Accuracy of Genomic Predictions Across Breeding Lines of Chickens

C. Andreescu[1], D. Habier[1], R. L. Fernando[1], A. Kranis[2], K. Watson[2], S. Avendano[2] and J.C.M. Dekkers[1]

Introduction

Genomic Selection (GS), introduced by Meuwissen et al. (2001), allows the use of high density marker data to estimate genomic breeding values (GEBVs) on selection candidates without phenotypic information. Marker effects are estimated in a reference population (the training dataset) containing individuals with both marker genotypes and trait phenotype information, and then GEBVs of any genotyped individual (the validation dataset) can be calculated using these estimates. Simulation and real data studies have shown that this method can predict GEBVs with high accuracy, with lower costs and shorter generation intervals than traditional methods, as long as training datasets are sufficiently large. Typically the reference population is a single breed population (or single line), and individuals in the validation dataset come from the same breed, though possibly from a different generation. A large reference population from the validation dataset breed is however not always available, and combining breeds in a training dataset may be advantageous. On the other hand, the predictive ability of GS relies on markers accounting for the effect on the trait of QTLs in linkage disequilibrium (LD) with them, and the accuracy of prediction may be reduced because of inconsistent LD across breeds. Simulation studies have shown that using a multi-breed reference population can improve the accuracy of GEBVs (de Roos et al. 2009, Ibanez-Escriche et al. 2009), especially if breeds are closely related or marker density is very high, suggesting that the increase in training dataset size more than compensates for the loss of homogeneity. Multi-breed training sets may also allow for higher accuracies because the markers selected in the GS model are likely to be in LD with a QTL across breeds and hence be more tightly linked to the QTL (de Roos et al. 2009).

Against this background, the objective of this study was to evaluate the predictive ability of Genomic Selection across multiple populations in a real dataset of broiler chicken lines.

Material and methods

Datasets. SNP genotype, phenotype and pedigree data from ten broiler chicken breeding lines from one major global breeding company (Aviagen Ltd.), coded Line 1 to Line 10, were used. The lines evaluated were representative of lines used in a commercial broiler

[1] Department of Animal Science, Iowa State University, Ames, Iowa 50010
[2] Aviagen Ltd., Newbridge, Edinburgh, EH28 8SZ, UK

breeding program and were closed populations which have undergone multiple generations of selection. Selection pressure was different for each line to the extent that considerable differences in key traits now exist. A study of previous generations of the same lines (Andreescu et al 2007) found that LD in these lines extends over shorter distances than reported in other livestock species but was consistent between lines at short distances, with correlations of LD measured by r greater than 0.9 for closely related lines. A total of 154 to 201 individuals from each line that were representative of males used for breeding for several consecutive generations were used. Phenotypes were sire progeny means (over 5 to 836 progeny) adjusted for systematic environmental effects and the estimated breeding value of the dam for a body weight trait of moderate heritability. The correlation of sire means with BLUP estimated breeding values exceeded 0.95 in each line. Genotype data for 12046 fairly equally distributed SNPs were available for each sire. All SNPs were used for analysis, although several are fixed in one or more lines but we don't expect this to affect the analysis. Because the number of genotyped individuals per line was limited, lines were pooled to create cross-validation datasets. In datasets cv1 to cv10, the training dataset was composed of 9 lines and the remaining line was used for validation. In datasets cvmix1, cvmix2, cvmix3 and cvmix4 each line was present in both the training and the validation dataset. Datasets cvmix1 and cvmix2, were created by randomly assigning individuals from each line to either the training or validation datasets, in a proportion of about 17:1. For datasets cvmix3 and cvmix4, individuals were assigned (using a program provided by David Habier) such that the relationship between individuals in training versus validation datasets was low.

Statistical analyses. Two related models were used for analysis.
The first model is: $y_{ij} = \mu_i + \Sigma b_k * g_{ijk} + e_{ij}$
where y_{ij} is the adjusted progeny mean of sire j from line i, μ_i is a line-specific mean, b_k is the effect of marker k, g_{ijk} the genotype of sire j in line i at marker k (0,1,2 or two times the allele frequency in the line when missing), and e_{ij} are residuals distributed $N(0,V/n_{ij})$, where V is the residual variance and n_{ij} the number of progeny. The second model also included in a polygenic term, $p_{ij,}$, assumed distributed $N(0, \mathbf{A}V_p)$ where **A** is the relationship matrix based on a 3-generation pedigree and V_p is the polygenic variance.

BayesC, a variation of the BayesB method introduced by Meuwissen et al. (2001) was used to fit these models. In BayesC, effects for markers included in the model were assumed to come from a normal distribution with common variance Va. This method gives results similar to BayesB but converges faster. The Gibbs sampler was run for 100,000 iterations of which 50,000 were burn-in. Priors for V, V_p, and V_a were inverted-X^2 with parameters chosen such that means were 200, 10, and $Vg/\pi\underline{2pq}$, where Vg is 10 and $\underline{2pq}$ the average value of 2pq over all segregating markers. Other choices for scale parameters did not affect results significantly. Small values were chosen for degrees of freedom for both the environmental (df=10) and the genetic variance (df=4). Polygenic effects were sampled using the method of Garcia-Cortes and Sorensen (1996). Parameter π was set equal to 0.90. Other values of π were also used but did not affect the predictive ability of the model unless extremely high. Results of the BayesC analyses were compared with those from three other methods: BayesB of Meuwissen et al. (2001) with π=0.90, GBLUP, and BayesCπ. GBLUP is BayesC with π = 0. BayesCπ is a variation of BayesC where π is treated as unknown with uniform(0,1) prior. Because the latter model converged much slower, 200,000 iterations with 150,000 burn-in were used.

The GEBVs for individuals in the validation data were computed using estimated marker and polygenic effects from models 1 and 2. Predictive ability was computed as the correlation of GEBVs with the adjusted progeny mean phenotypes in the validation dataset.

Results

The prediction of BVs for a line using a training dataset made up of the other 9 lines (datasets cv1 to cv10), resulted in correlations between GEBVs and phenotypes in the validation data sets ranging from -0.03 to 0.26 when BayesC was used (Table 1). The other three methods resulted in a similar range of correlations, although there were sizeable differences between methods for some datasets. The posterior mean of π from BayesCπ was low, ranging from 0.24 to 0.39. The posterior mean was independent of the initial value of π but the posterior distribution was very diffuse, suggesting that the accuracy of the BayesCπ method may be underestimated due to a lack of convergence.

Table 1: Correlations between GEBVs and progeny means in validation datasets separated by line and estimates of the proportion of non-zero SNP effects (π) from BayesCπ.

Dataset	cv1	cv2	cv3	cv4	cv5	cv6	cv7	cv8	cv9	cv10
BayesC	0.13	0.05	0.18	0.08	-0.03	0.06	0.00	0.06	0.16	0.20
BayesB	0.10	0.01	0.17	0.07	0.06	0.12	0.12	0.01	0.25	0.18
BayesCπ	0.09	0.06	0.14	0.08	0.00	0.07	0.01	0.10	0.10	0.21
GBLUP	0.09	0.06	0.13	0.08	0.01	0.08	0.00	0.11	0.09	0.21
π	0.26	0.31	0.39	0.26	0.34	0.35	0.35	0.35	0.29	0.24

Validation correlations were much higher for some datasets than others but this was not consistently associated with the degree of relatedness of the validation line to one or more lines in the training dataset. However, in datasets cv1 to cv10, the relationship between individuals in training and validations datasets was always low because lines were separated by at least several generations. When individuals from every line were present in both the training and the validation dataset (cvmix 1 to cvmix4), validation correlations were substantially greater (Table 2). As expected, this increase was more marked for datasets cvmix1 and cvmix2, where validation and training set individuals could be more closely related, than for datasets cvmix3 and cvmix4 where these relationships were constrained to be less than .5.

Table 2: Correlations between GEBVs and progeny means in mixed validation datasets.

Dataset	BayesC	BayesB	BayesCπ	GBLUP
cvmix1	0.63	0.58	0.65	0.66
cvmix2	0.61	0.55	0.62	0.62
cvmix3	0.33	0.41	0.31	0.31
cvmix4	0.43	0.47	0.45	0.46

Discussion

Correlations between GEBVs and progeny means were low when validation individuals came from a line that was not represented in the training dataset, although LD at short distances was fairly consistent across lines. The considerable differences in selection pressure and environments between lines may be partly responsible for the lack of accuracy, as well as the limited marker density used in this study. The GEBVs of individuals can be predicted with high accuracy when the validation lines were represented in the training dataset, even when the number of individuals of the same line in the training dataset was low. These results are consistent with studies in cattle (Hayes at al. 2009) that have found that using a training dataset of one breed to predict GEBVs of individuals from a different breed results in correlations close to 0, but combining breeds in the training dataset increased accuracy. The dependence of accuracy of GEBV with level of relatedness of validation and training individuals suggests a large proportion of the predictive ability of GS comes from its prediction of additive relationships among individuals. This pattern was replicated whether we used BayesB, BayesC, or BayesCπ.

Choice of priors had little impact on the accuracy of GEBVs. This was true for both BayesB and BayesC. The difference in accuracies between BayesC, BayesB, BayesCπ and GBLUP were negligible, although the accuracy of BayesCπ may be underestimated due to lack of convergence.

Conclusions

Genomic Selection methods have low accuracy when predicting GEBVs of individuals from lines not represented in the training data set. Including even a small number of individuals from the validation lines in the training data can increase accuracies appreciably, especially when these individuals are highly related to validation individuals. The results are independent of the Genomic Selection algorithm used.

Acknowledgements

Financial support was provided by Aviagen Ltd and USDA-NRI grant # USDA-NRI-2007-35205-17862.

References

Andreescu, A., Avendano, S., Brown, S.R. *et al.* (2007). *Genetics*, 177:2161–2169.

De Roos, A.P.W., Hayes, B.J., and Goddard, M.E. (2009). *Genetics*, 183:1545–1553.

Garcia-Cortes, L.A., and Sorensen, D. (1996). *Genet. Sel. Evol.*, 28:121-126.

Ibanez-Escriche, N., Fernando, R.L., Toosi, A. *et al.*(2009). *Genet. Sel. Evol.*, 41:12.

Hayes, B.J., Bowman, P.J., Chamberlain, A.C. *et al.* (2009*). Genet. Sel. Evol.*, 41:51.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). *Genetics*, 157:1819–1829.