GENOME-WIDE SELECTION IN DAIRY CATTLE: USE OF GENETIC ALGORITHMS IN THE ESTIMATION OF MOLECULAR BREEDING VALUES

R.E. Crump[1,2], B. Tier[1,2], G. Moser[1], J. Sölkner[1,3], R.J. Kerr[1,4], A.F. Woolaston[1,2], K.R. Zenger[1,3], M.S. Khatkar[1,3], J.A.L. Cavanagh[1,3], and H.W. Raadsma[1,3]

[1]CRC for Innovative Dairy Products, Level 1, 84 William Street, Melbourne, VIC 3000
[2]Animal Genetics and Breeding Unit*, University of New England, NSW 2351
[3]ReproGen, Faculty of Veterinary Science, The University of Sydney, Camden, NSW
[4]PlantPlan Genetics, School of Plant Sciences, University of Tasmania, TAS

SUMMARY

A procedure has been developed for the prediction of genetic merit and the simultaneous assessment of multiple genotypes for subsequent use in gene detection. The system utilises a large volume of genotype information but ignores pedigree. With a simple additive model of inheritance, high correlations between estimates of molecular breeding value and highly reliable progeny test estimated breeding values were observed (0.70–0.77).

INTRODUCTION

Marker genotype information may be utilised both for detecting genes and predicting genetic merit. Genome-wide selection (GWS) can be performed by incorporating genotype information on relatively small numbers of selected (not random) markers (around 20) alongside pedigree information in a best linear unbiased prediction analysis. However, given sufficient markers and sufficient phenotypes for calibration it should be unnecessary to use pedigree information at all, as the summation of marker effects (molecular breeding value, MBV) will be a good predictor of genetic merit. For prediction it is desirable that the markers are well spread across the genome, but further characterisation is not required.

With on-going development of genotyping technology, in particular single nucleotide polymorphisms (SNPs), it is now possible to generate genotypes for many markers on any individual. Consequently there may often be many more genotypes than observations. To estimate the SNP effects jointly from these over-parameterised data sets, it is necessary to either explore the parameter space (as will be discussed further here), or to use dimension reduction techniques such as partial least squares (Moser 2007).

MATERIAL AND METHODS

A set of 1,546 progeny tested dairy bulls from Genetics Australia Co-operative Ltd., with records from the Australian Dairy Herd Improvement Scheme were genotyped for 15,036 SNP markers. Bulls were born between 1955 and 2001 and for 1138 bulls the sire was also in the genotyped group. The genotyping was carried out on the commercial Parallele–Affymetrix platform utilising 10,410 public domain SNPs (98% intronic) and 4,626 proprietory SNP markers (all exonic). The proprietary markers were selected to cover regions in the genome predicted to be marker-sparse, known QTL regions, or contain candidate genes. After editing, 10,715 informative SNPs remained for use in the GWS

*AGBU is a joint venture of NSW Department of Primary Industries and the University of New England

analyses, 0.8% of the genotypes were missing for these SNPs. Estimated breeding values (EBV) with high reliability were used as a proxy for true genetic merit for the index Australian Profit Ranking (APR).

Statistical model. A simple multiple regression model was used; $y_i = \mu + \sum_{j=1}^{s} \alpha_j g_{ij}$ where y_i is the EBV for bull i, μ is the intercept, g_{ij} is the genotype of bull i for SNP j (0, 1 or 2 copies of one of the alleles), and α_j is the additive effect of SNP j. Unknown genotypes were assumed to be heterozygotes.

Genetic algorithm. A genetic algorithm (GA) was used to explore the vast set of possible models and find an approximate best model. Each individual in the GA population was a representation of a multiple regression model in terms of the SNPs included, and a Bayesian Information Criterion (BIC, Schwarz 1978) was used as the fitness criterion. Assuming Normally distributed errors $\mathrm{BIC} = n \ln(\mathrm{RSS}/n) + k \ln n$, where n is the number of observations, k is the number of parameters fitted and RSS is the residual sum of squares. Taking logs effectively makes BIC contrasts scale-independent.

At the beginning of a GA run, records to be analysed were randomly allocated to one of five internal cross-validation groups. For each model, five regression analyses were performed excluding data from each of these internal cross-validation groups in turn and predicting these omitted data. The RSS relating to prediction across this 5-fold internal cross-validation was used in calculating the BIC.

The GA generates results both for the best model tried and as weighted averages across models, for SNP effects as well as MBV. The results of models with higher fitness receive a higher weighting than less fit models. The weighting applied to the results for model i was $(\mathrm{BIC}_w - \mathrm{BIC}_i) / (\mathrm{BIC}_w - \mathrm{BIC}_b)$, where BIC_b, BIC_w and BIC_i were the BIC for the best, worst and i^{th} models, respectively.

Analyses. The APR data were analysed 5 times using the GA. In addition to the internal cross-validation 200 bulls were randomly selected to form an external validation set for the run, being excluded from all regression analyses but having their MBV predicted using the results at various stages of the procedure.

Table 1. Comparison of molecular breeding values (MBV) and estimated breeding values (EBV) for the 200 animals in the external validation set for each of 5 analyses of APR

Analysis	EBV variance	MBV variance		Correlation with EBV	
		W[A]	B[B]	W	B
1	2520	1276	1837	0.748	0.730
2	2383	1468	2313	0.751	0.729
3	2768	1105	1687	0.698	0.659
4	2677	1096	1658	0.729	0.713
5	2485	1065	1414	0.773	0.735

[A]MBV from all weighted SNP effect solutions. [B]MBV from best model SNP effect solutions.

RESULTS AND DISCUSSION

Burnham and Anderson (2004) suggest the use of across-model weighted estimates of parameters in particular when prediction is the goal. This is supported by the correlation estimates in Table 1, the

weighted estimates of the MBV having higher correlations with EBV than those based on the best model SNP effect estimates. The variance of the weighted MBV estimates is lower than the variance of best model MBV estimates this may be the result of complementary action among the SNPs. Deviation of the correlations from unity may result from incomplete linkage disequilibrium (LD) between genotyped SNPs and active genes, the presence of epistatic effects and variation of EBV from the true genetic merit. Analysis of haplotypes or tag SNPs could improve the LD situation and simplify the parameter space.

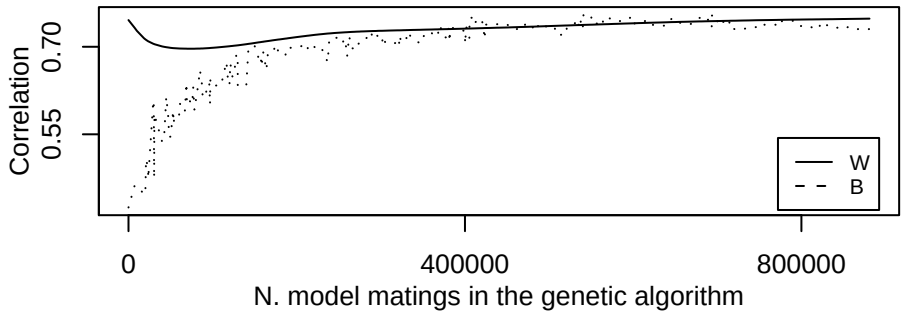

Figure 1. Correlations between molecular breeding values (MBV) and estimated breeding values for the 200 animal external validation set for analysis 1. MBV were estimated using all weighted SNP effect estimates (W) and best model SNP effect estimates (B).

Weight functions and random samples of models. In Figure 1 it can be seen that the correlation of weighted estimates of MBV from the initial random population of models was at a similar level to that at the end of the GA but less inbetween. This may imply that a suitably large random selection of models would be adequate for prediction purposes. Weighted estimates consistently outperform those from the best model across analyses. Other weighting functions may enhance the procedure and should be explored.

Table 2. Variance associated with the best model SNPs from each analysis

Analysis	N. SNPs in best model	$\sum 2p(1-p)\alpha^2$	Range of $2p(1-p)\alpha^2$
1	60	724.9	3.26–40.15
2	60	869.2	2.59–62.59
3	43	698.0	2.59–39.72
4	53	757.2	4.58–44.12
5	54	795.3	3.83–38.79

SNP selection. The process allows for the selection of subsets of SNPs that may be further utilised to genotype young bulls and cows more cheaply than a whole genome scan. In addition, SNPs that may be of interest in gene detection can also be found with much less tendency to overestimate the

size and significance of effects than is associated with single-marker approaches such as t or F-tests. The 5 analyses resulted in best models containing 60, 60, 43, 53 and 54 SNPs, respectively. These 270 SNP 'slots' were filled by 185 SNPs, only 22 of which were present in the best model from 3 or more analyses. While these analyses are not ideal for illustrating SNP selection; the omission of a random 200 observations from each analysis prevents them being true replicates of each other, the differences between the best models from each analysis (see Table 2) still illustrate the problem of inconsistent model selection associated with model selection via a GA. The problem is not restricted to our scenario (see for example, Zhu and Chipman 2006), but is likely to be exacerbated by the presence of linkage disequilibrium. Selection of SNP subsets from the GA would be more robust using a procedure like that proposed by Zhu and Chipman (2006) in which the results of multiple truncated GA runs are combined based on the frequency of regressors in the populations of models. In a random sample of models as mentioned above, SNP selection would be based upon the accumulated weight for each SNP.

Across the analyses, 16 chromosomes were represented by one or more SNPs in all of the best models in the 5 analyses. No chromosomes had no SNPs in any of the 5 best models. There are SNPs in the best models that do not have a chromosome assigned in build 3 of the bovine genome. Genetic effects on APR, and other traits not reported here, are distributed throughout the whole genome.

Over-fitting. In Figure 1 the correlation between weighted estimates of MBV and EBV appears stable at the end of analysis 1, similar patterns were observed in analyses 2 to 4. Decreasing correlations towards the end of analysis would be symptomatic of over-fitting. This was clearly observed in earlier analyses when the sums of squares of residuals across cross-validation sets was used as the model fitness criterion instead of the BIC. Across the 5 current analyses, it appears that there may be a small amount of over-fitting still occurring. This may be addressed either by alternative model fitness criteria or adjustment to the GA's stopping criterion.

CONCLUSIONS

Prediction of genetic merit from multiple SNP genotypes is a viable procedure. Genome-wide selection can be performed using these predictions without recourse to pedigree information and BLUP evaluation. The procedure has been used in the prediction of the genetic merit of young bulls prior to selection for progeny testing. Further improvements to the prediction are expected as the research progresses.

ACKNOWLEDGEMENTS

This work is part of the Cooperative Research Centre for Innovative Dairy Products project 1.4a. The Australian Dairy Herd Improvement Scheme provided the EBV data. Other colleagues in the Dairy CRC have been involved in data preparation and discussions during the development.

REFERENCES

Burnham, K.P. and Anderson, D.R. (2004) *Sociological Methods and Research* **33**:261.
Moser, G. (2007) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **17**:227.
Schwarz, G. (1978) *Annals of Statistics* **6**:461.
Zhu, M. and Chipman, H.A. (2006) *Technometrics* **48**:491.