Proc. Assoc. Advmt. Anim. Breed. Genet. 20:298-302

ACCURACY OF GENOMIC PREDICTION FOR RESIDUAL FEED INTAKE IN A MULTI-BREED CATTLE POPULATION

M. Khansefid[1,2,3], J.E. Pryce[1,2], S.P. Miller[4] and M.E. Goddard[1,2,3]

[1]Department of Environment and Primary Industries, AgriBio, 5 Ring Road, Bundoora, VIC, 3083, Australia
[2]Dairy Futures Cooperative Research Centre (CRC), AgriBio, 5 Ring Road, Bundoora, VIC, 3083, Australia
[3]Melbourne School of Land and Environment, The University of Melbourne, Parkville, VIC, 3010, Australia
[4]Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario, N1G 2W1, Canada
email: m.khansefid2@student.unimelb.edu.au and majid.khansefid@depi.vic.gov.au

SUMMARY

Combining information from different cattle breeds is a potential way to improve the accuracy of genomic estimated breeding values (GEBVs) by increasing the size of the reference population. However, the phase of linkage disequilibrium between SNPs and quantitative trait loci for traits such as residual feed intake (RFI) may vary from one breed to another, which would erode the value of combining breeds. RFI is a selection criterion for feed efficiency and is the difference between actual intake and expected intake for maintenance and production. The aim of this research was to evaluate the accuracy of GEBVs when RFI records were combined from 5,614 animals of different breeds including 842 Holstein heifer and 2,009 Australian beef cattle (1,134 Angus, 217 Herford, 79 Murray Grey and 579 Shorthorn) and 2,763 Canadian beef cattle (534 Angus, 384 Charolais and 1,845 mixed synthetic breed) and their genotypes (606,096 SNPs) were used. We estimated the variance explained by the SNPs and the variance explained by SNP x breed interactions. The model with the highest likelihood was when SNP effects within two groups of breeds in addition to pedigree was fitted. The first group comprised Holsteins and the Angus cattle from the Trangie Research Station in NSW, Australia and the second group included all the other cattle. The difference between these two groups is that the cattle in group 1 were measured for RFI on a pelleted diet shortly after weaning while those in group 2 were measured on a feedlot diet at >1 year of age. According to the best model, the SNP effects were not significantly different between the two breeds fed a similar diet and measured at a similar age. However, the SNP effects differed between groups that were fed different diets and measured at different ages. The GEBVs of the validation animals were calculated using their SNP genotypes and the estimated SNP effects and correlated with their actual RFI phenotypes to estimate the accuracy of the GEBV. The average accuracy was 0.31 which was near to expected from the BLUP equations (0.34). Thus an across breed reference population appears to be promising for genomic prediction of RFI provided the animals are at about the same age and on a similar diet. However, there is only a small increase in accuracy by adding animals of another breed because the relationships between animals in different breeds are low. The BLUP equations correctly predict this limited increase in accuracy.

INTRODUCTION

Residual feed intake is an important trait relevant to feed efficiency in beef and dairy cattle but it is difficult to improve genetically because it is expensive to measure (Arthur *et al.* 2004). It is hoped that genomic selection using DNA markers might be used to achieve genetic improvement in RFI. Since the introduction of genomic selection (Meuwissen *et al.* 2001) there has been much research into the accuracy with which genomic estimated breeding values (GEBVs) predict true breeding values. The most common method to estimate the accuracy of GEBV has been to put aside a proportion of the population (a validation group) and not use them in the estimation of SNP effects. Then the estimated SNP effects are used to calculate GEBVs for the excluded animals which are then correlated with their phenotypic records. This correlation is the accuracy with

which the GEBVs predict new phenotypes. This method has several disadvantages. For instance, accuracies (r) or reliabilities (r^2) are not available for individual animals. When conventional BLUP is used to predict breeding values, the reliabilities of individual EBVs are calculated from the BLUP equations and it would be useful if this could also be done for GEBV but to date this approach is not well accepted. Theory and experimental results show that the reliability of GEBVs depends mostly on the precision of phenotypic data and number of genotyped animals in the reference population (VanRaden 2009). One way of increasing the number of individuals with phenotypes and genotypes is using a multi-breed reference population. However, the gain in accuracy from multi-breed reference populations has been found to be low, although a convincing explanation for this finding has not been offered. Three possible explanations are: 1. the effect of a quantitative trait locus (QTL) varies from breed to breed (*i.e.* breed x QTL interaction). This could be due to a true interaction between breed and the QTL or to an interaction between QTL and the way the trait was measured in different breeds (*e.g.* at different ages). 2. the linkage disequilibrium (LD) between the QTL and the SNPs that are assayed varies between breeds. 3. the across breed LD is low and limited to SNPs very close to the QTL so that there is limited information which can be transferred across breeds. The first two reasons result in a breed x SNP interaction. The LD between SNPs and QTL is only likely to be consistent across breeds for SNPs very close to the QTL and therefore we need very dense markers. In this research we have used around 700,000 SNPs which should be dense enough because LD phase is conserved across breeds at distances of 5 kb (deRoos *et al.* 2009). The aim of this research is to explain the accuracy of GEBV for RFI using a multi-breed reference population and to assess if using prediction error variances (PEVs) of GEBVs from the BLUP equations can correctly predict the accuracy.

MATERIAL AND METHODS

Cattle and RFI measurement. RFI records of 5,614 animals including 842 Holstein heifer, 2,009 beef cattle of Australia and 2,763 Canadian beef cattle were available for analysis. The Australian beef cattle included different breeds, 1,134 Angus, 217 Herford, 79 Murray Grey and 579 Shorthorn) and RFI data of Canadian beef consisted of 534 Angus, 384 Charolaise and 1,845 mixed synthetic breed (average breed compositions were formed by Angus (45.9%), Simmental (20.7%), Piedmontese (5%), Gelbvieh (4.2%), Charolais (2%) and Limousin (1.4%). The Holstein heifers were fed with cubed alfalfa *ad libitum* (Pryce *et al.* 2012) and the Angus cattle from Trangie Research Station were fed a pelleted diet *ad libitum* shortly after weaning. The other beef cattle used in this study were fed a feedlot diet at > 1 year of age. Residual feed intake phenotypes for the animals were obtained from 3 different studies (Australian dairy cattle: Pryce *et al.* 2012; Australian beef cattle: Bolormaa *et al.* 2013; Canadian beef cattle: Montanholi *et al.* 2009).

SNP data. The SNP marker data was from Illumina HD Bovine SNP chip, with 777,963 SNPs for Holstein heifers or imputed from lower density SNP chips (7K, 10K and 50K) to HD (800K) with BEAGLE (Browning and Browning 2009) for beef cattle. The genotypes passed quality control procedures including Illumina Genetrain (GC) score greater than 0.6 and rare minor allele frequencies higher than 0.5 % (Pryce *et al.* 2012). In order to construct genomic relationship matrix (GRM) for genomic evaluation (Yang *et al.*, 2010), common SNPs (606,096 SNPs) in the 3 datasets (Holstein heifers, Australian beef cattle and Canadian beef cattle) were used.

Statistical analysis. There were two types of GRM in the analyses: 1.using all estimated genomic relationships between all animals in the data and 2. where genomic relationships between animals of different breeds were set to zero to indicate the lack of relationship between animals of different breeds. A pedigree relationship matrix was also added to some of the models to see whether adding a polygenic term improved the log likelihood. The statistical model when the fixed effects and all three random terms were used in the analysis was:

(1) $\mathbf{y} = \mathbf{Xb} + \mathbf{Z_1u_1} + \mathbf{Z_2u_2} + \mathbf{a} + \mathbf{e}$

where, **y** is the vector of RFI records, **X** and $\mathbf{Z_{1,2}}$ are design matrixes relating phenotypes to their corresponding fixed effects and random effects, **b** is the vector of fixed effects including dataset (source of data), herd, feed management group prior to and on trial, contemporary group, cohort, month of birth, sex and age, $\mathbf{u_1}$ are SNP effects ~ $N(0,I\ \sigma^2_{SNP})$, $\mathbf{u_2}$ are SNP effects within breed ~ $N(0,I\ \sigma^2_{SNP*breed})$ and **a** are polygenic effects ~ $N(0,A\ \sigma^2_{polygenic})$. In order to fit this model, an equivalent model was used, that is:

(2) $\mathbf{y = Xb + g_1 + g_2 + a + e}$

where, $\mathbf{g_1=Z_1u_1}$ ~ $N(0, \mathbf{Z_1Z_1'\sigma^2_{SNP}})$, $\mathbf{g_2=Z_2u_2}$ ~ $N(0, \mathbf{Z_2Z_2'\sigma^2_{SNP*breed}})$ $\mathbf{Z_1Z_1'}$ is the GRM and $\mathbf{Z_2Z_2'}$ is the GRM within breed, that is all relationships between animals in different breeds have been set to zero. To test the significance of the $\mathbf{g_1}$ and $\mathbf{g_2}$ terms, the log of likelihood of the model was calculated using the full model and after dropping either $\mathbf{g_1}$ or $\mathbf{g_2}$ from the model. To test the significance of a change in log of likelihoods, two times the difference in log of likelihoods was compared to Chi squared with 1 degree of freedom. To find the best GRM within breed in the model, some breeds were treated as part of the one "super breed" in the analysis. Murray Grey and Australian Angus cattle were always grouped together and treated as one breed due to the small number of Murray Grey animals. Conversely, the Trangie Angus animals were treated as a separate breed to other Angus because RFI was measured at a younger age and using different feed at Trangie. In order to calculate the accuracies of GEBVs in a genotyped population without phenotypes, 5 subsets of the main population were generated. The animals of subsets were selected randomly but for each validation no animals with common sires were allowed to be present in both validation and reference groups. The phenotypes of each validation group were removed and after estimating GEBVs by BLUP, the correlation between GEBVs and phenotypes adjusted for fixed effects in the validation population was calculated which was divided by the square root of estimated heritability to form the empirical accuracy of estimated breeding values in each validation population.

(3) $\text{Empirical Accuracy} = r_{\text{GEBVs, Corrected\_Phenotypes}} / \sqrt{h^2_{\text{Pedigree}}}$

The empirical accuracies were compared to theoretical accuracies calculated without a validation population directly from the mixed model equations. The empirical accuracies were correlations within breed and to be consistent the theoretical accuracies were also calculated within breed. To do this, the prediction error variances for the animal effects were calculated from the mixed model equations in the standard way and used to predict the theoretical accuracy of GEBVs in the validation population.

RESULTS AND DISCUSSION

After fitting a model with an overall effect of the SNPs ($\mathbf{g_1}$) instead of the polygenic term (**a**), the log of likelihood improved significantly ($P < 0.01$) and adding SNP x breed ($\mathbf{g_2}$) further improved log of likelihood ($P < 0.01$). The results indicated that keeping the relationship between Holstein and Trangie Angus while setting the relationship between them and non-Trangie Angus and other breeds to zero (model 6) improved the log of likelihood ($P < 0.01$). However, model 6 was not significantly better than model 7 in which only Trangie Angus and Holstein relationships were kept and the relationships between different breeds were set to zero (Table 1). One of the main differences of Trangie cattle compared with the other beef animals in the experiment was their age at RFI measurement time, it seems that the effect of age is more important than the effect of breed in RFI evaluation because by treating Trangie cattle and Holstein heifers as a super breed a better log of likelihood was achieved. Therefore, the best model was reached by applying 3 relationship matrixes; an overall GRM, super breed GRM when keeping relationships between Trangie beef cattle and Holstein heifers and setting all other breed by another breed relationships to zero and pedigree relationship matrix. In this model (model 9) the genetic variance was almost

entirely explained by the overall GRM (SNP effect) and within breed GRM (SNP x breed effect). The accuracies of GEBVs were also estimated with this model. The average accuracy for RFI in 5 validations was 0.31 which was near to expected from the BLUP equations (0.34). It seems that an across breed reference population can be used provided the animals are measured for RFI at about the same age and on a similar diet. However, there is only a small increase in accuracy by adding animals of another breed because the relationships between animals in different breeds are all low. The BLUP equations correctly predict this limited increase in accuracy (about 2%).

Table1. Application of different models to find the best fitted one (highest log of likelihood)

Model	Log of Likelihood	σ^2_{SNP}	$\sigma^2_{SNP*Breed}$	$\sigma^2_{polygenic}$	σ^2_e	h^2
1. $\mathbf{Xb + g_1}$	-2853.50	0.3010	-	-	0.7024	0.3000
2. $\mathbf{Xb + g_{2\_superbreed1}}$	-2853.95	-	0.3280	-	0.6730	0.3277
3. $\mathbf{Xb + g_{2\_superbreed2}}$	-2850.83	-	0.3197	-	0.6832	0.3188
4. $\mathbf{Xb + g_{2\_superbreed3}}$	-2852.61	-	0.3312	-	0.6702	0.3308
5. $\mathbf{Xb + a}$	-2901.49	-	-	0.3023	0.7022	0.3010
6. $\mathbf{Xb + g_1 + g_{2\_superbreed2}}$	-2849.18	0.1237	0.1959	-	0.6832	0.3187
7. $\mathbf{Xb + g_1 + g_{2\_superbreed3}}$	-2847.93	0.1537	0.1790	-	0.6693	0.3320
8. $\mathbf{Xb + g_1 + g_{2\_superbreed2} + a}$	-2848.21	0.1246	0.1697	0.0522	0.6569	0.3453
9. $\mathbf{Xb + g_1 + g_{2\_superbreed3} + a}$	-2847.33	0.1523	0.1588	0.0421	0.6495	0.3522

a=pedigree relationship matrix

$\mathbf{g_1}$=(DD+TT+NT+MG+HH+SS+AA+CC+XX); $\mathbf{g_{2\_superbreed1}}$=DD,TT,(NT+MG),HH,SS,AA,CC,XX

$\mathbf{g_{2\_superbreed2}}$=(DD+TT),(NT+MG+HH+SS+AA+CC+XX);

$\mathbf{g_{2\_superbreed3}}$=(DD+TT),(NT+MG),HH,SS,AA,CC,XX

* In each model the relationships between the breeds in the same brackets were kept while relationships of the breed with another breed were assigned to zero. (DD=Holstein heifers; Australian beef cattle: NT=Non-Trangie Angus, TT=Trangie Angus, MG=Murray Grey HH=Herford, SS=Shorthorn; Canadian beef cattle: AA=Angus, CC=Charolaise, XX= Mixed synthetic breed)

CONCLUSIONS

According to the best fitting model, it seems the SNP effects were not significantly different between Holstein and Trangie cattle, fed a similar diet and measured at a similar age. However, the SNP effects probably differed between groups fed different diets and measured at different ages. So, it is important to consider feed and age at measurement time in RFI evaluations.

REFERENCES

Arthur P.F., Archer J.A. and Herd R.M. (2004) *Aust. J. Exp. Agric.* **44**:361.

Bolormaa S, Pryce J.E., Kemper K., *et al* (2013) *J. Anim. Sci.* Published online before print May 8.

Browning, B.L. and Browning S.R. (2009) Am. J. Hum. *Genet.* **84**:210.

deRoos A.P.W., Hayes B.J. and Goddard M.E. (2009) *Genet.* **183**:1545.

Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001) *Genet.* **157**:1819.

Montanholi Y.R., Swanson K.C., Palme R., Schenkel F.S., McBride B.W., Lu D., and Miller, S.P. (2009) *Anim.* **4**:692.

Pryce J.E., Arias J., Bowman P.J., Davis S.R., Macdonald K.A., Waghorn G.C., Wales W.J., Williams Y.J., Spelman R.J. and Hayes B.J. (2012) *J. Dairy Sci.* **95**:2108.

VanRaden, P.M., Van Tassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F., and Schenkel F.S. (2009) *J. Dairy Sci.* **92**:16.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E. and Visscher P.M. (2010) *Nat. Genet.* **42**:565.