Prediction of Genomic Breeding Values in a Commercial Pig Population

M.A. Cleveland[*], S. Forni*, D.J. Garrick[†‡], and N. Deeb*

Introduction

Prediction of genomic estimated breeding values (GEBVs) from dense, genome-wide single marker polymorphisms (SNPs) shows promise for more accurate prediction of a pig's breeding value, particularly at a young age when selection decisions are often made. For traits related to female production there is only parental average breeding values with which to compare candidates for selection. An increase in the accuracy of a candidate's estimated breeding value (EBV) will increase the rate of genetic gain, particularly for lowly heritable traits. Training populations of individuals with high accuracy EBVs in place of phenotypes have been used to estimate SNP effects (e.g. in dairy cattle; Van Raden *et al.* 2009) and calculate GEBVs with accuracy exceeding that of parent average. Pig populations generally lack sufficient numbers of animals with EBV accuracies approaching those in dairy training datasets (>0.85), but sampling individuals across generations with only moderately lower accuracies than in dairy cattle may yield GEBVs that are more predictive than EBVs. The objective of this work was to calculate GEBVs for two female reproduction traits in pigs and to evaluate the potential accuracy of these predictions through cross-validation.

Material and methods

Sample selection. Samples were selected (N=3000) from a PIC Landrace-origin pig population for genotyping on the Illumina PorcineSNP60 BeadChip. Candidates for selection included genetic nucleus male and female pigs born between 2000 and 2009. Individuals were selected to be representative of the population with a particular focus on maximizing the accuracy of EBVs for the total number of pigs born in a litter (TB), which are routinely available for all pigs. Sampling multiple members of full-sib families was avoided.

Data. Genotypes were available for 64232 SNPs, where >90% were polymorphic in this dataset. Additional filtering was performed to exclude SNPs based on minor allele frequency (<0.05), X chromosome linkage, individuals with high proportions of missing genotypes (>20%) or extreme values for the Pearson chi-squared test statistic (>499), yielding 46741 informative SNPs. Missing genotypes were filled-in using genotype probabilities from a

[*] Genus/PIC 100 Bluegrass Commons Blvd., Suite 2200, Hendersonville, TN 37075 USA
[†] Department of Animal Science, Iowa State University, Ames, 50011 USA
[‡] Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

segregation analysis and a three generation pedigree (Kerr and Kinghorn 1997), with still unresolved genotypes filled based on genotypic frequencies.

The dependent variables were de-regressed BLUP EBVs (Garrick *et al.* 2009) for TB and percentage of stillborn pigs in a litter (SB). Mean accuracies ($\hat{r}_{II}$) for TB and SB BLUP EBVs were 0.826 and 0.831, respectively. Accuracy distributions are depicted in Figure 1.

Analysis. Effects were estimated in a training dataset for all SNPs using a Bayesian analysis as implemented in GenSel (Fernando and Garrick 2008) and GEBVs were calculated for a validation dataset based on these additive effects. Three cross-validation approaches were used to determine the mean accuracy of GEBVs for TB and SB. In the first approach (Random) 66% or 90% of the samples were randomly assigned to the training set, while remaining samples (or subset of remaining samples) were assigned to the validation set (Table 1). In the second approach (Age) pigs were assigned to two groups based on birthdates to better model selection in practice, where the oldest animals were candidates for the training set (N=2750) and pigs born in the last two years were potential validation samples (N=250). Pigs from each of the two groups were then randomly selected for training (N=2000) or validation (N=100). The third approach (Age-Sub) was identical to the second approach except that training and validation were performed using a subset of 384 SNPs selected on their frequency of inclusion in the model, for each trait, using the so-called BayesC (pi = 0.95; Kizilkaya *et al.* 2010) and a training set consisting only of the older animals. For each approach, the training and validation sets were sampled 100 times and BayesA performed on each sample using a chain length of 11000, with the first 1000 chains discarded. Accuracy of the GEBVs was calculated as the mean correlation between the GEBVs and EBVs in the validation set.

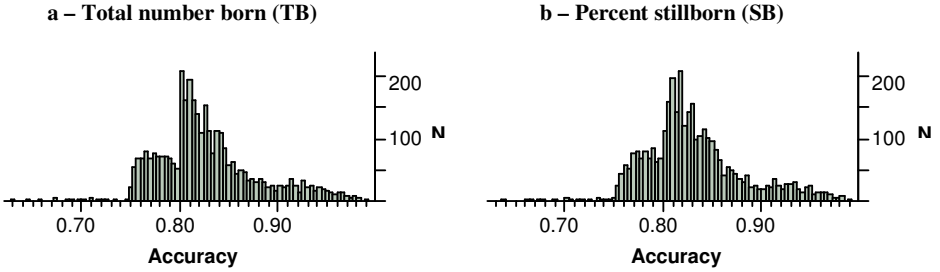

Figure 1: Distribution of accuracy for total number born (a) and percent stillborn (b) estimated breeding values used for genomic training and cross-validation

Results and discussion

Accuracies for GEBVs for TB and SB in cross-validation scenarios are in Table 1. Accuracies were larger when the training set was selected at random and for TB compared to SB in the same scenarios. The random cross-validation using 90% of the samples for training had the highest accuracy for both traits, while reducing the size of the training set to 66% of the samples resulted in a small reduction in accuracy. Cross-validation based on age resulted

in a considerable drop in accuracy for both traits, whereas reducing the number of SNPs used for training (Age-Sub) from 46741 to 384 had only a small impact on the accuracy for TB.

A decrease in accuracy with reduction in training set size was expected as less information is available to estimate SNP effects, but the observed decrease was quite small and suggests that little would have been gained from genotyping the additional samples. The difference in accuracy between the traits should be small considering that the inflated heritabilities of the two traits are equivalent (when using de-regressed EBVs instead of phenotypes), however a difference exists and suggests that the current training size may not be large enough, or SNP coverage dense enough, to appropriately characterize SNP effects for SB. The decrease in accuracy in the cross-validation by age compared to random was expected as the random approach represents an ideal situation where prediction of contemporary individuals is desired and the training and validation sets contain equivalent information (family structure, linkage disequilibrium, etc.). In practice, prediction of GEBVs for individuals in future generations is generally sought and thus there is some information lost between the training set and prediction sets in subsequent generations. The Age scenario represents this situation where a sample of breeding animals in a population are used for training and GEBVs are calculated for descendents of these animals in the most recent generations. In this case (using all SNPs) accuracies were 0.646 and 0.495 for TB and SB, respectively, which are expected to offer improvement over parent average for traits such as these with low heritability (h^2 = 0.16 for TB and SB) and no individual performance data at the time of selection (where males without daughters have no data at all). These values are similar to those reported in dairy cattle for comparable training set sizes over a range of traits (Hayes *et al.* 2009ab; Moser *et al.* 2009; Van Raden, *et al.* 2009).

In many situations it is desirable to select a subset of SNPs for genotyping on a reduced-density panel. The Age-Sub scenario represented the genotyping of selection candidates for 384 SNPs believed to be associated with the single trait of interest. The decrease in accuracy from training on the full SNP set to training on the subset was very small for TB (0.009), indicating that a trait-specific targeted selection of SNPs may be a viable strategy to reduce costs without sacrificing accuracy. The decrease for SB (0.163) was much larger, however, suggesting that the genetic architecture of these traits is different and that SB may be influenced by relatively more genomic regions (genes, QTL, etc.) of smaller effect size than TB, thus 384 SNPs were not sufficient to capture a large proportion of variance in SB. The accuracy decrease was similar to that shown for dairy cattle lifetime net merit when selecting a similar size subset of SNPs with large effect (Weigel *et al.* 2009). An alternative approach for implementing GEBVs for these, and likely other, traits may be needed (e.g., the imputing of high-density genotypes using a low-density approach suggested by Habier *et al.* 2009).

The accuracies of EBVs used for training in this pig population (Figure 1) were generally lower than those used in cattle training sets which consist mostly of sires with large progeny numbers. Likewise, the EBV accuracy of validation animals was such that those EBVs likely deviate from their true value. This reduces the information available to estimate SNP effects and to evaluate the true relationship between the effects and the value of an individual as a breeding animal. The GEBV accuracies reported here for TB and SB (Table 1) then can be viewed as the “worst-case”. Accuracy should improve when predicting GEBVs for as-yet un-

genotyped individuals with increases in EBV accuracy in the training set over time and the appropriate characterization of GEBV accuracy for individual pigs in the validation set.

Table 1: Accuracy of genomic breeding values for total number born (TB) and percent stillborn (SB) in different cross-validation scenarios

Trait	Scenario[a]	N Training	N Validation	Accuracy[b]
TB	Random	2700	300	0.821±0.002
	Random	2000	100	0.801±0.003
	Age	2000	100	0.646±0.005
	Age-Sub	2000	100	0.637±0.005
SB	Random	2700	300	0.683±0.003
	Random	2000	100	0.667±0.006
	Age	2000	100	0.495±0.006
	Age-Sub	2000	100	0.332±0.007

[a]Random: samples were randomly assigned to the training and validation sets; Age: samples were assigned as training or validation based on birth year; Age-Sub: samples were assigned based on birth year where training and validation used only a subset (N=384) of the full SNP panel.

[b]Mean correlation (s.e.) between genomic breeding values and pedigree-based estimated breeding values

Conclusion

Genomic breeding values were calculated for two female reproduction traits in a purebred commercial pig population with accuracies similar to those reported for a range of traits in dairy cattle when using similar-sized training populations. A training dataset in pigs as described here, combined with a strategy for continued high-density genotyping of individuals selected for breeding and a strategy for low-density genotyping selection candidates will enable the implementation of genomic breeding values to increase accuracies of selection at young ages for reproduction in a pig breeding program.

References

Fernando, R.L. and Garrick, D.J. (2008). *http://taurus.ansci.iastate.edu/genesel*. Accessed Feb. 19, 2010.

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

Habier, D., Fernando, R.L. and Dekkers, J.C.M. (2009). *Genet.* 182:343-353.

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

Hayes, B.J., Bowman, P.J., Chamberlain, A.C., *et al.* (2009b). *J. Dairy Sci.* 92:433-443.

Kerr, R.J. and Kinghorn, B.P. (1996). *J. Anim. Breed. Genet.* 113:457-469.

Kizilkaya, K., Fernando, R.L. and Garrick, D.J. (2010). *J. Anim. Sci.* 88:544-551.

Moser, G., Tier, B., Crump, R.E., *et al.* (2009). *Genet. Sel. Evol.* 41:56.

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

Weigel, K.A., de los Campos, G., Gonzalez-Recio, O. (2009). *J. Dairy Sci.* 92:5248-2092.