Technical note: Prediction of breeding values using marker-derived relationship matrices

B. J. Hayes*† and M. E. Goddard*†

*Department of Primary Industries Victoria, Attwood, Victoria 3049, Australia; and †Land and Food Resources, University of Melbourne, Parkville, Victoria 3010, Australia

ABSTRACT: In livestock populations, estimation of breeding values for selection requires a matrix describing the additive relationship between individuals in the population. This matrix can be derived from pedigree information. In some livestock populations, pedigree information may be unavailable, incomplete, or in error. Here we use simulated data to demonstrate that marker-derived relationship matrices can be used to predict breeding values and estimate additive variance components, provided the markers are sufficiently dense. The approach is demonstrated for an Angus data set with 9,323 SNP markers genotyped.

Key words: breeding values, marker-derived relationship matrix

©2008 American Society of Animal Science. All rights reserved. J. Anim. Sci. 2008. 86:2089–2091 doi:10.2527/jas.2007-0733

INTRODUCTION

Breeding programs for livestock species depend on accurate estimates of genetic parameters for breeding value prediction. In outbred populations, estimation of breeding values requires a matrix describing the additive relationship between individuals in the population (e.g., Henderson, 1984). If pedigree information has been collected over multiple generations, the additive relationship matrix can be constructed from this information. However, in many livestock populations, this information may be unavailable, incomplete, or contain errors. In beef or sheep breeding schemes, accurate pedigree information may be difficult to collect due to the practice of multiple sire matings (more than one sire per paddock). Error rates in pedigrees, even where they are collected, can be considerable. For example, Visscher et al. (2002) estimated the error rate in the UK dairy herd to be approximately 10%. Even minimal error rates in pedigrees can significantly reduce the accuracy of breeding values, and, therefore, the rate of genetic gain (Sanders et al., 2006).

An alternative to constructing the additive relationship matrix from pedigrees is to use marker information to infer relationships. Attempting to do this from a limited number of markers can result in bias and inaccurate estimates of genetic parameters (Wilson et al., 2003). However, if the markers are sufficiently dense, such an approach could potentially be more accurate than using pedigree information. The marker information should capture past relationships not contained in the pedigree and should not be subject to pedigree errors. In several livestock species including cattle, chicken, and pig, tens of thousands of SNP markers are now available.

The aim of this paper was to demonstrate that additive relationship matrices can be constructed from dense SNP data, and these matrices can be used to accurately predict breeding values and estimate additive variances. A data set of 9,323 SNP genotyped in an Angus cattle population, as well as simulated data, was used.

METHODS

Institutional Animal Care and Use Committee approval was not obtained for this study because the data were obtained from an existing experiment (Arthur et al., 2001).

Simulated Data Set

The simulation approach is fully described in Hayes and Goddard (2003). Briefly, a diploid population of \( n = N_e = 1,000 \) was simulated for 1,000 generations. Each individual in the population contained 29 pairs of chromosomes, and was either male or female (probability 0.5). Each chromosome was 100 cM long, and had 301 marker loci and 300 QTL loci. To create an offspring, a pair of parents of different sex was randomly chosen from the population. For each parent in a mating

1Corresponding author: ben.hayes@dpi.vic.gov.au
Received November 15, 2007.
Accepted April 3, 2008.
pair, a gamete was formed from its chromosome pairs by sampling the number of crossovers for each chromosome pair from a Poisson distribution, with mean of 1. Crossover points were randomly positioned along chromosome pairs. The haploid gametes were mutated at a rate of $1.7 \times 10^{-4}$ per locus per gamete per generation at the markers and $6 \times 10^{-6}$ at the QTL. If a locus was mutated, a new allele was added. If the locus was a QTL, the effect of the new QTL allele on the quantitative trait following mutation was sampled from a gamma distribution, scale parameter 5.4 and shape parameter 0.42, and with an equal probability of favorable or unfavorable effect, as described by Hayes and Goddard (2001). The marker mutation rate was chosen to give an average marker heterozygosity at mutation-drift balance of

$$H_{\text{markers}} = \frac{4N_e \mu}{4N_e \mu + 1} \approx \frac{4 \times 1,000(1.7 \times 10^{-4})}{4 \times 1,000(1.7 \times 10^{-4}) + 1} = 0.4.$$  

The QTL mutation rate was chosen to give approximately 200 segregating QTL across the genome, with allele frequencies following the distribution $f(p) = K/p(1-p)$, where $K$ is a constant and $p$ is the frequency of one allele, where $1/2 < p < 2N_e/1$. The observed values of $H_{\text{markers}}$, number of QTL, and distribution of QTL allele frequencies closely matched their expectations.

The genetic value of individual $i$ was

$$g_i = \sum_{j=1}^{8700} \left(p_{i,j} + q_{i,j} \right),$$

where $p_{i,j}$ is the effect of the paternal allele inherited by progeny $i$ at QTL $j$, and $q_{i,j}$ is the effect of the maternal allele inherited by progeny $i$ at QTL $j$. No dominance effect was simulated. Allele frequencies at the QTL were the result of mutation-selection-drift balance, with an average heterozygosity of

$$H_{\text{QTL}} = \frac{4N_e \mu}{4N_e \mu + 1} = \frac{4 \times 1,000 \times 6 \times 10^{-6}}{4 \times 1,000 \times 6 \times 10^{-6} + 1} = 0.023.$$  

After generation 1,000, 50 males and 200 females were selected for 5 generations on phenotype, and 1,000 offspring were produced each generation. In generation 1,005 of the simulation, phenotypes of individuals were generated by adding a random residual to their genetic value. These phenotypes were used in the prediction of breeding values and estimation of heritability. The genetic variance among individuals was determined in generation 1,000. Then a residual was sampled for individuals in generation 1,005 from

$$N(0, \sigma_e^2),$$  

where

$$\sigma_e^2 = \frac{\left(1 - 0.33\sigma_g^2 \right)}{0.33},$$

such that the (narrow sense) heritability was 0.33. The phenotypic records and average additive relationship matrix between individuals were used to estimate the additive and residual genetic variance with ASREML (Gilmour et al., 2006). Either pedigree or markers were used to build the average additive relationship matrix. The additive relationship was constructed from the markers among individuals in generation 1,005 as follows. For a given single locus, a similarity index $S_{xy}$ between 2 individuals $x$ and $y$ is calculated, where

$$S_{xy} = 1 \text{ when genotype } x = kk \text{ (i.e., both alleles at locus 1 are identical) and genotype } y = kk. S_{xy} = 0.5 \text{ when } x = kk \text{ and } y = kl, S_{xy} = 0.25 \text{ when } x = kl \text{ and } y = km, \text{ and } S_{xy} = 0 \text{ when the 2 individuals have no alleles in common at the locus (Eding and Meuwissen, 2001).}$$

The similarity index was averaged over loci. Each element of the matrix was transformed as

$$S_{xy} = (S_{xy} - \text{min}) / (1 - \text{min}),$$

where $\text{min}$ is the minimum relationship in the matrix. Ten replicate populations were simulated.

**Angus Data Set**

Three hundred seventy-nine Angus animals were selected from a research project based at Trangie Agricultural Research Centre in New South Wales, Australia. All animals were of Angus breed with sire and dam pedigree records, and animals born from 1993 to 2000 had been selected for high or low postweaning residual feed intake, a measure of feed efficiency. The original project design has been reported by Arthur et al. (2001). Approximately equal numbers of the extreme highest and lowest residual feed intake animals were selected for SNP genotyping. A set of 9,323 SNP randomly distributed across the genome was genotyped in the Angus animals. Genotyping was performed by Parallele or Affymetrix (San Diego, CA). These SNP were largely discovered as a result of the bovine genome sequencing project (http://www.ncbi.nlm.nih.gov/projects/genome/guide/cow/); other SNP were discovered as the result of assembly of expressed sequence tags (Hawken et al., 2004).

An additive model was fitted to the trait fat thickness at the P8 site (mm):

$$\text{fatp8}_j = \mu + \text{contemporary group}_j + a_j + e_{ij},$$

where $\text{fatp8}_j$ is the record of animal $j$ in the ith contemporary group [herd || sex || test group || management group (Arthur et al., 2001)], $\mu$ is the mean, $a_j$ is a polygenic breeding value for animal $j$, and $e_{ij}$ is a random residual genetic variance with ASREML (Gilmour et al., 2006). Either pedigree or markers were used to build the average additive relationship matrix. The additive relationship was constructed from the markers among individuals in generation 1,005 as follows. For a given single locus, a similarity index $S_{xy}$ between 2 individuals $x$ and $y$ is calculated, where

$$S_{xy} = 1 \text{ when genotype } x = kk \text{ (i.e., both alleles at locus 1 are identical) and genotype } y = kk. S_{xy} = 0.5 \text{ when } x = kk \text{ and } y = kl, S_{xy} = 0.25 \text{ when } x = kl \text{ and } y = km, \text{ and } S_{xy} = 0 \text{ when the 2 individuals have no alleles in common at the locus (Eding and Meuwissen, 2001).}$$

The similarity index was averaged over loci. Each element of the matrix was transformed as

$$S_{xy} = (S_{xy} - \text{min}) / (1 - \text{min}),$$

where $\text{min}$ is the minimum relationship in the matrix. Ten replicate populations were simulated.
animals from either pedigree or markers, and $\sigma_A^2$ is the additive genetic variance.

**RESULTS AND DISCUSSION**

In the simulated data sets, estimates of additive heritability were closer to the true heritabilities when the average genetic relationship was based on markers rather than on pedigree (Table 1). Accuracy of the heritability estimates from pedigree declined as fewer generations of pedigree were used. One partial explanation for the results may be that the average relationship matrix derived from pedigree cannot capture Mendelian sampling effects. That is, the relationship coefficient between 2 full sib individuals will always be 0.5 when estimated from pedigree. However, due to sampling during meiosis, 2 full sib individuals may actually share more or less than 50% of their chromosomes. The marker-derived relationship matrix will capture these Mendelian sampling effects (Visscher et al., 2006).

Reducing the number of markers used in the simulated data sets to derive the additive genetic relationship matrix resulted in decreased heritability estimates, particularly when the number of markers was below 5,000. The results indicate on the order of 9,000 markers are necessary for accurate estimates of genetic parameters using marker-derived relationship matrices in these pedigrees and populations.

The relationship coefficient estimated from markers was plotted against the relationship coefficient estimated from pedigree for the Angus data set (Figure 1). The correlation between marker and pedigree additive average relationship coefficients was high, 0.69. One way to assess the accuracy of our marker-estimated relationships is to compare the variability of relationship for the full sibs in the Angus data set with the variability expected due to Mendelian sampling. If the variability of marker-estimated relationship for full sibs is within the expected value, we can be confident the marker estimated relationships are accurate. For a genome of length 30 Morgan, the expected standard deviation of genome-wide IBD sharing between full sibs is approximately 0.04 (Visscher et al., 2006). The standard deviation of relationship of full sibs estimated by

---

**Table 1. True and estimated heritabilities from simulated data**

| Method                                      | Value (mean from 10 replicates ± SE) |
|---------------------------------------------|--------------------------------------|
| True heritability                           | 0.33                                 |
| One generation pedigree-derived relationship matrix | 0.22 ± 0.04                          |
| Five generation pedigree-derived relationship matrix | 0.26 ± 0.03                          |
| Marker-derived relationship matrix (9,000 markers) | 0.32 ± 0.03                          |
| Marker-derived relationship matrix (5,000 markers) | 0.30 ± 0.03                          |
| Marker-derived relationship matrix (1,000 markers) | 0.21 ± 0.03                          |

1Estimates were made using relationship matrices derived from pedigree or from markers.

**Figure 1.** Average additive relationship coefficients estimated from markers vs. average additive relationship coefficients estimated from pedigree for Angus animals.
the markers was 0.02, well within the range expected by Mendelian sampling, although the number of full sibs in both data sets was limited.

A higher estimate of fatp8 heritability was obtained using an average relationship matrix derived from marker data than with the pedigree-derived matrix (0.57 and 0.38, respectively). The estimated breeding values for the 379 animals in this data set were highly correlated with the breeding values from a larger data set with 1,279 records, with pedigree used to derive the average relationship matrix ($r = 0.86$).

In conclusion, there are several situations in which marker-derived relationship matrices will be valuable. When there is limited or no pedigree recorded in a population, marker genotypes may be the only source of information available to build relationship matrices. For example, in livestock, there are many traits that can only be recorded in animals that are not candidates for selection, such as meat quality. If there is no recorded pedigree linking selection candidates and commercial animals on which the trait is recorded, marker-derived relationship matrices could be used in estimation of breeding values for selection candidates. Another example is livestock populations where multiple sires are used in the same paddock of dams, such that recording of pedigree is impossible. In this situation marker-derived relationship matrices could be used in the prediction of breeding values for progeny resulting from multiple sire matings. Another valuable application of marker-derived relationships will be to avoid spurious associations in genome-wide association or QTL mapping experiments where relationships between animals are poorly identified.

**LITERATURE CITED**

Arthur, P. F., J. A. Archer, D. J. Johnston, R. M. Herd, E. C. Richardsson, and P. F. Parnell. 2001. Genetic and phenotypic variance and covariance components for feed intake, feed efficiency, and other postweaning traits in Angus cattle. J. Anim. Sci. 79:2805–2811.

Eding, H., and T. H. E. Meuwissen. 2001. Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J. Anim. Breed. Genet. 118:141–159.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2006. ASReml User Guide Release 2.0. In VSN International Ltd., Hemel Hempstead, UK.

Hawken, R. J., W. C. Barris, S. M. McWilliam, and B. P. Dalrymple. 2004. An interactive bovine in silico SNP database (IBISS). Mamm. Genome 15:819–827.

Hayes, B. J., and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33:209–229.

Hayes, B. J., and M. E. Goddard. 2003. Evaluation of marker assisted selection in pig enterprises. Livest. Prod. Sci. 81:197–211.

Henderson, C. R. 1984. Applications of linear models in animal breeding. Can. Catal. Publ. Data, Univ. Guelph, Guelph, Ontario, Canada.

Sanders, K., J. Bennewitz, and E. Kalm. 2006. Wrong and missing sire information affects genetic gain in the Angeln dairy cattle population. J. Dairy Sci. 89:315–321.

Visscher, P. M., S. E. Medland, M. A. Ferreira, K. I. Morley, G. Zhu, B. K. Corneis, 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.

Visscher, P. M., J. A. Woolliams, D. Smith, and J. L. Williams. 2002. Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. J. Dairy Sci. 85:2368–2375.

Wilson, A. J., G. McDonald, H. K. Moghadam, C. M. Herberger, and M. M. Ferguson. 2003. Marker-assisted estimation of quantitative genetic parameters in rainbow trout, Oncorhynchus mykiss. Genet. Res. 81:145–156.