Use of marker information in PIGBLUP v5.20

Ron Crump and Bruce Tier

Animal Genetics and Breeding Unit, a joint venture of NSW Department of Primary Industries and The University of New England.

Introduction

Many traits of economic importance are controlled by a large number of genes (polygenes) acting in concert. Selection on estimated breeding values (EBVs) based on the infinitesimal model using Best Linear Unbiased Prediction (BLUP) has proven to be very effective for traits that are easily measured. The infinitesimal model assumes that there are an infinite number of genes, each having a small effect upon the trait.

Recently, much research effort has been applied to finding genes with a large effect on quantitative traits – so-called quantitative trait loci (QTL). QTL and/or markers linked to QTL have been discovered for most livestock species. Given the effectiveness of selection based on current methods (e.g. BLUP) there is a general consensus that QTL are only going to be of use when the traits are expensive to measure, they are expressed later in life or they are sex-limited. In these cases significant improvements in the accuracy of estimated genetic merit are to be expected from genotyping individuals for the QTL.

Tests for QTL are becoming available for many livestock species (e.g. marbling in beef cattle). However, only a relatively small number of animals, as a proportion of the population, are genotyped. This will continue until the cost of genotyping reduces. In the meantime genotyping will be limited to animals of importance, which is mainly current and/or prospective parents.

Methods for performing marker assisted selection (MAS) have been suggested by various authors (e.g. Fernando and Grossman 1989). These methods generally assume that marker data are available for all individuals. Missing marker data can be inferred during the evaluation process (Hoeschele 2001) but the statistical methods required for this are currently impractical for routine evaluation of large populations. Alternatively, methods exist for inferring genotype probabilities for ungenotyped individuals given the pedigree of the population and the known genotypes. These methods only provide certainty for progeny of homozygous parents and for heterozygous parents with progeny carrying both alleles. Furthermore, variances differ between genotyped and ungenotyped individuals when the QTL has a significant effect. The effect of moderately sized QTL acting additively has been found to be incorporated into the polygenic effects (Tier and Henshall, 2000).

Tier and Bunter (2003) investigated methods for estimating genetic merit of animals when the genotype data in the population is incomplete. The approaches looked at by Tier and Bunter were (1) to modify EBVs after BLUP evaluation which ignores genotypic information, and (2) BLUP evaluation assuming different residual variances for different classes of individuals, according to the genotype information available.

This paper outlines their work and then describes the incorporation of their post-BLUP modification approach into PIGBLUP.

Tier and Bunter, 2003

A series of populations were simulated, modelled on a sheep population structure with 200 ewes mated to 8 sires in each of 10 years. Replacement parents were randomly chosen from among the progeny, with 50% of sires and 25% of dams replaced each year. Different levels of QTL effects (5% and 10% of the total variance) and different polygenic proportions (10% and 33% of the total variance) were simulated. Data (y) were generated using a model: $y=b+a+q+e$, where b is the mean of the contemporary group, a is the polygenic effect, q is the effect of the QTL and e is a residual. Data records were generated for all non-founder individuals – founder parents were unobserved. For the lower heritability option data were also limited to female parents. The QTL acted additively, was inherited according to mendelian sampling and founder alleles had an equal chance of being A or B. There were three genotype classes AA, AB and BB with effects of ψ, 0 and $-\psi$ respectively. All parents were genotyped. Genotypes were inferred where possible, ie for progeny of homozygous parents. The genotype probabilities of progeny from other crosses are shown in Table 1. Different proportions (10, 30 or 100%) of the remaining, uninferrable, progeny were randomly chosen to be genotyped. 100 replicates for each combination of effects were simulated.

Table 1 Genotype probabilities and within-family means and variances resulting from all possible crosses in a two-allele locus.

Parental genotypes		Within-family genotype probabilities			Within-family statistics	
Parent 1	Parent 2	AA	AB	BB	Mean	Variance
AA	AA	1.0	-	-	ψ	0
AA	AB	0.5	0.5	-	$\psi/2$	$\psi^2/4$
AA	BB	-	1.0	-	0	0
AB	AB	0.25	0.5	0.25	0	$\psi^2/2$
AB	BB	-	0.5	0.5	$-\psi/2$	$\psi^2/4$
BB	BB	-	-	1.0	$-\psi$	0

These data were analysed with three different methods. The first method (labelled 'Infinitesimal' in Table 2) used a typical BLUP (infinitesimal) model to evaluate the animals. Both polygenic and QTL genetic effects were included in a single breeding value ($u=a+q$). The variance of the true breeding value was the sum of the variances due to the polygenic and QTL effects (Var(u)=Var(a)+Var(q)). The second method (labelled 'Deregressed' in Table 2) adjusted the EBVs obtained from the first method using the formula EBV*=EBV+(1-r^2)q*, where r is the accuracy of the EBV derived from the first model and q* is the effect of the QTL determined by the animal's genotype if known or its parents' genotypes (the within-family mean in Table 1) otherwise. The third method (labelled 'Heterogeneous' in Table 2) fitted the polygenic and QTL effects independently. Different mean effects and different residual variances were used depending upon the status of the marker information. Data were pre-adjusted for the QTL effect if known, otherwise according to the family means shown in Table 1. The variance of the polygenic effects was the simulated value (Var(a)). The residual

variance was augmented by the appropriate within family variance when the QTL genotype was unknown. These effects and variances within families resulting from the QTL are shown in Table 1. The value of the QTL effect was added to the polygenic EBV to give an estimate of each animal's genetic merit. Estimates of genetic merit from the three evaluation methods were compared with simulated values.

Table 2 Correlations (×100) between estimated and simulated genetic merit for different models, levels of heritability, additive QTL effects, potential phenotypes and proportions of progeny genotyped in he sample populations. Means ± empirical standard errors of 100 replicates.

Data descriptors						
Var(q)	**0.05**			**0.1**		
Var(a)	0.33	0.1	0.1	0.33	0.1	0.1
Phenotypes	All progeny	All progeny	Female parents	All progeny	All progeny	Female parents
Evaluation method:						
Infinitesimal	72 ± 3	57 ± 5	30 ± 8	74 ± 3	62 ± 4	26 ± 8
10% of ambiguous progeny genotyped						
Deregressed	73 ± 3	65 ± 4	55 ± 5	76 ± 3	71 ± 3	60 ± 4
Heterogeneous	73 ± 3	66 ± 4	58 ± 4	76 ± 2	72 ± 3	66 ± 3
30% of ambiguous progeny genotyped						
Deregressed	73 ± 3	66 ± 4	59 ± 4	76 ± 3	73 ± 3	66 ± 3
Heterogeneous	73 ± 3	67 ± 4	60 ± 4	76 ± 2	74 ± 3	69 ± 3
All progeny genotyped						
Deregressed	74 ± 3	70 ± 3	64 ± 4	77 ± 2	77 ± 2	73 ± 3
Heterogeneous	73 ± 3	71 ± 3	66 ± 3	76 ± 2	79 ± 2	76 ± 2

Table 2 contains correlations between the approximations of the total genetic merit from the three approaches used and the simulated (that is, true) genetic merit. Higher correlations indicate that the approach is a better predictor of the total genetic merit. Results in Table 2 show that increasing amounts of genotypic information increases the correlation between simulated and predicted genetic merit. This is true when deregressing the EBVs predicted in ignorance of the QTL (deregressed), and when evaluating the QTL and polygenic effects independently (heterogeneous). When the whole population is considered little is gained from genotyping more than 10% of the uninferred population although, with all parents genotyped and the gene frequency at 0.5, the genotypes of approximately 50% of all progeny can be inferred. In any case, when only the most recent cohort is considered, the benefit of genotyping more progeny (not shown) approaches significance.

At the higher level of polygenic variance there is little to choose between these two methods, both of which are only slightly better than simply using the infinitesimal model. At the lower polygenic variance, when all progeny have phenotypes, estimates of genetic merit derived from the deregressed method are slightly, but not significantly, less correlated with true merit than those derived from the model that fits the genetic effects separately. Both methods provide more accurate and less variable estimates of the animals' genetic merit than the infinitesimal model. The benefit of using either deregressed or heterogenous methods compared with the infinitesimal method is much

more pronounced when data are only available on the female parents. The benefit of using methods that use genotype data increases with the size of the QTL effect.

When data are only available on dams, the correlation between simulated and predicted genetic merit for the infinitesimal model is 30% when Var(q) is 0.05 and 26% when Var(q) is 0.1. This is the only instance when the QTL with a smaller effect induces a higher correlation between estimated and simulated merit than the QTL with the larger effect, under otherwise similar conditions. While this difference is small it is usually in the other direction. This, and the greater variation of the results, suggests that the infinitesimal model is not efficient when the QTL is generating a large proportion of the total genetic variation.

It is unlikely that a different sized population or mating structure will produce radically different results. The effect of selection is likely to lead to an increase in the proportion of homozygous parents, and a consequent increase in the quantity of progeny whose genotypes can be inferred. The effect of selection and alternative modes of gene action on the predictions of genetic merit by deregressing EBVs generated ignoring any genotypic information are yet to be tested. Similarly, alternative strategies for analysing populations for multiple traits and with different genotyping strategies – such as all or current sires only – need consideration.

Implementation in PIGBLUP

The work of Tier and Bunter showed that the post-BLUP modification of EBVs can be as useful an indicator of the total genetic merit as estimates derived from an analysis model using heterogeneous variances to accommodate differences in the available genotypic information, at least under simple additive modes of genotype action.

Post-BLUP modification of EBVs is also far simpler to implement than BLUP analysis with heterogeneous variances. Therefore post-BLUP evaluation modification of EBVs has been adopted as the method to be used in PIGBLUP.

It is not intended that this will be the ultimate solution for this type of analysis within PIGBLUP. As the number of genotype tests being marketed increases and the level of uptake grows, PIGBLUP development in this area will continue.

Since the EBVs are modified post-BLUP, the module to do this (PBMARKER) did not need to be directly included in the main PIGBLUP program. Therefore, like the PIGBLUP Selection and Mate Allocation (PBSAMA) module introduced in PIGBLUP version 5.10, PBMARKER is housed in a standalone DLL that is called from the main program's 'Post-Analysis' menu. This allows updates to this module without the main program being affected in any way. In addition, apart from the presence of the PBMARKER item on the 'Post-Analysis' menu, PIGBLUP clients that do not currently wish to utilise genotype information will not see any difference in the way PIGBLUP functions due to the PBMARKER module.

The PBMARKER module is programmed so that multiple markers may be included, each marker can have an unlimited number of alleles (therefore genotypes), and each marker may affect multiple traits. However, it must be noted that the method used has

not yet been tested beyond the simple case of one two-allele marker affecting a single trait.

1. Inputs to the PBMARKER module

Inputs to the PBMARKER module fit into three categories:

- Population parameters relating to the genotype information.
- Genotype information on animals.
- Estimated breeding values, their accuracies and pedigree from the PIGBLUP data file.

The first of these consists of the size of the effects (that is, ψ in Table 1). These parameters are only input (or changed) when setting up (or modifying) the process, not every time the module is used. It is not possible to provide defaults for these as is done for the covariance matrices used by the main PIGBLUP program since we have no control over what genetic markers are used, what they may affect or how large the effects may be. This information must be obtained from the group providing the genotype information, or the research group supporting them if this is different. Without population and marker specific analysis, it will be assumed that the QTL variance is fully included in the existing estimates of the additive genetic variance. In this case modification of PIGBLUP covariance parameters is not necessary.

The genotype information will be obtained intermittently from the genotyping company, and imported into the PBMARKER module.

Estimated breeding values and accuracies are read automatically from the appropriate PIGBLUP output files for the breed being analysed by the PBMARKER module. The PIGBLUP run must make use of appropriately modified variance components. The pedigree of the animals is read from the PIGBLUP data file.

2. Setting up the PBMARKER module

The Animal Genetics and Breeding Unit has no control over what markers are used by clients, who provides the genotyping service or how genotype information is presented. To try to accommodate this, the module only assumes that animals are identified by an identifier that can be matched directly to the identifier used in the PIGBLUP data file, and that the information on marker genotypes is laid out in columnar format with the genotype of any given marker in a single column. This is conducive with the data being received from the genotyping service in a spreadsheet, such that the animal identifier is in one column and the genotypes of different marker(s) are in one column per marker.

No restriction is placed on the codes used for genotypes, that is ‘Resistant’, ‘Intermediate’ and ‘Susceptible’ and ‘AA’, ‘Aa’ and ‘aa’ are equally valid sets of codes for a marker with two alleles. Therefore, it is necessary to know which codes correspond to which effect. To help in this, setting up of the analysis must be done as data is imported. In this way the module can detect all the codes present and allow the user simply to provide the effect associated with them.

3. Importing Genotype information

The module will read comma-separated value (commonly known as CSV) files, which are a common export format from spreadsheets and databases. In addition, tab or space delimited text files or fixed format text files can be read.

Columns from the imported data can then be declared to contain identification or genotype information, or be ignored. In this way extraneous information can be ignored and changes in the layout of the data received can be coped with.

As genotyping results are received over time, a number of data files will be accumulated. The results can be re-imported each time or the necessary information can be stored in a simple database file by the PBMARKER module and new data simply appended to this.

If genotyping results are available from multiple sources, multiple results files will be incoming for the same animals. The module supports multiple files, with varying formats across files.

If the codes for genotypes change over time, the PBMARKER module will detect that the new data does not correspond with previous data and the user will have to map the new codes to the old ones within the PBMARKER module.

4. EBV modification

This is a rapid process, it involves a single read through the EBV results. If the genotype is unknown but the parents are both homozygous, progeny genotype is inferred. For animals with actual or inferred genotypes the simple modification calculation is performed.

5. Saving of results for further analysis

The original EBV and accuracy files for the analysis are backed up, then overwritten by the modified versions created by the PBMARKER module. In this way the results can be imported into herd management systems and used in the PIGBLUP Selection and Mate Allocation module.

References

Fernando R.L., and M. Grossman (1989). *Genet. Sel.Evol.* **21**:467-477.

Hoeschele, I. (2002). In "Handbook of Statistical Genetics", p. 599. eds. Balding, D.J., Bishop, M. and Cannings, C. (2001) Wiley. New York.

Tier, B., and K. Bunter (2003). "Estimating genetic merit when genotype data are incomplete." *Animal Science* **65**: 291-298.

Tier, B., and J.M. Henshall (2001). *Genet. Sel. Evol.* **33**:587-603.