Relation Between Accuracies of Genomic Predictions and Ancestral Links to the Training Data

M.S. Lund[a*], G. Su[a], U.S. Nielsen[b] and G.P. Aamand[b]

[a] Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, Aarhus University, DK-8830, Tjele, Denmark

[b] Danish Agricultural Advisory Service, DK-8200 Aarhus N, Denmark

[*] Corresponding author. Email: Mogens.Lund@agrsci.dk; Tel: 45 89991222; Fax: 45 89991300

Abstract

The aim of this study is to investigate the impact of ancestral links to training data on direct genomic estimated breeding value (GEBV) and on the index blending GEBV and parent average EBV (PA). The data in the analysis included 3,330 Nordic Holstein bulls with both published conventional EBV and single nucleotide polymorphism (SNP) markers (genotyped using Illumina Bovine SNP50 BeadChip). Two training dataset were created. One included the sires of predicted animals; the other excluded the sires. The traits under analysis were fertility, protein and udder-health. The response variables to estimate SNP effects were the official EBV with weighting factor of 1/(1-reliability of EBV). Reliability of GEBV was assessed using squared correlation between GEBV and EBV ($r^2_{DGV,EBV}$), and expected reliability from model, based on a 5-fold cross validation. When sires of predicted animals were in training data, $r^2_{GEBV,EBV}$ were 0.412, 0.412 and 0.435, expected reliability were 0.566, 0.528 and 0.557, correlation between GEBV and PA were 0.709, 0.584 and 0.679 for fertility, protein and udder-health, respectively. In the scenario that sires of predicted animals were excluded from training data, $r^2_{GEBV,EBV}$ were 0.326, 0.367 and 0.335, expected reliability were 0.487, 0.493 and 0.486, correlation between GEBV and PA were 0.536, 0.466 and 0.520 for fertility, protein and udder-health, respectively. Blending GEBV and PA using an index $I = b_1GEBV + b_2PA$, the difference in reliability of the index between the two scenarios was small, dependent on reliability of PA. These results indicate that ancestral links to the training data have a strong effect on accuracy of GEBV and the correlation between GEBV and PA. The dependency of correlation between GEBV and PA on ancestral links to the training data should be taken into consideration when calculating an index blending GEBV and PA.

Keywords: genomic estimated breeding value, genomic selection, parent average EBV

1. Introduction

Several studies based on real data from dairy cattle have reported that accurate breeding value can be predicted using genome-wide dense markers (e.g., Hayes *et al.,* 2009; Su *et al,* 2009; VanRaden *et al.,* 2009). In a genomic prediction setting, the accuracy of genomic selection of young candidates can be enhanced by including parent average (PA) information. It is well-known that the accuracy of genomic estimated breeding value (GEBV) is dependent on linkage disequilibrium (LD) between markers and the QTLs affecting the traits. However, Habier *et al.* (2007) reported that markers can capture genetic relationships among genotyped animals, thereby affecting accuracies of GEBV. It can be hypothesized that genetic links between predicted animals and the animals in training data have an impact on 1) the accuracy of GEBV of the predicted animals, the correlation between GEBV and parent average EBV, 3) the gain by including PA.

The objective of this study is to investigate the impact of ancestral links to the training data on genomic prediction, using training datasets including or excluding sires of predicted animals, based on the data from Nordic Holsteins.

2. Materials and Methods

2.1 Data

Danish and Swedish Holstein bulls from 258 half-sib families, born during years from 1986 to 2004 were genotyped using Illumina Bovine

SNP50 BeadChip (Illumina, San Diego, CA). Published conventional EBV provided by Danish Cattle Federation (2006) were used as response variables to estimate SNP effects for genomic prediction. After the editing, there were 3,330 bulls and 38,134 SNP (single nucleotide polymorphism) markers available. Two training dataset were created. One included the sires of predicted animals; the other excluded the sires. The traits under analysis were fertility, protein and udder-health.

2.2 Statistical model

SNP effects were estimated using conventional EBV as response variables weighted by a factor of 1/(1-reliability of EBV), applying the following model:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{m} \mathbf{X}_i \mathbf{q}_i v_i + \mathbf{e}$$

where $\mathbf{y}$ is the vector of published EBV, μ is the intercept, m is the number of SNP markers, $\mathbf{q}_i$ is the vector of scaled SNP effects (scaled by standard deviation) of marker i with $\mathbf{q}_i \sim N(\mathbf{0}, \mathbf{I})$, v_i ($v_i > 0$) is a scaling factor (standard deviation) for SNP effects of marker i, and $\mathbf{e}$ is the vector of residual with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. The effects of SNP alleles of marker i are the products of v_i and $\mathbf{q}_i$.

Scaling factors v_i were assumed to have a common prior distribution, which leads to a slight or moderate differentiation between small and large effects of markers. It was assumed to be a positive half-normal distribution,

$$v_i \sim TN(\mathbf{0}, \sigma_v^2), \; v_i > 0$$

The prior distributions of μ and σ_v^2 were assumed to be improper uniform distributions. The genomic estimated breeding value (GEBV) for individual k was defined as the sum of predicted effects of SNP over all markers

$$GEBV_k = \hat{\mu} + \sum_{i=1}^{m} \mathbf{x}_{i(k.)} \mathbf{q}_i v_i$$

Detailed description of the algorithm were presented by Villumsen *et al.* (2009) and Meuwissen and Goddard (2004)

2.3 Cross validation of reliability of GEBV

The accuracy of GEBV were evaluated using a 5-fold cross validation. In the cross validation, 134 half-sib families which have at least one bull born after 1993 were divided into 5 test datasets according to according to birth-year for the most of half-sibs. In each fold cross validation, the whole data excluded one test dataset to form a training dataset which was used to estimate marker effects and predict genomic breeding values of the "left out" animals. See detailed description on cross-validation in Su *et al.* (2009).

Accuracy of direct GEBV was assessed using two measurements. One was squared correlation between GEBV and published conventional EBV ($r^2_{GEBV,EBV}$) in test datasets, where both GEBV and EBV were adjusted for birth-year mean to account for genetic trend, i.e., within-year squared correlation. The other was expected genomic reliability, obtained from prediction error variance (PEV) which was measured as posterior variance of each GEBV. To be consistent with real life scenario, the bulls which had sons or grandsons in the training data were excluded from the validation.

2.4 Index combing GEBV and parent average EBV (PA)

An index combining GEBV and EBV was constructed as,

$$I = b_1 GEBV + b_2 PA$$

b_1 and b_2 were estimated using the following equation system

$$b_1 V_{GEBV} + b_2 V_{GEBV,PA} = V_{GEBV}$$
$$b_2 V_{GEBV,PA} + b_2 V_{PA} = V_{PA}$$

Replace variance/covariance by reliability (R^2_i) and correlation coefficient ($r_{GEBV,PA}$),

$b_1 + b_2 r_{GEBV,PA} R^2_{PA}/R^2_{GEBV} = 1$
$b_2 r_{GEBV,PA} R^2_{GEBV}/R^2_{PA} + b_2 = 1$

In the case that $r_{GEBV,PA} > R_{PA}/R_{GEBV}$, b_1 was fixed at 1 and b_2 at 0. Reliability of the index is $R^2_I = b_1 R^2_{GEBV} + b_2 R^2_{PA}$

3. Results and Discussion

3.1 Accuracy of direct GEBV

Sire status (being in or out of training data) have a significant impact on accuracy of GEBV (Table 1). When sires of predicted animals were in training data, $r^2_{DGV,EBV}$ were 0.412, 0.412 and 0.435, expected reliability were 0.566, 0.528 and 0.557 for fertility, protein, udder-health, respectively. In the scenario that sires of predicted animals were excluded from training data, $r^2_{DGV,EBV}$ were 0.326, 0.367 and 0.335, expected reliability were 0.487, 0.493 and 0.486.

3.2 Correlation between direct GEBV and PA

Table 2 presents the correlations of GEBV with sire EBV, maternal grandsire (MGS) EBV and PA, where PA was calculated as PA = $0.5EBV_{sire} + 0.25EBV_{MGS}$. Including sires in training data increased the correlation between GEBV of candidates and EBV_{sire}, consequently increased the correlation between GEBV and PA, more profound for fertility and udder-health than for protein. The correlations between GEBV and PA increased from 0.536, 0.466 and 0.520 in scenario that sires were excluded from training data to 0.709, 0.584 and 0.679 in scenario that sires were included in training data, for fertility, protein and udder-health, respectively.

3.3 Reliability of index (R^2_I) combing GEBV and PA

Due to different correlations between GEBV and PA, the gains of genetic prediction by including PA were different in the scenarios with or without sires in training data (Table 3a-d). When average reliability of EBV of bulls in genotyped data were used as reliability of sire EBV and MGS EBV, no gain can be obtained by including PA in scenario with sire in training data, except for protein. This resulted in a small difference in R^2_I between the two scenarios. When using reliability of sire EBV and MGS EBV published in March 2009 (which the calculation of correlation between GEBV and PA based on), the gain by including PA was obtained in both scenarios, more in scenario without sires in training data. Consequently, no difference in R^2_I between the two scenarios can be observed.

4. Discussion

The results from this study indicate that genetic ties between training dataset and test dataset have a strong influence on the accuracy of GEBV. In the scenario where sires are in training data, genomic breeding value are estimated using both information of linkage disequilibrium (LD) in whole population and sire genetic information, while in the scenario where sires are not in training data, GEBV are obtained only (or almost) from LD information. Sire in training dataset provides more information for genomic prediction of the sons, consequently, higher accuracy of GEBV than the situation where sire is absent from training data. In a study on genomic prediction in mice, Legarra *et al.* (2008) presented the correlation between observations and predictions when 50/% of full-sibs kept in training data was two times as high as the correlations when no full-sibs were in training data.

Accuracy of genomic prediction can be enhanced using an index combining direct GEBV and PA (VanRaden *et al.,* 2009). The gain by including PA information depends on reliability of GEBV and PA as well as correlation between GEBV and PA. Because genetic information of relatives in training data have a contribution to GEBV of a candidate, correlation between GEBV and PA is dependent on genetic link between predicted animals and the animals in training data. The stronger link, the higher correlation, the less gain in genetic evaluation can be obtained by including PA information in genomic selection frame.

It was found that the influence of sire status on accuracy of GEBV and correlation between GEBV and PA is larger for fertility and udder-health than that for protein. It implies that sire genetic information of relatives is relatively more important for low heritability traits than for high heritability trait with regard to predicting genomic breeding value. Correspondingly, the exclusion of sire genetic information from GEBV results in a larger reduction in correlation between GEBV and PA for low heritability trait than high heritability trait.

5. Conclusion

The existence of sire in training data, on one hand, increases accuracy of direct GEBV, on the other hand, decrease the gain by including PA information due to high correlation between GEBV and PA. Consequently, there is a small difference between scenarios where sire is in or out of training data with regard to the reliability of an index combining GEBV and PA. The importance is that the dependency of correlation between GEBV and PA on genetic link to training data should be taken into account in calculation of such an index.

Reference

Habier, D., Fernando, R.L. & Dekkers, J.C.M. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics 177,* 2389–2397.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci. 92,* 433-443.

Legarra, A., Robert-Granie, C., Manfredi, E. & Elsen, J.M. 2008. Performance of Genomic Selection in Mice. *Genetics 180,* 611–618.

Meuwissen, T.H.E. & Goddard, M.E. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol. 36,* 261–279.

Su, G., Guldbrandtsen, B. & Lund, M.S. 2009. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein Population. (Submitted to *J. Dairy Sci.*).

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

Villumsen, T.M., Janss, L. & Lund, M.S. 2009. The importance of haplotype lenght and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet. 126,* 3-13.

Table 1. Squared correlation between GEBV and EBV ($r^2_{GEBV,EBV}$) for all bulls in the test data, and expected reliability obtained from prediction error variance (posterior variance of each GEBV).

Trait	Sires in training data		Sires out of training data	
	Exp. reliability	$r^2_{GEBV,EBV}$	Exp. reliability	$r^2_{GEBV,EBV}$
Fertility	0.566	0.412	0.487	0.326
Protein	0.528	0.412	0.493	0.367
Udder-health	0.557	0.435	0.486	0.335

Table 2. Correlation of GEBV with sire EBV, maternal grandsire EBV and parent average EBV(PA).

Trait	Sires included in reference data			Sires excluded from reference data		
	Cor_sire	Cor_mgs	Cor_pa	Cor_sire	Cor_mgs	Cor_pa
Fertility	0.618	0.360	0.709	0.410	0.387	0.536
Protein	0.522	0.234	0.584	0.395	0.232	0.466
Udder-health	0.625	0.281	0.697	0.392	0.298	0.500

Table 3a. Reliability of index blending GEBV and PA (PA = 0.5EBV_{sire} +0.25 EBV_{mgs}), using average reliability of genotyped bulls as reliability of EBV_{sire} and EBV_{mgs}, and using $r^2_{GEBV,EBV}$ as reliability of GEBV.

Data	Trait	R^2_{PA}	b_{gebv}	b_{pa}	R^2_I
Sire in training data	Fertility	0.216	0.981	0.040	0.413
	Protein	0.292	0.876	0.392	0.475
	Udder-health	0.237	0.953	0.101	0.438
Sire out of training data	Fertility	0.216	0.866	0.430	0.375
	Protein	0.292	0.887	0.537	0.482
	Udder-health	0.237	0.869	0.483	0.406

Table 3b. Reliability of index blending GEBV and PA (PA = 0.5EBV_{sire} +0.25 EBV_{mgs}), using average reliability of genotyped bulls as reliability of EBV_{sire} and EBV_{mgs}, and using expected reliability from model as reliability of GEBV.

Data	Trait	R^2_{PA}	b_{gebv}	b_{pa}	R^2_I
Sire in training data	Fertility	0.216	1	0	0.566
	Protein	0.292	0.879	0.309	0.555
	Udder-health	0.237	1	0	0.557
Sire out of training data	Fertility	0.216	0.885	0.288	0.493
	Protein	0.292	0.857	0.481	0.563
	Udder-health	0.237	0.863	0.382	0.510

Table 3c. Reliability of index (r^2_I)blending GEBV and PA (PA = 0.5EBV_{sire} +0.25 EBV_{mgs}), using reliability of EBV_{sire} and EBV_{mgs} published in March 2009 (which the calculation of correlation between GEBV and PA based on), and using $r^2_{GEBV,EBV}$ as reliability of GEBV.

Data	Trait	R^2_{PA}	b_{gebv}	b_{pa}	R^2_I
Sire in training data	Fertility	0.302	0.910	0.247	0.449
	Protein	0.310	0.882	0.406	0.489
	Udder-health	0.303	0.907	0.243	0.468
Sire out of training data	Fertility	0.302	0.930	0.482	0.449
	Protein	0.310	0.900	0.544	0.499
	Udder-health	0.303	0.920	0.516	0.465

Table 3d. Reliability of index blending GEBV and PA (PA = 0.5EBV_{sire} +0.25 EBV_{mgs}), using reliability of EBV_{sire} and EBV_{mgs} published in March 2009 (which the calculation of correlation between GEBV and PA based on), and using expected reliability from model as reliability of GEBV.

Data	Trait	R^2_{PA}	b_{gebv}	b_{pa}	R^2_I
Sire in training data	Fertility	0.302	0.973	0.055	0.568
	Protein	0.310	0.875	0.333	0.565
	Udder-health	0.303	0.953	0.099	0.561
Sire out of training data	Fertility	0.302	0.864	0.412	0.545
	Protein	0.310	0.859	0.495	0.577
	Udder-health	0.303	0.860	0.455	0.556