all founders were genotyped (e.g., in a nucleus scheme); in this case, the system is fully coherent. For practical purposes, $\mathbf{A}_p$ 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 $\mathbf{u}_2 = \mathbf{Z}\mathbf{a}$. Then

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

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

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

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

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

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 $\mathbf{A}_p$ 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 $\mathbf{A}\begin{bmatrix}\mathbf{t}_1\\ \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{11}\mathbf{t}_1\\ \mathbf{A}_{21}\mathbf{t}_1\end{bmatrix} = \begin{bmatrix}\mathbf{z}\\ \mathbf{y}\end{bmatrix}$, 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).

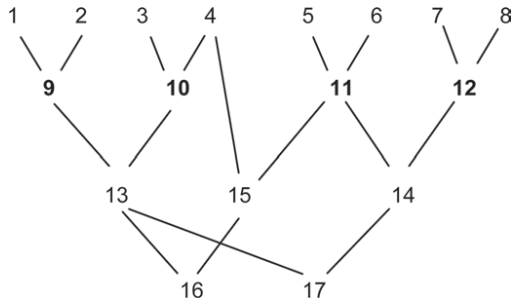

Figure 1. Example pedigree. Genotyped animals are in bold.

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.

Appendix C

Derivation of the inverse for the combined relationship matrix[1]

Let the inverse of the numerator relationship matrix ($\mathbf{A}$) be:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix},$$

where animals are partitioned into 2 groups with group 2 denoting genotyped animals. To derive an inverse for the combined relationship matrix of Legarra et al. (2009), using the properties of the inverse of partitioned matrix, useful identities from $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ are:

$$\mathbf{A}^{11}\mathbf{A}_{11} + \mathbf{A}^{12}\mathbf{A}_{21} = \mathbf{I}, \tag{C1}$$

$$\mathbf{A}^{21}\mathbf{A}_{12} + \mathbf{A}^{22}\mathbf{A}_{22} = \mathbf{I}, \tag{C2}$$

$$\mathbf{A}^{11}\mathbf{A}_{12} + \mathbf{A}^{12}\mathbf{A}_{22} = \mathbf{0}, \tag{C3}$$

$$\mathbf{A}^{21}\mathbf{A}_{11} + \mathbf{A}^{22}\mathbf{A}_{21} = \mathbf{0}, \text{ and} \tag{C4}$$

$$\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right) = \mathbf{A}^{11} \tag{C5}$$

using (C1) through (C4) and multiplying the whole-population matrix

[1]Developed by D. L. Johnson (Livestock Improvement Corp., Hamilton, New Zealand)

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}-\mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G}-\mathbf{A}_{22}) \\ (\mathbf{G}-\mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}-\mathbf{A}_{22} \end{bmatrix}$$

by

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

gives $\mathbf{H}^{-1}\mathbf{H} = \mathbf{I}$.

A direct approach to getting $\mathbf{H}^{-1}$ comes from the distribution function. Based on the conditional distribution

$$\mathbf{u}_1|\mathbf{u}_2 \sim N(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})$$

and (C1) through (C5), the full distribution can be written as

$$\begin{aligned} p(\mathbf{u}_1,\mathbf{u}_2) &= p(\mathbf{u}_1,\mathbf{u}_2|\mathbf{u}_2)p(\mathbf{u}_2) \\ &= p(\mathbf{u}_1|\mathbf{u}_2)p(\mathbf{u}_2) \\ &\propto \exp[-0.5\,(\mathbf{u}_1-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)'\mathbf{A}^{11}(\mathbf{u}_1-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)]\exp[-0.5\mathbf{u}'_2\mathbf{G}^{-1}\mathbf{u}_2] \\ &= \exp\left(-0.5\begin{bmatrix}\mathbf{u}'_1 & \mathbf{u}'_2\end{bmatrix}\begin{bmatrix}\mathbf{A}^{11} & -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11} & \mathbf{G}^{-1}+\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{bmatrix}\right) \\ &= \exp\left(-0.5\begin{bmatrix}\mathbf{u}'_1 & \mathbf{u}'_2\end{bmatrix}\begin{bmatrix}\mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{G}^{-1}+\mathbf{A}^{22}-\mathbf{A}_{22}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{bmatrix}\right) \end{aligned} \tag{C6}$$

The matrix in (C6) is the inverse of the variance matrix of the full distribution. Therefore

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

Appendix D

Decomposition of joint predictions

To illustrate the role of λ and decomposition of joint predictions in PA, genomic prediction (GP), and pedigree prediction from the subset of genotyped relatives (PP_{22}), consider, $\mathbf{H}^{-1}$ after including λ:

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} + \lambda(\mathbf{G}^{-1} + \mathbf{A}_{22}^{-1}) \end{bmatrix}. \tag{D1}$$

Denote $\mathbf{H}^{-1}$ as $\{h^{ij}\}$, $\mathbf{G}^{-1}$ as $\{g^{ij}\}$, and $\mathbf{A}_{22}^{-1}$ as $\{a_{22}^{ij}\}$. Consider the equation for breeding value u_i of individual i without records or progeny, in the spirit of VanRaden and Wiggans (1991); k indicates genotyped individuals (in $\mathbf{A}_{22}$), and j indicates all individuals (in $\mathbf{A}$):

$$\sum_{\mathrm{j}} h^{ij} u_j = 0$$

and

$$\lambda \sum_k g^{ik} u_k + (1-\lambda) \sum_k a_{22}^{ik} u_k + \sum_j a^{ij} u_j - \sum_k a_{22}^{ik} u_k = \lambda \sum_k g^{ik} u_k - \lambda \sum_k a_{22}^{ik} u_k + \sum_j a^{ij} u_j = 0$$

Thus, for $\lambda = 0$, only contributions from pedigree relationships remain. Consider more specifically young animal i without records or progeny. The equation with inbreeding ignored is

$$-u_s - u_d + 2u_i + \lambda \sum_j (g^{ij} - a_{22}^{ij}) u_j = 0,$$

where s and d correspond to sire and dam, respectively. Then

$$
\begin{aligned}
u_i &= \frac{u_s + u_d + \lambda \sum_{j,\, j \neq i} (a_{22}^{ij} - g^{ij}) u_j}{2 + \lambda (g^{ii} - a_{22}^{ii})} \\
&= \left(\frac{u_s + u_d}{2} \right) \left(\frac{2}{2 + \lambda (g^{ii} - a_{22}^{ii})} \right) + \left(\frac{\lambda \sum_{j,\, j \neq i} a_{22}^{ij} u_j}{2 + \lambda (g^{ii} - a_{22}^{ii})} \right) - \left(\frac{\lambda \sum_{j,\, j \neq i} g^{ij} u_j}{2 + \lambda (g^{ii} - a_{22}^{ii})} \right) \qquad \text{(D2)} \\
&= 2(w) PA + \lambda(w)\, g^{ii}\, GP - \lambda(w) a_{22}^{ii} PP_{22},
\end{aligned}
$$

where

$$
PA = \left(\frac{u_s + u_d}{2} \right),
$$

$$
GP = \frac{- \sum_{j,j \neq i} g^{ij} u^j}{g^{ii}},
$$

and

$$
PP_{22} = \frac{- \sum_{j,j \neq i} a_{22}^{ij} u^j}{a_{22}^{ii}},
$$

that is, parent average, genomic prediction, and subset pedigree prediction, with weights summing to one. These are the same sources of information as in VanRaden et al. (2009) except that they are estimated jointly. Note that PP_{22} might be different from PA because 1) both parents might not be genotyped and 2) only the subset of genotyped animals is considered if PP_{22} is computed independently (as in $PedGenM_{04}$). If $\lambda = 0$, only PA remains; if $\lambda = 1$, then weighting of the 3 sources of information depends on the elements $a^{ii} = 2$, g^{ii}, and a_{22}^{ii}, which measures the precision of the 3 information sources relative to other breeding values. That approach is similar to the reliabilities used to combine the 3 information in VanRaden et al. (2009).