Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Evaluation of the use of a Meta-population for Genomic Selection in Spanish Beef Cattle Breeds

***E.F. Mouresan*[1], S. Munilla[1,2], C. Diaz[3], A. González Rodríguez[1], J. Piedrafita[4], C. Avilés[5], J. A. Baro[6], C. Moreno[1], and L. Varona[1].**
[1]Universidad de Zaragoza, Zaragoza, Spain, [2]Universidad de Buenos Aires, Buenos Aires, Argentina,
[3]INIA, Madrid, Spain, [4]Universitat Autònoma De Barcelona, Bellatera (Barcelona), Spain,
[5]Universidad de Córdoba, Córdoba, Spain, [6]Universidad de Valladolid, Palencia, Spain,

ABSTRACT: A total of 116 triplets from 5 Spanish beef cattle populations were genotyped using the BovineHD BeadChip. After the quality control, the phases of the parental chromosomes were established. From them, a base population for each breed was defined and 3 generations of 500, 1000 and 1000 individuals were simulated. Phenotypes and true genomic breeding values for a trait with heritability 0.3 were simulated. Purebred and admixed populations were used as training sets for the genomic evaluation, and purebred populations were used for validation. The within-breed evaluation yielded the highest accuracies (0.759 - 0.735). Moreover, the predictive ability of the admixed ×2 populations ranged between 0.681 and 0.628 for the populations included in the mixture. Further, the predictive ability of the admixed ×5 population ranged from 0.558 to 0.475.
Keywords: beef cattle; genomic selection; multiple populations

Introduction

The advances in the area of molecular genetics have allowed the development of genotyping with SNP chips that provide information throughout the genome. Along with the molecular advances, new statistical methods have been developed with the purpose of predicting the genomic breeding values of candidates to selection (Meuwissen et al. (2001)). The potential application of these methods have been tested through simulation (Meuwissen et al. (2001)) and through cross-validation techniques in different species such as mice (Legarra et al. (2008)), dairy cattle (Luan et al. (2009)), aquaculture (Sonesson and Meuwissen (2009)) and poultry (González-Recio et al. (2009)). Currently, genomic selection is a reality in dairy cattle (Hayes et al. (2009)).

Nevertheless, the implementation of such methods in the beef cattle industry is still questionable. The main limitation is the lack of association studies (Garrick et al. (2009)), due to the limited census of the beef populations, the great variability of the production systems and the limited use of artificial insemination. To overcome these constraints, several authors (De Roos et al. (2009); Toosi et al. (2010); Kizilkaya et al. (2010)) have made efforts to increase the precision of the genomic predictions, in simulation studies, by using phenotypic and genomic information provided by several populations. Their results indicate that the use of a meta-population is more helpful when the populations involved have diverged for a small number of generations, for populations of reduced size, and for traits of low heritability if high density genotypes are available.

The Spanish autochthonous cattle breeds have a *Bos taurus* ancestral origin and it is estimated that they have a recent common origin (Beja-Pereira et al., (2003)). Moreover, the small size of these populations and the limited economic resources available for genotyping suggest that using a meta-population for genomic selection is more appealing than the classic application of within-breed evaluation. Thus, the objective of this study is to evaluate the efficiency of the potential implementation of multi-breed genomic selection in the Spanish beef cattle populations from a meta-population.

Materials and Methods

Animals and sample size. A total of 116 sire/dam/offspring triplets were collected from five Spanish beef cattle populations, including Asturiana de los Valles (AV, n = 25), Avileña – Negra Ibérica (ANI, n = 24), Bruna dels Pirineus (BP, n = 25), Pirenaica (Pi, n = 24) and Retinta (Re, n = 18) breeds. The selected parents were chosen as unrelated as possible.

SNP genotyping and phasing. Genomic DNA was extracted by standard protocols. High density SNP genotyping was performed by using the BovineHD BeadChip (Illumina Inc, USA) designed to genotype 777,962 SNPs, according to the protocol of the manufacturer at a commercial laboratory (Xenética Fontao, Lugo, Spain). The SNPs kept for the study belonged to autosomal chromosomes and were not in repeated positions. Additional requirements were Mendelian error rate < 0.05, SNP and individual call rate > 0.95 and MAF > 0.01. The quality control was made using PLINK software (Purcell et al. (2007)) and retained 706,704 SNPs, covering 2,510,606 kb, with one marker each 3.553 kb on average. The haplotypes of the parental chromosomes were established by means of Beagle software (Browning and Browning, (2009)).

Simulation. The simulation structure tries to mimic the linkage disequilibrium structure of the analyzed populations. Thus, for each breed, we defined a base population from the available paternal haplotypes. Thereinafter, for each population, the 706,704 SNP markers of the individuals of three discrete generations of 500, 1000 and 1000

individuals were simulated by gene-dropping and assuming a map distance of 1cM every Mb. The parents of each generation were selected randomly from the previous generation and ignoring their sex. In order to simulate the causative mutations of a trait, 3% of the SNP markers of each chromosome were randomly selected and they were attributed an additive effect sampled from a Gaussian distribution with zero mean and a standard deviation of one. For every individual, true genomic breeding values (TGBVs) were calculated as the sum of the effects of their genotype. Moreover, phenotypes were simulated for all individuals summing to their TGBV, a trait mean (= 1000) and a residual drawn from a Gaussian distribution with appropriate variance to generate a trait with heritability 0.3.

Genomic evaluation. The genomic evaluation was performed with the software GS3 (Legarra et al., (2012)) with the GBLUP method (Meuwissen et al., (2001)) under the model:

$$y_i = \mu + \sum_{j=1}^{n} \boldsymbol{x_{ij}\alpha_j} + e_i \,,$$

where y_i is the phenotype of the i^{th} individual, μ is the trait´s mean, n is the number of SNP markers, $\boldsymbol{x_{ij}}$ is the vector of genotypes of the i^{th} individual and the j^{th} marker coded as 0, 1 and 2, $\boldsymbol{\alpha_j}$ is the vector of the substitution effects of the markers and $\boldsymbol{e_i}$ is the residual of the i^{th} individual. The SNP markers selected as causative mutations were excluded from the marker panel during the genomic evaluation.

Three different scenarios were considered, depending upon the type of training sets used for the genomic evaluation. The different types of training sets were:

- The 5 purebred populations, with 2500 individuals each.
- 10 admixed populations (admixed ×2) of 2,500 individuals, composed from information of individuals randomly selected from 2 purebred populations each time, with a ratio of 1:1 between them.
- 1 admixed population (admixed ×5) of 2,500 individuals as well, constructed from information of one fifth of the individuals randomly chosen from all breeds.
-

Additionally, the same purebred populations at the time of evaluation and after three generations served as validation sets for all scenarios. The accuracy of the predictions was calculated as the correlation between the estimated genomic breeding values (EGBVs) and the TGBVs. The results shown below are the average of 5 replicates of the simulation.

Results and Discussion

Single-breed evaluation. In the first scenario, the SNP marker effects were estimated within each breed. Then, they were used to obtain the predictive ability within and across breeds. Table 1 shows the results of the accuracies obtained. Within-breed accuracies were the highest, ranging from 0.735 (BP) to 0.759 (Pi) and they were very close to the results of other simulation studies (Toosi et al. (2010)). As pointed out by these authors, within-breed ability of prediction is the highest. On the other hand, the across-breed accuracies resulted very low, with the highest value obtained when training in Pi to predict over AV (0.150), and the lowest when training in Re to predict over AV (0.056). These results confirmed the postulate of Harris et al. (2008) indicating that there is evidence that training in one population and validating in another is not effective. However, it is remarkable that all the average estimates are positive and the results are coherent with the studies of persistence of LD phase by Cañas et al. (2014) in the same populations.

Table 1. Accuracies (s.e) of the predictions within and across purebred populations.

Val* / Tr*	AV	ANI	BP	Pi	Re
AV	0.758	0.110	0.097	0.114	0.056
	(0.014)	(0.031)	(0.021)	(0.032)	(0.013)
ANI	0.089	0.754	0.105	0.084	0.125
	(0.028)	(0.004)	(0.025)	(0.035)	(0.029)
BP	0.102	0.136	0.735	0.146	0.131
	(0.032)	(0.038)	(0.010)	(0.014)	(0.014)
Pi	0.150	0.083	0.134	0.759	0.102
	(0.029)	(0.037)	(0.021)	(0.010)	(0.012)
Re	0.104	0.080	0.086	0.089	0.754
	(0.016)	(0.029)	(0.027)	(0.013)	(0.016)

*Val = validation set, Tr = training set

Admixed ×2 populations. The training sets in this scenario were set up by mixing information from two purebred populations with equal proportion of each. All possible combinations were considered which resulted in 10 different admixed populations. Table 2 contains the results of the predictive ability of these populations over the purebred. When the purebred population was included in the mixture of the training set, the accuracies ranged from 0.681 (AV+ANI over AV) to 0.628 (AV+BP over BP). When the purebred population was not included in the mixture, the accuracies obtained ranged from 0.150 (AV+Pi over BP) to 0.081 (AV+BP over Re and AV+Re over BP). These results are somewhat smaller than the ones reported by Toosi et al. (2010) when simulating only one chromosome and slightly higher to the ones presented by Mouresan et al. (2013) in a simulation study with a 20 chromosomes genome but with a much smaller marker density.

Table 2. Accuracies (s.e) of the predictions of 10 admixed ×2 populations over purebred populations.

Val* / Tr*	AV	ANI	BP	Pi	Re
AV+ANI	0.681	0.650	0.114	0.093	0.098
	(0.014)	(0.006)	(0.014)	(0.030)	(0.022)
AV+BP	0.671	0.134	0.628	0.137	0.081
	(0.018)	(0.041)	(0.015)	(0.009)	(0.023)
AV+Pir	0.675	0.090	0.150	0.667	0.104
	(0.020)	(0.018)	(0.013)	(0.017)	(0.009)
AV+Re	0.673	0.114	0.081	0.110	0.657
	(0.013)	(0.018)	(0.023)	(0.016)	(0.021)
ANI+BP	0.091	0.662	0.644	0.129	0.146
	(0.036)	(0.007)	(0.017)	(0.022)	(0.022)
ANI+Pi	0.118	0.663	0.110	0.668	0.125
	(0.037)	(0.008)	(0.029)	(0.011)	(0.017)
ANI+Re	0.085	0.665	0.103	0.083	0.666
	(0.024)	(0.005)	(0.019)	(0.015)	(0.016)
BP+Pi	0.104	0.124	0.630	0.672	0.121
	(0.031)	(0.046)	(0.019)	(0.019)	(0.018)
BP+Re	0.125	0.108	0.641	0.119	0.654
	(0.018)	(0.029)	(0.007)	(0.015)	(0.022)
Pi+Re	0.122	0.099	0.136	0.671	0.677
	(0.021)	(0.034)	(0.024)	(0.013)	(0.014)

*Val = validation set, Tr = training set

Admixed ×5 population. The last training set used for the genomic evaluation was constructed by combining information of 500 individuals from each of the 5 purebred populations. Its predictive ability over the purebred populations is shown in Table 3. The accuracies were between 0.558 (over Pi) and 0.475 (over BP). The results were slightly lower than the ones from the Admixed ×2 populations, being lower for the BP. This population is geographically located in a peripheral region with respect to the other populations and shares just a small common region with one of the other populations (http://feagas.com/index.php/es/razas/bovino).

Table 3. Accuracies (s.e) of the predictions from the admixed ×5 population over purebred populations.

Val* / Tr*	AV	ANI	BP	Pi	Re
×5	0.535	0.534	0.475	0.558	0.537
	(0.016)	(0.014)	(0.022)	(0.025)	(0.015)

*Val = validation set, Tr = training set

Predicting over the next 3 generations. Additionally, 3 generations of 1,000 individuals each, were created for every purebred population and the estimates of the SNP marker effects from the different types of training sets were used to predict their genomic breeding values (results not shown). In the case of single-breed evaluation, the within-breed accuracy decreased 23.5% in the first generation, 35.4% in the second and 41.7% in the third. In the case of admixed ×2 meta-populations, for the populations included in the mixture, the decline was 26.4%, 37% and 43% respectively. Finally, in the case of the admixed ×5 meta-population the decline was of 27.2%, 39% and 44% respectively in all purebred population. As described by Ibáñez-Escriche and Blasco (2011), the decline in accuracy with time, caused by the loss of linkage disequilibrium between markers and genes, requires the re-estimation of the marker-effects every 2-3 generations.

Final Remarks. The results obtained in this study indicate that the use of a meta-population can be beneficial for all populations as it provides reasonable accuracies. However, in this study the selection of the individuals was done randomly and the weights for each population were equal.

Conclusion

The results of this study indicate that the use of a meta-population composed by a small number of individuals from all the available populations can be beneficial for all populations giving reasonable accuracies, which can be improved increasing the size of the meta-population, with a reduced cost of genotyping and data collection in each population. Moreover, further research must be done about the sampling strategy of individuals and the weighting of information from each population.

Aknowledgements

The financial support of Spanish AGL2010-15903 and UE FP7-289592-GENE2FARM grants made possible this research. A. González-Rodríguez acknowledges the financial support given by the fellowship BES-2011-045434.

Literature Cited

Beja-Pereira, A., Alexandrino, P., Bessa, I., et al. (2003). *J. Hered.,* 94:243-250.

Browning, B. L. and Browning, S. R. (2009). *Am. J. Hum. Genet.* 84(2):210-223.

Cañas-Alvarez, J. J., Mouresan, E.F., Díaz, C., et al. (2014). 10th WCGALP (Submitted)

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

Garrick, D. J., Golden B. L. and Benyshek, L. L. (2009). *J. Anim. Sci.* 87:E3-E10.

González-Recio O., Gianola, D., Rosa, G. J. M., et al. (2009). *Genet. Sel. Evol*. 41:3.

Harris, B. L., Johnson, D. L. and Spelman, R. J. (2008). *Proc. ICAR 36th Session,* pp. 325-330.

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

Ibáñez-Escriche, N. and Blasco, A. (2011). *J. Anim. Sci.* 89(3):661-668.

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

Legarra, A., Ricard, A., Filangi, O. (2012). GS3: Genomic Selection, Gibbs Sampling, Gauss Seidel. http://snp.toulouse.inra.fr/~alegarra/

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

Luan, T., Woolliams, J. A., Lien, S., et al. (2009). *Genetics* 183:1119-1226.

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

Mouresan, E. F., González-Rodríguez,A., Moreno, C., et al (2013). *64th Ann. EAAP Meeting.* Nantes, France.

Purcell, S., Neale, B., Todd-Brown, K. et al. (2007). *Am. J. Hum. Genet.* 81:559-575.

Sonesson, A. K. and Meuwissen, T. H. E. (2009). *Genet. Sel. Evol.* 41:37.

Toosi, A., Fernando, R. L. and Dekkers, J. C. M. (2010). *J. Anim. Sci.* 88:32–46.