J. Dairy Sci. 97:3930–3942
http://dx.doi.org/10.3168/jds.2013-7769

# Are evaluations on young genotyped animals benefiting from the past generations?

**D. A. L. Lourenco,*[1] I. Misztal,* S. Tsuruta,* I. Aguilar,† T. J. Lawlor,‡ S. Forni,§ and J. I. Weller#**
*Department of Animal and Dairy Science, University of Georgia, Athens 30602
†Instituto Nacional de Investigacion Agropecuaria, Las Brujas 90200, Uruguay
‡Holstein Association USA Inc., Brattleboro, VT 05302
§Genus PIC, Hendersonville, TN 37075
#Agricultural Research Organization, The Volcani Center, Bet Dagan 50250, Israel

## ABSTRACT

Data sets of US Holsteins, Israeli Holsteins, and pigs from PIC (a Genus company, Hendersonville, TN) were used to evaluate the effect of different numbers of generations on ability to predict genomic breeding values of young genotyped animals. The influence of including only 2 generations of ancestors (A2) or all ancestors (Af) was also investigated. A total of 34,506 US Holsteins, 1,305 Israeli Holsteins, and 5,236 pigs were genotyped. The evaluations were computed by traditional BLUP and single-step genomic BLUP, and computing performance was assessed for the latter method. For the 2 Holstein data sets, coefficients of determination ($R^2$) and regression ($\delta$) of deregressed evaluations from a full data set with records up to 2011 on estimated breeding values and genomic estimated breeding values from the truncated data sets were computed. The thresholds for data deletion were set by intervals of 5 yr, based on the average generation interval in dairy cattle. For the PIC data set, correlations between corrected phenotypes and estimated or genomic estimated breeding values were used to evaluate predictive ability on young animals born in 2010 and 2011. The reduced data set contained data up to 2009, and the thresholds were set based on an average generation interval of 3 yr. The number of generations that could be deleted without a reduction in accuracy depended on data structure and trait. For US Holsteins, removing 3 and 4 generations of data did not reduce accuracy of evaluations for final score in Af and A2 scenarios, respectively. For Israeli Holsteins, the accuracies for milk, fat, and protein yields were the highest when only phenotypes recorded in 2000 and later were included and full pedigrees were applied. Of the 135 Israeli bulls with genotypes (validation set) and daughter records only in the complete data set, 38 and 97 were sons of Israeli and foreign bulls, respectively. Although more phenotypic data increased the prediction accuracy for sons of Israeli bulls, the reverse was true for sons of foreign bulls. Also, more phenotypic data caused large inflation of genomic estimated breeding values for sons of foreign bulls, whereas the opposite was true with the deletion of all but the most recent phenotypic data. Results for protein and fat percentage were different from those for milk, fat, and protein yields; however, relatively, the changes in coefficients of determination and regression were smaller for percentage traits. For PIC data set, removing data from up to 5 generations did not erode predictive ability for genotyped animals for the 2 reproductive traits used in validation. Given the data used in this study, truncating old data reduces computation requirements but does not decrease the accuracy. For small populations that include local and imported animals, truncation may be beneficial for one group of animals and detrimental to another group.

**Key words:** single-step genomic BLUP, pedigree depth, genomic selection, dairy cattle

## INTRODUCTION

Quantitative genetics theory postulates that accuracy of genetic evaluations increases if all known ancestors are included in construction of the relationship matrix (Henderson, 1984), provided that the analysis model corresponds to reality. However, models used in practice are only approximations of "true" models. For example, definitions of traits change over time, accounting for selection may be incomplete, and nonadditive genetic effects are ignored in the model. Also, the contributions of distant generations decay with time. Although parents can explain up to 50% of the genetic variation in an animal, this fraction is divided by 4 with each previous generation. Therefore, the effect of distant ancestors on the accuracy of the youngest animals can be small or even negative. Furthermore, larger data sets require more computing resources.

Mehrabani-Yeganeh et al. (1999) studied the selection response in a simulated population. The accuracy of evaluation for the most recent generation was the same

Received November 26, 2013.
Accepted February 11, 2014.

[1] Corresponding author: danilino@uga.edu

regardless of whether all 9 or only the last 2 generations of data were used. The mean simulated breeding value of the selected animals was the same in both scenarios, but mean inbreeding of selected animals was lower for the truncated data set.

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 a real population of broiler chickens, that decay was faster than initially expected but still slower than in traditional BLUP (Wolc et al., 2011); the rate of decay changed only slightly for different methods, with lesser decay with bivariate genomic BLUP (**GBLUP**) and BayesCπ (Habier et al., 2011) than with univariate GBLUP. If the decay in accuracy is faster than expected, the contributions of older generations may be overestimated with genomic selection.

Recently, Misztal et al. (2013) studied possible biases with unknown parent groups (**UPG**) in a single-step genomic evaluation (**ssGBLUP**). In this method, calculation of unbiased GEBV requires scaling the genomic relationship matrix ($\mathbf{G}$) to make this matrix compatible with the numerator relationship matrix for the genotyped animals ($\mathbf{A}_{22}$) (Chen et al., 2011; Vitezica et al., 2011). Too-small $\mathbf{G}$ causes downward bias for the genotyped animals relative to all the animals, and too-large $\mathbf{G}$ causes upward bias. The additive relationships for the young animals depend on the length of their pedigrees. Because scaling of $\mathbf{G}$ is for an average of $\mathbf{A}_{22}$, GEBV for young animals may be biased up or down depending on the length of the pedigree, with a corresponding decrease in accuracy. A partial solution for this problem is to delete pedigree and phenotypic data of older generations. In this case, missing information from the eliminated pedigrees does not bias evaluations.

The purpose of this study was to evaluate the effect of deleting phenotypic and pedigree data on the accuracy of young genotyped animals in several populations and different traits.

## MATERIALS AND METHODS

Three different data sets were analyzed in this study: US Holstein final score data provided by Holstein Association USA Inc. (Brattleboro, VT); Israeli Holstein 305-d milk, fat, and protein yields and fat and protein percentage data provided by Israel Cattle Breeders Association (Caesaria, Israel); and pig reproductive traits from purebred and crossbred lines, provided by PIC (a Genus company, Hendersonville, TN). For all data sets, variance components were estimated based on the full data using phenotypes and pedigree. Multiple species and a range of population structures were included to give this study a broad application. Animal Care and Use Committee approval was not obtained for this study, because the data were obtained from existing databases.

### US Holsteins

***Data.*** Initially, 2 data sets were prepared for US Holsteins. The full data set contained 10,944,571 final score records up to 2011 for 6,586,605 cows born from 1951 to 2009, and a reduced data set (**TR**) included 10,167,064 records up to 2007 for 6,012,441 cows born from 1951 to 2006. Records were deleted from the reduced data set to exclude data of cows born before 5 different thresholds. The thresholds set according to an approximate average generation interval of 5 yr in dairy cattle were T1980, T1985, T1990, T1995, and T2000. Thus, T1980 comprised data of cows born from 1980 to 2006 with records up to 2007, with the same procedure applied for the other 4 thresholds. Two scenarios for constructing the numerator relationship matrix ($\mathbf{A}$) were used. The first scenario included relatives of phenotyped animals traced back 2 generations (short pedigree = **A2**); the second scenario included all known relatives of phenotyped animals (deep pedigree = **Af**). The number of animals included in $\mathbf{A}$ and the number of phenotypes available for each data set are shown in Table 1.

After a general quality control analysis, genotypes on 42,503 SNP markers from the BovineSNP50K BeadChip (Illumina Inc., San Diego, CA) were available for 34,506 bulls.

***Model.*** A single-trait animal model was used for evaluation of final score (Tsuruta et al., 2002). The heritability for this trait is 0.31 (Table 2). Unknown parent groups were assigned for missing parents according to year of birth and sex. Traditional evaluations (BLUP) were performed for all data sets, whereas genomic evaluations were not performed for the full data set. Pedigree, genotypes, and phenotypes were analyzed by ssGBLUP (Aguilar et al., 2010). In this method, the inverse of matrix $\mathbf{A}$ is replaced by the inverse of matrix $\mathbf{H}$ in the mixed model equations. $\mathbf{H}$ inverse is as follows:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{G}$ was constructed as in VanRaden (2008), using current allele frequencies; $\mathbf{A}_{22}^{-1}$ is the inverse of pedigree-

**Table 1.** Data structure[1]

| | | A2 | | Af | | |
|---|---|---|---|---|---|---|
| Data | Threshold[2] | Animals (no.) | Animals removed (%) | Animals (no.) | Animals removed (%) | Records (no.) |
| US Holstein | Full | 9,106,249 | 0.00 | 9,602,031 | 0.00 | 10,944,571 |
| | TR | 9,106,249 | 0.00 | 9,602,031 | 0.00 | 10,167,064 |
| | T1980 | 7,369,426 | 0.19 | 8,588,711 | 0.11 | 7,530,770 |
| | T1985 | 6,305,553 | 0.31 | 7,388,211 | 0.23 | 5,741,868 |
| | T1990 | 5,025,481 | 0.45 | 6,175,923 | 0.36 | 3,808,580 |
| | T1995 | 3,594,892 | 0.61 | 4,692,257 | 0.51 | 2,295,754 |
| | T2000 | 2,392,963 | 0.74 | 3,363,998 | 0.65 | 1,123,896 |
| Israeli Holstein | Full | 826,653 | 0.00 | 829,398 | 0.00 | 1,543,830 |
| | TR | 826,653 | 0.00 | 829,398 | 0.00 | 1,205,801 |
| | T1990 | 731,141 | 0.12 | 748,714 | 0.10 | 930,429 |
| | T1995 | 588,165 | 0.29 | 637,624 | 0.23 | 607,876 |
| | T2000 | 434,931 | 0.47 | 516,574 | 0.38 | 267,486 |
| PIC | Full | 681,907 | 0.00 | 682,764 | 0.00 | 2,176,298 |
| | TR | 681,907 | 0.00 | 682,764 | 0.00 | 1,750,226 |
| | T1991 | 677,081 | 0.01 | 679,329 | 0.01 | 1,736,607 |
| | T1994 | 655,511 | 0.04 | 658,984 | 0.03 | 1,650,003 |
| | T1997 | 628,034 | 0.08 | 632,780 | 0.07 | 1,550,846 |
| | T2000 | 611,206 | 0.10 | 617,079 | 0.10 | 1,498,259 |
| | T2003 | 553,573 | 0.19 | 561,299 | 0.18 | 1,271,455 |
| | T2006 | 472,154 | 0.31 | 482,811 | 0.29 | 930,152 |

[1] A2 = short pedigree (including only 2 generations of ancestors) and Af = deep pedigree (including all ancestors).

[2] Full = complete data set and contains data up to 2011 for US and Israeli Holsteins and up to 2012 for PIC data (pig data from Genus, Hendersonville, TN); TR = reduced data set and contains data up to 2007 for US Holsteins, 2006 for Israeli Holsteins, and 2009 for PIC data; the other thresholds contain data from the indicated year up to 2007, 2006, and 2009 for the 3 data sets, respectively.

based relationship matrix for genotyped animals; $\tau$ and $\omega$ are scaling factors for $\mathbf{G}^{-1}$and $\mathbf{A}_{22}^{-1}$, respectively. Both factors were investigated by Misztal et al. (2010) for final score for US Holsteins. Changes in $\tau$ had little effect on accuracy and prediction bias, but $\omega < 1$ helped to reduce the inflation of GEBV. The values chosen in this study were 1.0 for $\tau$ and 0.7 for $\omega$ (Tsuruta et al., 2011). Weights for $\mathbf{G}$ ($\alpha = 0.95$) and $\mathbf{A}_{22}$ ($\beta = 0.05$) can avoid singularity problems and slightly improve predictions (VanRaden, 2008).

***Validation.*** The validation method was chosen based on VanRaden et al. (2009). The following regression model was used to assess prediction bias of evaluations:

$$\mathrm{DD_{full}} = \mu + \delta X_{\mathrm{Tx}} + e,$$

where $\mathrm{DD_{full}}$ are deregressed evaluations of 2,232 genotyped bulls born after 2003 and with no daughters in the reduced and threshold data sets but with ≥20 daughters in the full data set; $\mu$ is mean; δ is a regression coefficient and will be referred to as prediction bias; $X_{\mathrm{Tx}}$ is a bull's parent average (**PA**) or GEBV based on the reduced or threshold data sets (TR, T1980, T1985, T1990, T1995, and T2000); and $e$ is the residual. Values of δ close to 1 indicate a 1:1 ratio in changes in the evaluation and in the trait (Wiggans et al., 2011). The linear regression model was weighted by reliability of $\mathrm{DD_{full}}$. According to VanRaden (2008), $\mathrm{DD_{full}}$ can be obtained as follows:

$$\mathrm{DD_{full}} = \frac{\mathrm{EBV_{full}} - \mathrm{PA_{full}}}{R_{\mathrm{full}}} + \mathrm{PA_{full}},$$

**Table 2.** Heritabilities of all evaluated traits

| Data | Trait | Heritability |
|---|---|---|
| US Holstein | Final score | 0.31 |
| Israeli Holstein | Milk (yield) | |
| | Parity 1 | 0.39 |
| | Parity 2 | 0.29 |
| | Parity 3 | 0.27 |
| | Fat (yield and %) | |
| | Parity 1 | 0.42 |
| | Parity 2 | 0.38 |
| | Parity 3 | 0.34 |
| | Protein (yield and %) | |
| | Parity 1 | 0.34 |
| | Parity 2 | 0.29 |
| | Parity 3 | 0.27 |
| PIC[1] | Trait 1 | 0.14 |
| | Trait 2 | 0.11 |
| | Trait 3 | 0.15 |
| | Trait 4 | 0.09 |

[1] Pig data from PIC (Genus, Hendersonville, TN).

where the fraction is the deregressed Mendelian sampling; $R_{\text{full}}$ is the deregression factor obtained by:

$$R_{\text{full}} = \frac{\text{DE}_{\text{animal}}}{\text{DE}_{\text{animal}} + \text{DE}_{\text{PA}} + 1},$$

where $\text{DE}_{\text{animal}}$ is equivalent daughter contributions from the animal and its progeny and $\text{DE}_{\text{PA}}$ is equivalent daughter contributions from parent averages. The subscript "full" indicates that the values were calculated from the full data set.

The coefficient of determination ($R^2$) of this model was used to quantify the validation reliability of PA and GEBV. Regression of $\text{DD}_{\text{full}}$ on PA was the benchmark used to compare the gain in predictive ability due to genomics, and regressions of $\text{DD}_{\text{full}}$ on reduced or threshold data sets were used to compare the response in predictive ability due to the exclusion of old data (T1980 to T2000). Although DD is not the only response variable that can be used for validation, it is easier to obtain and has been widely adopted (VanRaden et al., 2009; Tsuruta et al., 2011; Wiggans et al., 2011).

### Israeli Holsteins

***Data.*** For Israeli Holsteins, the full data set contained 305-d milk, fat, and protein (yield traits) and fat and protein percentages (percentage traits), for cows born from 1982 to 2010, with 713,686 records for parity 1, 503,827 records for parity 2, and 326,317 records for parity 3. The cows calved from 1985 through 2011. The reduced data set included only production records through 2006 for 563,870 cows, with records for parity 1, 391,977 records for parity 2, and 249,954 records for parity 3. The cows were born from 1982 to 2005. From this reduced data set (TR), 3 different thresholds for data truncation were applied: T1990: cows born before 1990 were deleted from the reduced data and pedigree; T1995: cows born before 1995 were deleted; and T2000: cows born before 2000 were deleted. The 2 scenarios with respect to the depth of the pedigree (A2 and Af) were also applied. The numbers of animals in the pedigree relationship matrices and phenotypes available for each data set are shown in Table 1. Heritabilities for all traits and parities are in Table 2.

A total of 1,305 bulls were genotyped for the Illumina BovineSNP50 BeadChip (Illumina Inc.), which includes approximately 54,000 markers. After quality control, 30,359 SNP remained in the genotype file.

***Model.*** A multiparity animal model has been used in Israel for routine evaluation of each one of the production traits (Weller and Ezra, 2004). The same model was used in this study to compute traditional genetic evaluation for parities 1 through 3 as correlated traits. The model was analyzed with and without UPG; UPG were defined based on year of birth, sex, and which parents were missing. A small fraction of the ancestor bulls were not Holsteins and additional groups were defined for these animals based on breed. When pedigrees were truncated, UPG assignments were left intact in retained pedigrees, whereas base animals generated by deletion of their ancestors were set to a common group.

For genomic evaluations, the **H** matrix was constructed and scaled in the same way as for US Holsteins evaluations.

***Validation.*** The same validation method (VanRaden et al., 2009) was used for US and Israeli Holstein data sets. However, for Israeli data, $\text{DD}_{\text{full}}$ are deregressed evaluations of 135 genotyped bulls born after 2001 and with no daughters in the reduced and threshold data sets, but with ≥20 daughters in the full data set; $X_{\text{Tx}}$ is a bull's PA or GEBV based on the reduced (TR) or threshold data sets (T1990, T1995, and T2000). Of the validation bulls, 38 were sons of Israeli sires and 97 were sons of foreign sires. All dams were local cows with records in the Israeli database. The Israeli sires all had genetic evaluations based on progeny tests in Israel, whereas the foreign sires generally did not.

### PIC Pigs

***Data and Model.*** The PIC pig data set consisted of phenotypes collected in purebred and crossbred animals for 4 reproductive traits: litter size and number of stillborn for purebreds (traits 1 and 2, respectively) and crossbreds (traits 3 and 4, respectively). The heritabilities for the 4 traits ranged from 0.09 to 0.15 (Table 2). The full data set included at least 2,176,298 records for 655,037 animals born from 1971 to 2012. The reduced data set contained at least 1,750,226 records for 468,486 animals born from 1971 to 2009. From the reduced data set, 6 thresholds were established according to an average generation interval of 3 yr in pigs: T1991, T1994, T1997, T2000, T2003, and T2006. Similar to the previous data sets, T1991 included data of animals born from 1991 to 2009 with records up to 2009, whereas T2006 included only data from 2006 to 2009. Similarly, the data sets were analyzed including all known ancestors (Af) and only 2 generations (A2). A 4-trait animal model with permanent environmental effect was used to evaluate this data set. The numbers of animals in the pedigree relationship matrices and phenotypes included in each data set are shown in Table 1.

Genotypes from the Illumina PorcineSNP60K chip (~64,000 SNP; Ramos et al., 2009) were available for 5,236 animals. After quality check procedures, only 35,324 SNP were retained for analysis. For genomic evaluations, the H matrix was constructed in the same way as for Holsteins, but with different scaling. Values for $\alpha$, $\beta$, $\tau$, and $\omega$ were 0.7, 0.3, 0.7, and 0.8, respectively.

***Validation.*** The validation method adopted for the PIC data set was chosen based on Christensen et al. (2012); therefore, it was different from the validation used for the Holsteins. Predictive ability of evaluations using full and recent data sets was defined as the correlation between breeding value and phenotypes corrected for fixed and random effects other than genetic additive and residual:

$$r = \mathrm{cor}(X_{\mathrm{Tx}}, Y_{\mathrm{c\_full}}),$$

where $r$ is interpreted as accuracy of evaluations; $Y_{\mathrm{c\_full}}$ are corrected phenotypes from full data set of 1,034 genotyped pigs born in 2010 and 2011 and with no data in the reduced and threshold data sets; $X_{\mathrm{Tx}}$ is EBV or GEBV based on the reduced (TR) or threshold data sets (T1991, T1994, T1997, T2000, T2003, and T2006). Because young genotyped animals were purebred and had no phenotypes on traits 3 and 4, validations were not performed for these traits.

## RESULTS AND DISCUSSION

Figure 1 presents $R^2$ and regression coefficients ($\boldsymbol{\delta}$) of $DD_{\mathrm{full}}$ on PA and GEBV from reduced and threshold data sets for US Holstein in A2 and Af scenarios. The $R^2$ for PA did not change significantly with the removal of any quantity of the historical data, whereas $R^2$ for GEBV declined after T1995 when all pedigree data were included and at T2000 with A2. We also observed a decline with A2 when very old data were included. The presence of older pedigree data when older phenotypes were removed reduced the accuracy slightly. The regressions for PA were stable, except for a decline at T2000. The regressions for GEBV were slightly increasing. They were higher than those for PA because of the $\omega$ parameter (Tsuruta et al., 2011). Summarizing, eliminating the data before 1990 or perhaps even 1995 did not decrease the accuracy. This period included 12 to 17 yr or 2 to 3 generations.

Plots for Israeli Holsteins (Figure 2) are presented for models with and without UPG. When the model did not contain UPG, nearly all $R^2$ for milk, fat, and protein increased when more phenotypes were eliminated. In fact, the highest $R^2$ was for T2000. The opposite occurred for fat and protein percentages, although the relative change was much smaller. In general, $R^2$ were lower with truncated pedigrees for milk, fat, and protein, and similar or slightly higher for fat percentage and protein percentage.

As noted, most of the Israeli validation bulls were sons of foreign bulls. The ancestors of the foreign bulls generally did not have records in Israel. Therefore, $R^2$ in the model with UPG that accounted for different genetic origins was higher and followed a similar pattern as in the model without UPG (Figure 2). Thus, for the yield traits, the highest $R^2$ values were obtained with all phenotypes before 2000 deleted. Conversely, the lowest $R^2$ values were obtained for the concentration traits with all phenotypes before 2000 eliminated. These apparently conflicting results can be explained in 2 ways. First, adjustments in the older data were detrimental to the accuracy of some, but not all, traits. The other possibility is that the influence of past generations is smaller for traits under strong selection.

Lower accuracies with truncated pedigrees could be due to the difficulty of determining UPG that correctly reflect reality for a relatively small population; UPG are important in traits under selection. For the concentration traits that were not under major selection, UPG may not increase the accuracy but add an estimation error. However, almost all regressions increased with deletion of more phenotypic records (Figure 3), whether including UPG or not.

To investigate the effect of different origins on the accuracy of evaluation, $R^2$ were calculated separately for the 38 bulls with Israeli sires and the 97 bulls with foreign sires (Figure 4). For milk, fat, and protein, higher levels of truncation decreased $R^2$ for bulls with Israeli sires and increased $R^2$ for bulls of foreign sires. The highest $R^2$ in the latter case were those with the least data (T2000). Although 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, who generally did not have high evaluation reliability within the Israeli population. In general, the accuracy of the genomic selection for a young animal depends on the relationship of that animal to the training population (Habier et al., 2010). In an ideal model, the addition of old data should not diminish the $R^2$ for any group of young animals. Refining the model to better account for different genetic origins requires further study.

Regression of the deregressed evaluations of 135 genotyped bulls with no daughters with records in 2006, but with $\geq$20 daughters with records in 2011 on parent averages computed separately by origin of the bulls' sires are given in Figure 5. Results were consistent across truncation levels, except for T2000. For these data sets, regressions decreased for sons of local sires and increased for sons of foreign sires. Because regressions of unity indicate unbiased evaluations, these trends resulted, in nearly all cases, in less-biased regressions for the sons of foreign bulls and more-biased regressions

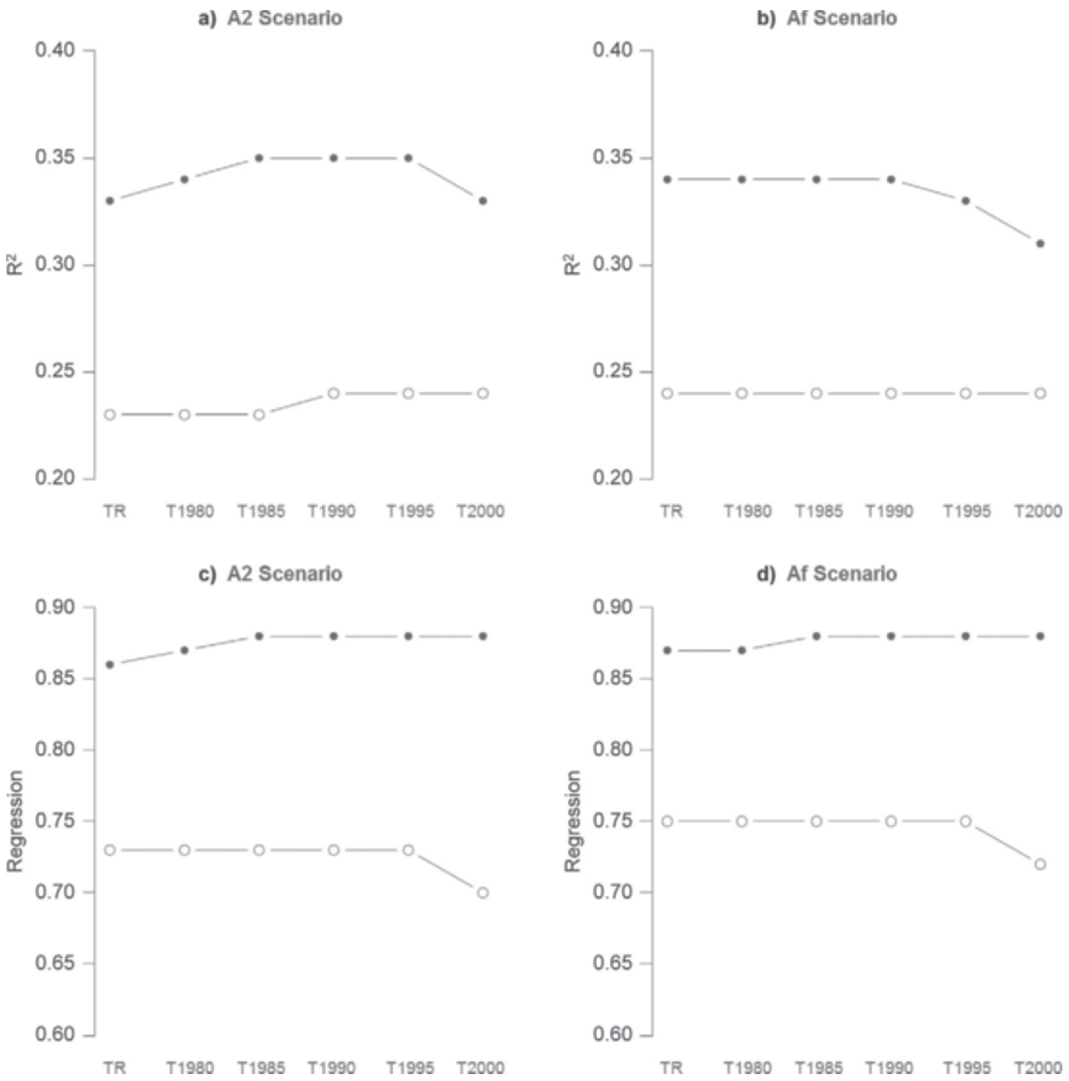

**Figure 1.** Coefficients of determination ($R^2$) (a, b) and regression (c, d) of deregressed evaluations for final score of 2,232 US Holstein genotyped bulls with no daughters with records in 2007, but with ≥20 daughters with records in 2011 on parent averages (○) or genomic EBV (●), for A2 (a, c) and Af (b, d) scenarios. A2 = including only 2 generations of ancestors; Af = including all ancestors; TR = reduced data set; T1980, T1985, T1990, T1995, and T2000 = threshold for exclusion of records before 1980, 1985, 1990, 1995, and 2000, respectively.

for the sons of local bulls. Again, deletion of historical data improved the evaluations only for sons of foreign bulls. The negative effects of inclusion of pedigree data for bulls with foreign sires may be due to the fact that bulls selected to sire sons in Israel are a highly selected sample of all foreign Holstein bulls.

In a study involving a simplistic simulation and multistep genomic method, Neuner et al. (2009) analyzed the effect of a pedigree 3 and 4 generations deep in a simulated population and found differences on variance components estimation, but no influence was observed on accuracy of predictions. Our study indicates that depth of pedigree had a very small influence over validation reliability of genomic evaluations in US Holstein data and hardly affected predictions for Israeli Holstein and PIC data.

Lower values of $R^2$ for yield compared with percentage traits have been reported for a small set of 117 validation bulls in a Brown Swiss dairy population (Wiggans et al., 2011). However, the reliabilities increase when the number of genotyped animals also increases (VanRaden et al., 2009).

Correlations between corrected phenotypes and EBV or GEBV for the 2 traits in pigs are given in Figure 6. For all truncation levels, accuracy of evaluations for the genotyped animals was almost identical, with a slight improvement at T2003. Differences in graphs with full or truncated pedigrees were very small or none. Thus,

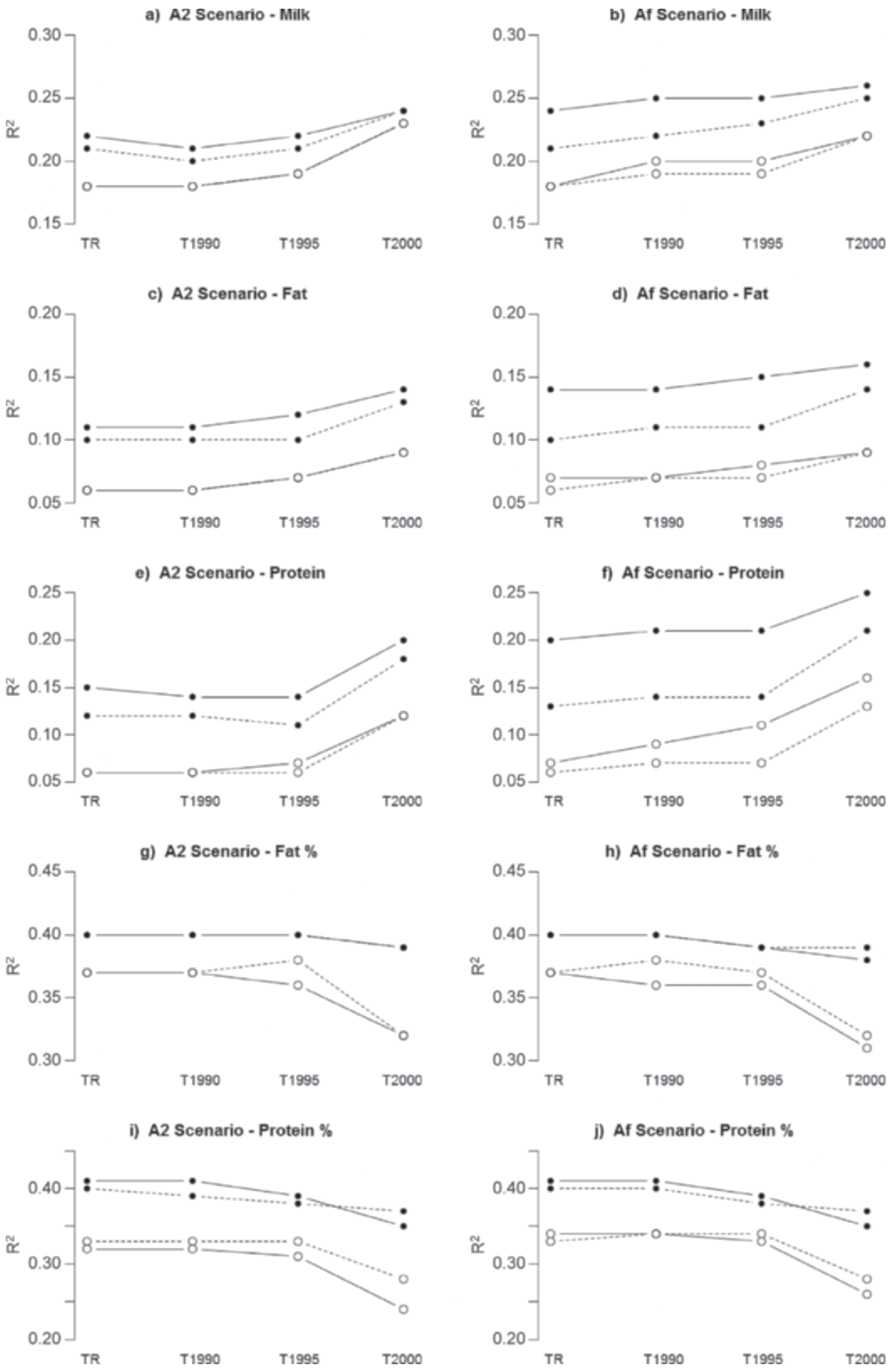

**Figure 2.** Coefficients of determination ($R^2$) of deregressed evaluations of 135 Israeli Holstein genotyped bulls with no daughters with records in 2006, but with ≥20 daughters with records in 2011 on parent averages (○), or genomic EBV for parity 1 (●), for A2 (a, c, e, g, i) and Af (b, d, f, h, j) scenarios for milk yield (a, b), fat yield (c, d), protein yield (e, f), fat percentage (g, h), and protein percentage (i, j). Results are presented for models that include unknown parent groups (UPG; —) or do not include UPG (----). A2 = including only 2 generations of ancestors; Af = including all ancestors; TR = reduced data set; T1990, T1995, T2000 = threshold for exclusion of records before 1990, 1995, and 2000, respectively.

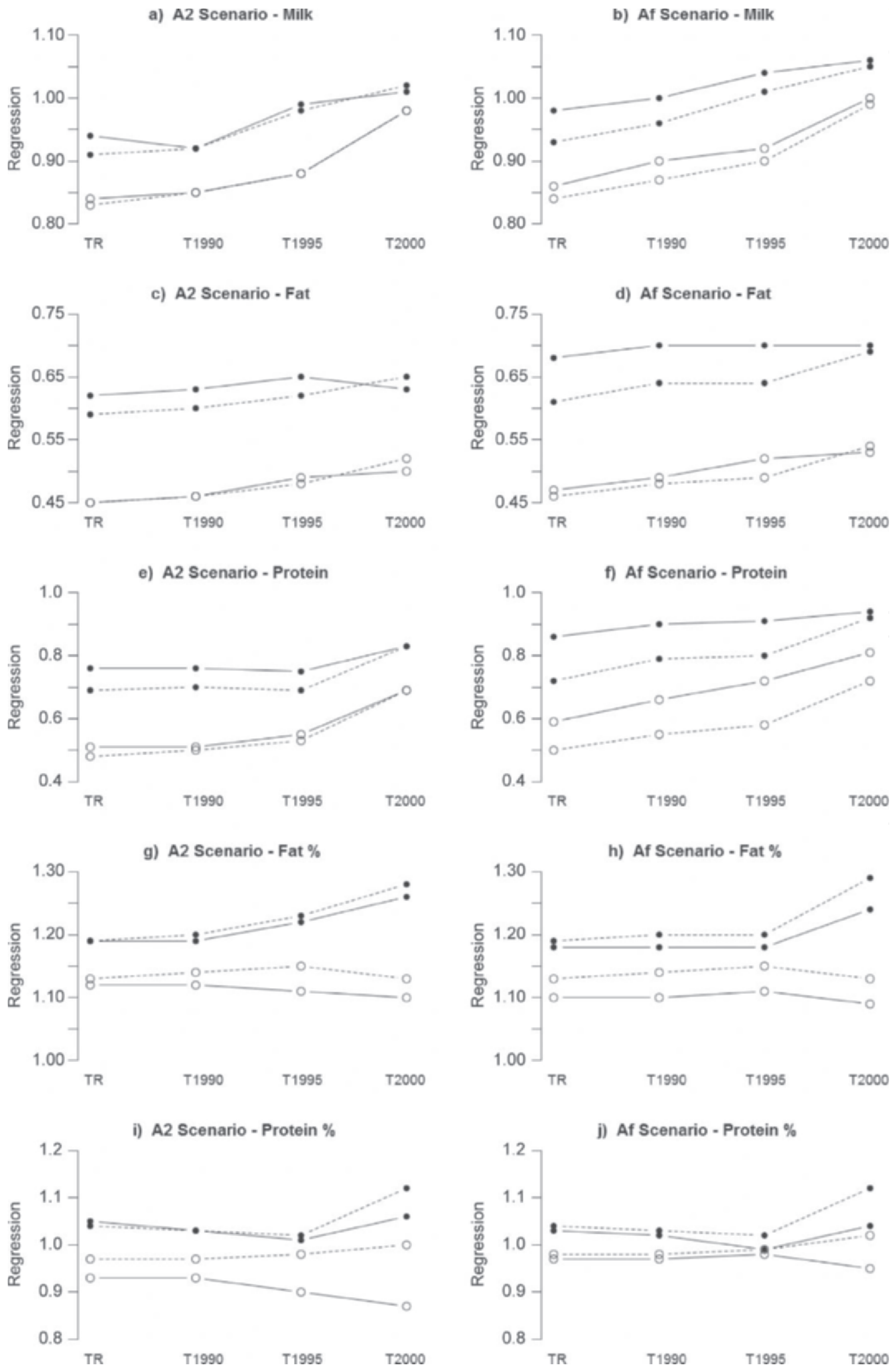

**Figure 3.** Regression of deregressed evaluations of 135 Israeli Holstein genotyped bulls with no daughters with records in 2006 but with ≥20 daughters with records in 2011 on parent averages (○) or genomic EBV for parity 1 (●), for A2 (a, c, e, g, i) and Af (b, d, f, h, j) scenarios for milk yield (a, b), fat yield (c, d), protein yield (e, f), fat percentage (g, h), and protein percentage (i, j). Results are presented for models that include unknown parent groups (UPG; —) or do not include UPG (----). A2 = including only 2 generations of ancestors; Af = including all ancestors; TR = reduced data set; T1990, T1995, T2000 = threshold for exclusion of records before 1990, 1995, and 2000, respectively.

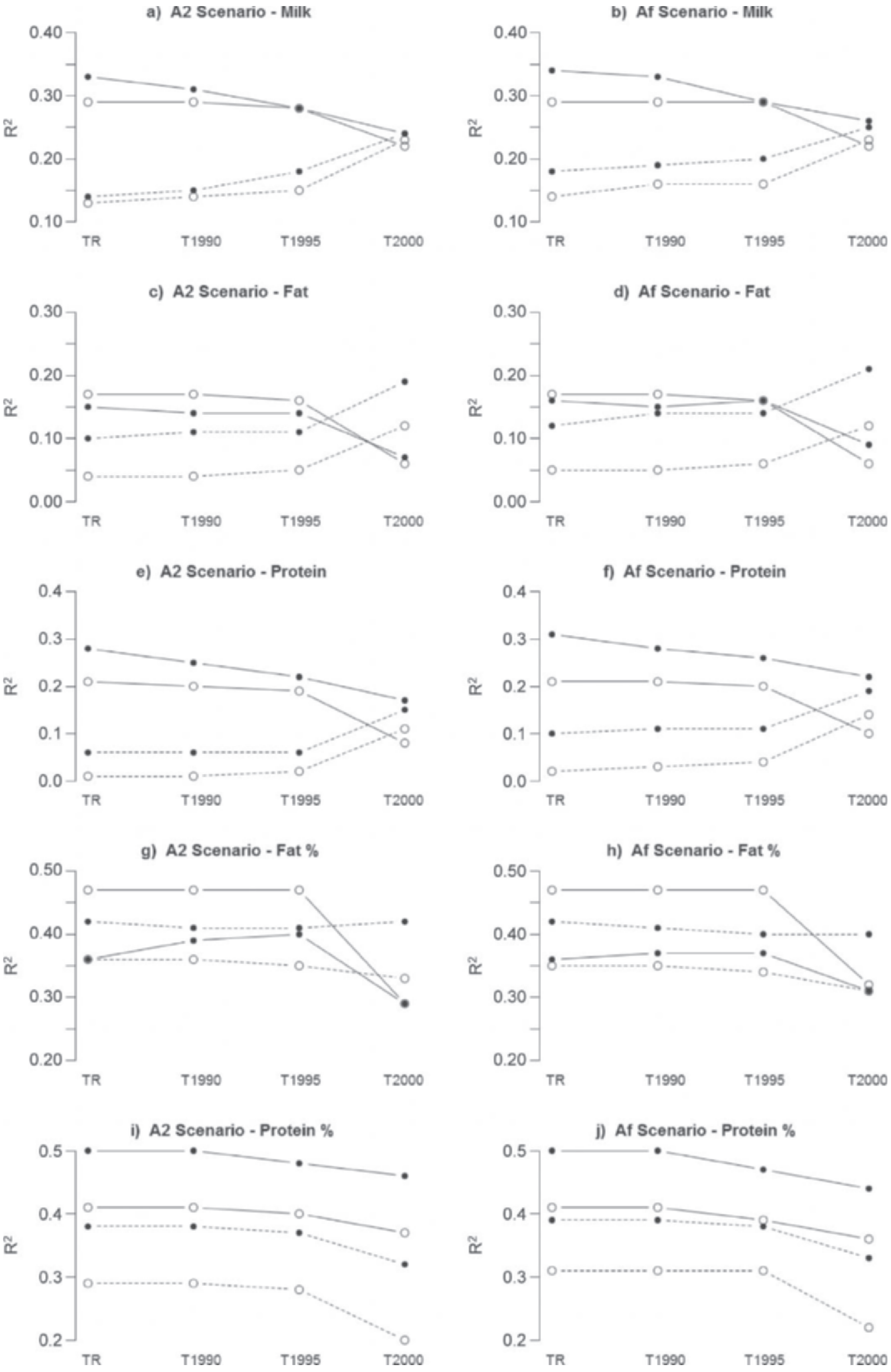

**Figure 4.** Coefficients of determination ($R^2$) of deregressed evaluations of 135 Israeli Holstein genotyped bulls with no daughters with records in 2006, but with ≥20 daughters with records in 2011 on parent averages (○), or genomic EBV for parity 1 (●), for A2 (a, c, e, g, i) and Af (b, d, f, h, j) scenarios for milk yield (a, b), fat yield (c, d), protein yield (e, f), fat percentage (g, h), and protein percentage (i, j). Results are presented separately for 38 bulls with Israeli sires (—) and 97 bulls with foreign sires (----). A2 = including only 2 generations of ancestors; Af = including all ancestors; T1990, T1995, T2000 = threshold for exclusion of records before 1990, 1995, and 2000, respectively.

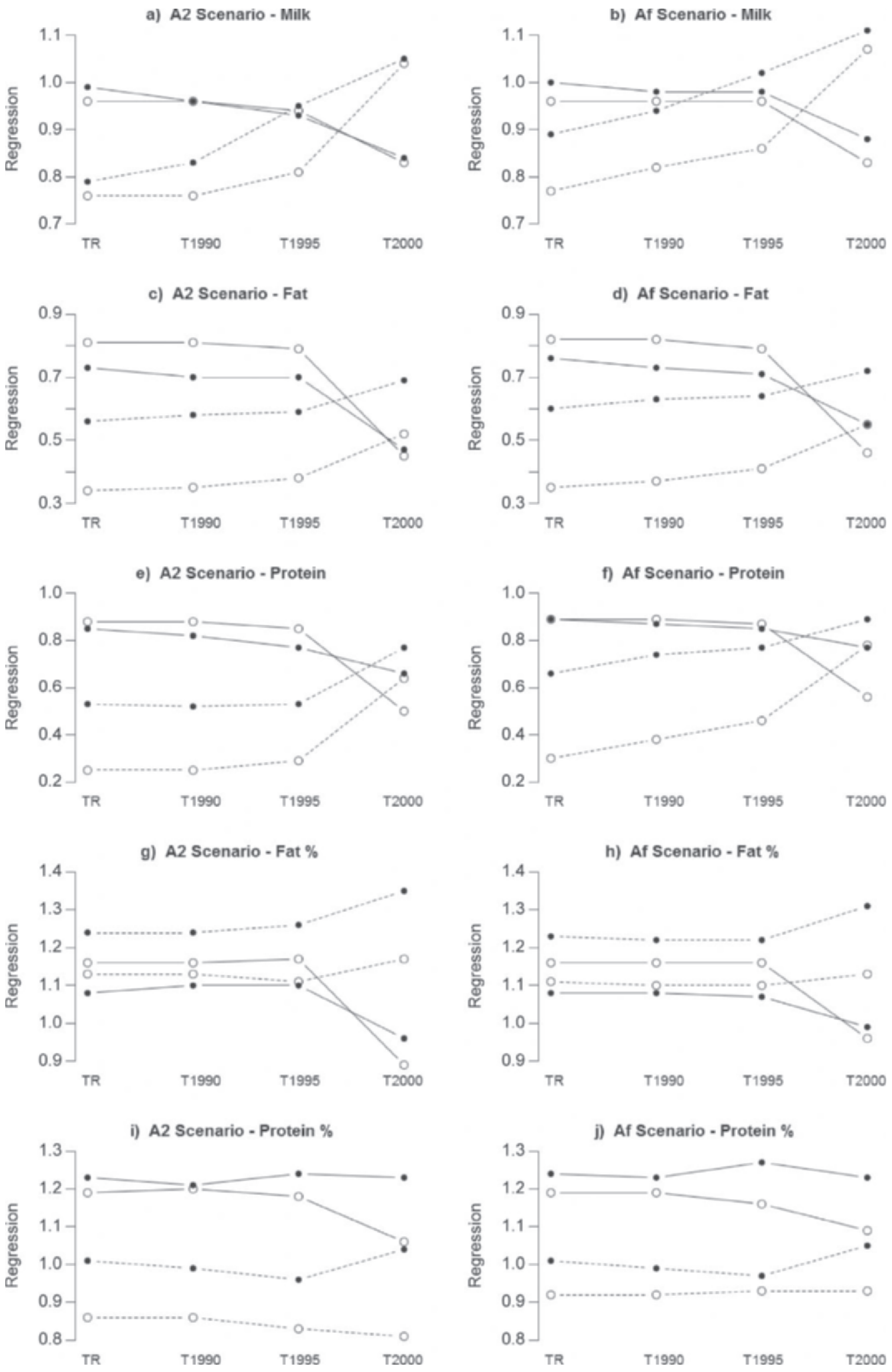

**Figure 5.** Regression of deregressed evaluations of 135 Israeli Holstein genotyped bulls with no daughters with records in 2006, but with ≥20 daughters with records in 2011 on parent averages (○), or genomic EBV for parity 1 (●), for A2 (a, c, e, g, i) and Af (b, d, f, h, j) scenarios for milk yield (a, b), fat yield (c, d), protein yield (e, f), fat percentage (g, h), and protein percentage (i, j). Results are presented separately for 38 bulls with Israeli sires (—) and 97 bulls with foreign sires (----). A2 = including only 2 generations of ancestors; Af = including all ancestors; TR = reduced data set; T1990, T1995, T2000 = threshold for exclusion of records before 1990, 1995, and 2000, respectively.

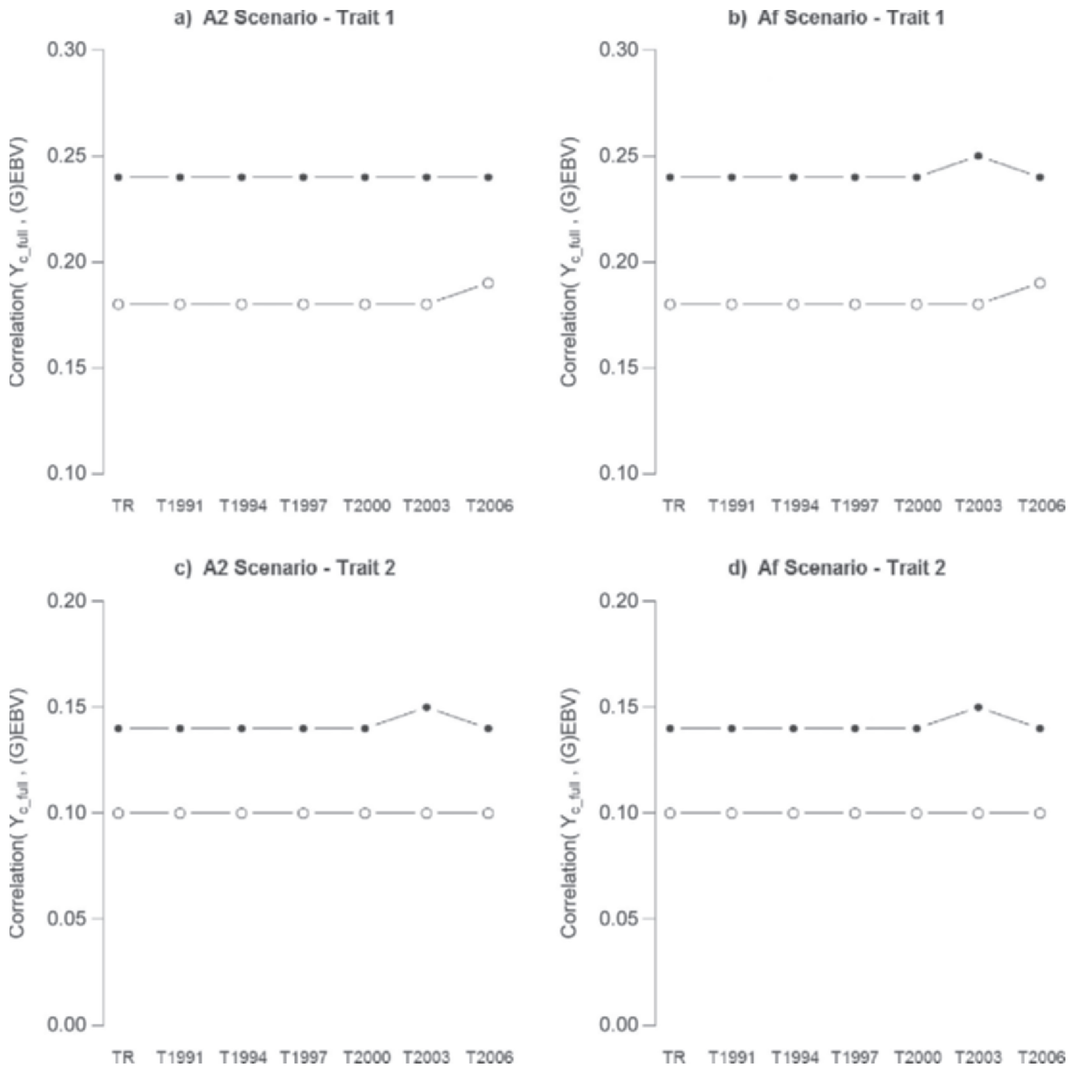

**Figure 6.** Correlations between corrected phenotypes ($Y_{c\_full}$) and EBV (○) or genomic EBV (●) for 1,034 genotyped pigs born in 2010 and 2011, in A2 (a, c) and Af (b, d) scenarios for litter size (trait 1; a, b) and numbers of stillborn (trait 2; c, d). A2 = including only 2 generations of ancestors; Af = including all ancestors (Af); TR = reduced data set; T1991, T1994, T1997, T2000, T2003, and T2006 = threshold for exclusion of records before 1991, 1994, 1997, 2000, 2003, and 2006, respectively.

more than 7 yr of historical data, or about 2 generations, did not improve the accuracy.

### Computing Costs

Computing costs of genomic evaluations are shown in Table 3. With less data, time per round was generally smaller; discrepancies were due to inaccuracy of a timing routine. When the phenotypes were deleted, the number of iterations stayed level or increased when all pedigree data were included and stayed level or decreased when the pedigrees were reduced. When all pedigree data were retained, limiting the records to 2 generations (T1995 in US and Israeli Holsteins or T2003 in pigs) decreased computation time by 59% for US Holsteins, by 33% for Israeli Holsteins, and by 27% for pigs. When the pedigrees were limited to 2 generations before animals with records, the computations decreased by 70, 36, and 38%, respectively. With a large number of genotypes, computing time will be more dependent on the number of genotypes. However, the inverse of $\mathbf{A}_{22}$ will be more sparse with smaller pedigrees.

Different data sets and models used in our study helped us to better understand the influence of previous generations in the evaluation of young genotyped animals. The effect of removing old generations on predictive ability appears to be data-structure driven. However, it was possible to remove at least a few generations from all data sets used here, and for almost all traits,

**Table 3.** Average computing performance at convergence for single-step genomic BLUP (ssGBLUP) evaluations[1]

| | | A2 | | | Af | | |
|---|---|---|---|---|---|---|---|
| Data | Threshold[2] | Iterations | Time per iteration (s) | Total time (s) | Iterations | Time per iteration (s) | Total time (s) |
| US Holstein | TR | 495 | 8.73 | 4,321 | 509 | 9.75 | 4,963 |
| | T1980 | 473 | 6.52 | 3,084 | 512 | 7.61 | 3,896 |
| | T1985 | 465 | 5.81 | 2,702 | 539 | 6.44 | 3,471 |
| | T1990 | 398 | 4.04 | 1,608 | 591 | 5.10 | 3,014 |
| | T1995 | 382 | 3.42 | 1,306 | 659 | 3.14 | 2,069 |
| | T2000 | 325 | 1.93 | 627 | 636 | 2.24 | 1,425 |
| Israeli Holstein | TR | 435 | 1.63 | 709 | 432 | 1.61 | 696 |
| | T1990 | 453 | 1.41 | 639 | 434 | 1.18 | 512 |
| | T1995 | 450 | 1.01 | 455 | 491 | 0.89 | 437 |
| | T2000 | 489 | 0.61 | 298 | 589 | 0.69 | 406 |
| PIC | TR | 873 | 2.93 | 2,558 | 874 | 2.87 | 2,508 |
| | T1991 | 825 | 3.02 | 2,492 | 863 | 2.90 | 2,503 |
| | T1994 | 815 | 2.73 | 2,225 | 875 | 2.72 | 2,380 |
| | T1997 | 717 | 2.59 | 1,857 | 860 | 2.77 | 2,382 |
| | T2000 | 682 | 2.59 | 1,766 | 877 | 2.73 | 2,394 |
| | T2003 | 676 | 2.35 | 1,589 | 879 | 2.10 | 1,846 |
| | T2006 | 691 | 1.87 | 1,292 | 820 | 1.79 | 1,468 |

[1] A2 = short pedigree (including only 2 generations of ancestors) and Af = deep pedigree (including all ancestors).

[2] TR = reduced data set and contains data up to 2007 for US Holsteins, 2006 for Israeli Holsteins, and 2009 for PIC data (pig data from Genus, Hendersonville, TN). The other thresholds contain data from the indicated year up to 2007, 2006, and 2009, respectively, for the 3 data sets.

without reducing validation reliabilities and improving the computing performance of the evaluations. The option for data reduction will depend on the objectives of the evaluation. According to Jamrozik and Schaeffer (1991), the inclusion of all data in genetic studies is important if the interest is in estimating genetic trends over time. In contrast, if the interest is to investigate the selection response, the use of the last 2 (discrete) or 4 (overlapping) generations had no significant effect in traditional evaluations of a simulated chicken population (Mehrabani-Yeganeh et al., 1999). Furthermore, if the objective is to predict genomic breeding values for young genotyped animals, the addition of one extra generation of pedigree did not improve the prediction accuracies in simulated data (Neuner et al., 2009).

## CONCLUSIONS

Retaining only 2 or 3 generations of phenotypic records and an extra 2 generations of pedigree records did not decrease the accuracy of evaluations for young genotyped animals and did decrease computing costs. When the population is small and contains a mix of local and external animals, additional generations of phenotypic records can increase the accuracy for progeny of local animals and decrease the accuracy for progeny of imported animals. Analyzing realized accuracy with various levels of data truncation may uncover problems with the analysis model and lead to model refinements.

## ACKNOWLEDGMENTS

This research was supported by grants from the Israel Dairy Board, the Chief Scientist of the Israeli Ministry of Agriculture and Rural Development, and Binational Agricultural Research and Development Fund (BARD) Research Project IS-4394-11R. Support from the Pig Improvement Company (Genus, Hendersonville, TN) and the Holstein Association (Brattleboro, VT) is gratefully acknowledged. The authors thank H. Wang (Genus PIC, Hendersonville, TN) and A. Schein and N. Avidan (both from Pharmacogenetics and Translation Medicine Center, The Rappaport Institute for Research in the Medical Sciences, Technion, Haifa, Israel) for helping with the data sets.

## REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752.

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673–2679.

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565–1571.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5.

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. Can. Catal. Publ. Data. Univ. Guelph, Guelph, ON, Canada.

Jamrozik, J., and L. R. Schaeffer. 1991. Procedures for updating solutions to animal models as data accumulate. J. Dairy Sci. 74:1993–2000.

Mehrabani-Yeganeh, H., J. P. Gibson, and L. R. Schaeffer. 1999. Using recent versus complete pedigree data in genetic evaluation of a closed nucleus broiler line. Poult. Sci. 78:937–941.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Misztal, I., I. Aguilar, A. Legarra, and T. J. Lawlor. 2010. Choice of parameters for single-step genomic evaluation for type. J. Dairy Sci. 93(Suppl. 1):533. (Abstr.)

Misztal, I., Z. G. Vitezica, A. Legarra, I. Aguilar, and A. A. Swan. 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130:252–258.

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.

Neuner, S., C. Edel, R. Emmerling, G. Thaller, and K. Gotz. 2009. Precision of genetic parameters and breeding values estimated in marker assisted BLUP genetic evaluation. Genet. Sel. Evol. 41:26.

Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z. Hu, H. H. Kerstens, A. S. Law, H. Megens, D. Milan, D. J. Nonneman, G. A. Rohrer, M. F. Rothschild, T. P. L. Smith, R. D. Schnabel, C. P. Van Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. A. M. Groenen. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS ONE 4:e6524.

Tsuruta, S., I. Misztal, I. Aguilar, and T. J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J. Dairy Sci. 94:4198–4204.

Tsuruta, S., I. Misztal, L. Klein, and T. J. Lawlor. 2002. Analysis of age-specific predicted transmitting ability for final scores in Holsteins with a random regression model. J. Dairy Sci. 85:1324–1330.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

VanRaden, P. M., C. P. Van Tassel, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb.) 93:357–366.

Weller, J. I., and E. Ezra. 2004. Genetic analysis of the Israeli Holstein dairy cattle population for production and nonproduction traits with a multitrait animal model. J. Dairy Sci. 87:1519–1527.

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94:3202–3211.

Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet. Sel. Evol. 43:23.