Genomic Evaluations in Spanish Dairy Cattle

J. A. Jiménez-Montero[1*], *O. González-Recio*[2], *R. Alenda*[1]

[1]Departamento de Producción Animal, E.T.S.I. Agrónomos – Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain
[2] Departamento de Mejora Genética Animal, INIA, Crta. La Coruña km. 7.5, 28040 Madrid, Spain

ABSTRACT

The aim of this study was to analyze the recorded genotypes and genomic evaluation methodologies for the Spanish Holstein population as an initial step toward the first official national genomic evaluation. This study presents different descriptors of the genomic structure of the genotypes of progeny tested bulls in Spain and compares different genomic evaluation methodologies.

Two different Bayesian linear regressions, Bayes-A and Bayesian-Lasso and a machine learning algorithm, Random-Boosting, were compared with regard to accuracy, bias, mean square error and the regression coefficients. Five different traits that are currently included in the Spanish genetic evaluation were used: milk yield, fat yield, protein yield, fat percentage and udder depth. In total, genotypes for these traits from 1797 accurate progeny tested bulls were included. The training set was composed of bulls born before 2005; 1576 bulls were included for production and 1562 bulls were included for type. The testing sets contained 221 and 196 younger bulls for production and type, respectively. Routine evaluations from January 2009 were used as the dependent variables. Note that bulls in the test set did not have progeny test proofs at that time. The December 2011 progeny proofs were used to compare the predicted response of the sires in the test set.

Genomic evaluations were more accurate than the traditional pedigree index at predicting future progeny proofs for young bulls. The increment of Pearson correlation between observed and predicted response depended on the trait and ranged between 0.093 and 0.389. The different methodologies implemented showed similar results. Results averaged across traits showed that Bayesian-Lasso had the highest accuracy showing an increment between 0.010 and 0.001 points in Pearson correlation regarding Bayes-A and Random-Boosting. Bayes-A showed the lowest bias (0.006 and 0.057 s.d. units below Bayesian-Lasso and Random-Boosting estimates). However, the Random-Boosting algorithm predicted the genomic merit of individuals with mean square error estimates 4.03% and 4.29% lower than Bayesian-Lasso and Bayes-A respectively, and their regression coefficients on the adjusted progeny proofs were closer to unity.

The observed predicted ability obtained with these methods was within the range of values expected for a population of a similar size. These methods and the described reference population are a good start point for the implementation of genome-assisted evaluations in the Spanish dairy cattle.

Keywords: genome-assisted evaluation, machine learning, dairy cattle, predictive ability

[*] Corresponding author: joseantonio.jimenez.montero@upm.es

INTRODUCTION

Over the last decade the Spanish breeding program has provided competitive bulls for the national and international markets due to a strong milk-recording scheme. Special care has been taken in recording morphologic traits. GS has revolutionized dairy cattle breeding since 2009. Taking advantage of this technology is necessary to maintain the program's viability.

Different approaches are currently used for estimating genomic values. It is important to evaluate the performance of diverse methodologies and to identify the methodology that has a higher predictive accuracy for routine Genomic Selection (**GS**) evaluations in a given population. One of the alternatives for dealing with these situations is machine learning methods (Long et al., 2007). Machine learning methods usually compare favorably to Bayesian regression (e.g., Moser et al., 2009; González-Recio and Forni, 2011). These non-parametric methods can be implemented in both regressions on markers (e.g., Boosting) and/or building a (co)variance structures such as RKHS (Gianola and van Kaam. 2008). The boosting algorithm is one of the most appropriate machine learning methods for dealing with genomic-assisted evaluation problems (Ogutu et al., 2011) and provide higher accuracies and lower biases than other methods (González-Recio et al., 2010). A more efficient estimation of DGVs in dairy cattle can be obtained through some modifications to the algorithm; this modified algorithm, called Random-Boosting (**R-Boost**) was described by González-Recio (Personal communication)

The aim of this study was to validate the recorded data and genomic selection methodologies for the Spanish Holstein population as an initial step towards the first official national genomic evaluation.

MATERIAL AND METHODS

Genotyped Bulls

A total of 1797 progeny-tested sires were genotyped. Using the BovineSNP50.v1 and v2 assays (Illumina, San Diego, CA). The genotyping was performed in two different labs.

Phenotypes

The January 2009 progeny proofs were used as a dependent variable, resulting in 1797 bulls for production traits and 1758 for type. The production and type data were collected between 1980 and 2008. Production data existed for 1,414,347 daughters of genotyped bulls, while 969,567 of them had also type data available.

Five traits were examined, including milk yield (**MY**), fat yield (**FY**), protein yield (**PY**), fat percentage (**FP**) and udder depth (**UD**).

SNP Editing

SNPs with a greater than 5% incidence of missing genotypes across individuals and SNPs with a minimum allele frequency (**MAF**) lower than 5% were discarded, leaving only 39,714 SNPs for testing.

Training and Validation Data Sets

Training and validation data sets were generated based on the bull's year of birth. A total of 1576 bulls with proofs in January 2009 and born before 2005 were used for the production training set and 1562 bulls were used for the type training set. Bulls born between 2005 and 2007 were used as the validation set, 221 for production traits and 196 for UD. The validation bulls had their proof in December 2011 with at least 60 daughters.

Genomic Evaluation Model

The general structure for the models in linear form is

$$\mathbf{y} = \mu\mathbf{1}_n + \sum_j \mathbf{X}_j\mathbf{g}_j + \mathbf{e}$$,

where **y** is the vector of phenotypic records, μ is the overall mean, $\mathbf{1}_\mathrm{n}$ is a vector of *n* ones, $\sum_\mathrm{j}$ is a summation over all markers, $\mathbf{g}_\mathrm{j}$ is the vector of the effects for each marker, $\mathbf{X}_j$ is a design matrix of genotype codes and **e** is a vector of residuals.

Method 1 Bayes-A

Bayes-A was introduced by Meuwissen et al. (2001). The Gibbs sampler was run for 10,000 cycles with the first 1000 cycles of burn-in discarded.

Method 2 Bayesian-Lasso

The Bayesian counterpart of the LASSO model (Park and Casella 2008; de los Campos et al. 2009) was also used to estimate the SNP coefficients in the training population. A single chain of Gibbs sampling was run using 25,000 iterations and a burn-in period of 15,000.

Method 3 Random-Boosting

The boosting algorithm is a machine learning technique that combines different predictors and some shrinkage factor (Friedman, 2001). Boosting iteratively adds basis functions such that each addition further reduces the selected loss function (Hastie et al., 2009). In this study, the ordinary least square estimation was chosen as basis function and was successively applied to the residuals of the previous estimation in a sequential manner. The MSE of the prediction was used as the loss function to minimize. R-Boost propose to sample mtry covariates at random out of the p SNPs at each iteration, and select the SNP among the mtry that minimizes the given loss function.

The R-Boost algorithm would flow as follows:

(Initialization): Given data $\Psi = (\mathbf{y}, \mathbf{X})$, let the prediction of phenotypes be $\hat{F}_0 = \hat{\mu}$.

Then, for *m* in {1 to *M*}, with *M* being large proceed as:

Step 1. Draw *mtry* out of *p* covariates from the original training set to construct a reduced training covariate matrix $\Psi^{(b)} = (\mathbf{y}, \mathbf{X}_{mtry})$ to train the algorithm in iteration *m*.

Step 2. Calculate the loss function $L\left(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i; \mathbf{x}_i, mtry_m)\right)$ for all mtry SNPs and select that minimizing $\sum_{i=1}^{n} L\left(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i; \mathbf{x}_i, mtry_m)\right)$ in the tuning set at iteration *m*, with $h(y_i; \mathbf{x}_i, mtry_m)$ being the prediction of the observation *i* in the tuning set using the learned parameters or coefficients of $h(\cdot)$ on the SNP $mtry_m$. These parameters or coefficients are learned using the training set as in the original algorithm.

Step 3. Updated predictions at iteration *m* in the form $F_m(\mathbf{x}_i) = F_{m-1}(\mathbf{x}_i) + v \cdot h(y_i; \mathbf{x}_i, mtry_m)$ with *v* being some shrinkage factor, e.g. *v*=0.10.

Step 4. Update the residuals to be used in the next iteration as $y_i = y_i - F_m(\mathbf{x}_i)$.
Repeat steps 1 to 4 a large number of times (*M*).

Following their results, v was set to 0.10 for the production traits and 0.20 for UD and the percentage of SNPs selected at each iteration (*mtry*) was set to 0.05, 0.01, 0.05, 1.00 and 0.05 for MD, FD, PD, FP and UD, respectively.

The main advantage of this approach is that the covariates (SNPs) are randomly sampled to compute the loss function, thereby decreasing the computation time while maintaining similar or better predictive ability than the original Gradient Boosting.

Criterion for Comparisons

January 2009 progeny proofs and the genotypes from progeny tested bulls in the training set were used to estimate marker coefficients. DGVs were then predicted for sires in the testing set. The accuracy of the genomic evaluation was computed as the Pearson correlation between the predicted DGV and the December 2011 progeny proofs. The pedigree index for sires in the testing set was used as benchmark. It was calculated as 50% of the sires EBV, 25% of the maternal grand sires EBV and 12.5% of the maternal great-grand sires EBV.

The average difference between the 2011 proofs and the predicted DGVs in the testing population provided a measure of the bias in the genomic predictions; this bias estimate was standardized. The regression coefficients of the realized December 2011 proofs on the estimated DGVs were also calculated because this parameter is commonly used as a measurement of the prediction bias in genomic evaluations (Mäntysaari et al., 2010). Finally, the MSE of the predictions was also estimated.

RESULTS AND DISCUSSION

Quality Control

After filtering, the distribution of MAF was nearly uniformly distributed with a mean of 0.28. The average distance of adjacent SNPs was 0.06 Mb. The remaining SNPs had a heterozygosity of 0.286%. The linkage disequilibrium, measured as r^2, between adjacent SNPs was 0.24. All of those values were in the range of previously reported values for other Holstein populations (Wiggans et al., 2009; Banos and Coffey, 2010). The descriptors of the genomic structure of the population used in this study showed that the Spanish population is similar to other Holstein dairy cattle populations, as expected. Based on this similarity, genomic evaluations of genotyped animals for recorded traits included in the milk recording scheme should be feasible.

Accuracy

The accuracy, standardized prediction bias, regression coefficients and MSE of the estimated DGVs were calculated for each approach (Bayes-A, B-Lasso and R-Boost) in the testing set. The results are shown in Table 1. Traditional PI accuracies ranged from 0.386 to 0.460. The predicted DGVs showed higher accuracies than the PIs for all considered traits, with an average increment of 50%, ranging from 24% for UD to 90% for FP. Similar results have been previously reported in other studies using Holstein populations (VanRaden et al., 2009; Moser et al., 2010). Consequently, the selection of young animals based on genomic values may be preferable to selection based on traditional pedigree information as expected.

Table 1. Accuracy, standardized bias, regression coefficients and mean square error (MSE) for the genomic estimation of different evaluation methodologies indexed for five traits of economic interest in dairy cattle.

Methods[1]	Milk Yield (MY)	Fat Yield (FY)	Protein Yield (PY)	Fat Percentage (FP)	Udder Depth (UD)
Accuracy					
P. Index	0.386	0.411	0.452	0.400	0.460
B-Lasso	**0.590**	**0.655**	**0.583**	0.755	0.562
Bayes-A	0.558	0.651	0.582	0.734	0.570
R-Boost	0.572	0.649	0.545	**0.788**	**0.584**
Standardized Bias					
B-Lasso	0.078	**-0.002**	0.002	**0.000**	**0.000**
Bayes-A	**0.047**	-0.003	**0.001**	**0.000**	**0.000**
R-Boost	0.083	-0.102	0.051	-0.051	-0.049
Regression coefficient					
B-Lasso	0.72	0.89	0.70	1.08	0.61
Bayes-A	0.71	**0.97**	0.77	**0.96**	0.63
R-Boost	**0.84**	1.06	**0.85**	1.18	**0.82**
MSE					
B-Lasso	172328.60	**273.47**	143.38	**0.02**	0.69
Bayes-A	179098.70	275.50	**136.31**	0.03	0.66
R-Boost	**167063.40**	282.36	141.84	**0.02**	**0.58**

In bold: The preferred method within trait and comparison criteria.

[1]Methods: P. Index (Traditional pedigree index), B-Lasso (Bayesian Lasso), Bayes-A, and R-Boost (Random Boosting)

B-Lasso showed higher Pearson correlations for MY (0.590), FY (0.655) and PY (0.583), whereas R-Boost showed larger values for FP (0.788) and UD (0.584). Bayes-A performed intermediately in three cases and had a lower accuracy for MY and FP.

Bias

The DGV of the bulls for the five considered traits showed an average deviation over the realized progeny proofs of 0.031 genetic s.d. across methods, ranging from 0.016 (UD) to 0.069 (MY) s.d. units. Increasing the reference population size is recommended to address this problem because it is supposed to results in less biased GS estimates (Lund et al., 2011).

The standardized bias showed higher differences across methods than the Pearson correlations did. B-Lasso averaged 0.016 genetic s.d. across traits, while Bayes-A and R-Boost presented s.d. of 0.010 and 0.067, respectively. R-Boost showed predictions with larger bias for the five considered traits. B-Lasso achieved lower biased for FY while Bayes-A predictions were less biased for MY, PY. B-Lasso and Bayes resulted almost unbiased for FP and UD.

Regression Coefficients

Regression coefficients close to 1 indicate that the evaluations are successful for predicting the actual magnitude of differences among the animals. The results showed that the regression coefficients for the December 2011 progeny proofs on the genomic predictions ranged from 0.61 for the B-Lasso (UD) to 1.18 for the R-Boost (FP).

With respect to unity, the R-Boost coefficients showed closer estimates for MY (0.84), PY (0.85), UD (0.82) and for the averaged across traits. Bayes-A had low coefficients for FY (0.97), and FP (0.96).

MSE

The three considered methods performed similarly in terms of MSE. Bayes-A showed higher MSE averaged across traits in comparison to B-Lasso (0.38%) and R-Boost (4.29%). In particular for UD, B-Lasso showed 5% and 16% higher MSE than the Bayes-A and R-Boost approaches, respectively. Bayes-A was preferred for PY and B-Lasso for FY and FP. The MSE for R-Boost estimates was lower for MY, UD and FP (equal to B-Lasso) and similar to the preferred for the other traits.

The differences between the methods were more remarkable in terms of estimated MSE than of accuracy. The estimated MSE for MY, and PY were larger for B-Lasso despite the fact that this method was the most accurate. In a previous study, Verbyla et al. (2009) showed similar MSE when Bayesian approaches are compared with Genomic BLUP. Their results showed similar but still larger MSE than results of the Spanish population for FY, PY and FP even when they used Bayes-A. The reason of these differences could be related to their smaller reference population size (1098 progeny tested bulls). Different methodologies, including non-parametric, implemented in this study showed similar predictive ability, although the best method was trait dependant. Further research is needed to determine the relationship between the type of trait and the most suitable method for evaluation. B-Lasso showed to be preferable in terms of Pearson correlation. However, the methods that presented the highest Pearson correlation also showed large biases. This should be considered in the model comparison when deciding the method with better predictive ability. MSE may be a more convenient comparison criterion than Pearson correlation.

With the aim of improving the selection efficiency in both IA centers and commercial farms, GS has been implemented in the Spanish Holstein breeding program. Identification of superior animals is therefore expected to be more accurate and feasible at younger ages than was previously possible. Research will continue on the reported traits and will be extended to the remaining traits included in the Spanish genomic evaluations. The collaboration of the EUROGENOMICS consortium, which includes a reference population with over 20,000 progeny-tested bulls, is expected to substantially increase the accuracy of genomic evaluations and reduce the predicted bias.

ACKNOWLEDGMENTS

The authors acknowledge funds from the project CDTI-P080250866 UPM and INIA-CC10-046, to CONAFE, IA Centers, and Farmers for providing biological samples and phenotypes used in this study and to “Dirección General de Producciones y

Mercados Agrarios" and "Laboratorio Central de Veterinaria del Ministerio de Agricultura, Alimentación y Medio Ambiente" for support on the genotyping process.

REFERENCES

Banos, G., and M. P. Coffey. 2010. Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. J. Dairy Sci. 93:2775–2778.

De los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Posterior predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics 182:375–385.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29:1189–1232.

Gianola, D., J. B. C. H. M. Van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303.

González-Recio, O., K. A. Weigel, D. Gianola, H. Naya, and G. J. M. Rosa. 2010. **L2-**Boosting algorithm applied to high-dimensional problems in genomic selection. Genet.Res. (Camb). 92(3):227-37.

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet. Sel. Evol. 43:7.

Hastie, T. J., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. 2nd ed. Springer. New York, NY.

Long, N., D. Gianola, G. J. M. Rosa, S. Weigel, and S. Avendano. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality inbroilers. J. Anim. Breed. Genet. 124:377–389.

Lund, M. S., S. P. W. de Ross, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet. Sel.Evol.43:43.

Mäntysaari, E., Z. Liu, and P. M. VanRaden. 2010. Interbull Validation Test for Genomic Evaluations InterbullBull. 41:17-22.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41:56.

Moser, G., M. S. Khatkar, B. Hayes, andH. W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol.42:37.

Ogutu, J. O., H. P. Piepho, and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings 5(Suppl 3):S11.

Park, T., and G. Casella. 2008. The Bayesian Lasso. J. Amer. Stat. Soc. 103:681–686.

VanRaden, P. M., C. P. Van Tassell, G. R.Wiggans,T. S. G. Sontegard, R. D. Schnabel, J. F. Taylor, and F. S.Schenkel. 2009. Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci.92:16–24.

Verbyla, K. L., B. J. Hayes, P. J. Bowman and M. E. Goddard. 2009. Accuracy of genomic selection using stochastic variable selection in Australian Holstein Friesian dairy cattle. Genet.Res. (Camb). 91:307-311.

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431–3436.