Evaluating the Impact of Including Residual Polygenic Effects in Dairy Genomic Evaluations using Bayesian Methods

***R. Mrode*[1]*, K Moore*[1]*, Marco Winters*[2] *and M. Coffey*[1]**
[1]Scottish Agricultural College, Roslin Institute Building, Easter Bush EH25 9RG
Midlothian, UK, [2]DairyCo, Agriculture and Horticulture Development Board Stoneleigh Park
Kenilworth, Warwickshire, CV8 2TL, UK

Abstract

The impact of different levels of residual polygenic effect was studied for milk yield and somatic cell count (SCC) using GBLUP, BayesA and BayesB in a Holstein/Friesian population. The reference population consisted of 8605 and 7092 bulls with at least 10 EDCs and reliability of 69% for milk yield and SCC respectively. Corresponding bulls in the validation set were 4090 and 2448 respectively. A linear model was used for the estimation of SNP effects with fixed mean effect and random residual polygenic (RP) and SNP effects. The RP levels were set at 0, 5, 10, 15, 20 and 25% of the total genetic variance. In the case of BayesA and BayesB, analyses were carried out with no restriction on the percentage of the total genetic variance accounted for by the polygenic effect. The regressions of direct genomic breeding values (DGV) on de-regressed sire proofs (DSP) increased (0.47 to 0.98) with increasing levels of polygenic effect for both traits. The regressions for milk yield were similar for GBLUP and BayesB but higher than estimates from BayesA, indicating slightly poorer predictions from BayesA. Similarly, higher regressions were obtained for SCC from GBLUP compared to BayesA. In general correlations between DGV and DSP were generally higher with GBLUP compared with the Bayesian methods at each level of RP fitted. The correlations of the polygenic solutions with DGV were always higher for SCC at every level of RP compared with milk yield indicating the higher impact of the RP effect for traits of lower heritability. Contributions from parent average accounted for 68% and 89% of the RP contributions for bulls with at most 15 or 100 EDCs with GBLUP at 10% RP for milk yield and SCC respectively. There was a decreasing trend in the mean of SNP solutions for SNPs with alleles of medium, and high frequencies as the percentage of polygenic increased with all three methods for milk yield. However such a trend was not observed for SCC. In general mean SNP variances for milk yield and SCC declined with increasing levels of polygenic effects. When no constraint was imposed on the level of RP, estimates of the polygenic variance were unexpectedly high from BayesA and BayesB which were difficult to interpret.

Key words: Genomic evaluations, polygenic effects, Bayesian methods

Introductions

Recent developments in molecular biology have resulted in the emergence of low cost genotyping technology for Single Nucleotide Polymorphism (SNP). Due to linkage disequilibrium between SNPs and quantitative trait loci for traits of economic value, genomic breeding values can now be computed directly for animals on the basis of SNP effects.

The genomic BLUP model used to estimate SNP effects in most dairy populations are based on SNPs obtained from the Illumina Bovine 54K chips and it is usually assumed that these SNPs explain all the genetic variation for the traits analysed. However, fitting a residual polygenic effect (RP) may account for the fact that SNPs may not explain all the genetic variance and has also been found to render SNP effects less biased (Solberg *et al.*, 2009). Liu *et al.* (2011) has demonstrated that the optimum level of RP may differ for traits of different heritabilities. As far as the authors are aware, the impact of fitting a RP effect in the model using Bayesian methods has not previously been demonstrated. Thus this study examines the accuracies of predictions at different levels of RP using GBLUP, BayesA and BayesB. In addition, the impact of different levels of RP was examined on mean SNP effects and variances for SNPs with alleles at different frequencies. The two traits analysed were milk yield and $\log_e$ somatic cell count (SCC) in Holstein/Friesian, representing traits of high and low heritabilities respectively.

Materials and Methods

Data from 11480 bulls with 50k genotypes were used in the analysis. However, 600 of these bulls were genotyped with the Illumina 800K chip but only the corresponding SNPs on the 50k chip were extracted and used for these bulls. These genotypes are from a collaboration through DairyCo with the North American Cooperative Dairy DNA Repository (CDDR), ANAFI (Italy), UK AI industry and SAC. Minor allele frequency was set to 0.05 and call rate for animals at 95%. In total 41866 SNPs were selected for genomic evaluations after these various edits. Genotypes were coded as 0 and 2 for the homozygotes and 1 for the heterozygotes. De-regressed sire proofs (DSP) from the UK official April 2011 run and MACE proofs were used as input variables in the genomic evaluations. A linear model was used for the estimation of SNP effects with fixed mean effect and random polygenic and SNP effects. The analysis was carried out using GBLUP, BayesA and BayesB. However, BayesB was only applied for the analysis of milk yield due to time constraints. The RP levels were set at 0, 5, 10, 15, 20 and 25% of the total genetic variance. In the case of BayesA and BayesB, analyses were also carried out with no restriction on the percentage of the total genetic variance accounted for by the polygenic effect. For the Bayesian methods, the MCMC chains were run for 80000 cycles, the first 24000 of which were discarded as the burn in period. In the case of BayesB, 20 Metropolis-Hastings cycles were undertaken within each MCMC chain and the proportion of SNPs with non-zero variance was fixed at 0.70.

Genotyped bulls born before and in 2004 with an EDC of at least 10 and a reliability of at least 69% were used as the reference population and were used to estimate the SNP effects. This consisted of 8605 and 7092 bulls for milk yield and SCC respectively. A total of 4090 and 2448 bulls with at least 10 EDCs and reliability of 69% for milk yield and SCC respectively and born after 2004 were used for the purposes of validation. The effects of different levels of RP were examined by computing correlations between direct genomic values (DGV, which is combination of SNP solutions and the polygenic component) and DSP for bulls in the reference and validation sets. In addition, correlations between SNP and polygenic solutions with DGV were computed. The contribution from the polygenic effect was partitioned to contributions from DSP and parent average (Mrode and Swanson, 2004) and examined for bulls with different number of daughters in the reference populations. The mean SNP effects and variances for SNPs of different allele frequencies; low (< 0.4), medium (0.4 – 0.6) and high (> 0.6) were calculated for the different levels of RP. The numbers of observations SNP in the low, medium and high frequency categories were 12874, 14888 and 14104 respectively.

Results and Discussion

The regressions of DGV on DSP and the correlation between both variables in the validation set are shown in Table 1 for milk yield and Table 2 for SCC. In general, the regressions increased with increasing levels of polygenic effect for both traits and this is similar to the results of Liu *et al.* (2011) who studied more traits. The regressions for milk yield were similar for GBLUP and BayesB but higher than estimates from BayesA, indicating slightly poorer predictions from BayesA. Similarly, higher regressions were obtained for SCC from GBLUP compared to BayesA. In general correlations between DGV and DSP were higher with GBLUP compared with the Bayesian methods at each level of RP fitted.

Table 1. Regressions (Reg) and correlations (Corr) between DGV and DSP in the validation set for milk yield.

Polygenic level (%)	Milk Yield					
	Gblup		BayesA		BayesB	
	Reg	Corr	Reg	Corr	Reg	Corr
0	0.852	0.68	0.625	0.53	0.753	0.61
5	0.876	0.68	0.684	0.55	0.801	0.62
10	0.889	0.67	0.700	0.56	0.823	0.63
15	0.899	0.67	0.719	0.56	0.833	0.63
20	0.899	0.67	0.734	0.57	0.833	0.62
25	0.912	0.67	0.718	0.56	0.828	0.62
	-	-	0.695	0.51	0.794	0.59

Table 2. Regressions (Reg) and correlations (Corr) between DGV and DSP in the validation set for somatic cell count.

Polygenic level (%)	Somatic cell count					
	Gblup		BayesA		BayesB	
	Reg	Corr	Reg	Corr	Reg	Corr
0	0.852	0.68	0.625	0.53	0.753	0.61
5	0.876	0.68	0.684	0.55	0.801	0.62
10	0.889	0.67	0.700	0.56	0.823	0.63
15	0.899	0.67	0.719	0.56	0.833	0.63
20	0.899	0.67	0.734	0.57	0.833	0.62
25	0.912	0.67	0.718	0.56	0.828	0.62
	-	-	0.695	0.51	0.794	0.59

The correlations of solutions for SNP effects with DGV decreased while the correlations for polygenic solutions increased with increasing levels of RP in a similar manner for both traits (Table 3). However the correlations of the polygenic solutions with DGV were always higher for SCC at every level of RP. This would indicate a higher impact of the RP for traits of lower heritability. Partitioning the contribution of the RP into contributions from DSP and PA, indicated that contributions from PA accounted for 45% and 80% of the RP contributions for bulls with at most 15 EDCs with GBLUP at 5% RP for milk yield and SCC respectively. However this contribution was only 5 percent when bulls have at least 100 EDCs for milk yield. This implies that DGV of young bulls with few daughters will change over time as the contributions from PA decreases as the bull accumulates more daughters and this change can be substantial for traits of lower heritability, like SCC. In the study of Liu *et al.* (2011), the choice of the optimum level of RP was based on the regression coefficients, favouring the regressions that were about unity. In this study, all the regressions were less than unity and approached unity with increasing levels of RP but gain in predicative ability after a particular RP level are not substantial, for instance 10% for milk or 20% for SCC. The choice of the optimum level could be based on the basis of the regressions, correlations and mean square error variance and the awareness that PA contribution can be substantial at lower levels of RP when bulls have few daughters.

There was a decreasing trend in the mean of SNP solutions for SNPs with medium, and high allele frequencies as the level of RP increased for milk yield with all three methods (Table 4).

Table 3. Correlations between SNP or polygenic (RP) contributions with DGV for the GBLUP in the reference population.

Polygenic%	**Milk yield**		**Somatic cell count**	
	Contribution to DGV		Contribution to DGV	
	SNP	RP	SNP	RP
5	0.99	0.69	0.99	0.70
10	0.98	0.71	0.97	0.74
15	0.97	0.72	0.97	0.78
20	0.95	0.73	0.96	0.81
25	0.94	0.77	0.94	0.83

Table 4. Mean of SNP solutions for milk yield at different levels of polygenic effect and SNPs with alleles of low, medium (Med) and high frequencies.

Polygenic level (%)	Mean SNP effects milk yield								
	Gblup			BayesA			BayesB		
	Low	Med	High	Low	Med	High	Low	Med	High
0	-0.280	0.059	0.432	-1.124	-0.313	0.289	-0.371	-0.094	0.239
5	-0.260	0.019	0.368	-0.395	-0.140	0.187	-0.464	-0.087	0.191
10	-0.245	0.015	0.348	-0.522	-0.043	0.120	-0.335	-0.116	0.139
15	-0.241	0.008	0.337	-0.500	-0.183	0.168	-0.365	-0.131	0.168
20	-0.235	0.005	0.328	-0.441	-0.101	0.066	-0.304	-0.138	0.100
25	-0.222	0.014	0.318	-0.347	-0.147	0.092	-0.249	-0.166	0.160
Not Fixed	-	-	-	-0.118	-0.109	-0.004	-0.202	-0.100	0.057

Table 5. Mean of SNP solutions for SCC at different levels of polygenic effect and SNPs with alleles of low, medium (Med) and high frequencies.

Polygenic %	Gblup			BayesA		
	Low	Med	High	Low	Med	High
0	-0.009	0.262	-0.027	-1.439	-0.426	0.438
5	-0.014	0243	-0.016	-0.576	-0.280	0.413
10	-0.017	0.224	-0.012	-0.547	-0.231	0.367
15	-0.016	0.207	-0.008	-0.522	-0.228	0.347
20	-0.016	0.192	-0.006	-0.513	-0.192	0.322
25	-0.016	0.179	-0.002	-0.535	-0.205	0.340
Not Fixed	-	-	-	-0.529	-0.205	0.318

The decrease in mean SNP effects at 25% RP in the model varied from 26 and 76% for milk yield compared with 0% RP. In contrast, an increasing trend was observed for mean SNP solutions of low allele frequency. Again the increase in mean SNP effects at 25% RP in the model varied from 21 and 69% for milk yield. However, the same consistent trends for mean of SNP solutions were not observed for SCC (Table 5). While there was a decreasing trend for SNP solutions for alleles of medium frequencies with GBLUP, the opposite tended to be the case for results for BayesA. In fact, the trends were generally opposite for mean SNP solutions for GBLUP and BayesA for SCC at the different level of RP fitted.

In general mean SNP variances for milk yield declined with increasing levels of polygenic effects in the models (Table 6) with either BayesA or BayesB. However, the rate of decline was more pronounced for SNPs with alleles of high frequency for BayesA. For instance the rate of decline at 25% RP was 71% compared with no RP in the model for SNPs with alleles of high frequencies. The corresponding figure for SNPs of low or medium frequency was about 41%. A similar declining trend was observed for SCC for SNPs at low and medium frequencies (Table 7) but the variances were in general higher (16 – 28%) for SNPs of higher frequency when RP was fitted. It seems that the inclusion of RP effects in the model resulted in the decline of the variance of SNPs. This was more pronounced when no constraint was imposed on the polygenic effects resulting in substantial decrease in SNP variances. When no constraint was imposed on the polygenic variance, the proportions of total genetic variance due to polygenic effects were unexpectedly high at about 65% and 49% from BayesA for milk yield and SCC respectively. It could be that the inclusion of a numerator relationship matrix for the polygenic effect makes more difficult for the model to partition the variances appropriately. The very high polygenic variance when no constraint is imposed is reflected in the poorer predictive ability of this model for milk yield (Table 2) but the predictive ability was not affected in the case of SCC.

Table 6. Mean of SNP variances for milk yield at different levels of polygenic effect and SNPs with alleles of low, medium (Med) and high frequencies.

Polygenic level (%)	Mean SNP variances for milk yield					
	BayesA			BayesB		
	Low	Med	High	Low	Med	High
0	0.816	0.950	1.128	1.049	0.884	0.801
5	0.823	0.793	0.471	0.905	0.796	0.695
10	0.784	0.666	0.498	0.751	0.763	0.609
15	0.683	0.626	0.382	0.652	0.699	0.598
20	0.655	0.586	0.267	0.718	0.624	0.551
25	0.485	0.520	0.338	0.634	0.615	0.561
Not fixed	0.185	0.272	0.126	0.491	0.542	0.422

Table 7. Mean of SNP variances for SCC at different levels of polygenic effect and SNPs with alleles of low, medium (Med) and high frequencies.

Polygenic %	Mean SNP variances for SCC		
	BayesA		
	Low	Med	High
0	0.419	0.414	0.163
5	0.181	0.204	0.226
10	0.176	0.196	0.219
15	0.172	0.189	0.215
20	0.171	0.184	0.209
25	0.169	0.181	0.205
Not fixed	0.164	0.169	0.195

Conclusions

On the basis of the results in the validating set, the predictive ability of GBLUP was generally higher than those for Bayesian methods for both milk yield and SCC at the different levels of the RP fitted. The contribution of parent average to the polygenic solutions could be substantial at low levels of RP effects especially for bulls with few daughters and traits of lower heritability. The choice of optimum RP level may involve a careful consideration of the regression coefficient and other factors such as mean square error, correlations in the validation set and accounting for possible effects of parent contribution. The fitting of RP had different effects on SNPs depending on their allele frequencies and the methodology applied to estimate SNP effects. In general, increasing levels of RP resulted in a decreasing trend in SNP variances which was independent on the allele frequencies of the SNP. When no constraint was imposed on the level of RP, estimates of the polygenic variance were unexpectedly high and difficult to interpret.

Acknowledgements

Funding of implementation of genomic evaluations by DairyCo and access to the combined CDDR, ANAFI and UK genotypes is gratefully acknowledged. Contributors of programmes to Clubware are thanked as these were modified to carry out this study. The UK HD genotypes were made available by SAC as part of the UK Ruminant Genetic Improvement Network (funded by Department of Environment, Food and Rural Affairs, UK)

References

Liu, Z., Seefried, F.R., Reinhardt, F., Rensing, S., Thaller, G. & Reents , R. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genetic Selection Evolution, 43:*19.

Mrode, R. & Swanson, G.T.L. 2004. Calculating cow and daughter yield deviations and partitioning evaluations under a random regression model. *Livest. Prod Sci. 86,* 253-260.

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Ødegard, J. & Meuwissen, T.H.E. 2009. Persistence of accuracy of genomic-wide breeding values over generations when including a polygenic effect. *Genetic Selection Evolution, 41:*53.