*Breeding program design including MAS*

# COMBINING ESTIMATES OF SNP EFFECTS WHEN THEY ARE SUBPOPULATION SPECIFIC

**P. R. Amer and G. M. Payne**

AbacusBio Limited, Dunedin, New Zealand

## SUMMARY

Three methods for combining subpopulation specific coefficients linking numerous genetic markers to phenotypic trait performance are compared using simulation. Resulting combined coefficients are used to predict the genetic merit of selection candidates. Sub populations varying in size, true QTL effects, and their degree of similarity in the extent of linkage disequilibrium between markers and quantitative trait loci to a test population of selection candidates are simulated. The methods of combining estimates that were considered differ in the way that weighting is placed on coefficients from the subpopulations. At one extreme, only coefficients from the sub population from which the target population is derived are used. At the other extreme, weighting is placed jointly on the standard errors of estimates of coefficients, as well as the similarity of coefficients between the subpopulation from which the test population was derived, and other subpopulations. An example is shown whereby the weighted by correlation method outperforms the other methods. The role of population specific coefficients in motivating investment in genotyping and trait recording at the subpopulation level is discussed in the context of these results.

## INTRODUCTION

A major revolution in dairy cattle breeding programmes is currently underway as predictions of genetic merit of young bulls can now be made at moderate accuracy (e.g. Spelman *et al.* 2007), with huge scope for circumvention of progeny testing (Schaeffer 2006). There is considerable investment (at the level of industry and commercial service company level) currently under way with a view to this technology being applied in other industries, including sheep. Application in these industries may be more problematic for several reasons including: less genetically uniform populations with many breeds and subpopulation structures; less opportunity to shorten generation intervals; an absence of recording of profitable traits linked to genetic variation (e.g. maternal traits in adult breeding animals); and a lower commercial value of individual elite breeding animals (Dodds *et al.* 2007).

Because of current cost versus value trade-offs in sheep, it seems likely that a relatively modest subset of the total number of SNPs used in a discovery phase would be incorporated into a SNP key for whole genome selection (WGS). A likely application is that selection candidates that are genotyped for the SNP key can then have their genetic merit predicted directly from their genotype results, provided that robust and validated coefficients are available which translate genotype results into effects on traits. These SNP based predictions of genetic merit can then be integrated with phenotype based predictions of genetic merit (Estimated Breeding Values, EBV) generated from conventional genetic evaluation systems. Other potential applications can also be considered; for example, in theory both molecular and phenotypic based predictions of genetic merit could be computed in a simultaneous and co-ordinated way.

One key issue that must be addressed is the predictive performance of the SNP key in populations other than those in which the SNP/trait associations were discovered. Similarly, there may be considerable variability within subsets of the discovery population in the predictive performance of the SNPs. In these situations, the SNP key's ability to predict breeding values is related to the degree of linkage disequilibrium between SNPs and trait QTLs. Given the low SNP

density of the 60K SNP chip, the physical distance between a given SNP and QTL is relatively high. This translates to a high chance that phase associations between them can break down across populations or even among individuals within populations. This paper addresses the issue of how best to incorporate available SNP information from multiple subpopulations/breeds to increase accuracy of molecular breeding value prediction in target populations, where direct SNP effects are unknown.

## MATERIALS AND METHODS

**Context.** We assume that a subset of SNP markers with reasonably robust links to quantitative trait performance in at least a subset of subpopulations has been identified following a research discovery phase of technology development. This so-called SNP key then becomes validated on a range of new animals in each subpopulation. The validation process involves genotyping sires with estimated breeding values derived from significant numbers of progeny records. An output of the validation process is a set of subpopulation specific coefficients that are capable of predicting the phenotype of an animal from that subpopulation based on SNP test results. The predictive ability of the coefficients is assumed to be independent of the predictive ability that arises when SNP frequencies are confounded with population substructure.

**Simulation.** Three simulations were run, one each for 50, 100 and 200 SNPs on the SNP key. Each SNP was associated with a single QTL, with the coefficient defining linkage disequilibrium (LD) between SNP/QTL pairs (expressed as deviation of observed haplotype frequency from expected) randomly drawn from a uniform distribution with 0.4 and 0.7 as the minimum and maximum values. The minor allele frequencies were randomly drawn from a uniform distribution from 0.3 to 0.5 and 0.1 to 0.5 for SNPs and QTLs respectively. When simulated levels of LD were incompatible with QTL and SNP frequencies, the LD was adjusted down to the maximum feasible value. QTL effects at each locus were simulated as a base effect with a random additional increment drawn from a beta distribution with parameters of 1.5 and 5. QTLs were assumed to be completely independent, and SNP's were linked only to their corresponding QTL. True breeding values were simulated for sires as the sum of simulated QTL effects plus a normally distributed polygenic component that is independent of any SNPs. It was assumed that 60% of total genetic variance was explained by the simulated QTLs linked to SNPs; irrespective of how many SNPs were being simulated. Estimated breeding values for sires were also simulated by adding a normally distributed error term onto the true breeding value and rescaling the result to account for the trait phenotypic variance and assuming that the accuracies of the estimated breeding values were 0.8.

**Subpopulations.** Four reference populations were derived from the founder population. Each reference population deviated from the founder population in regards to the level of LD between SNP and QTL pairings and QTL effects. Parameters used to create differences between each reference population and the founder population are described in Table 1. New QTL effects were simulated for each reference population so that the correlations between effects in the founder population with effects in the reference population were as specified in Table 1. In populations two and three, the level of LD simulated in the founder population for each SNP – QTL pairing was scaled by multiplication by a random deviate sampled from a triangular distribution with a mode of 1 and lower and upper bounds as specified in Table 1. SNP coefficients were generated for each SNP in each reference population by regressing SNP genotypes on de-regressed, sire estimated breeding values. Associated standard errors of SNP coefficients were also computed.

A test population consisting of 1000 animals was derived with identical parameters to reference population one. SNP based predictions of merit were then created using SNP genotypes and SNP coefficients derived by three different methods and their correlation with simulated true breeding values for genetic merit computed. In method one, coefficients from reference population one were used ignoring completely the coefficients from the other populations. In method two, a weighted average of coefficients from all populations was used, whereby the weightings were taken as the reciprocal of the squared standard error of the regression coefficient. In method three, the weightings used in method two were further updated to account for the similarity of coefficient estimates among the reference populations. To do this, correlations (r) among coefficient estimates from the 4 subpopulations were computed and incorporated into the method 2 weighting factors as follows:

$w_{i,j} = \frac{se_{i,j}^{-2}}{\sum_j se_{i,j}^{-2}} \cdot \frac{r_j}{\sum_j r_j}$ giving method 3 coefficients $\tilde{b}_i = \sum_j \frac{w_{i,j} b_{i,j}}{\sum_j w_{i,j}}$ where method 3 weighting factors ($w$) and standard errors ($se$) correspond to SNP regression coefficients ($b$) for locus $i$ estimated from subpopulation $j$. Correlation estimates for coefficients between pairs of subpopulations incorporated their own weighting factors based on the inverse of squared coefficient standard errors. In order to determine the weighting applied to reference population one coefficients, relative to other populations (i.e. to estimate $r_1$), it was necessary to simulate a further validation set of animals from population one, and determine the correlation between coefficients from the two equally sized sets of animals from population one. Weightings were therefore higher for coefficients from subpopulations with a high correlation with population one's coefficients and higher for coefficients with a low standard error.

**Table 1. Parameters used to define differences between 4 reference populations and a founder population.**

| Parameter | Reference population | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Number of sires | 500 | 1000 | 500 | 500 |
| Lower bound for LD scaler | 1.0 | 0 | 0 | 1.0 |
| Upper bound for LD scaler | 1.0 | 1.2 | 1.2 | 1.0 |
| Average SNP-QTL LD | 0.5 | 0.36 | 0.36 | 0.5 |
| QTL effects correlation | 1.0 | 0.2 | 0.5 | 1.0 |

## RESULTS AND DISCUSSION

Average correlations (from 100 replicates) and their standard errors between true breeding values and SNP based estimates of true breeding values are presented in Table 2 for simulations with 50, 100 and 200 SNPs considered respectively. Because the proportion of genetic variance explained by QTLs linked to SNPs remained the same, irrespective of the number of SNPs simulated, individual QTLs had decreasing individual effects as the number of SNPs increased. With larger QTL effects, SNP coefficients can be estimated more accurately, and hence correlations are higher irrespective of methods used.

For the situation simulated, method 3 (incorporating standard errors and the general prediction of similarity between populations) performed better than the other methods. Method one performs poorly in many situations (not shown); an example is when the information from other

subpopulations is useful. Method two performs poorly when none of the sub populations have coefficients that have similar LD and similar QTL effects to the target population.

Further development of method three is required. In particular, its implementation for this simulation required an additional validation population for the reference subpopulation that matches the target population. In the real world, this would be a waste of resources, as it would be better to use the test population data to improve the coefficients for population one. A system of splitting data from subpopulation one into two sub populations of varying size and then extrapolating the correlations between the two subpopulations out to what they might be with the same number of animals as in population one is one approach that could be considered and tested. With future knowledge, it might also be possible to integrate other information about subpopulation similarity into the weighting factors for combining coefficients. Breed knowledge, and genetic distance information based on phylogeny type analysis are potential information sources that could be used. Such an approach would benefit from reformulating the problem in the context of Bayesian decision theory.

**Table 2. Accuracy of prediction of true genetic merit using 3 methods of aggregating SNP coefficients from subpopulation averaged over 100 replicates (standard errors in brackets)**

| SNP coefficient method | 50 SNPs | | 100 SNPs | | 200 SNPs | |
|---|---|---|---|---|---|---|
| 1. Population 1 only | 0.287 | *(0.004)* | 0.235 | *(0.003)* | 0.182 | *(0.003)* |
| 2. Weighted average | 0.295 | *(0.004)* | 0.256 | *(0.004)* | 0.212 | *(0.003)* |
| 3. Weighted by correlation | 0.325 | *(0.004)* | 0.276 | *(0.004)* | 0.220 | *(0.003)* |

The concept of customised coefficients might fit well in an industry context. It provides an incentive for breeders to invest in validation phenotyping and genotyping, while at the same time, creates a situation where there are modest mutual benefits from sharing data, without significant loss of intellectual property because coefficients are most relevant to the subpopulation they are estimated in.

## CONCLUSIONS

Development of population specific SNP coefficients may provide superior predictors of genetic merit for most animals and could lead to mechanisms for incentivising private investment by breeding companies and farming organizations to undertake SNP genotyping in animals recorded for novel phenotypes and for historically recorded animals. This work has identified a method which, after further refinement, could provide an opportunity for custom derivation and application of SNP prediction coefficients that use information from multiple subpopulations in a robust way.

## ACKNOWLEDGMENTS

We gratefully acknowledge funding for this work by Ovita Limited.

## REFERENCES

Dodds, K.G., Amer, P.R., Spelman, R.J., Archer, J.A. and Auvray B. (2007) *Proceedings of the New Zealand Society of Animal Production* **67**:162.

Schaeffer, L. R. (2006) *Journal of Animal Breeding and Genetics* **123**:218

Spelman, R. J., Arias, J., Keehan, M., Obolonkin, V., Winkelman, A., Johnson, D. and Harris, B. (2007) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **17**:471.