One-Step vs. Multi-Step Methods For Genomic Prediction In Presence Of Selection

Z.G. Vitezica[*], I. Aguilar[†] and A. Legarra[#*]

Introduction

With the increasing availability of marker information in livestock, genomic selection methods can allow breeders to preselect animals on genotypes early in life. The combination of pedigree, phenotypic and genotypic data will increase genetic progress by decreasing the generational interval and increasing the accuracy of genetic merit estimates.

Most proposed strategies for genomic selection are currently based on multiple-step procedure, like in US dairy cattle evaluation (VanRaden 2008; VanRaden *et al.*, 2009a). Recently, a single-step genetic evaluation was proposed based on the pedigree relationship matrix augmented with genomic information as described by Legarra *et al.* (2009), Misztal *et al*. (2009) and Christensen *et al.* (2010). Evaluations using the single-step procedure were compared to a multiple-step procedure in data from dairy cattle (Aguilar *et al*., 2010). Both methods were comparable in terms of accuracy and bias. However, simplicity (no need for several, possibly error-prone steps to compute DYD's), computational time (only slightly longer than traditional pedigree-based evaluation) and generality (straightforward extension to other models or species) advantages show the potential of single-step procedure to became the standard tool for genomic evaluation.

In dairy cattle, genomic prediction can be obtained by combining traditional genetic evaluation results with genotypic data. Daughter yield deviation (DYD), which is a deregressed variable, can be used to predict breeding values of genotyped young bulls. The use of these pseudo-observations may inflate genetic evaluation accuracy when they are computed from animals with small progeny numbers, as VanRaden *et al.* (2009b) suggested. Also, the fact that most genotyped animals have undergone strong selection is considered an issue. To overcome the problem of DYD, one strategy consists in analysing phenotypic records directly. The objective of this paper is to compare by simulation a 2-step procedure where DYDs computed from full record and pedigree data were used in genomic evaluation, against a single-step procedure using directly phenotype records as observations, in two different selection scenarios.

[*] UMR1289, INRA-ENVT-ENSAT TANDEM, 31 326 Castanet Tolosan, France
[†] INIA, Las Brujas, Uruguay
[#]INRA, UR 631 SAGA, 31326 Castanet Tolosan, France
[*]This work has been partly financed by ANR project AMASGEN. Project partly supported by Toulouse Midi-Pyrénées bioinformatic platform.

Material and methods

Simulations. QMSim (Sargolzaei and Schenkel, 2009) was used to simulate a historical and a recent population structure. In total 10 chromosomes of equal length (100 cM) were simulated in the genome. Bi-allelic markers (10,000) were distributed at random along the chromosomes with equal frequency in the historical population. Potentially, 250 QTLs affect the phenotype; QTL allelic effects were sampled from a Gamma distribution with a shape parameter of 0.4. The mutation rate of the markers (recurrent mutation process) and QTL was assumed to be 2.5×10^{-5} per locus per generation.

First, a base population consisting of 200 males and 2,600 females was generated by mutation and drift over 100 generations (t) in a historical population with an effective population size of 100 (from t = 1 to 95), and gradually expanded to 3000 offspring (t = 100). Then, 10 generations (t = 101 to 110) of selection for a sex-limited trait of 0.30 heritability (*i.e.*, milk yield) and phenotypic variance=1 were simulated. In each generation, 200 males were mated with 2,600 females to produce 2,600 offspring following random (P_Y) or positive assortative (P_{EBV}) designs. Animals were selected based on their phenotype (P_Y) or on estimated breeding values (EBVs) (P_{EBV}). We want to emphasize that in P_{EBV} EBV's were computed in each generation with data cumulated so far. Pedigree information was available for all ten generations (28,800 records), phenotypes were not available for the last generation (13,100 records), only sires of all generations (920) and 260 animals in generation 110 were genotyped. No fixed effects were simulated.

Genetic evaluations. For prediction, the 260 genotyped selection candidates in generation 10 were evaluated using genomic information. EBVs in the selection candidates were obtained under 3 analysis. First, the traditional evaluation with an animal model based on the pedigree relationship matrix (BLUP$_{PED}$) was computed. Second, a 2-step procedure was used where DYDs were computed from a regular genetic evaluation and then used for genomic evaluation (BLUP$_{DYD}$). In BLUP$_{DYD}$, 840 sires were included with DYD information from 14 daughters on average. DYD's were weighted by equivalent daughter contributions. In the third analysis, the full data set was directly used in single-step evaluation (BLUP$_{ONESTEP}$). Both BLUP$_{ONESTEP}$ and BLUP$_{DYD}$ used $\boldsymbol{G}=\boldsymbol{ZZ}'/2\sum_i p_i(1-p_i)$; $\boldsymbol{z_i}$ was coded as $-p_i a_i$ or $1-p_i a_i$ for the first or second allele respectively, where p_i is the allelic frequency of the second allele (Van Raden 2008).

Quality of prediction was checked computing the linear regression coefficient (b, measure of bias) and the coefficient of determination (R^2, measure of accuracy) between the TBV and the EBV of young animals in generation 10. Differences in the average EBV of the candidates were checked as well (measure of the ability to estimate the genetic trend). Results are the average of 20 replicates for each population.

Results and discussion

Genetic trend. Table 1 shows the means for the three prediction methods when the animals were selected by their own phenotype (P_Y) or by EBV computed by an animal model-BLUP (P_{EBV}). Selection P_{EBV} is quite strong. As expected, even with selection, the traditional evaluation based on pedigree relationship matrix predicts correctly the average EBV. The other methods underestimate average TBV in the last generation. However, single-step genomic predictions reduce the bias of the estimator up to 37%.

Table 1: Means (SD) of true (TBV) and estimated breeding values (EBV) in the selection candidates, computed with different prediction methods under individual selection (P_Y) and EBV selection (P_{EBV})

Prediction method	Py	PEBV
	TBV = 0.53 (0.03)	TBV = 2.01 (0.15)
BLUP$_{PED}$	0.54 (0.03)	2.05 (0.14)
BLUP$_{DYD}$	0.22 (0.02)	0.70 (0.05)
BLUP$_{ONESTEP}$	0.29 (0.02)	1.41 (0.17)

Inflation. The values of *b* and R^2 are presented in table 2. The 'best method' for prediction of young animals would have *b* close to 1 and R^2 as high as possible. The values of *b* for BLUP$_{DYD}$ show that this pseudo-observation induce an inflation in young animals predictions, even in the case of mass selection (P_y). This bias is considerably reduced and/or non-significant in BLUP$_{PED}$ and BLUP$_{ONESTEP}$, even with strong selection. This inflation is the reason that young bulls appear with an unfair advantage over older progeny-tested bulls (Aguilar *et al.*, 2010).

Accuracy. Concerning accuracy, compared to pedigree-based BLUP$_{PED}$, BLUP$_{ONESTEP}$ increased accuracy by about 0.2 (0.3) in EBV selection (mass selection). Even under strong selection, R^2 in BLUP$_{ONESTEP}$ was as good as those obtained with BLUP$_{DYD}$. Even with an R^2 comparable to BLUP$_{DYD}$ procedure, the BLUP$_{ONESTEP}$ has the advantage of provide an unified framework eliminating all the assumptions applied in multiple-step evaluations.

Table 2: Regression coefficient (*b*) and coefficient of determination (R^2) (with SD) with different prediction methods under individual selection (P_Y) and EBV selection (P_{EBV})

Prediction method	P_y		P_{EBV}	
	R^2	*b*	R^2	*b*
BLUP$_{PED}$	0.20 (0.04)	1.00 (0.01)	0.23 (0.06)	0.89 (0.01)
BLUP$_{DYD}$	0.46 (0.05)	0.70 (0.04)	0.52 (0.06)	0.70 (0.06)
BLUP$_{ONESTEP}$	0.54 (0.04)	0.98 (0.07)	0.47 (0.06)	0.86 (0.11)

It would seem, from tables 1 and 2, that two reasons exist for the differences in biases, inflations and accuracies. The first is that maternal information was not used in BLUP$_{DYD}$,

with loss of information. This could explain the inflation of BLUP$_{DYD}$ in the P_y situation. To avoid this loss of information, Van Raden *et al.* (2009a) included in their multi-step genetic evaluation a pedigree-based evaluation in addition to the pure genomic evaluation, weighted by their respective accuracies. We did not attempt so because of its complexity.

Effect of selection. The second reason, more obvious in table 2, is the genotyping of highly selected sires. Classical theory to model covariance among individuals assumes no selection. Thus, the covariances among TBV of the selected sires are no longer well described by any (genomic or pedigree-based) of its relationship matrices, unless all records used in selection are accounted for (as in BLUP$_{PED}$). Although all records are used in BLUP$_{ONESTEP}$, only genotyped (and mostly selected) animals are included in the genomic relationship matrix **G**. Because breeding values of ungenotyped animals are *a priori* conditioned on breeding values of genotyped animals (Legarra *et al.*, 2009) the inflation is alleviated, but not fully corrected. Aguilar *et al.* (2010) also found that giving different weights to genomic and pedigree-based relationship matrices, inflation was alleviated.

Conclusion

The conditions of this study are slightly favorable to the BLUP$_{DYD}$ procedure (exact computation of DYD and weights, no preferential treatments, no fixed effects like herd). Overall, the one-step genetic evaluation (BLUP$_{ONESTEP}$) is as accurate as the two-step procedure (BLUP$_{DYD}$), while being less biased, less inflated and more general. The BLUP$_{ONESTEP}$ has the smaller value for prediction error variance (PEV) and 73% less mean square error (MSE) than BLUP$_{DYD}$ under strong selection. These results clearly show that genomic and pedigree evaluation by the single-step approach is very promising for animal breeding allowing the pre-selection of animals on genotypes. More studies should check the effect, and possible methodological solutions, of the non-random genotyping (because bulls are strongly selected) on the quality of the prediction.

References

Aguilar, I., Misztal, I., Johnson D.L. *et al.* (2010). *J. Dairy. Sci.*, 93:743–752.

Christensen, O.F., Lund, M.S. (2010). *Gen. Sel. Evol.*, 42:2.

Legarra, A., Aguilar, I., Misztal, I. (2009). *J. Dairy. Sci.*, 92:4656–4663.

Misztal, I., Legarra, A., Aguilar, I. (2009). *J. Dairy. Sci.*, 92:4648–4655.

Sargolzaei, M., Schenkel, F. (2009*). Bioinformatics*, 25:680–681.

VanRaden, P.M. (2008). *J. Dairy. Sci.*, 91:4414–4423.

VanRaden, P.M., Van Tassell, C.P. *et al.* (2009a). *J. Dairy. Sci.*, 92:16–24.

VanRaden, P.M., Tooker, M.E., Cole, J.B. (2009b). *J. Dairy. Sci.*, 92 (E-Suppl. 1):314.