Use of different marker pre-selection methods based on single SNP regression in the estimation of Genomic-EBVs

Ezequiel Luis Nicolazzi[1], Riccardo Negrini[1], Corrado Dimauro[2]

[1]Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza, Italy

[2]Dipartimento di Scienze Zootecniche, Università di Sassari, Italy

Corresponding author: Ezequiel Luis Nicolazzi. Istituto di Zootecnica, Facoltà di Agraria, Università Cattolica del Sacro Cuore. Via Emilia Parmense 84, 29100 Piacenza, Italy - Tel. +39 0523 599205 - Fax: +39 0523 599276 – Email: ezequielluis.nicolazzi@unicatt.it

ABSTRACT - Two methods of SNPs pre-selection based on single marker regression for the estimation of genomic breeding values (G-EBVs) were compared using simulated data provided by the XII QTL-MAS workshop: i) Bonferroni correction of the significance threshold and ii) Permutation test to obtain the reference distribution of the null hypothesis and identify significant markers at P<0.01 and P<0.001 significance thresholds. From the set of markers significant at P<0.001, random subsets of 50% and 25% markers were extracted, to evaluate the effect of further reducing the number of significant SNPs on G-EBV predictions. The Bonferroni correction method allowed the identification of 595 significant SNPs that gave the best G-EBV accuracies in prediction generations (82.80%). The permutation methods gave slightly lower G-EBV accuracies even if a larger number of SNPs resulted significant (2,053 and 1,352 for 0.01 and 0.001 significance thresholds, respectively). Interestingly, halving or dividing by four the number of SNPs significant at P<0.001 resulted in an only slightly decrease of G-EBV accuracies. The genetic structure of the simulated population with few QTL carrying large effects, might have favoured the Bonferroni method.

Key words: Genomic Selection, SNP pre-selection, Bonferroni correction, Permutation test.

Introduction - The recent availability of high-density SNP panels for the bovine genome boosted fine-mapping QTL studies, association studies with functional traits, and the search for causative mutations. However, the highest expectation is in Genomic Selection (GS), which uses dense marker panels for predicting genomic estimated breeding values (G-EBVs) on young animals before phenotypic information is available (Meuwissen *et al.*, 2001). A major statistical and computational limitation to be solved in GS is the estimation of tens of thousands of marker effects based only on few thousands of phenotypes. The size of available SNP panels (54K in cattle) largely affects the dimension of matrices in the mixed model equations and the required computational resources for data storage and algorithm solving (Legarra and Misztal, 2008). To face these problems, an important issue is whether or not to include all the available SNPs in the predictive model (Gonzalez-Recio *et al.*, 2008). In spite of likely decreasing G-EBV accuracies, SNPs pre-selection will sensibly reduce the number of equations in the model. The choice of a suitable predictive model, able to combine adequate G-EBV accuracies with reasonable computing requirements, is another key issue. In simulated data, Meuwissen *et al.* (2001) using Bayesian MCMC methods obtained values of accuracies ranging from 6 to 11% higher than those obtained using BLUP. However, Bayesian methods require substantially longer computing time compared to BLUP. Moreover, early results on real data indicate that G-EBV accuracies obtained with BLUP are only 2-3% lower than those obtained with Bayesian methods (Harris *et al.*, 2008). Therefore, BLUP predictions based on pre-selected SNPs seem a reasonable compromise between loss of accuracy and computational effort. In this paper we tested two single marker regression based methods to re-

duce the number of equations in the model comparing the variations in G-EBV accuracies.

Material and methods - The simulated data set comprised 5,865 individuals structured in 7 generations. Pedigree relationships and genotypes at 6,000 SNPs evenly distributed across six chromosomes were available for all individuals, whereas phenotypic information was provided for the first 4 generations only. A total of 4,665 individuals from generation 0 to 3 were considered as training animals and 1,200 individuals from generations 4 to 6 as prediction young animals. True breeding values (TBV), calculated by summing QTL effects, were available for all animals. Although the SNP phases were known, all the analyses were performed by single markers because the level of linkage disequilibrium (LD) of the dataset (r^2=0.21 at 0.1 cM distance) greatly reduces the potential advantage of using haplotypes (Hayes *et al.*, 2007).

SNP pre-selection using Bonferroni correction was performed (Bolding, 2006) fixing an empirical threshold of $1.6E^{-6}$ (i.e., 0.01/6000) for the *P* values of the F test.

SNPs pre-selection by Permutation test was performed considering 1,000 iteration (a good compromise between statistical significance and computational time) and fixing two different significance thresholds: 0.01 and 0.001. Furthermore, two subsets comprising 50% and 25% of SNPs significant at 0.001 threshold were randomly assembled, to assess the effect of the number of SNPs on G-EBV accuracies. Random sampling procedure was iterated three times for each subset. Marker effects were estimated with the following mixed linear model:

$$y_{ijk} = \mu + SEX_i + GEN_j + \sum_{k=1}^{m} \mathbf{H}_k \mathbf{b}_k + e_{ijk}$$

where y is the trait value, μ is the overall mean, SEX is the fixed effect of sex (i=1, 2), GEN is the fixed effect of generation (j=0-6), b is a vector of genotype random effects for all m significant SNPs, H is the corresponding design matrix, and e is the random residual. An equal contribution of each locus to the genetic variance was considered (e.g.: σ^2_a*1/m), thus, λ was calculated as $\sigma^2_e/(\sigma^2_a$ /number of m significant SNPs). Moreover, no interaction effect between SNPs was assumed. G-EBVs for training and prediction generations were obtained as:

$$\text{G-EBV}_i = \mu + \sum_{k=1}^{m} \mathbf{h'}_k \hat{\mathbf{b}}_k$$

Variance components were calculated with the MTDFREML package, and accuracies were estimated by calculating the correlation between G-EBVs and TBVs.

Results and conclusions - The additive variance (σ^2_a) of the trait was 1.324 and the residual variance (σ^2_e) was 3.142. The heritability was 0.30.

The polygenic animal model for traditional EBV estimation produced accuracies of 71% for training and 33% for prediction generations.

Bonferroni correction method retained 595 out of 6,000 markers, whereas the permutation approach yielded 2,053 and 1,352 significant SNPs for 0.01 and 0.001 significance thresholds, respectively. All the Bonferroni-selected markers overlapped those selected with permutation test, with the exception of one marker at the 0.001 threshold.

High conservative Bonferroni correction showed its drawback failing to retain markers close to 8 small effect QTLs out of the 44 QTLs embedded in the dataset. However, Bonferroni-selected markers yielded higher accuracies in prediction generations (Table 1).

Conversely, permutation test was able to identify all QTLs but the cost for this sensitivity was a "background noise" - due to the higher number of false positives - that negatively affected G-EBV accuracies. Indeed, SNPs significant at 0.001 threshold performed better than those significant at 0.01 threshold, albeit only 2/3 of the markers were used in G-EBV estimation (81.11% *vs.* 79.37% ac-

curacy in prediction generations). Given the hard computation effort needed to further decrease the significance threshold (e.g., to 1/10,000), a lower number of markers was tested just creating subsets of randomly selected SNPs among those passing the 0.001 threshold. Interestingly, randomly halving the number of SNPs used in the estimation, G-EBV accuracies decreased only 2% on average. Indeed, many of the 1,352 SNPs were located nearby the 44 QTLs and the random selection of marker subsets still tagged all or most QTLs. When decreasing the number of markers fourfold, G-EBV accuracies decreased further (84.97 and 74.26 for training and prediction, respectively).

Table 1. Accuracies obtained with Bonferroni and Permutation methods.

			Bonferroni correction		Permutation test 0.01 threshold		Permutation test 0.001 threshold	
			Training	Prediction	Training	Prediction	Training	Prediction
4 training generations	All SNPs	"	89.00	82.80	89.20	79.37	89.19	81.11
	50% SNPs	%	-	-	-	-	87.79 (0.008)	78.40 (0.001)
	25% SNPs	"	-	-	-	-	84.97 (0.008)	74.26 (0.017)
1 training generation	All SNPs	"	84.85	71.15	83.12	64.55	83.93	68.46
	50% SNPs	%	-	-	-	-	83.25 (0.006)	65.79 (0.023)
	25% SNPs	"	-	-	-	-	81.20 (0.007)	61.67 (0.014)

The accuracy values obtained in this paper, combining pre-selection methods based on single marker regression and BLUP estimation of G-EBV, were lower than those reported in literature for Bayesian methods while higher than those obtained by the polygenic animal model. These results were also comparable with those reported in simulated data with similar marker density and models (Kolbedhari *et al.*, 2007; Muir, 2007).

In traits where few QTLs explain large proportions of genetic variance - as in this simulated data set - Bonferroni correction seems a better pre-selection method compared to Permutation test at 0.001 significance threshold.

The authors wish to thank Prof. N. Macciotta, Prof. P. Ajmone Marsan, and Dr. G. Gaspa for their contribution. Research funded by the Italian Ministry of Agricultural Policies (research project SELMOL).

REFERENCES - Bolding, D.J., 2006. A tutorial on statistical methods for population association studies. Nature Reviews Genetics. 7:782-791. **Gonzalez-Recio,** O., Gianola, D., Long, N., Weigel, K.A., Rosa, G.J.M., Avendano, S., 2008. **Nonparametric methods for incorporating genomic information into genetic** evaluations: an application to mortality in broilers. Genetics. 178:2305-2313. **Harris**, B.L., Johnson, D.L., Spelman, R.J., 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. Proc. 36th ICAR Biennial Session, Niagara Falls, USA, 11-16. **Hayes**, B.J., Chamberlain, A.J., McPartlan, H., Macleod, I., Sethuraman, L., Goddard, M.E., 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res. Camb. 89:215-220. **Kolbedhari**, D., Schaeffer, L.R., Robinson, J.A.B., 2007. Estimation of genome-wide haplotype effects in half-sibs design. J. Anim. Breed. Genet. 124:356-361. **Legarra**, A., Misztal, I., 2008. Technical note: Computing strategies in genome-wide selection. J. Dairy. Sci. 91:360-366. **Meuwissen**, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense markers maps. Genetics. 157:1819-1829. **Muir**, W.M., 2007. Comparison of genomic and traditional BLUP estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342-355.