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

Are Evaluations on Young Genotyped Dairy Bulls Benefiting from the Past Generations?

D. A. L. Lourenco[*], I. Misztal[*], S. Tsuruta[*], I. Aguilar[†], T. J. Lawlor[‡] and J. I. Weller[§]
[*]University of Georgia, Athens, GA, USA; [†]Instituto Nacional de Investigación Agropecuaria, Las Brujas, Uruguay; [‡]Holstein Association USA Inc., Brattleboro, VT, USA; [§]Agricultural Research Organization, the Volcani Center, Bet Dagan, Israel.

ABSTRACT: Datasets of US and Israeli Holsteins were used to evaluate the impact of older generations on ability to predict EBV of young genotyped animals in traditional and single-step genomic BLUP. Inclusion of two (A2) or all (Af) ancestor generations was also evaluated. A total of 34,506 US and 1,305 Israeli bulls were genotyped. Thresholds for data deletion were based on 5 years interval. The number of generations deleted without reduction in accuracy depended on data structure and trait. For US Holsteins, removing 3 and 4 generations of data did not reduce accuracy for final score in Af and A2, respectively. For Israeli Holsteins, the effect of removing older generations depended on the genetic origins of the bulls. Therefore, truncating older data does not decrease the accuracy of young genotyped bulls, while reducing computation requirements and helping to find problems in the population structure.
Keywords: ssGBLUP; genomic selection; pedigree depth

Introduction

While parents can explain up to 50% of the genetic variation in an animal, this fraction is divided by four with each previous generation. Thus the contributions of distant generations decay over time, and the impact of distant ancestors on the accuracy of the youngest animals can be small or even negative. Furthermore, larger datasets require more computing resources. Mehrabani-Yeganeh et al. (1999) studied the selection response in a simulated population. The accuracy of traditional evaluations for the most recent generation was the same regardless of whether all nine or only the last two generations of data were used. In initial predictions with genomic selection, the decay of accuracy for subsequent generations without phenotypes was much slower than with the traditional selection (Meuwissen et al. (2001)). Muir (2007) found that the decay of accuracy in genomic selection is much faster under strong selection.

In single-step genomic BLUP (ssGBLUP), calculation of unbiased genomic EBV (GEBV) requires scaling the genomic relationship matrix (**G**) to make it compatible with the numerator relationship matrix for the genotyped animals ($\mathbf{A_{22}}$) (Chen et al. (2011); Vitezica et al. (2011)). Too small **G** causes downward bias for the genotyped animals relative to all the animals, and too large **G** causes the upward bias. The additive relationships for the young animals depend on the length of their pedigrees. Since **G** is scaled for an average of $\mathbf{A_{22}}$, GEBV for young animals may be biased up or down dependent on the length of the pedigree, with a corresponding drop in accuracy (Misztal et al. (2013)). A partial solution for this problem is to delete pedigree and phenotypic data of older generations. Therefore, the purpose of this study was to investigate the effect of deleting phenotypic and pedigree data on the accuracy of young genotyped bulls in dairy populations evaluated for different traits.

Materials and Methods

Data. For US Holsteins, a full dataset contained 10,944,571 final score (FS) records up to 2011 for 6,586,605 cows born from 1951 to 2009, and a reduced dataset (TR) with 10,167,064 records up to 2007 for 6,012,441 cows born from 1951 to 2006. A total of 34,506 bulls were genotyped for 42,503 SNP. More details on this dataset are in Tsuruta et al. (2011). For Israeli Holsteins, the full dataset contained 305-d milk, fat, and protein yields and fat and protein percentages, for cows born from 1982 to 2010 with 713,686 records for parity 1. The TR included only production records through 2006 for 563,870 cows born from 1982 to 2005. A total of 1305 bulls were genotyped for 30,359 SNP. More details on this dataset are in Lourenco et al. (2014).

Thresholds for old data exclusion were set according to an approximate average generation interval of 5 years in dairy cattle. Five thresholds were applied for US Holsteins (T1980, T1985, T1990, T1995, and T2000), while only 3 thresholds were possible for Israeli (T1990, T1995, and T2000) due to the lack of older pedigree data. Two scenarios for constructing the numerator relationship matrix (**A**) were used. The first scenario included relatives of phenotyped animals traced back two generations (A2); in the second scenario all known relatives of phenotyped animals were included (Af).

Model. For US Holsteins, a single-trait repeatability animal model was used for evaluation of FS. Unknown parent groups (UPG) were assigned for missing parents according to year of birth and sex of unknown parent. For Israeli Holsteins, a multiparity animal model was used for the yield and percentage traits. This model considered parities 1 through 3 as correlated traits; however, only results for first parity were presented. The UPG were defined based on year of birth, sex and which parent was missing. A small fraction of the ancestor bulls were not Holsteins, and additional groups were defined for these animals based on breed. Traditional evaluations (BLUP) were performed

for all datasets, while genomic evaluations were not performed for the full datasets. Genomic evaluations were computed by ssGBLUP where pedigrees, genotypes, and phenotypes are jointly considered in a single analysis (Aguilar et al. (2010)).

Validation. The young genotyped bulls were considered the validation set. The set of US (Israeli) Holstein included 2232 (135) bulls born after 2003 (2001) with no daughters in the reduced and threshold datasets, but with ≥ 20 daughters in the full dataset. Most of the validation US bulls had US sires, while most of the Israeli bulls had foreign sires.

Coefficients of determination (R^2) and regression (δ) of deregressed evaluations (DD) from a full dataset with records up to 2011 on parent average (PA) or GEBV from the TR and truncations were computed. The regression models were weighted by reliability of DD. The regression of DD on PA was the benchmark used to compare the gain in predictive ability due to genomics, and regressions of DD on GEBV for truncated datasets were used to compare the response in predictive ability due to the exclusion of old data. While R^2 was used to quantify the validation reliability, δ was used to assess the prediction bias.

Results and Discussion

Table 1 presents R^2 and δ of DD on GEBV from reduced and threshold datasets for US Holstein in A2 and Af scenarios. Values of R^2 for PA did not change significantly with the exclusion of any quantity of the historical data (results not shown). R^2 for GEBV showed a decline at T1995 and at T2000 with Af and A2, respectively. There was also a decline with A2 when very old data were included. It seems that the presence of older pedigree data slightly reduces the accuracy when older phenotypes are removed. The δ for PA were stable, except for a decline at T2000 (results not shown). For GEBV, they slightly increased after removal of few generations. Summarizing, eliminating the data prior to 1990 or perhaps even 1995 did not decrease the accuracy. This period includes 12-17 years or about 2-3 generations.

Table 1. Coefficients of determination (R^2) and regression (δ) of deregressed evaluations on GEBV for final score of 2,232 US Holstein genotyped bulls with no daughters with records in 2007, but with >=20 daughters with records in 2011, for A2 and Af scenarios.

	A2		Af	
Threshold[1]	R^2	δ	R^2	δ
TR	0.33	0.86	0.34	0.87
T1980	0.34	0.87	0.34	0.87
T1985	0.35	0.88	0.34	0.88
T1990	0.35	0.88	0.34	0.88
T1995	0.35	0.88	0.33	0.88
T2000	0.33	0.88	0.31	0.88

[1]TR is for reduced data set, T1980, T1985, T1990, T1995, and T200 are the thresholds for old data exclusion.

Values for R^2 and δ for Israeli validation bulls in scenarios A2 and Af are in Table 2. For the yield traits, the highest R^2 were obtained with all phenotypes prior to 2000 deleted. Conversely, the lowest R^2 were obtained for the percentage traits at the same threshold. Nearly all δ increased with deletion of more phenotypic records. As most of the Israeli validation bulls were sons of foreign bulls, R^2 were also calculated separately for the 38 bulls with Israeli sires and 97 bulls with foreign sires (Table 3). For yield traits, higher levels of truncation decreased R^2 for bulls with Israeli sires and increased R^2 for bulls with foreign sires. While more data benefited animals that were descendants of animals with phenotypic records in the population, old data reduced the accuracy for bulls with foreign sires, which generally did not have high reliability within the Israeli population. The effect of imported animals on accuracy was also present in US Holsteins, but at a much lower level because the number of imported sires is much smaller.

Table 2. Coefficients of determination (R^2) and regression (δ) of deregressed evaluations on GEBV for yield and percentage traits for 135 Israeli Holstein genotyped bulls with no daughters with records in 2006, but with >=20 daughters with records in 2011, for A2 and Af scenarios.

Trait		A2		Af	
	Threshold[1]	R^2	δ	R^2	δ
Milk	TR	0.22	0.94	0.24	0.98
	T1990	0.21	0.92	0.25	1.00
	T1995	0.22	0.99	0.25	1.04
	T2000	0.24	1.01	0.26	1.06
Fat	TR	0.11	0.62	0.14	0.68
	T1990	0.11	0.63	0.14	0.70
	T1995	0.12	0.65	0.15	0.70
	T2000	0.14	0.63	0.16	0.70
Protein	TR	0.11	0.62	0.20	0.86
	T1990	0.14	0.76	0.21	0.90
	T1995	0.14	0.75	0.21	0.91
	T2000	0.20	0.83	0.25	0.94
Fat %	TR	0.40	1.19	0.40	1.18
	T1990	0.40	1.19	0.40	1.18
	T1995	0.40	1.22	0.39	1.18
	T2000	0.39	1.26	0.38	1.24
Protein %	TR	0.41	1.05	0.41	1.03
	T1990	0.41	1.03	0.41	1.02
	T1995	0.39	1.01	0.39	0.99
	T2000	0.35	1.06	0.35	1.04

[1]TR is for reduced data set, T1990, T1995, and T200 are the thresholds for old data exclusion. Results are for the first parity.

Table 3. Coefficients of determination (R^2) of de-regressed evaluations on GEBV for yield and percentage traits for 135 Israeli Holstein genotyped bulls with no daughters with records in 2006, but with >=20 daughters with records in 2011, for A2 and Af scenarios. Values are split for 38 bulls with Israeli sires (ISR) and 97 bulls with foreign sires (FOR).

Trait		A2		Af	
	Threshold[1]	ISR	FOR	ISR	FOR
Milk	TR	0.33	0.16	0.34	0.18
	T1990	0.31	0.15	0.33	0.19
	T1995	0.28	0.18	0.29	0.20
	T2000	0.24	0.24	0.26	0.25
Fat	TR	0.15	0.10	0.16	0.12
	T1990	0.14	0.11	0.15	0.14
	T1995	0.14	0.11	0.16	0.14
	T2000	0.07	0.19	0.09	0.21
Protein	TR	0.28	0.06	0.31	0.10
	T1990	0.25	0.06	0.28	0.11
	T1995	0.22	0.06	0.26	0.11
	T2000	0.17	0.15	0.22	0.19
Fat %	TR	0.36	0.42	0.36	0.42
	T1990	0.39	0.41	0.37	0.41
	T1995	0.40	0.41	0.37	0.40
	T2000	0.29	0.42	0.31	0.40
Protein %	TR	0.50	0.38	0.50	0.39
	T1990	0.50	0.38	0.50	0.39
	T1995	0.48	0.37	0.47	0.38
	T2000	0.46	0.32	0.44	0.33

[1]TR is for reduced data set, T1990, T1995, and T200 are the thresholds for old data exclusion. Results are for the first parity.

Limiting records to the two last generations in A2 (Af) scenario, the computations decreased by 70% (59) and 36% (33) for US and Israeli Holsteins, respectively. Computing time will be more dependent on the number of genotyped animals when it increases. However, the inverse of $\mathbf{A_{22}}$ will be more sparse with smaller pedigrees. Therefore, cutting pedigrees may be helpful in recursions for $\mathbf{A_{22}^{-1}}$ when the number of genotyped animals surpass the current computing limit of 100k.

Conclusion

Different datasets and models used in our study helped to better understand the influence of previous generations in the evaluation of young genotyped animals. It seems that the impact in predictive ability by removing old generations is data-structure driven. Retaining only 2 or 3 generations of phenotypic records and 2 generations of pedigree records does not decrease the accuracy of evaluations for young genotyped bulls while decreasing computing costs. In addition, evaluating realized accuracy with various levels of data truncation might uncover problems due to population structure. However, the option for data deletion will depend on the objectives of the evaluation.

Literature Cited

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

Chen, C.Y., Misztal, I., Aguilar, I. et al. (2011). *J. Anim. Sci.* 89:2673–2679.

Lourenco, D. A. L., Misztal, I., Tsuruta, S. et al. (2014). *J. Dairy Sci.* 97:1742–1752.

Mehrabani-Yeganeh, H., Gibson, J. P., and Schaeffer, L. R. (1998). *Poultry Sci.* 78:937–941.

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

Misztal, I., Vitezica, Z., Legarra, I. et al. (2013). *J. Anim. Breed. Genet.* 130:252–258.

Muir, W. M. (2007). *J. Anim. Breed. Genet.* 124:342–355.

Tsuruta, S., Misztal, I., Aguilar, I. et al. (2011). *J. Dairy Sci.* 94:4198–4204.

Vitezica, Z. G., Aguilar, I., Misztal, I. et al. (2011). *Genet. Res.* 93:357–366.