Accuracy of Genomic Breeding Values Predicted Within and Across Breeds in Pig Populations

***A. M. Hidalgo*[*,†], J. W. M. Bastiaansen[*], M. S. Lopes[*,‡], B. Harlizius[‡], M. A. M. Groenen[*] and D. J. de Koning[†].**
[*]Animal Breeding and Genomics Centre, Wageningen University, the Netherlands, [†]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden, [‡]TOPIGS Research Center IPG BV, the Netherlands

ABSTRACT: Data were available from animals of two commercial dam populations: 1070 Dutch Landrace-based and 1389 Large White-based. Four traits were analyzed: age at first insemination (AFI), total number of piglets born (TNB), litter birth weight (LBW), and litter variation (LVR). Deregressed estimated breeding values (DEBV) were used as the response variable. The accuracy of genomic estimated breeding values (GEBV) was the correlation between GEBV and DEBV. Scenarios were divided into two groups: within- and across-breed prediction. Accuracies for within-breed prediction for the four traits ranged from 0.14 to 0.70, indicating a modest to good predictive ability. Accuracies for across-breed prediction for AFI and TNB were not significantly different from zero, whereas accuracies for LBW and LVR ranged from 0.16 to 0.26. These results suggest that the accuracies are trait-dependent, but in general across-breed prediction was not effective in this data set.
Keywords: genomic selection; pigs; reproduction traits

Introduction

Genomic selection (GS) is a very useful tool that capitalizes on the linkage disequilibrium (LD) between markers and the quantitative trait loci (QTL) (Meuwissen et al. (2001)). A training population is geno- and phenotyped and used to estimate the marker effects that will be used to compute the genomic breeding value (GEBV) of other genotyped animals in the prediction population. GS implementation depends on having accurate GEBV. GEBV accuracy depends on four factors: the level of LD between the single nucleotide polymorphism (SNP) and the QTL; the distribution of the QTL effects; the size of the training data set; and the heritability (h^2) of the trait (Hayes et al. (2009a)). Increasing the training data set by adding animals from a different breed could potentially improve the accuracies, but only if linkage phase between markers and QTL as well as the genetic architecture of the trait are similar between the populations. Hence, the objective of our study was to assess the GEBV accuracy when using within- and across-breed training populations.

Materials and Methods

Data. Phenotypes and genotypes were available from animals of two pig populations born between 2006 and 2012: 1070 Dutch Landrace-based and 1389 Large White-based. Four female reproduction traits were analyzed: age at first insemination (AFI), total number of piglets born (TNB), litter birth weight (LBW), and litter variation (LVR). AFI was the age at the second estrus, at which time the first insemination was performed. TNB was the sum of the number of piglets either born alive or stillborn. LBW was the sum of individual birth weights of all piglets born in the same litter. Finally, LVR was the standard deviation of individual birth weight of the piglets from the same litter.

Deregressed estimated breeding values (DEBV) were used as the response variable for each trait under study. The EBVs were deregressed using the methodology proposed by Garrick et al. (2009). The EBV of each animal was obtained from routine genetic evaluation performed by TOPIGS using MiXBLUP (Mulder et al. (2012)) in a multi-trait model. The h^2 used for deregression were the same as those used in the routine genetic evaluation and were estimated using a pedigree-based relationship matrix. The h^2 of the traits were: 0.30 for AFI, 0.11 for TNB, 0.38 for LBW, and 0.14 for LVR.

All individuals were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos et al. (2009)). SNPs with GenCall<0.15, unmapped SNPs and SNPs located on either the X or Y chromosome, according to the Sscrofa10.2 assembly of the reference genome (Groenen et al. (2012)), were excluded. Quality control was performed in all populations simultaneously, which involved excluding SNPs based on call rate (<0.95), minor allele frequency (<0.01), and deviations from Hardy-Weinberg equilibrium (χ^2>600). After quality control, 42,139 SNPs remained out of the initial 64,232 SNPs. Individuals with missing genotype frequency >0.05 were removed as well. Missing genotypes of the remaining animals were imputed using BEAGLE 3.3.2 (Browning and Browning (2007)).

Statistical analyses. SNP effects were estimated using the bigRR package (Shen et al. (2013)) implemented in R (R Development Core Team (2013)). We used the ridge regression - best linear unbiased prediction (GBLUP) model (Meuwissen et al. (2001)) to estimate the SNP effects. Traits were assumed to be affected by additive effects only. To estimate the GEBV accuracy, after estimating the SNP effects from the training data, the GEBV for animal *k* in the prediction data was calculated as:

$$\mathrm{GEBV_k} = \mu + \sum_{i=1}^{s=42139} x_{ik}\,\hat{g}_i$$

where x_{ik} is the genotype and $\hat{g}_i$ is the estimated SNP effect at locus *i*, and *s* is the total number of SNPs. The correlation between GEBV and DEBV of the same animal in the prediction data was computed, resulting in the GEBV accuracy. We compared our realized GEBV accuracy to the expected accuracy according to the formula derived by Daetwyler et al. (2010).

Scenarios. Four scenarios were investigated. These can be divided into two groups according to the training and prediction data sets:

Table 1. GEBV accuracies using GBLUP for scenarios 1-4 for age at first insemination (AFI), total number of piglets born (TNB), litter birth weight (LBW) and litter variation (LVR).

Trait	Heritability	Scenario	Training	Prediction	Accuracy	Std. Error	Expected accuracy[5]		
							50	250	1000
AFI	0.30	1	DL[2]	DL[2]	0.17[1]	0.02[3]	0.93	0.74	0.57
		2	LW[2]	LW[2]	0.14[1]	0.02[3]	0.94	0.79	0.64
		3	DL[2]	LW[2]	-0.05	0.03[4]	0.93	0.75	0.58
		4	LW[2]	DL[2]	-0.01	0.03[4]	0.94	0.79	0.64
TNB	0.11	1	DL[2]	DL[2]	0.24[1]	0.03[3]	0.83	0.56	0.39
		2	LW[2]	LW[2]	0.16[1]	0.02[3]	0.86	0.61	0.45
		3	DL[2]	LW[2]	-0.01	0.03[4]	0.84	0.57	0.40
		4	LW[2]	DL[2]	0.01	0.03[4]	0.87	0.62	0.45
LBW	0.38	1	DL[2]	DL[2]	0.63[1]	0.01[3]	0.94	0.78	0.62
		2	LW[2]	LW[2]	0.70[1]	0.01[3]	0.95	0.82	0.68
		3	DL[2]	LW[2]	0.26	0.03[4]	0.94	0.79	0.63
		4	LW[2]	DL[2]	0.23	0.03[4]	0.96	0.82	0.69
LVR	0.14	1	DL[2]	DL[2]	0.49[1]	0.02[3]	0.86	0.60	0.43
		2	LW[2]	LW[2]	0.46[1]	0.02[3]	0.89	0.65	0.49
		3	DL[2]	LW[2]	0.16	0.03[4]	0.87	0.61	0.44
		4	LW[2]	DL[2]	0.18	0.03[4]	0.89	0.66	0.50

[1] Estimates obtained by 40 randomly generated training-testing evaluation
[2] DL = Dutch Landrace, LW = Large White.
[3] Standard error of the accuracies' mean
[4] Standard error of the correlation coefficient (accuracy)
[5] According to Daetwyler et al. (2010), using N_{QTL}= 50, 250 and 1000

- Scenarios 1-2: training and prediction data were subsets from the same population, i.e. prediction was within breed. These scenarios determined an upper limit of the accuracies and how good the model fits the within-breed prediction.
- Scenarios 3-4: One population was used for training to predict the other population, i.e. prediction was across breeds. These scenarios determined how well across-breed predictions perform.

A total of 40 randomly generated training-testing evaluations were performed for scenarios 1-2. We randomly set aside part of the population as the prediction population (N=50) and used it in a later step to determine the accuracy of prediction. In these cases, the prediction population was sampled 40 times, generating 40 groups in which the accuracy was estimated. We used all available animals from a given population for scenarios 3-4 resulting in one accuracy.

Results and Discussion

For scenarios 1-2, accuracies for within-breed predictions for the four traits ranged from 0.14 to 0.70, indicating a modest to good predictive ability (Table 1). The expected accuracies according to Daetwyler et al. (2010) ranged from 0.39 to 0.95. Scenarios 3-4 that predicted performance in one population from another population performed poorly for AFI and TNB, where accuracies were not significantly different from zero. Accuracies for LBW and LVR ranged from 0.16 to 0.26. The expected accuracies were about three times higher ranging from 0.40 to 0.96 for these four traits (Table 1).

According to the accuracies that were estimated for the four traits across all the scenarios, two groups can be formed: 1) AFI and TNB, 2) LBW and LVR. In general, AFI and TNB had lower accuracies than LBW and LVR. Additionally, for across-breed scenarios, GEBV of AFI and TNB could not be predicted at all, whereas LBW and LVR still demonstrated some predictive ability. In the within-breed prediction, the accuracies of all traits were greater than zero. The h^2 of the traits was expected to be a grouping factor as low-heritable traits have phenotypes that are less determined by genetics, and are envisaged to be less predictable via genomic selection than high-heritable traits (Resende et al. (2012)). Unexpectedly, the h^2 does not separate the groups (AFI=0.30, TNB=0.11, LBW=0.38, LVR=0.14). We propose that the source of genetic influences is a possible reason for the grouping, with the first group, AFI and TNB, being the sow-dependent traits, and the second group, LBW and LVR, the traits where genetics of the piglets makes a contribution to the expression of the traits. Another possible explanation, at least for the differences observed in the across-breed predictions, is that there are more QTLs and SNP markers in the same linkage phase for LBW and LVR than for AFI and TNB, as the former two traits yielded higher accuracies. This grouping of traits in low and moderate accuracies of prediction shows that the GEBV accuracies may be more dependent on the (genetic architecture of) trait, than on the level of h^2. For the utility of data from

other breeds in the application of genomic selection, each trait needs to be studied separately.

Within-breed prediction. For scenarios 1-2, all traits had modest to good predictive ability. Realized accuracies LBW and LVR were within the range of the expected accuracies considering that the trait has between 250 and 1000 QTLs (Table 1). Realized accuracies for AFI and TNB, however, were lower than expected. These discrepancies may occur because there is incomplete LD between SNPs and QTL for these traits, a few genotypes are imputed or read wrongly (VanRaden et al. (2009)), or these traits are more polygenic (i.e. are affected by a larger number of QTL) than LBW and LVR.

In dairy cattle, Luan et al. (2009) found accuracies concordant to ours, when studying seven traits with h^2 ranging from 0.01 to 0.28 resulting in observed accuracies between 0.20 and 0.61. Within-breed prediction was also performed in dairy cattle by Hayes et al. (2009b) and their realized accuracies ranged from 0.49 to 0.71 among the five traits in study, which agreed with their expected accuracies. For LBW and LVR we found realized accuracies that met the expectations, similar to Hayes et al. (2009b), but not for AFI and TNB.

Across-breed prediction. Some predictive ability was found when predicting across breeds for LBW and LVR, whereas for AFI and TNB all the accuracies were null. The accuracies did not reach the expected values, presumably because the prediction was across breeds, while the expectations are based on within-breed prediction.

In the study of Harris et al. (2008), the prediction across Holstein-Friesian and Jersey breeds of cattle was also studied. Predictions were not accurate, ranging from -0.1 to 0.3 over 25 traits. In another study, Hayes et al. (2009b) predicted the GEBV of Jersey animals using a Holstein population as training data and vice-versa, resulting in accuracies ranging from -0.06 to 0.23 for five traits. Our results were very similar, ranging from -0.05 to 0.26.

A simulation study (De Roos et al. (2009)) indicated that across-breed prediction is substantially less accurate than within- or multiple-population prediction. These lower accuracies were due to differences in phase between markers and QTL in the two populations. A marker may be in high LD with QTL in only one population, which results in poor predictions for other populations. In addition, other QTL variants and/or different QTL might segregate in different populations and the genetic background in the other line might change the effect of a specific QTL allele.

Conclusion

Within-breed prediction yielded modest to high accuracies, whereas across-breed prediction yielded null or low accuracies depending on the trait. The possible advantage of increasing the training population with animals from other breeds or using a different breed as training population needs to be evaluated for each data set in a trait-specific way.

Acknowledgements

AMH benefited from a joint grant from the European Commission and TOPIGS Research Center IPG within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". DJK acknowledges support from the Mistra Biotech project, a research program financed by Mistra – Stiftelsen för miljöstrategisk forskning and SLU.

References

Browning S. R. and Browning B. L. (2007). *Am. J. Hum. Genet.* 81:1084-97.

Daetwyler H. D., Pong-Wong R., Villanueva B. et al. (2010). *Genetics.* 185:1021-31.

De Roos A. P. W., Hayes B. J. and Goddard M. E. (2009). *Genetics.* 183:1545-53.

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

Groenen M. A. M., Archibald A. L., Uenishi H. et al. (2012). *Nature.* 491:393-8.

Harris B. L., Johnson D. L. and Spelman R. J. (2008). In *Proc Interbull Meet.* 325-330

Hayes B. J., Bowman P. J., Chamberlain A. J. et al. (2009a). *J. Dairy Sci.* 92:433-43.

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

Luan T., Woolliams J. A., Lien S. et al. (2009). *Genetics.* 183:1119-26.

Meuwissen T. H. E., Hayes B. J. and Goddard M. E. (2001). *Genetics.* 157:1819-19.

Mulder H. A., Lidauer M., Strandén I. et al. (2012). MiXBLUP Manual. *Anim Breed Genomics Centre, Wageningen UR Livest Res*.

R Development Core Team. (2013). http://www.R–project.org/.

Ramos A. M., Crooijmans R. P. M. A., Affara N. A. et al. (2009). *PLoS One.* 4:1-13.

Resende M. F. R., Muñoz P., Resende M. D. V. et al. (2012). *Genetics.* 190:1503-10.

Shen X., Alam M., Fikse F. et al. (2013). *Genetics.* 193:1255-68.

VanRaden P. M., Van Tassell C. P., Wiggans G. R. et al. (2009). *J. Dairy Sci.* 92:16-24.