Proc. Assoc. Advmt. Anim. Breed. Genet. 20:340-343

PREDICTION OF GENOMIC BREEDING VALUES ACROSS GENETIC GROUPS

J.H.J. van der Werf[1,2], D.J. Brown[1,3] and A.A. Swan[1,3]

[1]Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW, 2351
[2]School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351
[3]Animal Genetics and Breeding Unit, Armidale, NSW, 2351

SUMMARY

An estimate of breeding value is generally made up of two components; a prediction of line or breed effect (genetic group) and a prediction of the deviation within genetic group. When merging conventional and genomic breeding values, bias can easily occur when these two components are not correctly identified, estimated and weighted. More work is needed to determine the best way to combine information from pedigree based groups with genomic information.

INTRODUCTION

Genomic selection has been introduced in various livestock industries. Genotypic data on a relatively small number of animals are usually merged with phenotypic and pedigree information from many animals that were not genotyped. Methods that are used to achieve this vary from the *ad-hoc* "blending" methods to the so–called "single step" (SS) method (Misztal *et al.* 2009). In most genetic evaluations the origins of the current cohort of selection candidates can be traced back to a number of different base populations. These can represent different breeds or strains within breed. We will refer to these different genetic origins as "genetic groups". This paper provides a discussion on handling genetic groups in genomic evaluation.

The variation that exists across genetic groups can be large and grouping strategies can have significant effects on the ranking of selection candidates. This is the case especially in sheep breeding programs where across breed evaluations are common, the pedigree is often not very deep, and seedstock flocks are sometimes not sufficiently linked. When genomic information is merged with phenotypic information, handling group effects appropriately can be a challenge. Genotypic data provides information about population substructures, and this could be utilized to estimate differences across groups in genetic evaluation procedures. For example, how well can we rank an animal on genetic merit, when it is genotyped but otherwise of unknown origin? This paper explores the procedures used to estimate genetic merit of animals across genetic groups, both with and without genomic selection, and proposes strategies that can be used in genetic evaluation. We will use the term EBV for an estimated breeding value based on pedigree and phenotypes, GBV for estimated breeding values based on genotypes and phenotypes and GEBV for combinations of those.

THEORY

Across-group EBVs based on pedigree and phenotypes. Best Linear Unbiased Prediction (BLUP) procedures are generally used for the prediction of breeding values. Quaas (1988) described the theory of using genetic groups. A mixed model containing genetic groups is

$$\mathbf{y} = \boldsymbol{Xb} + \boldsymbol{ZQg} + \mathbf{Za} + \boldsymbol{e} \qquad [1]$$

where the vector $\boldsymbol{y}$ contains the phenotypes, $\boldsymbol{b}$ contains fixed effects, $\boldsymbol{g}$ refers to group effects, $\boldsymbol{a}$ refers to animals' additive genetic effects within genetic groups and $\boldsymbol{e}$ are residual effects. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices relating data to fixed effects and animals, respectively. The matrix $\boldsymbol{Q}$ relates animals to groups and $\boldsymbol{ZQ}$ relates records to groups. We will consider both animal and group

effects as random. The across-group estimated breeding value $\hat{\boldsymbol{u}} = \boldsymbol{Q}\hat{\boldsymbol{g}} + \hat{\boldsymbol{a}}$. Genetic group effects can be estimated when sufficient phenotypes exist within groups, and if there is sufficient linkage between groups, i.e. animals from different groups with records in the same contemporary group. Quaas (1988) used mixed model equations based on Eqn.[1] but modified to

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{0} & \boldsymbol{X'Z} \\ \boldsymbol{0} & \alpha\boldsymbol{Q'A^{-1}Q} + \lambda\boldsymbol{I} & -\alpha\boldsymbol{Q'A^{-1}} \\ \boldsymbol{Z'X} & -\alpha\boldsymbol{A^{-1}Q} & \boldsymbol{Z'Z} + \alpha\boldsymbol{A^{-1}} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{g}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{0} \\ \boldsymbol{Z'y} \end{bmatrix} \qquad [2]$$

where $\boldsymbol{A}$ is the NRM among animals, α = var(e)/var(a) and λ= var(e)/var(g). The modified equations provide solutions for across-group EBVs ($\hat{\boldsymbol{u}}$), and the part of the equations relating the genetic groups can be seen as an augmentation of the inverse of the matrix $\boldsymbol{A}$. In fact, we can factor out the group equations by substitution. Since off-diagonals blocks with fixed effects are zero, this equates to absorbing group equations into animal equations, giving

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \alpha[\boldsymbol{A^{-1}} - \boldsymbol{A^{-1}Q}(\boldsymbol{Q'A^{-1}}\,\boldsymbol{Q} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Q'A^{-1}}\,] \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix} \qquad [3]$$

Eqn.[3] will give the same solutions as Eqn.[2] hence across-group EBVs are estimated in $\hat{\boldsymbol{u}}$ via regular mixed model equations without groups, but by using: $\boldsymbol{A^{-1}} - \boldsymbol{A^{-1}Q}(\boldsymbol{Q'A^{-1}Q} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Q'A^{-1}}$ rather than $\boldsymbol{A}^{-1}$.
Therefore, the inverse of this matrix can be seen as an 'across-group numerator relationships matrix (*NRM*)' which is:

$$\boldsymbol{G} = [\boldsymbol{A^{-1}} - \boldsymbol{A^{-1}Q}(\boldsymbol{Q'A^{-1}Q} + \lambda\boldsymbol{I})^{-1}\boldsymbol{Q'A^{-1}}]^{-1}.$$

Consider a simple example where we have phenotypes on 4 unrelated animals, two from each of two genetic groups. The matrix $\boldsymbol{G}$ will then have diagonals 1+ k, within-group off-diagonals k and across group off-diagonals 0, where k= var(g)/var(a). When var(g) is large in comparison to var(a), i.e. for large k, the group differences will mainly determine ranking on across-group EBVs whereas with small k, $\boldsymbol{G}$ is close to $\boldsymbol{A}$ and groups can be practically ignored. Hence, Eqn.[3] shows that across-group EBVs can be calculated when using the appropriate across-group NRM. The latter involves knowledge of the group structure (as defined in the matrix $\boldsymbol{Q}$), and knowledge of variance ratio k.

Across-group GBVs based on genotypes and phenotypes. Genomic relationship matrices (***GRM***) can be constructed, e.g. using VanRaden (2008). When a ***GRM*** is formed for multi-breed populations the diagonal elements will be larger and there will be larger off-diagonal elements within breed compared with a breed-specific *GRM*. Also, off-diagonals within breeds will be larger than across breeds (or groups). This is similar to $\boldsymbol{G}$ described in the previous paragraph, and the ***GRM*** can be considered as an 'across-group' relationships matrix if across breed allele frequencies are used. However, there are two differences. Firstly, $\boldsymbol{G}$ is trait specific, because k varies between traits. The ***GRM*** could be trait specific if genomic regions were differentially weighted according to their significance in explaining genetic variance, but not otherwise. The second difference between $\boldsymbol{G}$ and ***GRM*** is that the grouping structure in $\boldsymbol{G}$, as defined in $\boldsymbol{Q}$ and based on pedigree, is not necessarily the same as the grouping structure implied by the ***GRM***. Principal component analysis has been proposed to reveal population structure from genomic information (Price *et al*. 2006). The variance among "groups" in a genomic evaluation can be estimated by replacing $\boldsymbol{Q}$ with the eigenvectors relating to the most significant eigenvalues of the ***GRM*** and fitting these to the data. Brown *et al.* (2013) observed strong relationships between the principal components of the GRM and average flock EBVs in Merino sheep. On the other hand, Daetwyler *et al*. (2010) observed that large single sire families could explain more variation in the ***GRM*** than lowly represented breeds. Moreover, genetic distances between individuals do not

always reflect phenotypic differences. Nonetheless, partitioning of variance based on the ***GRM*** structure is likely to be an improvement.

Combining EBV and GBV. In a method where information from phenotypes and pedigree is combined with genomic information, we have:

$$GEBV = w_1.EBV + w_2.GBV,$$

and the weights (w_i) are derived from their respective *within group* accuracies. If both EBV and GBV were across-group estimated breeding values we can rewrite this blending formula as:

$$GEBV_{across} = w_1.[PGroup + EBV_{within}] + w_2.[GGroup + GBV_{within}],$$

where PGroup is the group solution for pedigree defined groups and GGroup is the solution for groups derived from the ***GRM***. The GGroup term can represent breed differences for the animals that were genotyped, or genetic groups within breeds. Often, breed differences are fitted in genomic analysis, but it may be harder to fit a more subtle group structure within breeds. For example, in the Australian genetic evaluation, Merino rams are grouped by flock of origin, but within the cohort of genotyped animals there may be limited information per flock to estimate these differences reliably. Hence, flock-groups are not fitted in the genomic analysis and the GBV_{within} term will contain genetic differences between flocks and within breed because the ***GRM*** is derived across those flock groups. This is illustrated in Fig. 1 where we plot flock averages of GBV and EBV for 1610 young Merino rams in 34 Australian flocks that were genotyped using the Illumina 50K ovine SNP chip. The EBVs were calculated based on the full *MERINOSELECT* genetic evaluation based on pedigree and phenotypes, including those of animals that were genotyped. In this analysis, genetic groups are allocated to groups on a flock basis. The GBVs were estimated based on an analysis of genotyped animals only, including these young rams and ~10,000 animals in the CRC multi-breed reference population (Moghaddar et al., 2013). In the latter model, no genetic groups were defined at the flock level but Fig. 1 demonstrates that there is a good concordance between average flock solutions for EBV and GBV, hence flock differences are estimated implicitly using an across flock GRM. This is not a surprise, given that the reference population is an important link between animals from the different groups in both types of analysis, with most flocks having a significant genomic relationship with the reference population. This also confirms the observation in this paper, that genomic analysis can accommodate a between group component, even if it is not explicitly fitted.

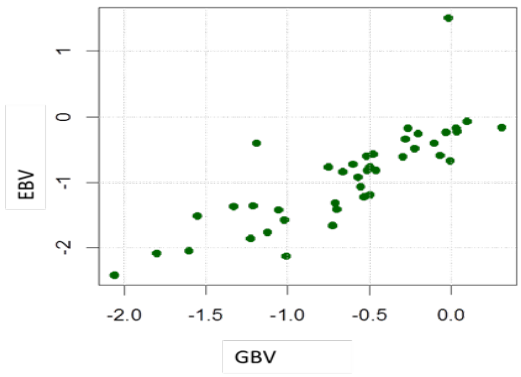

Figure 1: Average GBV for fibre diameter versus average EBV by flock (genetic group) based on 1610 Merino rams in 34 flocks in *MERINOSELECT*.

Although the example in Fig.1 suggests a good agreement between across group ranking of EBV and GBV, there are several pitfalls when blending these sources of information. If EBV and GBV contain between group differences, it would be appropriate to derive weights based on across group accuracy, or, better, to use different weights for 'between' and 'within'-group components of breeding value, because the reliability of estimating within-group differences might be very different from the reliability of estimating groups effects. With relatively few animals genotyped, the accuracy of estimating GGROUP effects is likely small. A second problem could occur if

genetic evaluation was across breed and breed differences maybe included in EBV but not in GBV. Breed effects or principal components are often fitted in genomic analysis, but not added back into GBV, hence only estimating $GBV_{within.}$ This would give:

$$GEBV_{across} = w_1.\,[PGroup + EBV_{within}] + w_2.GBV_{within},$$

which is biased as the between group differences are given insufficient weight ($w_1 < 1$). A solution could be to only blend the within group components and use the more reliable estimate of PGROUP to compare across groups:

$$GEBV_{across} = PGroup + w_1.EBV_{within} + w_2.GBV_{within}.$$

In this approach it is required that implicit group differences in GBV_{within} are not also included in the variation in PGROUP, but this may be difficult to avoid as illustrated in Figure 1. Finally, an obvious problem would occur in the blending procedure if there is a significant overlap in the data used to estimate EBV and GBV, respectively. A more coherent way to combine information is proposed in the SS method where information on genotype, pedigree and phenotypes are combined into one model (Misztal *et al*. 2009, Swan *et al*. 2011). Misztal *et al.* (2013) discussed how this procedure could account for genetic groups, and he proposes to use Eqn.[2] but with $\boldsymbol{A}^{-1}$ replaced by $\boldsymbol{H}^{-1}$, where $\boldsymbol{H}^{-1}$ is based on pedigree as well as genomic relationships. This procedure avoids double counting of information and the weighting of the various 'between' and 'within'-group components will be more likely correct. The main challenge is the genetic groups defined based on pedigree are not necessarily interpretable the same way as the genetic groups derived from the genomic data. More work is needed on how to best handle this problem in SS methods.

CONCLUSION

Both pedigree and genomic approaches can be used to retrieve information about an animal's breeding value from information across genetic groups. However, group definitions may differ and blending the information via ad-hoc methods can easily lead to bias. A SS approach should be able to handle this correctly, but methods to combine genomic and pedigree genetic groupings need more investigation.

REFERENCES

Brown D.J., Swan A.A. Gill J.S. and Banks R.B. (2013) *Proc. AAABG* **20**: (this publ.)
Daetwyler H.D., Kemper K., van der Werf J.H.J and Hayes B.J. (2012) *J. Anim. Sci*. **90**: 3375.
Misztal I., Legarra A., Aguilar I. (2009) J. Dairy Sci. **92**: 4648.
Misztal I., Vitezica Z.G., Legarra A., Aguilar I., and Swan A.A. (2013) *J. Anim. Breed. Genet.* doi: 10.1111/jbg.12025.
Moghaddar N., Swan A.A., and van der Werf J.H.J. (2013) *Anim. Prod. Sci.* (In Press)
Price A.L., Patterson N.J., Plenge R.M. *et al.* (2006) *Nature Genetics* **38**: 904.
Quaas R.L. (1988) *J. Dairy Sci.* **71**: 1338.
Swan A.A., Brown D.J., Tier B., and van der Werf J.H.J. (2011) *Proc. AAABG* **19**: 331.
VanRaden P.M. (2008) *J. Dairy Sci.* **91**: 4414.