Modelling missing pedigree with metafounders and validating ssGBLUP evaluation for Uruguayan Holstein dairy cattle

R.D. López-Correa[1,2*], A. Legarra[3] and I. Aguilar[4]

[1] Universidad de la República, Facultad de Agronomía, Montevideo, Uruguay; [2] Universidad de la República, Facultad de Veterinaria, Montevideo, Uruguay; [3] UMR GenPhySE, INRA Toulouse, Castanet Tolosan, France; [4] Instituto Nacional de Investigación Agropecuaria, Montevideo, Uruguay; [*]rlopez@fagro.edu.uy

Abstract

The objective of this study was to check the quality of single-step GBLUP (ssGBLUP) predictions in Uruguayan dairy cattle, with special focus on the modelling of missing pedigree defined as metafounders (MF). A second objective was to use different estimators to compute relationship matrix among MF (**Γ**). Five strategies followed to fit unknown parent groups (UPGs) in ssGBLUP were: i) MF using original Γ; ii) MF with modified Γ based on the median of elements in original Γ; iii) MF bounding base allelic frequencies of markers between 0 and 1 before computing Γ; iv) UPG in $\mathbf{A}^{-1}$ and v) UPG in $\mathbf{H}^{-1}$. Estimated Γ with restricted allele frequencies gave less biased and more accurate genomic predictions. Future studies are warranted in the definition of **Γ** and UPGs to improve quality of genomic predictions and understand the origin of bias.

Introduction

Uruguayan dairy herds are creators of genetic progress through selection but are also heavy importers of foreign genetics. Use of genomic predictions is important as producers start using it, so it is important to ascertain the quality of these predictions. Genomic predictions can be done by ssGBLUP but it is challenging to include UPGs due to their specific structure. The use of UPGs in genomic model evaluations may lead to bias in genomic enhanced breeding values (GEBV) and reduce convergence speed in ssGBLUP (Misztal *et al.*, 2013; Tsuruta *et al.*, 2013). The application of MF theory (Legarra *et al.*, 2015; García-Bacino *et al.*, 2017) could be an alternative to fit missing pedigree in a genomic evaluation model and help to overcome bias and convergence problems. However, Γ may give estimation problems, for instance when genotyped individuals have a weak bond with UPGs (Lourenco *et al.*, 2020). Another important issue in the Uruguayan genetic evaluation is the restricted capacity to perform large progeny tests. Therefore, we need to define statistics that measure bias and accuracy of genomic predictions. Predictability is a major problem in dairy cattle genomic evaluations, because large differences between GEBV and EBV after progeny test were reported (Bradford *et al.*, 2019). Thus, the linear regression method (Legarra and Reverter, 2018), commonly known as LR, could be a useful tool to assess the predictive ability and accuracy of genomic predictions. The objective of this study was to check the quality of ssGBLUP predictions in Uruguayan dairy cattle, with special attention to the modelling of missing pedigree defined as MF. A second objective was to use different estimators to compute gamma relationship matrix among MF.

Materials & Methods

Data. Data consisted of milk yield records from one to five lactations that were used for the Uruguayan Holstein genetic evaluation in April 2021. A total of 1,031,174 records were

available for 367,423 cows. The pedigree file included 578,172 animals and was created by tracing back three generations of ancestors of either cows with lactation records or genotyped animals. In total, 18 UPGs were defined based on sex, origin (foreign, national) and birth year. Genotypes for 5573 animals were available, 3499 were generated using the ICBF_IDBv3 50k panel and 2074 genotypes of bulls were provided by international cooperation. After quality control, 39,288 SNPs were analysed. Cows had the largest number of genotypes in the sample (3275) and more than 95% had at least a single lactation record. They were born between 1993 and 2019, and 263 of the cows had both parents genotyped whereas 1297 had one parent genotyped. Genotyped bulls (2298) were born between 1973 and 2019, and 36 bulls had both parents genotyped and 1181 bulls had one parent genotyped. Most of genotyped animals had both parents known.

Models. Milk adjusted 305-d yield was analyzed in a single-trait repeatability animal model to obtain GEBV using ssGBLUP (Aguilar *et al.*, 2010; Christensen and Lund, 2010). The basic model included fixed effects of herd year season (n=54,274) and age-parity-calving interval-dry period (n=118). The random effects were the additive genetic effect (n=578,190) and the permanent environmental effects (n=423,317). Data included up to five lactation records. Matrix $\mathbf{H}^{-1}$ was defined as in Aguilar *et al.* (2010). Genomic relationship matrix (**G)** was computed by the first method of VanRaden (2008); and blended with default parameters as $\mathbf{G}^{*}=0.95*\mathbf{G} + 0.05*\mathbf{A}_{22}$. To make **G** and $\mathbf{A}_{22}$ compatible, we used default tuning in BLUPF90 (Misztal *et al.*, 2018). To fit UPG five methodologies were applied: ssGBLUP with MF using original Γ (MFO model), ssGBLUP with MF using gamma-robust estimator (MFGrob model), ssGBLUP with MF using gamma-complete estimator (MFBou model), ssGBLUP with UPG in $\mathbf{A}^{-1}$ (PED model), and ssGBLUP with UPG in $\mathbf{H}^{-1}$ (EXA model). For MF models, each UPG was regarded as a MF, i.e., a different ancestral population, and original Γ was computed by generalized least squares using observed genotypes (García-Bacino *et al.*, 2017) and for each MF we calculated the base allelic frequencies. Additionally, the gamma-robust estimator designed Γ with two different values based on the original $\mathbf{\Gamma}$. The diagonal elements of $\mathbf{\Gamma}$ had the median of all self-relationships, whereas the off diagonals contained the median of all relationships across MF. The complete gamma estimator bounded base allelic frequencies of markers between 0 and 1 before computing $\mathbf{\Gamma}$. Values in Γ^{-1} were included in the numerator relationship matrix $\mathbf{A(\Gamma)^{-1}}$, as previously defined (Legarra *et al.* 2015). The PED model fitted UPGs as a fixed effect into ssGLBUP equations and they were included into the pedigree-based numerator relationship matrix $\mathbf{A}^{-1}$ by QP-transformation (Quaas and Pollak, 1981). The EXA model considered UPGs as a random effect into ssGLBUP equations and they were implemented in both $\mathbf{A}^{-1}$ and $\mathbf{G}^{-1}$ (Misztal *et al.*, 2013).

Validation of genomic predictions. To validate genomic predictions, we used 177 genotyped sires with at least 10 daughters with lactation records (valG). For those sires, we compared their GEBV predicted from the whole dataset ('w') with GEBV predicted from a partial dataset ('p'). The partial dataset removed phenotypes of cows recorded after 2014, so that records of daughters from those sires were not considered. GEBV of sires in valG with the whole dataset used 23,482 lactation records of 14,167 cows, 171 of which were genotyped. All GEBV were expressed in relation to the mean of GEBV for cows born in 2010 and with lactation records. GEBVs obtained with reduced data (GEBVp) and with whole data (GEBVw) were compared using statistics described in the 'LR' method (Legarra and Reverter, 2018). Predictions of GEBV and estimates of UPG effects were obtained by ssGBLUP using BLUP90iod2 program (Tsuruta *et al.*, 2001). The Γ was computed using gammaf90 program from BLUPF90 suite.

Results

Models

The original Γ showed higher extreme values than elements of Γ obtained with gamma complete estimator (Table 1). Using the gamma-robust estimator the self-relationship of MF (diagonal) was 0.776 and relationship between MF (off-diagonal) was 0.641.

Table 1. Statistics of estimated original Γ (MFO) and Γ with restricted allele frequencies (MFBou).

	MFO		MFBou	
	Diagonal values	Off-diagonal values	Diagonal values	Off-diagonal values
Minimum	0.661	-2.371	0.654	-0.259
Median	0776	0.641	0.762	0.5911
Maximum	6.302	1.354	1.478	0.709

Except for MFO, all models converged. However, EXA (average rounds: 86), MFBou and MFGrob (average rounds: 89) converged faster than PED (average rounds: 190).

Validation of genomic predictions.

In Table 2, statistics of LR are described for each model, except for MFO. MBou model gave the lowest bias (108). All models had a slope that was far from one and it ranged from 0.672 to 0.699 across models. Correlations between GEBVw and GEBVp were similar for models and doubled the accuracy of GEBVs when using whole data instead of partial data.

Table 2. Bias ($\widehat{\mu}_{w,p}$), slope ($\widehat{b}_{w,p}$) and ratio of accuracies ($\widehat{p}_{w,p}$) between GEBV estimated for each model using bulls in validation set (valG).

	ValG		
Model	$\widehat{\mu}_{w,p}$	$\widehat{b}_{w,p}$	$\widehat{p}_{w,p}$
MFGrob	118	0.672	0.520
MFBou	108	0.698	0.538
PED	296	0.629	0.487
EXA	122	0.699	0.514

Discussion

In this study, we assessed ssGBLUP for the genomic evaluation of Holstein in Uruguay and several methodologies were tested to include UPG in the genomic models. Bias was present in all models. MFBou, MFGrob and EXA models gave better genomic predictions than PED. In contrast, a simulated dairy cattle population, Bradford *et al.*,2019) obtained unbiased and accurate predictions with ssGBLUP and metafounders for the original **Γ**. Besides, Bradford *et al.* (2019) reported greatest bias and least accuracy for ssGBLUP with UPG for **H** in a moderate heritability trait. In our study a possible confounding among different UPGs could also happen because of a low number of animals contributing to estimate those UPGs (Misztal *et al.*, 2013).

Except for MFO, all models were robust and converged, but overdispersion was observed. The low values of slope could be partly explained by the small number of bulls in the validation set (177 genotyped bulls). Our study showed a greater variation of values in the original $\boldsymbol{\Gamma}$ than those estimated for dairy cattle (Legarra *et al.*, 2015; Bradford *et al.*, 2019). This implies that in the Uruguayan Holstein genetic evaluation some MF represented very different base populations. The gamma complete estimator computes different values for elements of $\boldsymbol{\Gamma}$. Thus, $\boldsymbol{\Gamma}$ would be a better proxy to reflect relationship between UPGs and we would be able to obtain less bias in genomic evaluations.

Conclusions

We found that ssGBLUP was successfully implemented in the Uruguayan Holstein genetic evaluation. Genomic models that used metafounders with the complete gamma estimator gave less biased and more accurate predictions. However, the origin of bias was not completely understood, so future studies are warranted in the definition of $\boldsymbol{\Gamma}$ and of UPGs to improve quality of genomic predictions.

References

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta *et al.* (2010). J Dairy Sci 93(2): 743-752. https://doi.org/10.3168/jds .2009-2730.

Bradford, H.L., Y. Masuda, P. M. VanRaden, A. Legarra and I. Misztal (2019). Modeling missing pedigree in single-step genomic BLUP. J Dairy Sci 102(3):2336-2346. https://doi.org/10.3168/jds.2018-15434.

Garcia-Baccino, C.A., A. Legarra, O.F. Christensen, I. Misztal, I. Pocrnic *et al.*(2017). Genet Sel Evol 49(1): 34. https://doi.org/10.1186/s12711-017-0309-2.

Legarra, A., O.F. Christensen, Z.G. Vitezica, I. Aguilar and I. Misztal (2015). Genetics 200(2): 455-468. https://doi.org/10.1534/genetics.115.177014 .

Legarra, A. and A. Reverter (2018). Genet Sel Evol 50(1): 53. https://doi.org/10.1186/s12711-018-0426-6.

Lourenco, D., Legarra,A., Tsuruta, S., Masuda, Y., Aguilar *et al.* (2020). Genes. 11, 790; doi:10.3390/genes11070790.

Misztal, I., S. Tsuruta, D.A. L. Lourenco, Y. Masuda, I. Aguilar, *et al.* 2018. Manual for BLUPF90 family of programs. Available at: http://nce.ads.uga.edu/wiki/

Misztal, I., Z.G. Vitezica, A. Legarra, I. Aguilar, and A.A. Swan (2013). J Anim Breed Genet 130(4): 252-258. https://doi.org/10.1111/jbg.12025.

Quaas, R. L. and E. J. Pollak (1981). J Dairy Sci 64(9): 1868-1872.

Tsuruta, S., I. Misztal, D.A.L. Lourenco, and T.J Lawlor. (2013). J. Dairy Sci. 97 :5814–5821. http://dx.doi.org/10.3168/jds.2013-7821.

Tsuruta, S., I. Misztal, and I. Stranden (2001). J Anim Sci 79(5): 1166-1172. doi:10.2527/2001.7951166x.

VanRaden, P.M. J. Dairy Sci. 2008, 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 772787.