Proc. Assoc. Advmt. Anim. Breed. Genet. 22:265-268

A MULTI-TRAIT APPROACH TO INCORPORATING FOREIGN PHENOTYPES AND GENOTYPES IN GENOMIC PREDICTIONS TO INCREASE ACCURACY AND REDUCE BIAS

B.J.Hayes[1,2], G. Nieuwhof[1,3], and M. HaileMariam[1]

[1]AgriBio, Department of Economic Development, Jobs, Transport and Resources, Victoria, Australia
[2]Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, University of Queensland, Queensland, Australia
[3]Datagene,Agribio,Bundoora,Australia

SUMMARY

We investigated improvements in reliability of genomic estimated breeding values (GEBV) for key dairy traits as a result of including a large number of genotypes of dairy bulls, with North American daughter performance, in the Australian genomic reference set. Two strategies for incorporating the North American information into Australian genomic evaluations were compared, a single trait approach, where the phenotypes used were de-regressed Interbull proofs (DRP), and a multi-trait approach, where North American performance and Australian performance (DYD of bulls based on Australian daughters) were treated as two, potentially correlated, traits. The two strategies were compared by assessing the correlation of GEBV and DYD from Australian daughters in a set of validation bulls, for milk, fat and protein, somatic cell count, survival, daughter fertility, stature and overall type. Including genotypes of bulls with North American daughter performance in Australian genomic evaluations improved the correlation of GEBV and DYD in the validation bulls for all traits, by between 3% and 7% for production, and up to 15% for fertility and survival. The single trait approach resulted in bias (GEBV over predicting DYD) for some traits including survival, somatic cell count and overall type, while the multi-trait approach gave unbiased GEBV for these traits.

INTRODUCTION

Reliability of genomic estimated breeding values (GEBV) for animals without a phenotype of their own or for daughters (eg young unproven bulls) is a function of the heritability of the trait, the proportion of genetic variance explained by the markers, the genetic diversity of the population, and the number of animals in the reference population where SNP effects are estimated (Daetwyler et al., 2008; Goddard, 2009). The reliability of GEBV is also a function of how closely related young genomic bulls are to the reference population. One way of improving reliabilities of GEBV would be to expand the reference set to include bulls with only overseas daughter information, including those that are sires or grandsires of young genomic bulls used in Australia. This requires a method that appropriately accounts for genotype by environment interaction between Australia and the other countries.

Here we investigate improvements in reliability of Australian genomic breeding values (ABVg) for key dairy traits that can be achieved for young, unproven bulls as a result of including a large number of genotypes of bulls with North American daughter performance into the Australian genomic reference set. Two strategies for incorporating the North American information into Australian genomic evaluations were compared,

1) a single trait approach, where the phenotypes used were de-regressed Interbull multiple across country evaluation (MACE) proofs, and

2) a multi-trait approach, where North American performance (daughter yield deviation, DYD, of bulls based on North American daughters) and Australian performance (DYD of bulls

based on Australian daughters) were treated as two, potentially correlated, traits.

The two strategies were compared by assessing the correlation of genomic estimated breeding values (GEBV) and DYD from Australian daughters) in a set of validation bulls (born in or after 2008). The increase in this correlation relative to a single trait approach with only the current Australian reference set was evaluated. The regression of DYD on GEBV was also evaluated, to determine if there was any bias (i.e. if the GEBV systematically over estimated or underestimated the proofs of top ranking bulls when the bulls had daughters). Traits investigated were production (milk, fat and protein kg), somatic cell count, survival, daughter fertility, stature and overall type (standardised traits with a mean of 100 and standard deviation of 5).

MATERIALS AND METHODS

Genotypes for 18,377 North American registered bulls with daughter records were extracted from the Northern American Cooperative Dairy DNA Repository (CDDR) database, and 13,072 bulls and cows from the ADHIS (Australian dairy herd improvement scheme) database. A set of 36,655 SNP common to the Australian evaluation and present in the North American genotypes was identified. Any missing genotypes were imputed using Beagle 3.2 (Browning and Browning 2009). The traits investigated were milk yield, fat yield, protein yield, somatic cell count (SCC), survival (longevity), stature, overall type (overall conformation score was the corresponding Interbull trait) and fertility (daughter pregnancy rate).

North American phenotypes were daughter yield deviations from the US, for milk yield, fat yield, protein yield, somatic cell count (SCC) and survival. For fertility, the North American PTA was de-regressed as suggested by Van Raden (pers comm). For type traits, de-regressed breeding values were used (de-regression removed the pedigree contribution of the EBV), where bulls had at least 50 daughters scored for the trait, using the procedure of Liu (2009). Australian phenotypes were daughter trait deviations (DTDs) for all traits.

The data were split based on year of birth into reference and validation sets. Bulls (either North American or Australian) born before 2008 were included in the reference set, used to calculate SNP effects, and bulls born in or after 2008 were used in the validation set. There were 275 bulls in the validation set, and only Australian daughter information was used in the validation

Three models were fitted to the data 1) Single trait model, Australian information only (de-regressed MACE proofs as phenotypes), 2) Single trait model, Australian and North American information (de-regressed MACE proofs as phenotypes) 3), Multi-trait model, using daughter trait deviations for bulls with daughters in Australia, and DYD for bulls with North American daughters (as described above). Where DYD were not available, de-regressed proofs were used.

The multi-trait model was $\begin{bmatrix}\boldsymbol{y_1}\\ \boldsymbol{y_2}\end{bmatrix}= \begin{bmatrix}\boldsymbol{I_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I_2}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix}+\begin{bmatrix}\boldsymbol{Z_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Z_2}\end{bmatrix}\begin{bmatrix}\boldsymbol{g_1}\\ \boldsymbol{g_2}\end{bmatrix}+\begin{bmatrix}\boldsymbol{e_1}\\ \boldsymbol{e_2}\end{bmatrix}$ where $\mathbf{y}_1$ and $\mathbf{y}_2$ are the vector of response variables (i.e. trait 1 is the DTD of Australian bulls and trait 2 are the DYD of bulls with North American daughters), $\mathbf{I}_1$ and $\mathbf{I}_2$ are identity matrices, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is the vector of intercepts of DTD and DYD, $\mathbf{Z}_1$ **and** $\mathbf{Z}_2$ are the design matrices that relate genomic breeding values with the individuals, $\mathbf{g}_1$ and $\mathbf{g}_2$ is the vector of genomic breeding values for DTD and DYD, and $\mathbf{e}_1$ and $\mathbf{e}_2$ are vectors of random residuals for DTD and DYD. It was assumed that $\begin{bmatrix}\boldsymbol{g_1}\\ \boldsymbol{g_2}\end{bmatrix} \sim N(0, \boldsymbol{G}\otimes\boldsymbol{T})$, where $\mathbf{T} = \begin{bmatrix}\boldsymbol{\sigma^2_{g1}} & \boldsymbol{\sigma_{g12}}\\ \boldsymbol{\sigma_{g12}} & \boldsymbol{\sigma^2_{g2}}\end{bmatrix}$, the variance-covariance matrix of DTD and DYD, and $\begin{bmatrix}\boldsymbol{e_1}\\ \boldsymbol{e_2}\end{bmatrix} \sim N(\boldsymbol{0}, \boldsymbol{I}\otimes\boldsymbol{R}), where\ \boldsymbol{R} = \begin{bmatrix}\boldsymbol{R_{11}} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R_{22}}\end{bmatrix}$, the residual variance-covariance matrix of DTD and DYD, with weights on phenotypes for bulls and cows according to Garrick et al. (2009), and **G** is the genomic relationship matrix. ASReml (Gilmour et al., 2009) was used to estimate variance

components, including genetic correlations between performance in North American and Australia.

The three different models were used to predict GEBV for the validation bulls, born 2008 and later. The following statistics were assessed: correlation between DTD and GEBV for bulls in the validation set, and the slope of the regression (b) of DTD on GEBV for validation bulls.

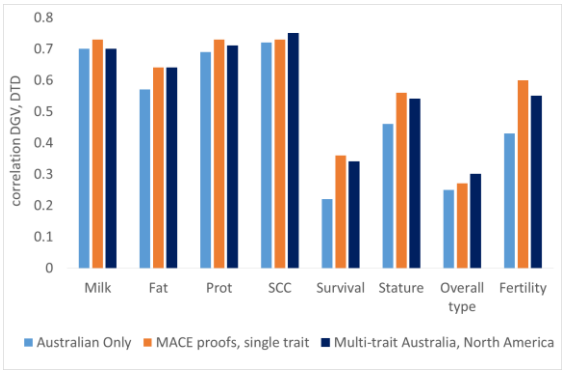
Note that to (considerably) simplify implementation in routine evaluations, if a bull has Australian daughters, only his Australian information is used. Bulls are included for the second country trait (eg North America) only if they have daughters in that second country and not in Australia. This means it is not necessary to consider residual correlations among the countries. A second step to simplify implementation was to pre-correct records in each country for the mean and sex effect. Then the solutions to the multiple trait model are (Ignoring fixed effects, and with t the elements of the inverted T matrix, eg t^{11} is the element in the first row and column of T^{-1}):

$$\begin{bmatrix}\widehat{\boldsymbol{g}_1}\\ \widehat{\boldsymbol{g}_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Z_1'R^{11}Z_1 + G^{-1}t^{11}} & \boldsymbol{G^{-1}t^{12}}\\ \boldsymbol{G^{-1}t^{12}} & \boldsymbol{Z_2'R^{22}Z_2 + G^{-1}t^{22}}\end{bmatrix}\begin{bmatrix}\boldsymbol{Z_1'R^{11}y_1}\\ \boldsymbol{Z_2'R^{22}y_2}\end{bmatrix}$$

RESULTS AND DISCUSSION

When de-regressed MACE proofs were used as the phenotype in a single trait analysis, the correlations r(GEBV,DTD) for production were relatively high, and improved by up to 7% (fat) with the addition of the North American data (Figure 1A).

A



B

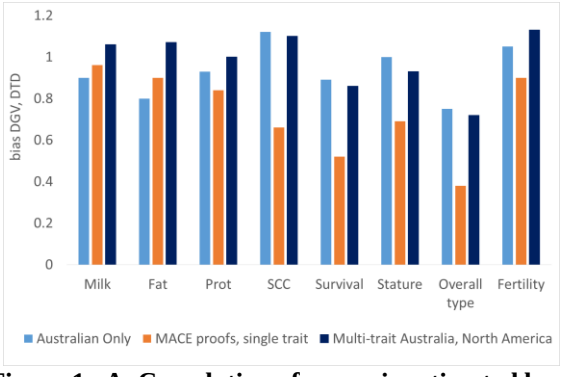


Figure 1. A. Correlation of genomic estimated breeding values (GEBV) for eight dairy traits, using three genomic prediction models. B. Regression of daughter yield deviation on GEBV for eight dairy traits using three genomic prediction models.

The multi-trait approach did result in regression coefficients (slopes) of DTD on GEBV in

validation bulls closer to one for some traits (Figure 1B). Particularly survival, somatic cell count and overall type were closer to one with the multi-trait approach.

As a result of running the multi-trait model, genetic correlations between North America and Australia were estimated for all traits considered. These were slightly higher than, but close to the Interbull reported correlations, Table 1. This is interesting, as the multi-trait model uses genomic information only, while the Interbull correlations are based on pedigree.

Table 1. Genetic correlations between Australia and the US, estimated either from the multi-trait genomic model, or from pedigree (Interbull reported correlations).

Trait	Multi-trait genomic estimate	Interbull*
Milk	0.81	0.77
Fat	0.81	0.76
Prot	0.75	0.75
SCC	0.73	0.77
Survival	0.75	0.69
Stature	0.95	0.89
Overall type	0.72	0.64

*http://www.interbull.org/index

Using either a multi-trait approach or a single trait approach (using de-regressed MACE proofs) to add North American daughter performance information to the reference population for calculating GEBV resulted in an increase in r(DTD,GEBV) for a set of validation bulls. The multi-trait approach resulted in slightly less bias (slope of DTD on GEBV for validation bulls) for some traits, and is therefore the preferred approach for these traits. Estimates of genetic correlations between North America and Australia derived from the genomic information were similar to, but slightly higher than, the published Interbull correlations.

ACKNOWLEDGMENTS

The authors are grateful to the CDDR, George Wiggans, Jay Weiker Gordon Doak, Kent Weigel, Jaques Chesnais, Daniel Abernethy and Matthew Shaffer for facilitating this research project.

REFERENCES

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. **84**:210–223.

Garrick, D. J., J. F. Taylor, and R. L. Fernando, 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. **41**:55.

Liu, Z. 2009. Deregressing MACE proofs for genomic evaluations. Proteje meeting, Brussels, Belgium.