Accuracy of Genomic Predictions in USA Holstein Cattle from Different Training-testing Designs

M.A. Pérez-Cabal[*†], A.I. Vazquez[‡¥], D. Gianola[‡§], G.J.M. Rosa[‡], and K.A. Weigel[‡]

Introduction

Since Meuwissen *et al.* (2001) proposed genomic selection as an alternative or complementary tool to traditional methods in animal breeding programs, many studies have addressed its effectiveness and how to make it economically viable for the industry because. Cross-validation has been shown to be useful to evaluate predictive ability and it is widely used (Goddard and Hayes, 2007). Ideally, cross-validation must be based on independent partitions of the data, which is often violated in animal breeding because individuals in the training and testing sets can be genetically related. Some studies have already investigated the impact of genetic relationships among animals in the cross-validation design on the accuracy of predictions. Using simulated data Calus and Veerkamp (2007) observed that the influence of information from relatives on accuracy was more important for low heritability traits. Even with a high heritability trait individuals with offspring in the training set had higher accuracies due to the genetic relationships (Habier *et al*., 2007). More recently, studies using real data in different species, such as those by Legarra *et al.* (2008) in mice, Luan *et al.* (2009) in Norwegian Red cattle, and Habier *et al.* (2010) in German Holsteins, have confirmed that accuracy is positively associated to heritability and to the number of related individuals in the reference group. The aim of this study was to analyze different training-testing designs of constant subsets (training and testing) sizes to assess the accuracy of genomic predictions of a high and a low heritability trait from a USA Holstein bull's population.

Material and methods

Data. This study used sire progeny-test PTAs for protein yield **(PY)** and for somatic cell score **(SCS)** obtained from the Animal Improvement Programs Laboratory, at the USDA-ARS Beltsville Agricultural Research Center (Beltsville, MD) and SNP genotypes derived from the Illumina® BovineSNP50 BeadChip (Bovine Functional Genomics Laboratory). The data set included 4,703 genotyped sires (3,305 in the training set and 1,398 in the testing set). Analyses were performed using high-dense SNP genotypes for 32,518 markers after edition, as described by Weigel *et al.* (2009).

[*] Department of Animal Production, Complutense University of Madrid, Avda. Puerta de Hierro, s/n, 28040, Madrid, Spain.

[†] Department of Animal Production, Polytechnic University of Madrid, Ciudad Universitaria s/n, 28040, Madrid, Spain.

[‡] Department of Dairy Science, University of Wisconsin, Madison 53706, USA.

[¥] Biostatistics Department, Section on Statistical Genetics, University of Alabama-Birmingham, USA.

[§] Department of Animal Sciences, University of Wisconsin, Madison 53706, USA.

Testing-training design. Two different ways of partitioning the data were compared: by generations and at random. In the partition by generations (**GENE**) as proposed in VanRaden *et al*. (2009), the models were trained with 3,305 sires born before 1999 using either the PTAs of the 2003 progeny-test evaluation (**GENE_0308**) or the 2008 progeny-test evaluation (**GENE_0808**), and the models were tested in a set with 1,398 sires using the PTAs of the 2008 evaluation. The GENE_0308 design has been used later in other studies (e. g. Weigel *et al*., 2009). The random design (**RAN_0808**) was performed to evaluate the effect of dependencies of the data, such as sires were distributed completely at random in the training and testing sets using 2008 progeny-test PTAs as response variable.

Statistical analyses. Standardized sire PTAs for PY and SCS were regressed on marker covariates in the training set. A Bayesian LASSO (Tibshirani, 1996) model was used to estimate markers effects, implemented via Gibbs sampling (de los Campos *et al.,* 2009) on R version 2.9.0 (R Development Core Team, 2009). The probability model of the Bayesian LASSO was defined in Weigel et al. (2009). The specifications assumed for prior distributions were: degrees of freedom for the inverted-chi-square distribution equals 1; an inverted scale parameter of 0.5; and shape and rate parameters of the gamma distribution equal 1.4. A chain of 70,000 samples was run in each analysis and the first 20,000 samples were discarded as burn-in, with convergence checked by visual inspection of trace plots. Posterior summaries were computed using a thinning rate of 10. The correlation between the predicted and the realized PTAs was used as a measure of the accuracy of the predictions.

Relationship measure. A measure of genetic independence between training and testing sets for each scenario was obtained from the additive relationship matrix. Let $\mathbf{A}$ be the additive relationship matrix, then $\mathbf{A}_{Tr}$ is the **A** sub-matrix including the sires in the training set; $\mathbf{A}_{Ts}$ is the sub-matrix including the sires in the testing set; $\mathbf{A}_{TrTs}$ is the sub-matrix between the sires included in the training and the testing sets. The similarities within and between the training and testing sets for each scenario were measured as the average entries of $\mathbf{A}_{Tr}$, $\mathbf{A}_{Ts}$, and $\mathbf{A}_{TrTs}$, that is a_{Tr}, a_{Ts}, and a_{TrTs}, respectively, such that the lower values of a indicate more genetically different populations.

Results and discussion

Animals in the testing set for the RAN_0808 design had four times (1,239 *versus* 455) more relatives in the training set than the GENE designs (Table 1) mainly because of the offspring, given the requirements imposed in the construction of scenarios GENE_0308 and GENE_0808. However, the differences in a_{TrTs} were small for GENE and RAN_0808 scenarios (0.0238 and 0.0248, respectively) as shown in Table 2. Average relationships within the training set were similar regardless the partition but the individuals within the testing set for the RAN_0808 design were less related than in the GENE designs (0.0254 and 0.0301, respectively).

Table 1: Number of ancestors and offspring present in the training set data for the partitions (training and testing sets) in the generational (GENE) and the random (RAN_0808) designs

	GENE		RAN_0808	
	Tr[1]	Ts	Tr	Ts
Sires	237	121	228	192
Maternal grandsires	218	193	184	160
Paternal grandsires	127	141	114	105
Offspring	2,590	0	1,981	782
TOTAL	3,172	455	2,507	1,239

[1]Tr: Training set (3,305 sires); Ts: Testing set (1,398 sires)

Table 2: Average relationships within and between the training and the testing sets (a_{Tr}, a_{Ts}, and a_{TrTs}, respectively) for the generational (GENE) and the random (RAN_0808) designs

	a_{Tr}	a_{Ts}	a_{TrTs}
GENE	0.0250	0.0301	0.0238
RAN_0808	0.0251	0.0254	0.0248

The accuracy from GENE_0308 and GENE_0808 was the same for both PY and SCS (Table 3). Recall that the only difference between these two scenarios was that in GENE_0808 the 2008 PTA were used for the training set instead of the 2003 PTA. Most of the training sires already had very accurate PTAs in 2003 (not shown) such that adding more information from their sons proven before 2008 did not affect much. As expected, accuracy was larger for PY than for SCS in agreement with previous studies showing that low heritability traits need more information to achieve the same level of accuracy than traits with high heritability or reliability (Calus and Veerkamp, 2007). Predictive accuracy was always higher for the RAN_0808 design regardless of the trait because bulls in the validation set had more information from close relatives in the training set and average relationship between reference and validation groups was higher, as Habier *et al.* (2010) found in German Holstein. The increase in accuracy for sires which had offspring among the relatives in the training set for the RAN_0808 was 13% for SCS and 9% for PY, as reported by Habier *et al.* (2007). Then, increasing the genetic relationships among individuals in the training and testing sets leads to higher accuracy of genomic breeding values, especially for low heritability traits.

Table 3: Accuracy for protein yield (PY) and somatic cell count (SCS) from the generational design using 2003 PTA for the training set and 2008 PTA for the testing set (GENE_0308), the generational design using 2008 PTA for both the training and the testing sets (GENE_0808), and the random design (RAN_0808)

			RAN_0808		
	GENE_0308 (n[1] = 1,398)	GENE_0808 (n = 1,398)	Total (n = 1,398)	Sires with offspring in training set (n = 129)	Sires without offspring in training set (n = 1,269)
PY	0.71	0.71	0.82	0.89	0.81
SCS	0.67	0.67	0.69	0.79	0.68

[1] n: Number of bulls used to compute accuracy.

Conclusion

This study evaluated the accuracy of genomic predictions for a high and a low heritability trait with real data for three designs of training-testing sets partitions depending on the level of relationships between the reference and the validation sets. The different partitions resulted in different prediction abilities, so this should be taken into account when planning a study. If the aim is to predict the genetic potential of animals at early ages using the observed performance of relatives from previous generations, the GENE partitions would be the most appropriate. However, sometimes it could be of interest to predict the performance of animals using information from contemporaries (either relatives or not). Then the RAN_0808 design would be helpful given that markers capture genetic relationships and even apparently unrelated animals can provide genetic information.

References

Calus, M.P.L., and Veerkamp, R.F. (2007). *J. Anim. Breed. Genet.*, 124: 362-368.

de los Campos, G., Naya, H., Gianola, D. *et al.* (2009). *Genetics,* 182: 375-385.

Goddard, M.E., and Hayes, B.J. (2007). *J. Anim. Breed. Genet.,* 124: 323-330.

Habier, D., Fernando, R.L., and Dekkers, J.C.M. (2007). *Genetics*, 177: 2389-2397.

Habier, D. Tetens, J., Seefried, F.R. *et al.* (2010). *Gen. Sel. Evol.* 42: 5, doi:10.1186/1297-9686-42-5.

Legarra, A., Robert-Granié, C., Manfredi, E. *et al.* (2008). *Genetics,* 180: 611-618.

Luan, T., Woolliams, J.A., Lien, S. *et al.,* (2009). *Genetics,* 183: 1119-1126.

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

R Development Core Team. (2009). R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Tibshirani, R. (1996). J. *R. Stat. Soc. Series B,* 58: 267-288.

VanRaden, 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., González-Recio, O., *et al.* (2009). *J. Dairy Sci.,* 92: 5248–5257.