Estimating breeding values for animals with genotype only when genetic group effects are important

A.A. Swan[1,3], P.M. Gurman[1,3], V. Boerner[1], D.J. Brown[1,3], S. Clark[2], K. Gore[1,3], T. Granleese[2,3] & J.H.J. van der Werf[2,3]

[1] Animal Genetics and Breeding Unit, University of New England, Armidale 2351, Australia
andrew.swan@une.edu.au (Corresponding Author)
[2] School of Environmental and Rural Science, University of New England, Armidale 2351, Australia
[3]Cooperative Research Centre for Sheep Industry Innovation, Armidale 2351, Australia

Summary

A method was developed to estimate breeding values for animals with genotype only by back-solving post analysis from the single step genomic BLUP model. A key feature of the method is a regression on genomic relationships to approximate genetic group contributions for animals without pedigree. In a test application it was found that flock genetic means for a range of key traits in Australian Merino sheep could be estimated with high accuracy from SNP genotypes from a sample of 20 animals (correlations usually exceeding 0.8). The genetic group approximation substantially improved accuracy for several traits, and helped to reduce bias in predicting flock genetic means across all traits.

Keywords: single step genomic BLUP, genetic groups, Merino sheep

Introduction

Large scale single step genomic BLUP (SS-GBLUP; Legarra *et al.*, 2014) has been used in genetic evaluations for the Australian sheep industry since 2016 (Brown *et al.*, 2018). An important characteristic of the Merino evaluation is that it includes a large number of genetic groups, defined at the levels of flocks and time periods within flocks. During the development of genomic analyses for Merinos it was observed that genomic predictions of flock mean genetic merit had high accuracy, essentially due to the presence of large variations in trait performance between genetic groups, and the ability of SNP genotypes to detect these groups (Swan *et al.*, 2014). This finding led to the development of the "Flock Profile" commercial service to benchmark commercial flocks against estimated breeding values from the routine genetic evaluation analysis (MERINOSELECT), using SNP genotypes collected in these commercial flocks. In this paper we describe efficient methods to estimate breeding values for animals with genotype only which can then be used to benchmark commercial flock genetic merit.

Material and methods

SS-GBLUP mixed model equations

A simplified form of the SS-GBLUP model used by the Australian sheep industry is:

$$y = X\beta + ZQg + Za + e$$

With y the vector of trait records, β the vector of fixed effects, g the vector of random genetic group effects, a the vector of random within genetic group breeding values for all animals of interest, and e the vector of random residuals. X and Z are design matrixes relating records to fixed effects and breeding values respectively, and Q is a matrix describing the genetic group content for all animals, defined by pedigree relationships (Quaas, 1988). For SS-GBLUP, variances for random effects are $var(g) = I \otimes \Sigma_g$, $var(a) = H \otimes \Sigma_a$, and $var(e) = R$, with Σ_g and Σ_a the covariance matrixes for genetic group effects and breeding values, H the SS-GBLUP combined pedigree and genomic relationship matrix, and $\otimes$ the matrix direct product operator.

Within this model we classify three types of animals: (1) animals in the evaluation with pedigree and potentially phenotypes but no genotype; (2) animals in the evaluation with pedigree and genotypes and potentially phenotypes; and (3) animals with genotype only. The SS-GBLUP mixed model equations require the matrix H^{-1}, which with these three classifications of animals is as follows:

$$H^{-1} = \begin{bmatrix} A^{11} & A^{12} & 0 \\ A^{21} & A^{22} + G^{22} - A_{22}^{-1} & G^{23} \\ 0 & G^{32} & G^{33} \end{bmatrix}$$

Where the A^{ij} are sub-matrixes of the A^{-1} matrix for all animals in the pedigree, and the G^{ij} are sub-matrixes of the inverse of the genomic relationship matrix for all genotyped animals. With the three groups of animals, y, Z, and Q can be partitioned column-wise as:

$$y' = [y_1' ; y_2'], Z' = [Z_1' ; Z_2' ; 0], Q' = [Q_1' ; Q_2' ; Q_3']$$

As noted above, the definition of Q is from pedigree information, and given that genotype only animals have unknown pedigree, Q_3 is also unknown.

With these definitions, the multi-trait mixed model equations are then as follows:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}ZQ & X'R_1^{-1}Z_1 & X'R_2^{-1}Z_2 & 0 \\ Q'Z'R^{-1}X & Q'Z'R^{-1}ZQ + I \otimes \Sigma_g^{-1} & Q_1'Z_1'R_1^{-1}Z_1 & Q_2'Z_2'R_2^{-1}Z_2 & 0 \\ Z_1'R_1^{-1}X & Z_1'R_1^{-1}Z_1Q_1 & Z_1'R_1^{-1}Z_1 + A^{11} \otimes \Sigma_a^{-1} & A^{12} \otimes \Sigma_a^{-1} & 0 \\ Z_2'R_2^{-1}X & Z_2'R_2^{-1}Z_2Q_2 & A^{21} \otimes \Sigma_a^{-1} & Z_2'R_2^{-1}Z_2 + (A^{22} + G^{22} - A_{22}^{-1}) \otimes \Sigma_a^{-1} & G^{23} \otimes \Sigma_a^{-1} \\ 0 & 0 & 0 & G^{32} \otimes \Sigma^{-1} & G^{33} \otimes \Sigma_a^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ g \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Q'Z'R^{-1}y \\ Z_1'R_1^{-1}y_1 \\ Z_2'R_2^{-1}y_2 \\ 0 \end{bmatrix}$$

Total estimated breeding values (EBVs) from this model are the sum of estimates of genetic group effects and within group breeding values, $\hat{u} = Q\hat{g} + \hat{a}$. We can see that the unknown genetic group allocations for genotype only animals in the matrix Q_3 have no influence on the mixed model equations, because they are eliminated by columns of zeros from the Z matrix (and rows of zeros from Z'). This means that the within group breeding value can be estimated for genotype only animals ($\hat{a}_3$) from the main analysis without knowledge of Q_3. However, to estimate the total breeding value, knowledge of genetic groups is required (i.e. $\hat{u}_3 = Q_3\hat{g} + \hat{a}_3$).

Back-solving within group EBVs for genotype only animals

Given the structure of the mixed model equations, the genotype only animals can be removed and their EBVs estimated post-analysis. The last row of the mixed model equations is:

$$(G^{32} \otimes \Sigma_a^{-1})a_2 + (G^{33} \otimes \Sigma_a^{-1})a_3 = 0$$

Which can be removed and solved post analysis as:

$$\hat{a}_3 = -(G^{33} \otimes \Sigma_a^{-1})^{-1}(G^{32} \otimes \Sigma_a^{-1})\hat{a}_2 = -G^{33^{-1}}G^{32}\hat{a}_2 \otimes I$$

In other words, the inverse genetic covariance matrix in a multi-trait analysis can be factored out, and EBVs can be back-solved trait by trait as:

$$\hat{a}_3 = G^{33^{-1}}G^{32}\hat{a}_2 \quad (1)$$

However, this equation still requires building and inverting the genomic relationship matrix for all animals genotyped, whether they are in the analysis or not. This could easily become a substantial computational overhead if there are large numbers of genotype only animals entering the system, in addition to a continually expanding database of genotyped animals which need to be included directly in the analysis.

Within group EBVs by back-solving SNP effects

The genomic relationship matrix constructed following VanRaden, (2008) or Yang *et al.*, (2010) can be written as:

$$G = WW'/m = m^{-1}\begin{bmatrix} W_2W_2' & W_2W_3' \\ W_3W_2' & W_3W_3' \end{bmatrix}$$

Where W is the animal by SNP marker matrix of marker genotypes coded 0 for the genotype which is homozygous for the first allele, 1 for the heterozygous genotype, and 2 for the genotype homozygous for the second allele, centred on allele frequencies, and m is a scaling factor dependent on the method used.

Given the equivalence of GBLUP and SNP-BLUP models (Strandén & Garrick, 2009) breeding values for a single trait may be expressed in terms of SNP effects as:

$$a_2 = W_2 s$$

Where s is a vector of SNP effects which can be estimated from an SS-GBLUP analysis following Wang *et al.*, (2012) as:

$$\hat{s} = m^{-1}W_2'G_{22}^{-1}\hat{a}_2 \quad (2)$$

EBVs for genotype only animals can then be re-constructed as:

$$\hat{a}_3 = W_3\hat{s} \quad (3)$$

And by substituting the equation (2) into equation (3) we find that:

$$\hat{a}_3 = m^{-1} W_3 W_2' G_{22}^{-1} \hat{a}_2 = G_{32} G_{22}^{-1} \hat{a}_2 \quad (4)$$

It can also be shown that equations (111) and (4) are equivalent according to the Woodbury matrix identity, but equation (4) is easier to implement in a routine evaluation because G_{22}^{-1} is already available from the main analysis and G_{32} can be calculated by row for individual animals, or in blocks for groups of animals of interest. That is, it is not necessary to build and invert the entire genomic relationship matrix for all genotyped animals.

Approximating genetic group effects for genotype only animals

As noted above, the total EBV for genotype only animals is $\hat{u}_3 = Q_3 \hat{g} + \hat{a}_3$. We have estimates of genetic group effects ($\hat{g}$) and within group EBVs ($\hat{a}_3$) but we do not know Q_3 because the pedigree is unknown, and under the model definition Q is a matrix of genetic group contributions constructed from pedigree. For the group 1 pedigree and group 2 genotyped animals, provided group 1 contains all base population ancestors of group 2, the genetic group contributions of the latter can be written as:

$$Q_2 = A_{21} A_{11}^{-1} Q_1$$

That is, Q_2 can be expressed as the regression of pedigree relationships between the two groups on Q_1. To approximate Q_3 for our example, an analogous form based on genomic rather than pedigree relationships is:

$$Q_3 \sim G_{32} G_{22}^{-1} Q_2$$

Using this approximation, and recalling equation (4), we can then conveniently estimate the total EBV for genotype only animals as:

$$\hat{u}_3 = G_{32} G_{22}^{-1} \hat{u}_2$$

Validation analyses

Our intended Flock Profile application aims to estimate the genetic mean of commercial flocks relative to the scale of EBVs from the MERINOSELECT genetic evaluation based on data collected in ram breeding flocks. This is a genetic benchmarking service available to commercial producers, in which SNP genotypes (12K density) are collected from a randomly sampled group of the youngest cohort of breeding ewes (approximately 20 animals). Breeding values on individual animals are estimated using the methodology above, and then averaged to establish the benchmark.

The method was validated using data from a routine Merino SS-GBLUP evaluation (Brown *et al.*, 2018), which included approximately 2.3 million animals in the pedigree, 544 genetic groups, 12.4 million observations on 53 traits, with 14,761 animals genotyped. To create a validation resource, SNP genotypes were sampled from 20 yearling ewe contemporaries in each of 35 MERINOSELECT ram breeding flocks. Single trait SS-GBLUP analyses were conducted on all available data for seven key traits, including yearling measurements of clean fleece weight (YCFW), fibre diameter (YFD), staple length (YSL), staple strength (YSS), body

weight (YWT), eye muscle depth (YEMD) and fat depth (YFAT). Analyses were then repeated removing data from each of the 35 flocks sequentially to calculate the average EBV for the 20 genotyped ewes using the back-solving method above. Average EBVs from the back-solving method were compared with average EBVs from the analyses with all data.

Results

Results from validation analyses are shown in Table 1. Correlations between flock means predicted using the back-solving method and means from the full analysis were high, ranging from 0.67 to 0.93 for the total EBV. Correlations for the within group EBV were lower than correlations for the total EBV for YCFW (0.68 versus 0.90) and YSS (0.48 versus 0.67) but were very similar for other traits. Regressing the full analysis EBVs on those obtained from the back-solving method suggest that the latter under predicted the variation in within group EBV while the total EBV was less biased (average regressions of 1.4 for $\hat{a}_3$ versus 1.1 for $\hat{u}_3$).

Table 1. Correlations and regressions between flock mean EBVs from full SS-BLUP analyses on EBVs based on reduced analyses with back-solving, for the within group ($\hat{a}_3$) and total ($\hat{u}_3$) EBVs.

	Correlation		Regression	
Trait	$\hat{a}_3$	$\hat{u}_3$	$\hat{a}_3$	$\hat{u}_3$
YCFW	0.68	0.90	1.45	1.09
YFD	0.90	0.93	1.51	1.12
YSL	0.86	0.85	1.51	0.81
YSS	0.48	0.67	1.12	1.51
YWT	0.79	0.81	1.19	0.87
YEMD	0.86	0.83	1.52	1.21
YFAT	0.77	0.77	1.48	1.20

Discussion

The method presented in this paper allows efficient computation of EBVs for animals with genotypes only, and in our Flock Profile application for Merino sheep in Australia we show that the prediction of commercial flock genetic mean can be highly accurate. This is possible due to a combination of the large variation between genetic groups in Merino sheep (Swan *et al.*, 2015), and the potential of SNP genotypes to explain the genetic group structure within the population (Gurman *et al.*, 2017). The trait with the lowest correlation in Table 1, YSS, has been previously shown to have the lowest level of variation between genetic groups (Swan *et al.*, 2015).

An important feature of the method is use of a regression on genomic relationships to approximate genetic group contributions for genotyped animals without pedigree. For this to be consistent with the usual pedigree based formulation of genetic groups, the base population implied in the genomic relationship matrix (as reflected in the allele frequencies used to construct G) should correspond with the base population of the pedigree. We note that the potential for inconsistency between genomic and pedigree relationships is often a concern in application of SS-GBLUP, and that the "Metafounders" approach (Legarra *et al.*, 2015) offers considerable promise in reconciling the two.

Implicit in the use of a genomic regression to predict breeding values for genotype only animals is that we are not accounting for a residual polygenic breeding value, as described by Fernando *et al.*, (2014). This would be of interest for further research.

Finally, in order to obtain accurate predictions of EBV for genotype only animals it is very important to have well designed genomic reference populations which cover the diversity of the breed for the traits of interest.

Acknowledgements

AAS, PMG, VB, and DJB were partly funded by Meat and Livestock Australia project LGEN.1704.

List of References

Brown, D.J., Swan, A.A., Boerner, V., Li, L., Gurman, P.M., McMillan, A.J., Van der Werf, J.H.J., Chandler, H., Tier, B., & Banks, R.G. 2018. Single step genetic evaluations in the Australian sheep industry. Proc 11th World Congr Genet Appl Livest Prod.

Fernando, R.L., Dekkers, J.C., & Garrick, D.J. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol., 46(1): 50.

Gurman, P.M., Swan, A.A., & Boerner, V. 2017. Use of genomic data to determine breed composition of Australian sheep. Proc Assoc Adv Anim Breed Genet (Vol. 22).

Legarra, A., Christensen, O.F., Aguilar, I., & Misztal, I. 2014. Single Step, a general approach for genomic selection. Livest. Sci., 166: 54–65.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I., & Misztal, I. 2015. Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. Genetics, 200(2): 455–468.

Quaas, R. 1988. Additive genetic model with groups and relationships. J. Dairy Sci., 71: 91–98.

Strandén, I., & Garrick, D.J. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci., 92(6): 2971–2975.

Swan, A.A., Brown, D.J., Daetwyler, H.D., Hayes, B.J., Kelly, M.J., Moghaddar, N., & Van der Werf, J.H.J. 2014. Genomic Evaluations in the Australian Sheep Industry. Proc 10th World Congr Genet Appl Livest Prod (p. 334). Vancouver, Canada.

Swan, A.A., Brown, D.J., & van der Werf, J.H.J. 2015. Genetic variation within and between subpopulations of the Australian Merino breed. Anim. Prod. Sci. Retrieved from http://dx.doi.org/10.1071/AN14560

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci., 91(11): 4414–4423.

Wang, H., Misztal, I., Aguilar, I., Legarra, A., & Muir, W.M. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res., 94(2): 73–83.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., & Visscher, P.M. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet., 42(7): 565–569.