THE EFFECTS OF USING MALE AND FEMALE GENOTYPES IN GENOMIC EVALUATIONS

D.L. Johnson

LIC, Private Bag 3016, Hamilton 3240, New Zealand

SUMMARY

The objective of this study was to quantify additional accuracy of genomic evaluation from the addition of female genotypes to a dairy cattle population. The basic training set consisted of 6,150 progeny-tested bulls born prior to 2007 and the validation set consisted of 350 progeny-tested bulls born 2007-2008. Additionally, 36,350 female genotypes were included in the training population. The phenotypes were deregressed breeding values for production traits. Ridge regression was used with two models: (1) common SNP effects fitted for both genders and (2) SNP effects depending on gender with an assumed correlation. Bayes methods B and Cπ were also fitted under scenario (1). The accuracy of genomic evaluation was increased by 5 to 10 percentage points with the inclusion of female genotypes, depending on breed and trait. There was little difference in accuracy among models and methods of analyses.

INTRODUCTION

Genomic breeding values are now being widely used for bull selection in the dairy industry. One factor influencing the accuracy of genomic predictions is the size of the reference or training population. The relationship between predictive ability and the size of reference population has been demonstrated in Daetwyler *et al*. (2008) and Goddard and Hayes (2009). An option to increase accuracy of genomic evaluation is to combine reference populations from different countries (EuroGenomics, David *e tal.* 2010). Another option to boost the reference population is to genotype females. Apart from bull dams, LIC has a program of genotyping daughters of young bulls in the sire proving scheme (SPS) to maintain integrity of bull proofs through parentage testing.

MATERIALS AND METHODS

As at the end of the 2012/2013 season, LIC had a total of 6,500 progeny-tested bulls genotyped on the Illumina BovineSNP50 Beadchip (Illumina Inc., San Diego, CA). The validation population was taken as the 350 bulls progeny tested over the last two seasons (born 2007-2008). The base reference population comprised the remaining 6,150 bulls born 2006 and earlier. From a larger pool of genotyped cows, a total 36,350 with lactation records were included in the reference population. The cows comprised SPS daughters and their contemporaries as well as cows genotyped for other research purposes but excluded daughters of young bulls in the validation population. Most of the cows were genotyped on the Illumina 50K panel with some genotyped on a lower density GGP panel (6.5K) and then imputed to 50K using Beagle 3.3.2 (Browning and Browning 2009). There were 38,808 SNP included in the analyses after removing SNP for low call rates, minor allele frequencies <2%, non-Mendelian inheritance, failed Hardy-Weinberg tests and low imputation accuracy. The bull population was multi-breed comprising mainly Holstein-Friesian (HF), Jersey (JE) and crossbred (FJ) bulls. The bull reference comprised 56% HF, 34% JE and and 7% FJ and the validation population was correspondingly 39%, 30% and 24% reflecting the development of crossbred bulls in recent years. The cow population was a similarly structured multi-breed population with 32% HF, 23% JE and 40% FJ.

The SNP effects were estimated using multiple-regression models where the marker effects are treated as random. The model can be written

$$\mathrm{y{=}Xb{+}Zs{+}e} \qquad [1]$$

where **y** is phenotype, **b** denotes fixed effects (in this case just an overall mean), **s** denotes SNP effects, **X** and **Z** are design matrices and $E(\boldsymbol{y}) = \mathbf{Xb}, \boldsymbol{var}(\boldsymbol{s}) = \mathbf{I}\boldsymbol{\sigma}_s^2, \boldsymbol{var}(\boldsymbol{e}) = \mathbf{R}\boldsymbol{\sigma}_e^2$ **.** The mixed model equations (MME) corresponding to [1] are

$$\begin{bmatrix} \mathrm{X'R^{-1}X} & \mathrm{X'R^{-1}Z} \\ \mathrm{Z'R^{-1}X} & \mathrm{Z'R^{-1}Z{+}\lambda I} \end{bmatrix} \begin{bmatrix} \hat{\mathrm{b}} \\ \hat{\mathrm{s}} \end{bmatrix} = \begin{bmatrix} \mathrm{X'R^{-1}y} \\ \mathrm{Z'R^{-1}y} \end{bmatrix}$$

where $\lambda = \sigma_e^2/\sigma_s^2$. This ridge regression with known λ is equivalent to GBLUP when solving directly for genomic breeding values g=Zs provided we have the relation $\sigma_g^2 = \sigma_s^2 \sum_j 2p_j\,(1 - p_j)$ between genetic variance and common SNP variance, with p_j denoting allele frequency of SNP *j*. This basic ridge regression model was used for the two reference sets; (1) bulls only and (2) bulls plus cows. In addition, for reference set (2), a mixture model approach was used for model [1]. A Bayes B model was fitted assuming that each marker has either a zero effect with known probability $\pi = 0.95$ or a non-zero effect with different λ values (Meuwissen *et. al.*, 2001). A Bayes Cπ model was also fitted where one assumes a common λ but unknown π (Habier *et. al.*, 2011).

For reference set (2), model [1] was extended to allow for different SNP effects depending on gender.

$$\mathrm{y{=}Xb{+}Z_1s_1} + \mathrm{Z_2s_2} + \mathrm{e} \qquad [2]$$

with

$$var\begin{bmatrix} \mathrm{s_1} \\ \mathrm{s_2} \end{bmatrix} = \begin{bmatrix} \mathrm{I} & \mathrm{I\rho} \\ \mathrm{I\rho} & \mathrm{I} \end{bmatrix} \sigma_s^2$$

where ρ denotes SNP correlation between genders and with common variance among SNP within gender and I is the identity matrix. The MME corresponding to model [2] are

$$\begin{bmatrix} \mathrm{X'R^{-1}X} & \mathrm{X'R^{-1}Z_1} & \mathrm{X'R^{-1}Z_2} \\ \mathrm{Z'_1R^{-1}X} & \mathrm{Z'_1R^{-1}Z_1} + \dfrac{\lambda \mathrm{I}}{(1-\rho^2)} & \dfrac{-\lambda\rho \mathrm{I}}{(1-\rho^2)} \\ \mathrm{Z'_2R^{-1}X} & \dfrac{-\lambda\rho \mathrm{I}}{(1-\rho^2)} & \mathrm{Z'_2R^{-1}Z_2} + \dfrac{\lambda \mathrm{I}}{(1-\rho^2)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{s_1} \\ \hat{s_2} \end{bmatrix} = \begin{bmatrix} \mathrm{X'R^{-1}y} \\ \mathrm{Z'_1R^{-1}y} \\ \mathrm{Z'_2R^{-1}y} \end{bmatrix}$$

Two values of ρ equal to 0.7 and 0.9 were assumed. The BLUP models were solved using a conjugate gradient method.

Phenotypes were the the deregressed BV for the three production traits, milk volume, fat and protein yield, hereafter referred to as milk, fat and protein. The deregression procedure was carried out as in Garrick *et. al.* (2009). The elements of the diagonal **R** matrix associated with the error structure were calculated as $\left(c + \frac{1-r_i^2}{r_i^2}\right)\sigma_g^2/\sigma_e^2$ where *c*=0.1 is the assumed fraction of genetic variance unexplained by the markers and the second component is associated with the error variance of the deregressed BV with reliability r_i^2 for individual *i*. The constant c^{-1} also acts as an upper bound for the weighting applied to phenotypes corresponding to highly proven sires.

The validation procedure involved the regression of deregressed BV on genomic BV for the young bulls within breed as per Interbull procedure (Mantysaari *et. al.* 2010). The correlations were summarized as well as the regression coefficients to assess accuracy and bias of prediction. The accuracy attained through selection of the top 20 bulls on genomic BV was also investigated.

RESULTS AND DISCUSSION

The correlations between genomic BV and progeny-test BV for milk, fat and protein, based on the validation population, are summarised within breed in Table 1. The first data column is based on the bull reference while all other results relate to the combined reference. The inclusion of cows in the reference has increased the correlations by an average of 0.09-0.10 for milk and fat and somewhat less at 0.05 for protein. Comparisons across columns of Table 1 indicate generally small differences among genomic evaluation methods using the combined reference population. In particular there appears little advantage to fitting marker effects by gender.

Table 1. Validation correlations for bull reference and combined (bull + cow) reference populations – RR=ridge regression, ρ is assumed correlation when fitting SNP effects by sex

Reference	bull	combined	combined	combined	combined	combined
Method	RR	RR	RR(ρ=0.9)	RR(ρ=0.7)	BayesB	BayesCπ
Fat						
HF	0.55	0.72	0.72	0.69	0.69	0.71
JE	0.62	0.64	0.66	0.66	0.61	0.63
FJ	0.50	0.60	0.62	0.62	0.60	0.60
Protein						
HF	0.50	0.54	0.56	0.55	0.54	0.56
JE	0.51	0.58	0.58	0.57	0.56	0.58
FJ	0.68	0.71	0.72	0.72	0.66	0.69
Milk						
HF	0.59	0.68	0.68	0.66	0.70	0.71
JE	0.54	0.70	0.67	0.64	0.69	0.69
FJ	0.74	0.76	0.77	0.77	0.76	0.77

The average reliabilities of the phenotypes for production traits were about 0.85 and 0.3 for bulls and cows, respectively. Based on the weighting formula, bulls would get an average weight of 3.6 and cows 0.4. Thus it takes about nine cows at that level of reliability to provide information equivalent to one progeny-tested bull and so 36,350 cows is equivalent to about 4,000 bulls. Based on the formula of Goddard and Hayes (2009), and assuming a heritability of 0.8 and effective population size of 100, the expected incremental change in the accuracy of genomic evaluation due to an increase of 4,000 bull equivalents above a base of 6,000 bulls is approximately 0.08. The results of this study suggest that the advantage of inclusion of the female data is close to expectation.

Table 2 summarises the regression coefficient of phenotype on estimated genomic BV for the ridge regression. The values represent a weighted average across breed. The expectation is unity and smaller values indicate some degree of inflation or bias in the genomic predictions. With the base reference set of bulls, the regressions are close to unity but decrease to about 0.8 for fat and protein when cows are included in the reference, suggesting some inflation. It is important to correct for this bias otherwise overestimation of genomic BVs will erode farmer confidence in genomic evaluations.

Table 2. Validation regression coefficient weighted across breed

Trait	Bull reference	Bull + cow reference
Fat	1.01±0.09	0.80±0.05
Protein	0.94±0.08	0.80±0.06
milk	0.97±0.07	0.95±0.05

The regression yields estimates of population parameters, however it is the animals at the top end of the distribution that are of interest. The top 20 bulls within breed were selected on genomic BV for the two reference populations. The change in average phenotype of these two groups of bulls (reference (2) minus reference (1)) is presented in Table 3 along with the number of bulls common to both groups. For each trait there was a positive change for two of the three breeds. The standard error of the difference (SED) between the averages of the two groups, assuming independence among bull proofs, is expected to be $\sigma_g\sqrt{2(n-m)(1/r-1)}/n$ where n=20 is the number of bulls selected, m is the number of bulls in common and r=0.75 is the daughter-proven reliability of an individual bull. Given genetic standard deviations of 329 litres, 13.6 kg, and 9.6 kg for milk, fat and protein, respectively, the SEDs are included in Table 3. Apart from fat, the evidence of significant improvement using data from the selected bulls is not as strong as that indicated by the population statistics however they are based on small numbers.

Table 3. Number of bulls intersecting the top 20 for genomic BV based on the two reference populations and difference in average deregressed daughter-proven BV

	Fat (kg)		Protein (kg)		Milk (litres)	
Breed	Bulls in common	difference	Bulls in common	difference	Bulls in common	difference
HF	12	6.1±1.6	9	0.4±1.3	9	20±45
JE	16	-1.6±1.1	13	1.0±1.0	15	30±30
FJ	13	2.6±1.5	14	-0.9±1.0	15	-60±30

CONCLUSIONS

There is some evidence to indicate that increasing the size of reference population through inclusion of cow data may lead to an improvement in the accuracy of genomic evaluation. The feasibility of including cow genotypes in a single-step method of evaluation (Aguilar *et. al.*, 2010), which combines information from genotyped and non-genotyped animals, is currently being investigated to confirm results of this study and check validation over a sequence of years. This will provide computational challenges in terms of inversion of the genomic relationship matrix which may become infeasible in the future as the number of genotyped animals increases. Reparameterisation of the MME in terms of marker effects instead of directly as BVs may be a better computational strategy.

REFERENCES

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

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

Daetwyler H.D., Villanueva B, and Woolliams J.A. (2008) *PLoS ONE* **3**:e3395.

David X., de Vries A., Feddersen E. and Borchersen S. (2010) *Interbull Bull.* **41**:77.

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

Goddard M.E. and Hayes B.J. (2009) *Nature Reviews Genetics* **10**:381.

Habier D., Fernando R.L., Kizilkaya K. and Garrick D.J. (2011) *BMC Bioinformatics* **12**:186.

Mantysaari E., Liu Z. and VanRaden P. (2010) *Interbull Bull.* **41**:17.

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