A METHOD TO TEST ALGORITHMS FOR INCORPORATING GENETIC MARKER DATA IN BLUP

B.E. Clarke[1,2] and B.P. Kinghorn[1]

[1]Department of Animal Science, University of New England, Armidale, NSW 2351
[2]CSIRO Division of Animal Production, Armidale, NSW 2350

SUMMARY

A method has been developed for simulating the gametic relationship matrix at a quantitative trait locus (QTL) for a pedigreed population with full or partial marker genotype information. This has provided a powerful basis for the comparison of different methods of genetic evaluation of individuals at the QTL. The method can be extended to cover multiple linked genetic markers. This work can be used to help ensure a rational basis for marker assisted selection breeding programs.

Keywords: Genetic marker, best linear unbiased prediction, quantitative trait loci

INTRODUCTION

Genetic markers provide information about the transmission of quantitative trait loci (QTL) alleles from parents to offspring. The use of this genetic information in the prediction of breeding values for domestic livestock has meant that additional additive gametic effects for individual QTL can be added to the usual mixed model equations to obtain best linear unbiased predictions (BLUP) of additive genetic effects. These extra gametic effects are included in a gametic relationship matrix (GRM) which contains probabilities of identity by descent of QTL alleles between gametes where paternally and maternally inherited alleles are considered separately.

Fernando and Grossman (1989) described rules to build the inverse of the GRM directly taking into account recombination rates between genetic markers and QTL. This involved assigning paternal and maternal origin to marker alleles. A recursive method to build the same GRM was developed by van Arendonk *et al.* (1994). However, use of either of these methods can result in loss of information when parental origin of marker alleles is unknown. Wang *et al.* (1995) and Bink and van Arendonk (1994) described how to build the GRM and its inverse when parental origin may not be known, with the method of Wang *et al.* including an inbreeding coefficient for a QTL conditional on observed marker genotypes. These methods differ in computational ease and the assumptions required to build the GRM and its inverse, and cover both knowledge of parental origin of marker alleles and no knowledge of parental origin.

The GRM contains probabilities that genetic marker alleles are identical by descent. This is an estimation of the relationship between individuals for QTL alleles. The *true* gametic relationship matrix would contain indicators (either 0 or 1) whether QTL alleles are identical by descent. However, without knowledge of QTL alleles carried by an individual, the GRM built using marker information is the best estimate of the true gametic relationships.

A method was developed to build the *true* gametic relationship matrix using simulated genotype information which is based on known genetic marker information. This method may be used to compare different methods of building the GRM when parental origin is either known or unknown. The effect of differing GRMs on estimation of breeding values is also illustrated.

QTL SIMULATED GRM

Given pedigree and genetic marker information the *true* GRM was simulated. QTL alleles linked to known marker alleles were simulated for individual animals and a GRM was built based on the QTL information. This could be identified as the *true* GRM as it was not only built using QTL instead of marker alleles, but consisted of only 0's and 1's for any one population, these being the known incidences of QTL alleles being identical by descent. This was possible through tracing the inheritance of simulated QTL alleles throughout the whole population.

A single autosomal marker locus was assumed. The marker locus and the QTL were both biallelic (alleles 1 and 2). Marker and QTL alleles linked with a recombination rate of 0.1 were simulated for each animal. In the case that simulated marker genotypes did not match known marker genotypes, the markers and the QTL alleles were resimulated. A QTL genotype and associated phenotype were simulated for each animal within each replicate population. This information could then be used for breeding value estimation.

Genotype values were simulated as a, d and -a for genotypes A_1A_1, A_1A_2 and A_2A_2. True breeding values were calculated as $2\alpha_1$, $\alpha_1 + \alpha_2$ and $2\alpha_2$, where: α is the average effect of a gene substitution ($\alpha = \alpha_1 - \alpha_2$) and $\alpha_1 = q\,(a + d\,(q - p))$, $\alpha_2 = -p\,(a + d\,(q - p))$, where p is the frequency of A_1. Phenotypes were simulated as a function of QTL genotype and environment, where z is a normally distributed random number $N(0,1)$:

$$\text{phenotype} = \text{genotype} + \sqrt{V_e} * z .$$

Using a fixed pedigree structure and known marker alleles (Table 1) many replicate populations were simulated with the same marker genotypes but varying linked QTL alleles. These GRMs were averaged over replicates and then averaged over total number of replicates to provide a simulated GRM (sGRM). Replicate populations continued to be generated until sGRM elements all changed by less than 10^{-6} with the addition of the last replicate.

GRM COMPARISON

Using the same pedigree and marker information as in Table 1, gametic relationship matrices were built using the methods of Fernando and Grossman (1989), van Arendonk *et al.* (1994), Wang *et al.* (1995) and the above mentioned simulated QTL method, labelled Method 1,2,3 and sGRM respectively. A comparison of these different methods required a qualitative assessment of individual elements in the relationship matrix and then an evaluation of the quantitative importance of resulting differences in estimation of breeding values. Breeding values were estimated according to the mixed model equations given by Fernando and Grossman (1989).

Table 1. Set pedigree and marker information

Animal	Sire	Dam	Marker Genotype
1	-	-	11
2	-	-	12
3	1	2	12
4	1	2	11
5	3	4	21
6	3	5	21

A six animal, three generation pedigree (Table 1) was used, with QTL alleles simulated according to the following parameter values: QTL allele frequency in the base population $p = 0.5$, $a = 1$, $d = 0$, $V_a = 2pq[a + d(q - p)]^2$, $V_e = 0.5$. Polygenic effects were assumed to be equal to zero. 100,000 replicate populations were simulated according to the pedigree and marker information in Table 1, therefore, giving a unique set of QTL genotypes and associated phenotypes for each replicate population. This allowed estimated breeding values (EBVs) to be predicted using the different gametic relationship matrices at each replicate.

Individual elements of the GRMs built using the 4 different methods were compared using a sum of squared differences (Equation 1) between equivalent elements (*i*) in two GRMs (*A and B*). The EBVs resulting from solving the mixed model equations which included the inverse of the GRMs were compared according to correlation coefficients of the estimated (X) versus the true ($\overline{X}$) breeding values. The higher the correlation coefficient represented the method that gave breeding values which were closest to the truth.

$$\text{difference} = \sum_{i=1}^{n*2} (a_i - b_i)^2 \qquad (1)$$

RESULTS

The methods of Fernando and Grossman (1989) and van Arendonk *et al.* (1994) gave the same GRM (Table 2). This assumed that the first marker allele given for each animal represented the marker inherited from the sire of the individual and that the second marker allele was from the dam. This assumption was valid for all animals other than animal number 6, whose parental origin of marker alleles is uncertain. The GRM built using the method of Wang *et al.* (1995) was almost identical to the sGRM. These two only differed from the first two methods in variances and covariances relating to animal number 6. The standard errors on the elements of the simulated GRM were extremely small, giving confidence in their values.

Table 2. Sum of squares of differences between elements in the four different GRMs

	Method 2	Method 3	s GRM
Method 1	0.000000	2.590268	2.592401
Method 2		2.590268	2.592401
Method 3			0.000014

Over 100,000 replicate populations the following means and standard deviations of correlations were produced (Table 3). Due to occasional lack of variance in the true breeding values, it was not possible to calculate correlations between the EBVs and true breeding values for some replicates. From the 100,000 replicates

77,963 provided information to calculate correlations. These results (Table 3) indicate significant differences in mean correlations between all methods.

Table 3. Means and standard deviations of correlations between estimated and true breeding values from 77,963 replicates of different methods of building the gametic relationship matrix

	Mean	Standard deviation
Method 1	0.6917	0.2127
Method 2	0.6889	0.2133
Method 3	0.6998	0.2124
sGRM	0.7021	0.2100

DISCUSSION

Although the relationship matrices built by Fernando and Grossman (1989) and van Arendonk *et al.* (1994) were identical, their methods to build inverses gave different results. The inverse of the GRM built using the rules of Fernando and Grossman did *not* give the correct inverse of the GRM for this particular pedigree. This explains the differences in correlations (Table 3). Fernando and Grossman's method of building the GRM and its inverse is far less computationally demanding than that of van Arendonk *et al.* However, the latter method may be more tractable for multiple QTL.

The method of Wang *et al.* (1995) appears to be most closely related to the *correct* gametic relationship matrix. This would suggest that this method should be used in preference to the other published methods. This is especially the case with the given pedigree which contained animals whose parental origin of marker alleles was unclear. In establishing the importance of such differences between elements within the GRM to estimating breeding values the results indicate that the consequences of error in using the less accurate methods may be relatively unimportant. Although this study was carried out using only a single marker locus linked to a quantitative trait locus, it may be used to compare new methods which involve any number of genetic markers linked to the loci of important traits.

The small pedigree was engineered with an appropriate loop to distinguish between GRM building methods. The sGRM is built using a gene drop method which aborts non-conforming pedigrees. This is a disadvantage of the method that it is unable to be used on larger pedigrees. However, the aim of the method is to investigate the properties of unusual GRMs which are not only a feature of larger pedigrees.

ACKNOWLEDGMENTS

This work was funded by a joint CSIRO/UNE Postgraduate Research Scholarship.

REFERENCES

Bink, M.C.A.M. and van Arendonk, J.A.M. (1994) *Proc. 5th World Cong. Genet. Appl. Livest. Prod.* **21**:233.

Fernando, R.L. and Grossman, M. (1989) *Genet. Sel. Evol.* **21**:467.

van Arendonk, J.A.M., Tier, B. and Kinghorn, B.P. (1994) *Genetics* **137**:319.

Wang, T., Fernando, R.L., van der Beek, S., Grossman, M. and van Arendonk, J.A.M. (1995) *Genet. Sel. Evol.* **27**:251.