Proc. Assoc. Advmt. Anim. Breed. Genet. 20:360-363

A GENOMIC PREDICTION CROSS-VALIDATION APPROACH COMBINING EWE REPEATED PHENOTYPES AND RAM DAUGHTER TRAIT DEVIATIONS

H.D. Daetwyler[1,2], S. Bolormaa[1,2], D. J. Brown[2,3], J.H.J. van der Werf[2,4] and B. J. Hayes[1,2,5]

[1]Biosciences Research Division, Department of Primary Industries, Bundoora 3083, Victoria, Australia, [2]Cooperative Research Centre for Sheep Industry Innovation, Armidale 2351, NSW, Australia, [3]Animal Genetics and Breeding Unit (AGBU), Armidale 2351, NSW, Australia
[4]School of Environmental and Rural Science, University of New England, Armidale 2351, NSW, Australia,[5]La Trobe University, Bundoora 3086, Victoria, Australia

SUMMARY

Reproduction traits generally are expressed later in life and have high economic value, which makes them good candidates for improvement through genomic selection. We derive genomic prediction equations for number of lambs weaned (NLW) from genotyped rams' daughter records and repeated ewe records for genotyped ewes to make the best use of all available information. In order to assess the accuracy of the genomic estimated breeding values (GEBV) for this trait, an across sire family cross-validation design is proposed, and was compared with fully random cross-validation. The accuracy of genomic BLUP (GBLUP) using both ram and ewe records is compared with GBLUP using only ewe records, as well as BLUP with no genomic information using both ram and ewe records. The combined approach resulted in higher accuracies than both GBLUP only using ewe records, and BLUP. The approach provides a way to make use of all data available to maximise accuracy of GEBVs for traits such as NLW.

INTRODUCTION

Although reproduction traits, such as number of lambs weaned (NLW) have a high economic value, they are difficult to genetically improve, due to low heritability (h^2) and because they are expressed later in life. Genomic prediction may increase genetic gain in these traits because it can predict a ram or ewe's performance early in life with, potentially, a higher accuracy than Best Linear Unbiased Prediction (BLUP).

Genomic prediction makes use of a reference population in which animals are both genotyped and phenotyped to predict genomic estimated breeding values (GEBV) of selection candidates (e.g. young rams) based only on genotype. The individuals available with both genotypes and phenotypes are primarily ewes with repeated NLW records in large research projects. In addition, genotyped industry rams have daughters with NLW records. Combining these two sources of information should make optimal use of the data available. However, each source needs to be properly weighted to account for the differences in phenotype accuracy.

The aim of this paper was to compare strategies for combining daughter information on genotyped rams with individual repeated observations on genotyped ewes in a reference population for NLW genomic predictions. Accuracy of the resulting GEBV was assessed with a number of different cross-validation strategies.

METHODS

Phenotypes, genotypes and estimates of heritability and repeatability. Proper weighting of ewe and ram records in this combined genomic prediction approach required the estimation of h^2 and repeatability (t). Data on NLW from years 1992 to 2012 were retrieved from the Australian Sheep Genetics database, giving 290,636 records on 244,672 ewes, where 111,572 ewes had a known sire (8,036 sires). The number of progeny per sire ranged from 1 to 339. The phenotype data included 53 breeds and each animal's breed proportions were calculated from the pedigree.

The main breeds were Merino (MER), Border Leicester (BL), Coopworth, Polled Dorset and White Suffolk. Values for h^2 and t were estimated using $\mathbf{y} = \mathbf{Xb} + \mathbf{Z_1Qq} + \mathbf{Z_1a} + \mathbf{Z_2pe} + \mathbf{e}$, where **y** is a vector of phenotypes, **X**, $\mathbf{Z_1}$, and $\mathbf{Z_2}$ are design matrices, **b** is vector of fixed effects, **q** is a vector breed effects**, a** is a vector of animal effects, **pe** is a vector of permanent environmental effects, and **e** is the vector of random errors. The following distributions were assumed: **a** ~ N (0, σ_a^2 **A**), **q** ~ N (0, σ_q^2 **I**), and **e** ~ N (0, σ_e^2 **I**), where **A** is the numerator relationship matrix, σ_a^2 is the genetic variance, σ_q^2 is the variance of breed effects, and σ_e^2 is the residual variance. Fixed effects included the mean, conception site, lambing site, year of lambing, age at lambing, conception method, and an indicator of whether a ewe lambed as a yearling.

A subset of the animals were genotyped using the Ovine50K SNP chip comprising a total of 54,977 single nucleotide polymorphism (SNP). The genotype quality control and imputation of sporadic missing genotypes is described in Daetwyler *et al.* (2012) and reduced the number of SNP to 48,599. The final multi-breed reference population included 4114 animals (of which 317 were rams), including 2,205 MER, 788 BL, 486 from terminal breeds, 185 other maternal breeds, 185 other MER type breeds, and 283 of unknown breed. The genotyped ewes originated from the Sheep CRC information nucleus and the SheepGENOMICS project (van der Werf et al. 2010; White et al. 2012). Results are reported for the MER and BL breeds and their crosses, as they made up the majority of genotyped animals.

Trait and daughter trait deviations. The phenotypes for NLW used in GBLUP analyses were trait deviations (TD) for ewes and Daughter Trait Deviations (DTD) for rams. The phenotype was corrected for fixed effects using the same model applied to calculate heritability and repeatability but excluding the animal effect. The residuals (corrected phenotype) were used to calculate TD and DTD. Trait deviations are calculated as TD = Σ(residual)/N, where N is the number of records. Rams DTD are calculated only from their ungenotyped daughters' TD: DTD =Σ(TD)/p, where p is the number of progeny per ram. Genotyped ewes are included in reference population and not used in their ram's DTD to avoid double counting. Rams with less than 3 progeny were removed. The number of records contributing towards genotyped ewe TD and ram DTD was 6066 and 9213, respectively, demonstrating that adding the sire DTD more than doubled the reference population. DTD contain only the genetic merit of the sire, thus DTD was doubled for analysis. The TD and DTD were weighted in the model to account for the differential accuracy of phenotypes using $\left(1-h^2\right)\left[ch^2+\left(\left(1+(n_i-1)t\left[n_i\right]^{-1}\right)-h^2\right)\right]^{-1}$ and $\left(1-h^2\right)\left[ch^2+(4-h^2)\left[p_i\right]^{-1}\right]^{-1}$, respectively, where n_i is the number of records for animal *i*, *c* is the proportion of the genetic variance *not captured* by the markers, and *p* is the number of progeny for ram *i* (Garrick et al. 2009). Five different values of the c (0.25, 0.35, 0.50, 0.60 and 0.75) were used for calculating weights of TD and DTD.

Genomic prediction analysis. Genomic BLUP (GBLUP) was used to predict GEBVs and BLUP was used for predict estimated breeding values (EBV). GEBVs were calculated based on the following model: $\mathbf{y^*} = \mathbf{1\mu} + \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$, where $\mathbf{y}^*$ is a vector of TD and DTD, **1** is a vector of ones, **μ** is the mean, **b** is vector of fixed effects and included sex, **X** and **Z** are design matrices, **e** is the vector of random errors, and **g** is a vector of either GEBVs or EBVs. In GBLUP, **g** it is distributed as N (0, σ_g^2 **G**), where σ_g^2 is the genetic variance explained by the markers and **G** is the genomic relationship matrix (Yang et al. 2010). All models were run in ASReml.

Measuring accuracy with cross-validation. Cross-validation, where the data is divided into a number of subsets and each subset is predicted once from the other subsets, was used to estimate accuracy of GEBVs. The six subsets for cross-validations were chosen either completely random or by random whole sire family. In random sire family cross-validation, sires were randomly

allocated to subsets. All progeny of a sire (both genotyped and ungenotyped) was then allocated to the same subset to ensure prediction was across sire families resulting in a conservative estimate of prediction accuracy. The division of subsets were the same for GBLUP and BLUP. The values presented are the mean of five replicated cross-validations, where new random subsets were chosen in each replicate.

Accuracies were calculated as the correlation of GEBV with **y*** for MER and BL breeds, where animals were assigned to breed groups according to sire breed. The accuracy of the true breeding value was approximated by dividing this correlation by the mean accuracy of the EBVs of sires and ewes in the reference population. The sires' EBV accuracy was calculated using only their non-genotyped daughters and ewe accuracies were from the BLUP model used to calculate h^2 and calculated as $\sqrt{\left(1-SE^2\right)/\sigma_a^2}$, where SE^2 is the standard error of prediction for the EBV. Potential bias of GEBV or EBV was investigated by regression of **y*** on GEBV or EBV.

RESULTS AND DISCUSSION

The h^2 and t estimated with full BLUP model for NLW was 0.06 and 0.08, respectively. The estimate for h^2 was consistent and the estimate for t was slightly lower than literature estimates (Safari et al. 2005). In cross validations, the accuracy of GBLUP was always higher than BLUP for both breeds (Table 1). BLUP accuracy was higher when cross-validation was random, due to prediction within families, but deteriorated significantly when cross-validation was across sire families, as expected. Adding the ram DTD into the analysis increased the accuracy in all scenarios showing a clear benefit of making use of all available data. The increase in GBLUP accuracy of combined data was greater when cross-validation was across sire families, indicating that combining data had a greater effect on animals less related to the reference. The average size of validation sets was 278 for MER and 99 for BL.

Table 1. Accuracy and slope of regression(y*, GEBV) of BLUP and GBLUP, when using only trait deviations of ewes (TD) or TD plus daughter trait deviations of rams (DTD)

		Accuracy			Slope		
Breed	Cross Validation	BLUP TD+DTD	GBLUP TD	GBLUP TD+DTD	BLUP TD+DTD	GBLUP TD	GBLUP TD+DTD
MER	Random	0.19	0.21	0.24	1.12	1.37	1.34
	Sire Fam	-0.05	0.06	0.12	-1.65	1.35	1.55
BL	Random	0.15	0.19	0.22	1.10	2.47	1.61
	Sire Fam	0.10	0.20	0.27	2.23	3.88	2.83

The slopes of the regression of **y*** on GEBV were variable. A slope > 1 (i.e. the GEBV under-estimated the phenotype) was observed in most analyses. Sire family cross-validation resulted in more bias than random subsets. BL slopes exhibited more upward bias than MER. It is of note that the slopes in BL based on TD alone showed a stronger upward bias than the combined data. One reason for this could be that the genotyped research ewes were actually BL/MER crosses exhibiting heterosis, whereas the daughters of the industry sires are expected to be more purebred BL. The bias in BL could also be due to the scaling of the multi-breed **G**, which can affect variance components. Further work will investigate fitting heterosis and scaling **G** for breed specific inbreeding and allele frequencies (Erbe et al. 2012). Using the current approach, the GEBVs would have to be blended with BLUP breeding values to be distributed to sheep breeders. A one-step approach (e.g. Aguilar et al. 2010) would also make use of all data and would reduce

issues related to blending. However, the combination of **A** and **G** matrices also needs to account for the proportion of the genetic variance not captured by the markers, just like *c* in our approach.

The weights on TD and DTD require an assumption on *c*, defined as the proportion of the genetic variance *not captured* by the markers (Garrick et al. 2009). The true value of *c* depends on the interplay between marker density and the effective population size (N_e) of the breed. A breed with a higher N_e (e.g. MER N_e 800, (Kijas et al. 2012)) would need a higher marker density than a breed with a lower N_e (e.g. BL N_e 150) to achieve the same *c*. The range of GEBV accuracy from different *c* values was small for both MER and BL (maximum 3%). This lack of a clear signal could be due to the limited size and multi-breed nature of the reference population and the low h^2 of NLW.

The accuracies achieved using GBLUP are encouraging and more genetic gain would be achieved for NLW through genomic selection than BLUP. Furthermore, making use of all data on both ewes and rams substantially increased the accuracy of prediction.

ACKNOWLEDGMENTS

The authors acknowledge funding from the CRC for Sheep Industry Innovation, MLA, and AWI Ltd. We thank K. Gore and K. Geenty for managing the CRC information nucleus database, A. Swan for comments and the many staff involved at the CRC sites across Australia.

REFERENCES

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. and Lawlor T.J. (2010) *J. Dairy Sci.* **93**:743.

Daetwyler H.D., Swan A.A, van der Werf J.H.J., and Hayes B.J. (2012) *Gen. Sel. Evol.* **44**:33.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., et al. (2012) *J. Dairy Sci.* **95**: 4114.

Garrick D., Taylor J. and Fernando R. (2009) *Gen. Sel. Evol.* **41**:55.

Kijas J.W., Lenstra J.A., Hayes B., Boitard S., Porto Neto L.R., San Cristobal M., et al. (2012) *PLoS Biol.* **10**:e1001258.

Safari E., Fogarty N.M. and Gilmour A.R. (2005) *Livest. Prod. Sci.* **92**: 271.

van der Werf J.H.J., Kinghorn B.P. and Banks R.G. (2010) *Anim. Prod. Sci.* **50**: 998.

White J.D., Allingham P.G., Gorman C.M., Emery D.L., Hynd P., Owens J., et al. (2012) *Anim. Prod. Sci.* **52**: 157.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., et al. (2010) *Nat. Genet.* **42**: 565.