J. Dairy Sci. 91:4414–4423
doi:10.3168/jds.2007-0980

Efficient Methods to Compute Genomic Predictions

P. M. VanRaden[1]
Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350

ABSTRACT

Efficient methods for processing genomic data were developed to increase reliability of estimated breeding values and to estimate thousands of marker effects simultaneously. Algorithms were derived and computer programs tested with simulated data for 2,967 bulls and 50,000 markers distributed randomly across 30 chromosomes. Estimation of genomic inbreeding coefficients required accurate estimates of allele frequencies in the base population. Linear model predictions of breeding values were computed by 3 equivalent methods: 1) iteration for individual allele effects followed by summation across loci to obtain estimated breeding values, 2) selection index including a genomic relationship matrix, and 3) mixed model equations including the inverse of genomic relationships. A blend of first- and second-order Jacobi iteration using 2 separate relaxation factors converged well for allele frequencies and effects. Reliability of predicted net merit for young bulls was 63% compared with 32% using the traditional relationship matrix. Nonlinear predictions were also computed using iteration on data and nonlinear regression on marker deviations; an additional (about 3%) gain in reliability for young bulls increased average reliability to 66%. Computing times increased linearly with number of genotypes. Estimation of allele frequencies required 2 processor days, and genomic predictions required <1 d per trait, and traits were processed in parallel. Information from genotyping was equivalent to about 20 daughters with phenotypic records. Actual gains may differ because the simulation did not account for linkage disequilibrium in the base population or selection in subsequent generations.
Key words: genomic selection, mixed model, computer program, relationship matrix

INTRODUCTION

Genomic selection increases the rate of genetic improvement and reduces cost of progeny testing by allowing breeders to preselect animals that inherited chromosome segments of greater merit (Meuwissen et al., 2001; Schaeffer, 2006). Single nucleotide polymorphism (**SNP**) markers can now cover the genome with high density and are inexpensive to obtain. Evaluations based on SNP genotypes can be computed as soon as DNA can be obtained, which allows selection in both sexes early in life. Application of genomic selection to dairy cattle has just begun (de Roos et al., 2007; van der Beek, 2007; Guillaume et al., 2008). Potential methods and strategies were compared by Meuwissen (2007).

Computer algorithms and programs are needed to incorporate genomic data into genetic evaluations and to process the rapidly expanding numbers of SNP genotypes. Previous algorithms for including markers often fit effects individually rather than simultaneously or fit additional polygenic effects because marker coverage of the genome was not yet complete (de Roos et al., 2007). Iterative algorithms such as Gauss-Seidel and preconditioned conjugate gradient can be used to estimate allele effects (Legarra and Misztal, 2008), but fewer numerical problems may result from direct inversion of variance matrices or mixed model equations (Lee and van der Werf, 2006). Genomic relationships can be included in multitrait derivative-free REML programs (Zhang et al., 2007).

Objectives of this research were 1) to develop computer methods to include genomic data in predictions, 2) to apply the methods to simulated data for actual Holstein and Jersey pedigrees, and 3) to estimate gains in reliability from genomic predictions.

MATERIALS AND METHODS

Predictions were computed by linear and nonlinear systems of equations. The linear predictions assumed that all markers contributed equally to genetic variation (no major genes). The nonlinear (Bayesian) predictions assumed that the prior distribution of marker or QTL effects was not normal. Genetic variance may not be equal across chromosomes or markers because, for example, major genes may exist on some chromosomes. The data vector in both linear and nonlinear predictions was modeled as a linear function of the unknown effects, but solutions for the unknown effects in the nonlinear predictions were nonlinear functions of the data vector. Nonlinear predictions may be better than

Received December 31, 2007.
Accepted June 26, 2008.
[1]Corresponding author: paul.vanraden@ars.usda.gov

the best linear predictions when data are not normally distributed (Henderson, 1963), in this case because of major genes.

Linear predictions were computed by selection index and by 2 equivalent sets of mixed model equations. One model estimated individual marker effects first and then summed those as in Meuwissen et al. (2001). The other model included a genomic relationship matrix in place of the traditional additive genetic relationship matrix (Garrick, 2007; VanRaden, 2007; Zhang et al., 2007). Regressions were on genotypes rather than haplotypes because haplotyping would increase computation time with little or no gain in accuracy at high marker numbers (Calus et al., 2008). Advantages from nonlinear predictions and from haplotyping were expected to be small because the allele effects simulated were not large.

Simulation

Data were simulated so that the experimental design and analysis methods could be assessed before actual genotyping was conducted. The simulated data provided a test for the computer programs, algorithms, time required, and reliability of predictions. The simulated genotypes differ from actual genotypes primarily in that all loci in the base population were in Hardy-Weinberg equilibrium and that selection was not practiced in the following generations, which could affect reported correlations and other statistics. Also, the distribution of QTL effects used in the simulation may not match the biology of actual traits.

Marker and QTL inheritance was simulated for 50,000 biallelic markers and 100 biallelic QTL on 30 equal-length chromosomes. Markers and QTL were randomly scattered across the genome by assigning the position of each by uniform distribution. None of the markers directly affected any trait; instead, the QTL had effects that were simulated with a normal distribution. Variance of QTL effects thus followed a chi-square distribution, with the largest effect accounting for about 10% of genetic variance. Allele frequencies were uniformly distributed between 0 and 1.

Predictions were tested with simulated genotypes for 2,967 Holstein bulls and 766 Jersey bulls. The Holstein bulls included 1,885 bulls born from 1995 through 1997, 290 ancestor bulls included in computing predictions, and 792 younger bulls born from 2001 through 2002 for testing predictions. The Jersey bulls included 563 older bulls to compute predictions plus 203 younger bulls to test predictions. Genotypes for all known ancestors born since 1950 also were simulated in age order for a total of 23,105 Holstein and 7,737 Jersey cows and bulls. Alleles in the earliest generation were in Hardy-Weinberg equilibrium. For subsequent generations, maternal or paternal chromosome segments were inherited with a mean of 1.7 crossovers simulated from Poisson distribution and located with uniform distribution across the chromosome. For actual data, mean crossovers per chromosome probably are closer to 1.0.

Genomic predictions can be obtained by combining traditional genetic evaluation results with genotypic data instead of reprocessing all phenotypic and pedigree data. Variables analogous to deregressed evaluations or daughter yield deviations (**DYD**) were simulated by adding an independent error to the simulated true breeding value. To mimic actual data as closely as possible, that DYD error variance was calculated from published USDA reliabilities for net merit and from reliabilities for parent averages from an actual set of bulls that are being genotyped by USDA (Van Tassell et al., 2007).

Reliability of net merit for older bulls was obtained from May 2003 evaluations when the younger bulls to be predicted were just 1 to 2 yr old. Parent averages were compared with genomic predictions for ability to predict either true breeding values of young bulls or simulated DYD with the same reliability as August 2007 evaluations. Observed reliability of genomic predictions was obtained by squaring the correlations of estimated with true breeding value. Regressions of true breeding value and of DYD on genomic predictions also were calculated. When DYD was the dependent variable, regressions were weighted by reliability from daughters, which was computed as total daughter equivalents minus daughter equivalents from parent average (VanRaden and Wiggans, 1991).

Genomic data sets include some missing genotypes and incorrect genotypes. Fractions of each were arbitrarily set to 1%. The number of Holstein genotypes generated was about a billion (23,105 animals in pedigree file × 50,000 SNP), but only 150 million (2,967 genotyped bulls × 50,000 SNP) were stored and used for genomic prediction. Seven replicates were formed by generating 7 independent sets of QTL effects and environmental errors. For the Jersey simulation, 10 replicates were formed. Time required for the Holstein simulation was only about 10 min, but initial memory usage was high at 8 gigabytes. Memory in the simulation program was reduced to <1 gigabyte by processing each chromosome separately and reusing the memory.

Genomic Relationships and Inbreeding

Let **M** be the matrix that specifies which marker alleles each individual inherited. Dimensions of **M** are the number of individuals (n) by the number of loci (m).

Equations can include marker information using $n \times n$ matrix $\mathbf{MM}'$ or $m \times m$ matrix $\mathbf{M}'\mathbf{M}$ (Legarra and Misztal, 2008). If elements of $\mathbf{M}$ are set to −1, 0, and 1 for the homozygote, heterozygote, and other homozygote, respectively, diagonals of $\mathbf{MM}'$ count the number of homozygous loci for each individual, and off-diagonals measure the number of alleles shared by relatives. In contrast, diagonals of $\mathbf{M}'\mathbf{M}$ count the number of homozygous individuals for each locus, and off-diagonals measure the number of times alleles at different loci were inherited by the same individual.

Let the frequency of the second allele at locus i be p_i, and let $\mathbf{P}$ contain allele frequencies expressed as a difference from 0.5 and multiplied by 2, such that column i of $\mathbf{P}$ is $2(p_i - 0.5)$. Subtraction of $\mathbf{P}$ from $\mathbf{M}$ gives $\mathbf{Z}$, which sets mean values of the allele effects to 0. Allele frequencies in $\mathbf{P}$ should be from the unselected base population rather than those that occur after selection or inbreeding. An earlier or later base population can lead to greater or fewer relationships and to more or less inbreeding. Subtraction of $\mathbf{P}$ gives more credit to rare alleles than to common alleles when calculating genomic relationships. Also, the genomic inbreeding coefficient is greater if the individual is homozygous for rare alleles than if homozygous for common alleles.

Genomic relationship matrix $\mathbf{G}$ can be obtained by at least 3 methods. The first uses the formula $\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$. Division by $2\sum p_i(1-p_i)$ scales $\mathbf{G}$ to be analogous to the numerator relationship matrix $\mathbf{A}$. The genomic inbreeding coefficient for individual j is simply $G_{jj} - 1$, and genomic relationships between individuals j and k, which are analogous to the relationship coefficients of Wright (1922), are obtained by dividing elements G_{jk} by square roots of diagonals G_{jj} and G_{kk}.

The second method for obtaining $\mathbf{G}$ weights markers by reciprocals of their expected variance instead of summing expectations across loci and then dividing: $\mathbf{G} = \mathbf{ZDZ}'$, where $\mathbf{D}$ is diagonal with $D_{ii} = \frac{1}{m[2p_i(1-p_i)]}$. That formula was proposed for human genetic studies (Leutenegger et al., 2003; Amin et al., 2007).

The third method for obtaining $\mathbf{G}$ does not require allele frequencies and instead adjusts for mean homozygosity by regressing $\mathbf{MM}'$ on $\mathbf{A}$ to obtain $\mathbf{G}$ using the model

$$\mathbf{MM}' = g_0\mathbf{11}' + g_1\mathbf{A} + \mathbf{E},$$

where g_0 is the intercept and g_1 is the slope. Matrix $\mathbf{E}$ includes differences of true from expected fractions of DNA in common plus measurement error because the full DNA sequences were not available and a subset of markers was genotyped instead. The regression was fit with $\mathbf{MM}'$ as dependent and $\mathbf{A}$ as independent variable instead of vice versa because $\mathbf{A}$ is the expected value of $\mathbf{G}$, not vice versa. However, that variable assignment required reversing the calculations after fitting the regression using the formula

$$\mathbf{G} = \frac{\mathbf{MM}' - g_0(\mathbf{11}')}{g_1}.$$

Estimates of g_0 and g_1 may not be obvious because the dependent and independent variables are matrices rather than vectors, but equations can be written with summation notation as

$$\begin{bmatrix} n^2 & \sum_j\sum_k A_{jk} \\ \sum_j\sum_k A_{jk} & \sum_j\sum_k A_{jk}^2 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \end{bmatrix} = \begin{bmatrix} \sum_j\sum_k (MM')_{jk} \\ \sum_j\sum_k (MM')_{jk}A_{jk} \end{bmatrix}.$$

Matrix $\mathbf{G}$ is positive semidefinite with the first 2 methods but can be singular if numbers of loci are limited or if 2 individuals have identical genotypes; $\mathbf{G}$ must be singular if $m < n$. Identical twins or clones can cause singularity even in $\mathbf{A}$. An improved, nonsingular matrix $\mathbf{G}_w$ can be obtained as the weighted (w) mean, $w\mathbf{G} + (1 - w)\mathbf{A}$ if numbers of markers are limited and $\mathbf{A}$ is nonsingular. Weights are reciprocals of the error variance in measuring true fractions of alleles shared.

Elements of $\mathbf{A}$ for noninbred full-sibs are 0.5 and have error variance of $(0.05)^2$ for predicting true fractions of shared DNA. An earlier study (VanRaden, 2007) assumed greater numbers of crossovers and incorrectly reported a standard deviation of 0.035 instead of 0.05. Exact distributions of shared alleles can be calculated (Stam, 1980). Within any diploid species, the percentage of DNA identical by descent for full-sibs is approximately $50 \pm \frac{50}{\sqrt{2(L+c)}}$, where L is the length of the genome in Morgans (about 30 for cattle) and c is the number of chromosome pairs (30 for cattle).

Elements of $\mathbf{G}$ have error variance of $\frac{0.125}{m}$. If minor allele frequencies are much less than 0.5, $4\sum p_i(1-p_i)$ can be substituted for m to reflect the information from markers more precisely. From the preceding algebra,

$$w = \frac{0.05^2}{\left(0.05^2 + \frac{0.125}{m}\right)},$$

and **G** should get more weight than **A** if $m > 50$ and nearly all the weight (>0.99) if $m > 5{,}000$. Most calculations in this study used **G** instead of $\mathbf{G}_w$ because m was 50,000. To test if any weight on **A** might be beneficial, analyses including $\mathbf{G}_w$ with $w = 0.90$, 0.95, and 0.98 were compared with those with $w = 1$.

Linear Predictions

If each individual is measured once for a trait and the inheritance of all alleles is known, then data vector **y** can be modeled as

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

where **Xb** is the mean and **e** is a random error vector with variance of $\mathbf{R}\sigma_e^2$. Matrix **R** is diagonal with elements $R_{ii} = \dfrac{1}{R_{dau}} - 1$, where R_{dau} is the bull's reliability from daughters with parent information excluded. Vector **u** contains the additive genetic effects that correspond to allele substitution effects for each marker. The sum **Zu** over all marker loci is assumed to equal the vector of breeding values (**a**).

Although the genotypic coefficient matrix has been labeled **X** in some recent studies (e.g., Meuwissen, 2007; Legarra and Misztal, 2008), the original labels of Henderson (1963) were retained with **Xb** for fixed effects (those with flat priors) and **Zu** for random effects (those with informative priors). With repeated records or other more complicated models, another incidence matrix to assign individuals to records would be required instead of the identity matrix implied here. In the current study, **y** included 1 daughter deviation variable for each bull.

Three different approaches that provide equivalent predictions were used to evaluate genotyped individuals with and without phenotypes. Mixed model estimates of **u** ($\hat{\mathbf{u}}$) were solved by iteration on data (Schaeffer and Kennedy, 1986); in this case genotypic data. Scalar λ is defined as the ratio σ_e^2 / σ_u^2, which equals the sum across marker loci $2\sum p_i(1-p_i)$ times the ratio σ_e^2 / σ_a^2, where σ_a^2 is total genetic variance. When phenotypic records are processed directly, λ is the ratio of error to additive genetic variance as usual. In this study, **R** accounts for heritability and differences in daughter numbers in this study so that σ_e^2 / σ_a^2 simplifies to 1. The EBV ($\hat{\mathbf{a}}$) were obtained as $\mathbf{Z}\hat{\mathbf{u}}$, and the resulting equations were

$$\hat{\mathbf{a}} = \mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{I}\boldsymbol{\lambda})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}). \quad [1]$$

The identity matrix **I** results from an assumption that marker effects in a large, randomly mating, unselected base population are uncorrelated. That assumption is true for the simulated QTL effects, but the markers themselves have only indirect effects by tracing the inheritance of linked QTL. Marker and QTL alleles may not be in equilibrium in the base population, and marker regressions are fit across rather than within families.

Selection index equations predict $\hat{\mathbf{a}}$ directly using genomic relationship matrix **G**, which is computed as $\dfrac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$. Selection index equations are constructed as the covariance of **y** and **a** multiplied by the inverse of the variance of **y** multiplied by deviation of **y** from $\mathbf{X}\hat{\mathbf{b}}$ or

$$\hat{\mathbf{a}} = \mathbf{G}\left[\mathbf{G} + \mathbf{R}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}). \quad [2]$$

Estimates of $\hat{\mathbf{u}}$ could also be obtained if needed using the selection index equations by substituting $\mathbf{Z}'$ for the leftmost **G** in [2] and then dividing by $2\sum p_i(1-p_i)$. This shows that $\hat{\mathbf{a}}$ is the sum $\mathbf{Z}\hat{\mathbf{u}}$ over all alleles that the individual inherited. Selection index and mixed model equations provide the same estimates of $\hat{\mathbf{a}}$ if the same estimates of $\mathbf{X}\hat{\mathbf{b}}$ are used (Henderson, 1963). Thus, [1] and [2] should be identical in many genomic analyses because DYD or deregressed evaluations are the data source and the fixed effects already have been removed.

A third solution strategy presented by Garrick (2007) could be more efficient than selection index because **G** can be inverted just once and then additional traits with differing heritability or **R** processed using iteration:

$$\hat{\mathbf{a}} = \left[\mathbf{R}^{-1} + \mathbf{G}^{-1}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}). \quad [3]$$

Matrix **G** may be singular, for example, if the number of markers does not exceed the number of individuals genotyped.

A main goal is to predict merit for individuals from genotypes before their phenotypes are measured, but

different procedures were required when [1], [2], or [3] were used. When [1] was used, predictions for younger individuals were computed as $\mathbf{Z}_2\hat{\mathbf{u}}$, with $\mathbf{Z}_2$ constructed from their genotypes. When [2] was used, predictions were computed by replacing the leftmost $\mathbf{G}$ in [2] by the genomic covariance matrix $\mathbf{C}$ constructed as $\frac{\mathbf{Z}_2\mathbf{Z}'}{2\sum p_i(1-p_i)}$. When [3] was used, predictions were computed by $\mathbf{C}\mathbf{G}^{-1}\hat{\mathbf{a}}$, which is a genomic regression on EBV of individuals with phenotypes.

Expected reliabilities for $\hat{\mathbf{a}}$ were computed from the matrices in [2] and [3] but were not computed from [1] because the 50,000 × 50,000 dense matrix was too large to invert. From [2], reliabilities were obtained from diagonals of

$$\mathbf{G}\left[\mathbf{G}+\mathbf{R}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\mathbf{G}$$

for individuals with phenotypes and from

$$\mathbf{C}\left[\mathbf{G}+\mathbf{R}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\mathbf{C}'$$

for those without. From [3], reliabilities were obtained similarly from

$$\mathrm{G}-\left[\mathbf{R}^{-1}+\mathbf{G}^{-1}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\left[\frac{\sigma_e^2}{\sigma_a^2}\right]$$

and from

$$\mathbf{C}\mathbf{G}^{-1}\mathbf{C}'-\mathbf{C}\mathbf{G}^{-1}\left[\mathbf{R}^{-1}+\mathbf{G}^{-1}\left(\frac{\sigma_e^2}{\sigma_a^2}\right)\right]^{-1}\mathbf{G}^{-1}\mathbf{C}'\left[\frac{\sigma_e^2}{\sigma_a^2}\right].$$

Genetic bases often set the mean evaluations of some recent group of cows to 0. Estimation of the genetic base is accomplished by 2 equivalent formulas that depend on whether equations are solved by inversion or iteration. With inversion, inverse of the variance of $\mathbf{y}$ ($\mathbf{V}^{-1}$) is available so that the mean of $\hat{\mathbf{b}}$ can be obtained directly as $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$, assuming that those equations are full rank. With iteration, updated EBV are available in each round so that the mean of $\mathbf{y}$ can be estimated iteratively as $\left(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{R}^{-1}\left(\mathbf{y}-\hat{\mathbf{a}}\right)$. With those algorithms, genomic predictions $\hat{\mathbf{a}}$ are solved as deviations from a genetic base, and then the genetic base is added back to the predictions.

Nonlinear Predictions

For nonlinear predictions, a genetic variance component can be estimated for each marker, marker bracket, or haplotype (Meuwissen et al., 2001). Alternatively, markers with smaller effects can be regressed further toward 0 and markers with larger effects regressed less to account for the nonnormal prior distribution. Let marker deviations ($\hat{\mathbf{d}}$) be defined as

$$\hat{\mathbf{d}}=\left[\mathrm{diag}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\right)\right]^{-1}\left[\mathbf{Z}'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{u}}\right)\right]+\hat{\mathbf{u}}. \quad [4]$$

Addition of $\hat{\mathbf{u}}$ is required because subtraction of $\mathbf{Z}\hat{\mathbf{u}}$ in [4] removes not only off-diagonal terms in $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$ but also the diagonal. Linear model solutions are obtained using linear regression on $\hat{\mathbf{d}}$:

$$\hat{\mathbf{u}}=\left[\mathrm{diag}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\right)+\mathbf{I}\lambda\right]^{-1}\mathrm{diag}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\right)\hat{\mathbf{d}}.$$

The optimal regression is nonlinear if distribution of $\hat{\mathbf{d}}$ is not normal. Of the total number of markers (m), only a subset (q) may be associated with QTL. If marker-QTL associations are not known a priori, distribution of $\hat{\mathbf{d}}$ is a mixture of 2 distributions. Elements of $\hat{\mathbf{d}}$ have a $1-\frac{q}{m}$ prior probability of containing only error variance and a $\frac{q}{m}$ prior probability of containing both error and QTL variance. Two normal density functions (f_{err} and $f_{QTL+err}$) were evaluated for each marker deviation in $\hat{\mathbf{d}}$ in each round of iteration to obtain converged posterior probabilities. To simplify calculations, marker solutions were first standardized by dividing $\hat{\mathbf{u}}$ by their standard deviations $\left(\sqrt{\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{m}}\right)$. Then, f_{err} was the standard normal density with a mean of 0 and variance of 1. Function $f_{QTL+err}$ had greater variance calculated as m/q, the reciprocal of the proportion of loci with effects. For extreme marker effects, computation of normal densities resulted in underflow, which was avoided by setting error density to zero for marker deviations of >15 standard deviations.

Nonlinear and linear regressions on SNP marker deviations are shown in Figure 1. Nonlinear predic-

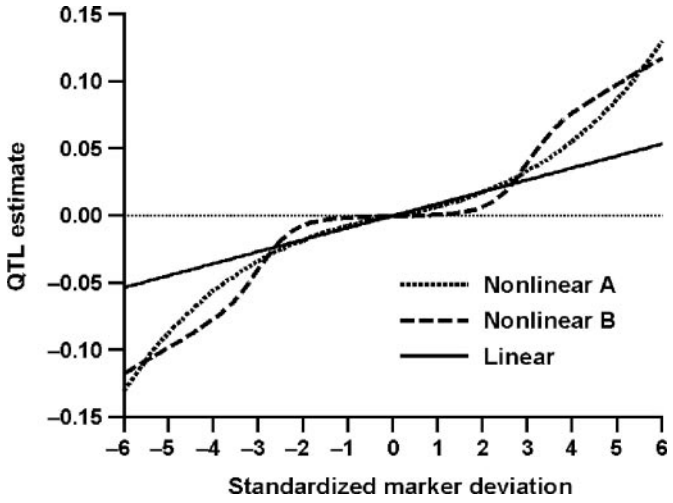

Figure 1. Nonlinear and linear regressions for marker allele effects.

tion A assumed that all markers had effects but with a simple, heavy-tailed distribution generated from a normal variable divided by $1.25^{\text{abs}(s-2)}$, where s is the number of standard deviations from the mean and 1.25 determines departure from normality. Instead of using a constant λ for all markers as in [1], nonlinear prediction A used individual λ_i for each locus, computed as $\lambda_i = \lambda/1.25^{\text{abs}(s-2)}$.

Nonlinear prediction B used:

$$\lambda_i = \lambda\left[\frac{q}{m} + \left(1 - \frac{q}{m}\right)\left(\frac{f_{err}}{f_{QTL+err}}\right)\right]. \quad [5]$$

Several markers may be needed for each actual QTL; therefore, the optimal choice for q was about 700 instead of the 100 QTL simulated. Nonlinear predictions A and B are analogous but not identical to Bayesian A and B methods of Meuwissen et al. (2001), and other prior distributions could fit actual data better. In this research, QTL effects were simulated with a normal distribution, and marker effects were estimated by iteration using an equation analogous to [1] but substituting λ_i from equation [5].

Computation

Allele frequencies in the base (founder) population were estimated with a linear model that solves for gene content of nongenotyped ancestors and descendants using pedigrees (Gengler et al., 2007). The known genotypes are treated as data, and the unknown genotypes of relatives are estimated using the inverse of the traditional relationship matrix and standard mixed model equations. Calculations are easy but not ideal because linear algebra is used instead of nonlinear probabilities. Equations without groups for unknown parents were presented by Gengler et al. (2007), and they also described and tested equations that included several groups. The current research included only 1 unknown-parent group because pedigrees were extremely complete and the goal was to estimate frequencies for the same population as the inbreeding base (1960).

Simple allele frequencies were also obtained as means of only the known genotypes. In the simple frequency estimates, summation was only over nonmissing genotypes so that missing genotypes did not contribute. Genomic predictions and inbreeding coefficients were computed using the simple estimate, the base population estimate, and also the true base frequencies to determine how sensitive the results were to different allele frequencies. Frequency estimation can bias the genomic inbreeding coefficients (Leutenegger et al., 2003).

Jacobi iteration was used in this study, but other methods may converge faster. First-order Jacobi iteration performed reasonably well with small data sets but converged too slowly with large data sets and unbalanced allele frequencies. A blend of first- and second-order Jacobi iteration performed better but required setting 2 relaxation factors.

Let the solution before relaxation from the current round be labeled $\mathbf{u}_c$, and let solutions after relaxation from the previous 2 rounds be $\mathbf{u}_1$ and $\mathbf{u}_2$. The solution after relaxation in the current round $\mathbf{u}_0$ was then computed as

$$\mathbf{u}_0 = \mathbf{u}_1 + relax_1(\mathbf{u}_c - \mathbf{u}_1) + relax_2(\mathbf{u}_1 - \mathbf{u}_2),$$

where $relax_1$ and $relax_2$ are relaxation factors. Optimal relaxation factors for this blend of first- and second-order Jacobi iteration were obtained by trial and error. The algorithms to estimate allele frequencies and allele effects both used this iterative technique, but the optimum relaxation factors differed.

Efficiency was greatly increased by summing elements of $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{u}}$ once per round of iteration and obtaining the off-diagonal sums for each row as

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{u}} - \text{diag}\left(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\right)\hat{\mathbf{u}}. \quad [6]$$

As genotypes are obtained, rows of $\mathbf{Z}$ are multiplied by $\hat{\mathbf{u}}$. Then, right-hand sides for each marker [elements of $\mathbf{Z}'\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)$] are adjusted for off-diagonal elements by subtracting the diagonal from the sum of all elements as in [6]. That procedure reduced computing

costs from quadratic with number of markers (Meuwissen et al., 2001) to linear. Similar computing strategies were discovered independently by Janss and de Jong (1999) and Legarra and Misztal (2008). As compared with the iterative methods of Legarra and Misztal (2008), the present algorithm requires much less memory by reading **Z** once per iteration instead of loading **Z** into memory, uses estimated allele frequencies instead of 0.5 to set coefficients of **Z**, accounts for missing genotypes, weights observations by including **R**, and is extended to nonlinear models.

Missing genotypes could be set to the gene content estimated from the algorithm of Gengler et al. (2007) or from a similar procedure. That approach would require recoding the data with fractional values instead of the simpler 0, 1, and 2 for known values and 5 to indicate unknown. Because ≤1% of genotypes were expected to be missing, simple genotype coding was retained for this study, but coefficients of **Z** were set to 0, which substitutes the mean allele frequency of the population for missing values. Other elements of **Z** were standardized to account for each animal's proportion of missing genotypes. That standardization was accomplished by replacing **Z** with **WZ**, where **W** is diagonal with

$$\mathbf{W}jj = \sqrt{\frac{\sum p_i(1-p_i) \text{ over all loci}}{\sum p_i(1-p_i) \text{ over only nonmissing loci}}}.$$

Elements of **WZ** for the 3 genotypes were then $-2p_i$, $1 - 2p_i$, and $2 - 2p_i$, each divided by $\mathbf{W}_{jj}$. That adjustment gave genomic relationship matrix $\frac{\mathbf{WZZ'W}}{\sum 2p_i(1-p_i)}$ an expected value of **A**, the traditional relationship matrix, when some genotypes were missing. Equations reduce to those of VanRaden (2007) when no genotypes are missing.

RESULTS AND DISCUSSION

Genomic predictions from several methods were all affordable, and timing tests on simulated data for 50,000 markers revealed that predictions for several traits of 3,000 bulls can be computed within a week. Genomic predictions offered much greater reliability for young animals than did traditional parent averages, even for traits affected by 100 QTL with unknown location.

Estimation of allele frequencies in the base population and gene content for 23,105 Holsteins in the pedigree file required about 400 iterations to converge to 5 digits of accuracy. Total time was 2 processor days, but actual clock time was reduced by processing loci on separate chromosomes in parallel. Time required for frequency estimation was proportional to the length of the pedigree file and number of markers. Optimal relaxation factors for the blended Jacobi iteration were 0.60 for the first factor and 0.88 for the second factor when used in the algorithm of Gengler et al. (2007). Simple frequency estimates required cycling through the known genotypes just once and <1 min of processing.

True allele frequencies were correlated with estimated frequencies in the base population by 0.98, but only by 0.94 with simple frequency estimates from the current population. The 2 estimates were correlated with each other by 0.97. Mean for base frequency estimates was 0.50 with a standard deviation of 0.27 compared with a mean of 0.49 and standard deviation of 0.30 for simple frequency estimates, both of which compared well with the mean of 0.50 and standard deviation of 0.29 for true allele frequencies simulated as uniformly distributed between 0 and 1. The largest difference between base frequency and simple frequency estimates was 0.41.

Genomic inbreeding coefficients from the first method for obtaining **G**, which weights summed allele effects by the sum of their expected variance, were more precise when base frequency rather than simple frequency estimates were used. Correlations between pedigree inbreeding coefficients from **A** and genomic inbreeding coefficients from **G** for the older Holstein bulls were 0.74 using true allele frequencies and 0.68 using estimates of base frequencies, but only 0.12 using simple allele frequency estimates. Corresponding correlations for the younger bulls were 0.66, 0.63, and 0.40. Genomic inbreeding coefficients were biased downward using either estimator of frequency. The mean of 7% was reasonable using true frequencies but was −4% with base frequency estimates and −2% with simple frequency estimates compared with 5% for traditional inbreeding coefficients of Holsteins from pedigree. Complete statistics including standard deviations and Jersey inbreeding coefficients are in Tables 1 and 2.

Inbreeding coefficients in **G** were less precise if **G** was estimated by weighting marker loci by reciprocals of their expected variance (second method) and more precise if **G** was obtained by adjusting for mean homozygosity by regression on **A** (third method) compared with the first method. Correlations of genomic inbreeding coefficients from the second method with pedigree inbreeding coefficients were about 0.06 lower than corresponding correlations for the first method when base allele frequencies were used. For the third method, which does not require allele frequencies, corresponding correlations were about 0.06 greater than for the first method. The first and third method had

Table 1. Correlations of pedigree with genomic inbreeding (F) coefficients computed with true or estimated allele frequencies

		Allele frequencies used in genomic F		
Breed	Birth year	True[1]	Base[2]	Simple[3]
Holstein	<2000	0.66	0.63	0.40
	2001 to 2002	0.74	0.68	0.12
Jersey	<2000	0.78	0.66	−0.26
	2001 to 2002	0.81	0.75	−0.28

[1]True frequencies in the base population.

[2]Estimated frequencies in the base population.

[3]Simple frequencies estimated by counting alleles in genotyped bulls.

nearly equal correlations when the first method used true rather than estimated base frequencies.

Subtraction of allele frequencies and weighting by variance may have theoretical appeal, but results from simulation indicated that simple counts of allele sharing and allele homozygosity can provide accurate measures of relationship and inbreeding. Comparisons using real data could give different results for the 3 methods because alleles in the base population may not be in equilibrium.

Construction of the genomic relationship matrix required 9 h for Holsteins and 1.5 h for Jerseys. Time increased by number of markers multiplied by number of bulls squared. Inversion of the matrix required about 12 min for Holsteins and 20 s for Jerseys. Time increased by number of bulls cubed. With that algorithm, traits with similar reliability such as milk, fat, and protein could be processed together with almost no additional cost.

Iteration on genotypic data took 1,000 rounds and 16 h to reach convergence for just 1 trait, but the cost for that algorithm did not increase as quickly as the cost of forming and inverting **G**. Time increased by number of iterations multiplied by number of markers multiplied by number of bulls multiplied by number of traits. In practice, clock times can be decreased by processing different traits in parallel and by using previously converged estimates of allele effects as starting values when additional genotypes are added. Rereading the genotypes increased time by 10% but decreased memory by >90% compared with storing the genotypes in memory. Optimal relaxation factors for the blended Jacobi iteration were 0.0002 for the first factor and 0.90 for the second factor. The first factor had greater optimums when numbers of markers were lower.

Predictions from iteration on data were correlated by 0.997 with those from direct inversion, which indicated agreement of [1] and [2]. With smaller data sets and more iteration, perfect correlations were achieved among predictions from [1], [2], and [3].

Reliability comparisons are in Table 3 and are means across replicates. Reliability averaged 0.66 for nonlinear predictions and 0.63 for linear predictions versus 0.32 for parent average for net merit of young Holstein bulls. Corresponding accuracies of selection obtained as square roots of those values were 0.81, 0.79, and 0.56. Thus, linear genomic predictions had reliabilities that were 0.31 greater than reliability for parent average for the younger Holstein bulls and 0.19 greater for the younger Jersey bulls. Simulated gains from genomic selection are much larger than estimates of Guillaume et al. (2008) because many more markers are now available.

Nonlinear predictions had a slight advantage over linear predictions (Table 3) of 0.04 for young Holstein bulls and 0.01 for young Jersey bulls. Reliability obtained from direct inversion of linear mixed model equations averaged 0.58, which is in reasonable agreement with the 0.63 obtained from the squared correlation of linear predictions with true breeding values. Reliability of older bulls and ancestors averaged 0.91 for predictions from both linear and nonlinear genomic models versus 0.90 from use of the traditional animal-model relationship matrix. Gains in reliability for the

Table 2. Means and SD of pedigree inbreeding (F) and genomic inbreeding coefficients computed with true or estimated allele frequencies

				Genomic F (%) using allele frequencies					
		Pedigree F (%)		True[1]		Base[2]		Simple[3]	
Breed	Birth year	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Holstein	<2000	4.6	2.4	6.7	3.9	−4.4	3.4	−2.1	5.4
	2001 to 2002	5.3	1.8	7.6	3.5	−3.6	3.0	−1.9	3.5
Jersey	<2000	5.1	3.4	7.2	4.8	0.0	4.0	−3.3	8.4
	2001 to 2002	7.1	2.7	9.2	4.3	2.1	3.7	−1.3	6.3

[1]True frequencies in the base population.

[2]Estimated frequencies in the base population.

[3]Simple frequencies estimated by counting alleles in genotyped bulls.

Table 3. Reliabilities of genomic predictions computed with true or estimated allele frequencies

		Reliability					
			Linear genomic model with frequencies				
Breed	Birth year	Animal model	True[1]	Base[2]	Simple[3]	0.5	Nonlinear model[2]
Holstein	<2000	0.896	0.915	0.913	0.912	0.905	0.913
	2001 to 2002	0.319	0.628	0.628	0.626	0.620	0.663
Jersey	<2000	0.797	0.817	0.817	0.817	0.814	0.816
	2001 to 2002	0.163	0.354	0.354	0.355	0.315	0.368

[1]True frequencies in the base population.
[2]Estimated frequencies in the base population.
[3]Simple frequencies estimated by counting alleles in genotyped bulls.

older bulls were small (about 0.01) because their reliabilities already were high and because the largest of the simulated QTL did not have very large effects. Gains from actual data could differ because of more or fewer QTL with non-normal distribution.

Reliability of predictions increased slightly for Holsteins when estimates of base allele frequency replaced simple estimates and nearly matched levels obtained when true allele frequencies were used. With iteration and coefficients of **Z** set to 0.5 and −0.5 as in Legarra and Misztal (2008), the system of equations converged more quickly, but predictions were less accurate. Realized reliabilities were about 0.007 greater for both younger and older Holstein bulls when simple estimates of allele frequency were used than when coefficients of 0.5 were used; corresponding gains for Jerseys were 0.040 and 0.003. With inversion, accuracy of predictions was similar to the use of true allele frequencies if elements of **G** were rescaled using

$$\frac{\mathbf{MM}' - g_0(\mathbf{11}')}{g_1}.$$

Thus, rescaling of **G** or of variance ratios seems useful, but predictions are not identical to those obtained by subtracting allele frequencies because elements of **G** are correlated by <1.

Matrix **G** was always nonsingular in this study because m was much larger than n, allowing use of equation [3], which requires $\mathbf{G}^{-1}$. Reliability of predictions increased very slightly when $\mathbf{G}_w$ replaced **G**, with maximum increase of only 0.0002 for $w = 0.95$. Reliability was lower for $w = 0.90$ than for $w = 1$. Thus, analyses may benefit slightly by including a small weight (5%) on traditional relationships.

Linear and nonlinear genomic predictions for younger Holstein bulls were correlated by 0.97 and had nearly equal means. The standard deviation was about 1.07 times larger for nonlinear predictions, which corresponds to the 0.04 greater reliability. Regressions of true breeding values on linear predictions were 0.99 ± 0.05 and on nonlinear predictions were 0.95 ± 0.05, only slightly less than the desired value of 1.0. Regressions of DYD on genomic predictions were similar, but with slightly larger standard error (0.06). Advantages of using base rather than simple estimates of allele frequency were small. Those extra gains in reliability were <0.01 for Holsteins and Jerseys.

Gains in reliability were converted to daughter equivalents by assuming a heritability of 20%, which is approximately the weighted mean of heritabilities of traits included in the net merit index (VanRaden and Multi-State Project S-1008, 2006). Total daughter equivalents equal daughter equivalents from parents, progeny, and own records (if available) plus additional daughter equivalents from genotyping. For both young and old bulls, information gained from genotyping was equivalent to including records from about 20 additional daughters. If similar gains occur with actual data, genomic evaluations should be released for young animals and for bulls with daughters and for cows with records instead of the traditional evaluations computed from only pedigree and phenotypic data. New methods are needed to explain genomic predictions (VanRaden and Tooker, 2007) because few scientists or breeders yet have the training or experience to understand how the new technology works.

During the review process for this study, the programs tested with simulated data were applied to actual genotypes of the Holstein bulls, and genomic predictions of merit were provided to North American breeders in April 2008. Primary advantages of simulated data were that true breeding values were known and statistical methods could be compared before actual data became available.

CONCLUSIONS

Computational methods for including genotypic data were developed and tested using simulation. Genomic

inbreeding coefficients required accurate estimates of allele frequencies in the base population; predictions of genetic merit were much less sensitive to allele frequency estimates. Genomic inbreeding and relationship coefficients that did not require allele frequencies were also obtained from simple counts of homozygous loci and alleles shared. Scaling these to match pedigree inbreeding and relationships was achieved by regression.

Three equivalent linear predictions were derived, each with different computational advantages. A nonlinear prediction produced slightly greater correlations with true breeding values and was solved using iteration on genotypic data and nonlinear regression on marker deviations. A blend of first- and second-order Jacobi iteration achieved reasonable convergence for estimating allele frequencies and effects, but other algorithms could be faster. Iterative algorithms are preferred because time increases linearly with number of genotypes. However, calculation of individual reliabilities required inverting mixed model equations that included genomic relationships. Calculation of predictions and reliabilities may require using more than one set of equations.

Tests on simulated data indicated that reliability for young animals could be >60% versus 32% from parent average. Genotypic data may add information worth about 20 daughter equivalents, and benefits should increase over time as more relatives are genotyped. However, actual genomic reliabilities may be affected by linkage disequilibrium in the base population and by subsequent selection, neither of which were simulated.

ACKNOWLEDGMENTS

The author thanks the following Animal Improvement Programs Laboratory (Beltsville, MD) staff: M. E. Tooker for performing many of the computations, G. R. Wiggans for assembling the lists of bulls to be genotyped, and S. M. Hubbard for technical editing of the manuscript and equations. Many ideas that were suggested by F. S. Schenkel (University of Guelph, Canada), I. Strandén (MTT Agrifood Research Finland, Jokioinen), Z. Liu (Vereinigte Informationssysteme Tierhaltung w.v., Verden, Germany), and 3 anonymous reviewers were helpful in improving descriptions of the procedures.

REFERENCES

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.

Calus, M. P. L., T. H. E. Meuwissen, A. P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561.

de Roos, A. P. W., C. Schrooten, E. Mullaart, M. P. L. Calus, and R. F. Veerkamp. 2007. Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. J. Dairy Sci. 90:4821–4829.

Garrick, D. J. 2007. Equivalent mixed model equations for genomic selection. J. Dairy Sci. 90(Suppl. 1):376. (Abstr.)

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

Guillaume, F., S. Fritz, D. Boichard, and T. Druet. 2008. Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet. Sel. Evol. 40:91–102.

Henderson, C. R. 1963. Selection index and expected genetic advance. NAS-NRC, Natl. Acad. Sci., Washington, DC.

Janss, L., and G. de Jong. 1999. MCMC based estimation of variance components in a very large dairy cattle data set. Interbull Bull. 20:62–67.

Lee, S. H., and J. H. J. van der Werf. 2006. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genet. Sel. Evol. 38:25–43.

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

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

Meuwissen, T. 2007. Genomic selection: Marker-assisted selection on a genome wide scale. J. Anim. Breed. Genet. 124:321–322.

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

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

Schaeffer, L. R., and B. W. Kennedy. 1986. Computing strategies for solving mixed model equations. J. Dairy Sci. 69:575–579.

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35:131–155.

van der Beek, S. 2007. Effect of genomic selection on national and international genetic evaluations. Interbull Bull. 37:111–114.

Van Tassell, C. P., L. K. Matukumalli, C. Taylor, T. P. L. Smith, T. S. Sonstegard, R. D. Schnabel, M. V. B. De Silva, G. R. Wiggans, G. Liu, S. Moore, and J. F. Taylor. 2007. Construction and application of a bovine high-density SNP assay. J. Dairy Sci. 90(Suppl. 1):421–422. (Abstr.)

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

VanRaden, P. M., and M. E. Tooker. 2007. Methods to explain genomic estimates of breeding value. J. Dairy Sci. 90(Suppl. 1):374. (Abstr.)

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

VanRaden, P. M., and Multi-State Project S-1008. 2006. Net merit as a measure of lifetime profit: 2006 revision. AIPL Res. Rep. NM$3(7–06). http://aipl.arsusda.gov/reference/nmcalc-2006.htm Accessed Nov. 27, 2007.

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

Zhang, Z., R. J. Todhunter, E. S. Buckler, and L. D. Van Vleck. 2007. Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. J. Anim. Sci. 85:881–885.