Iranian Society of Animal Science

Journal of Livestock Science and Technologies

Print ISSN: 2322-3553
Online ISSN: 2322-374X

Shahid Bahonar University of Kerman

Trend of bias in prediction of genomic estimated breeding values due to selective genotyping in genomic selection schemes in consecutive generations

Jabar Jamali[1], Seyyed Hasan Hafezian[1]*, Mohsen Gholizadeh[1] and Alireza Ehsani[2]

[1]Department of Animal Science, Faculty of Animal and Aquatic Science, Sari Agricultural Sciences and Natural Resources University, Sari, Iran.
[2]Department of Animal Science, Tarbiat Modares University, Tehran, Iran.

*Corresponding author,
E-mail address:
hassanhafezian@yahoo.com

Received: 30 May. 2019,
Accepted: 09 Sep. 2019,
Published online: 26 Dec. 2019.

Abstract The aim of this study was to investigate the trend of bias in genomic estimated breeding values (GEBVs) arising from selective genotyping of the candidate population in an ongoing selection scheme. The bias was calculated as the regression of true breeding values (TBVs) on GEBVs. A simulation study was performed under two scenarios with selection intensities (SI) of 0.798 and 1.755 for three traits with heritability (h^2) of 0.1, 0.25 and 0.4 in 10 consecutive generations. Regression of TBVs on GEBVs was close to one for the first generation when selective genotyping was random, and it continuously receded from one as selection shifted to choose animals with high EBVs from generations 2 to 10. Biasedness became larger with increased SI and decreased h^2. Further, biasedness increased over the generations but the rate of change in biasedness decreased dramatically after the second generation and became almost steady after generation 4 which may be due to Bulmer effect. The findings showed that scaling down the GEBVs, using a scale parameter, might help removing biasedness in generation 4 onwards.

Keywords: genomic selection, selective genotyping, bias

Introduction

Development of the genome wide dense markers has made it possible to implement the marker information in animal breeding programs (VanRaden et al., 2009). Genomic selection (GS) will result in greater response to selection mainly due to the shorter generation interval that achieve by GS than that obtain by traditional selection (VanRaden et al., 2009). Genomic selection exploits linkage disequilibrium (LD) between markers and quantitative trait loci (QTLs) (Meuwissen et al., 2001; Meuwissen and Goddard, 2004). Because the markers cover the whole genome, all casual variants of the traits are covered by the markers and this method is potentially capable of justifying all genetic variation. The availability of an appropriate model for estimating the effect of markers plays a key role in genomic selection programs and has a significant effect on accurate and unbiased prediction of genetic merits of young animals without phenotypic record (Ehsani et al., 2010). As a result of the recombination effect, linkage phase between markers and QTLs breakdowns over generations, it is essential to re-estimate the marker effect after several generations of genomic evaluation. Moreover, changes in variance components of subsequent generations due to selection (Van Grevenhof et al., 2012) represent one of the most important sources of biased predictions of genetic merits using the best linear unbiased prediction (BLUP). The basic assumption behind the BLUP is the random contribution of parental candidates to the next generation (Robinson, 1991). Mainly due to the cost of genotyping, two types of selection are applied in GS breeding schemes compared to the conventional progeny test scheme. Pre-

selection of animals for genotyping is mainly based on the estimated breeding values (EBVs) using information from their relatives and a final selection of elite animals based on the genomic estimated breeding values (GEBVs). The preselection of candidates for genotyping which is called selective genotyping (SG) violates the basic assumption behind BLUP and causes biased prediction of GEBVs before the final selection (Kuehn et al., 2007). Several studies have confirmed that bias arises as a subset of animals are selected for genotyping (VanRaden et al., 2009; Ducrocq, 2011; Patry and Ducrocq, 2011; Vitezica et al., 2011). Moreover, the statistical methods may influence the magnitude of bias (Vitezica et al., 2011). The source of information as response variable and different mixed model equations has been tested to evaluate their performance to overcome bias. Vitezica et al (2011) implemented a two-step procedure, including a conventional mixed model that uses pedigree relationship matrix and phenotypes, and a model in which daughter yield deviations (DYDs) were used as response variable, compared to three different single-step procedures (a model that uses the genomic and pedigree relationship matrix simultaneously, a model with genetic differences among genotyped and non-genotyped individuals corrected by considering the difference between pedigree and genomic relationships for genotyped animals, and a single-step method with corrected genomic relationship matrix (G) as proposed by Powell et al., 2010). They concluded that single-step procedures were unbiased and more accurate than two-step procedures. Moreover, they showed that a corrected G matrix was more comprehensive because it accounted for SG on prediction of GEBVs.

In spite of the fact that several studies have explored the bias in prediction of GEBVs for a given generation of selection, there is still a necessity to investigate the trend of bias in consecutive ongoing generations of selection. Therefore, this study aimed to explore the trend of bias in a population under selection for consecutive generations.

Materials and methods

Data

To explore the trend of bias in a population under selection, a non-overlapping population with equal sex ratio was simulated using QMSim software (Sargolzaei and Schenkel, 2009). Two selection intensity (SI) scenarios of 0.798 and 1.755 were simulated with two selection proportions (SP) of 50 (SP50) and 10 percent (SP10). To achieve mutation-drift equilibrium, a historical population was simulated that started with 200 individuals with equal number of males and females, mated at random and gradually increased to 4000 and 20000 individuals in generation 1000. The difference in the number of individuals for the two scenarios were to end up with equal number of 1000 males after selection in order to make it statistically possible for comparing their accuracy and bias. Because of the larger number of individuals in the second scenario reaching a mutation-drift, equilibrium was not achieved after 1000 generations. Therefore, the random mating continued to achieve a U-shaped distribution for allelic frequencies in all loci in generation 1500 (Figure 1). Allele frequencies were equal at the beginning of simulation in a bi-allelic model with recurrent mutation procedure for both markers and QTLs in all positions. Recombination rate was 1 percent per centimorgan with a mutation rate equal to 2.5e-5 at both marker and QTL positions. In every generation of historical population, females produced one progeny with equal probability of being male or female. To apply SG, following the construction of historical population, for the 10% SP, 1000 genotyped males were selected from a 10000 males, and for 50% SP from 2000 males originating from the previous generation based on their EBVs. Out of these 1000 genotyped males only 200 animals were selected based on their GEBVs to mate with 2000 and 10000 unselected females to produce the next generation. In the recent population of 10

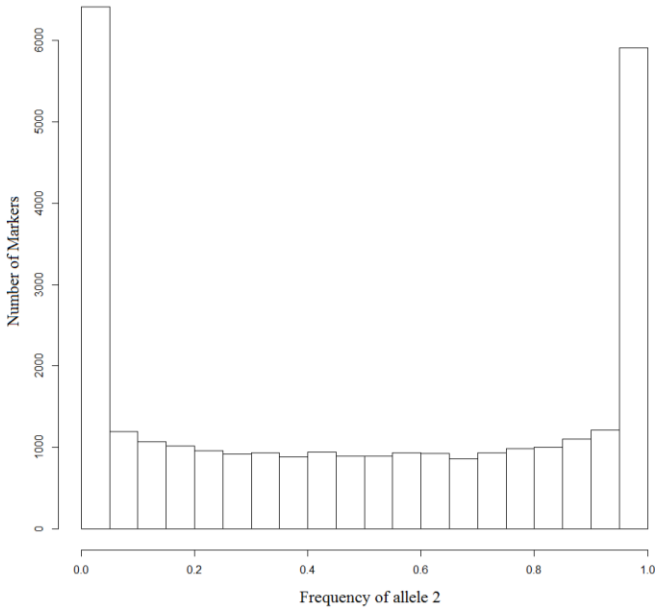

Figure 1. U-shaped distribution of allele frequencies for the second allele in a biallelic model for all positions (mutation-drift equilibrium) for the last historical generation (one generation ahead of artificial selection).

generations, females produced 2 offspring in both scenarios and in each generation to reproduce 4000 and 20000 animalsto keep all selection criteria constant over generations. For each scenario, three traits with heritability (h^2) of 0.1, 0.25 or 0.4 were simulated. Selection of male animals was random at the first generation and was based on higher estimated breeding value animals from generations 2 to 10. The selection of males for genotyping was based only on their EBVs. A genome consisting of 30 chromosomes, 100 cM in length each with 1000 markers and 50 QTLs were simulated for each animal. The total number of markers and QTLs were 30000 and 1500, respectively, spreading randomly across the genome. Each statistics presented here was a mean of 20 replications of simulated population.

Statistical analysis

Estimation of the breeding values for selective genotyping of the animals was calculated externally and introduced to QMSim as an external BV option. The BLUP predictor via an animal model was applied by the Henderson's mixed linear model (Henderson, 1975). The BLUP predictor has the smallest prediction error variance among all possible linear unbiased predictors. Two kinds of information were used: phenotypic records and pedigree data. The numerator relationship matrix (**A**) was used in the following mixed model equations to derive the BLUP of random additive effects for the QTLs:

$$\left[Z'Z + A^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right]\hat{a} = Z'y \quad (1)$$

in which, **y** is the vector of phenotypic records, **Z** is the incidence matrix relating the records to the random additive effects (**a**), σ_e^2 is the residual variance andσ_a^2 is the additive genetic variance. A gamma distribution for QTL effects with shape parameter equal to 0.4 was used. The mixed model equations were solved by the conjugate gradient method. For consecutive generations, all individuals from previous generations were included in the mixed model equations.

The statistical model included the random effects of SNPs, and was in fact a standard SNP-BLUP model. The proposed model was as follows:

$$y_i = \mu + Qg + e_i \quad (2)$$

in which, y_i, is the vector for the phenotypic values records for the quantitative traits, μ is the intercept, Q is the incidence matrix of the effects of SNPs, g is the unknown vector related to the effects of SNPs, and e_i is the residual effects. Genomic estimated breeding values were then estimated using the equation below:

$$y = \sum_n Qg \quad (3)$$

in which, y is the vector of GEBVs calculated from the sum of all n QTLs effects multiplied by their corresponding covariates coming from genotypes of each individuals in every given position.

There are different statistics to calculate the bias when predicting the genomic merit. The statistics developed are linear regression of true breeding values (TBVs) on estimated breeding values (EBVs), linear correlation coefficient between subsequent predictions, and variance of the genomic prediction differences (recent minus previous prediction) (Reverter et al., 1994). In a simulation study that the true QTL effects are known it is easy to use regression of true breeding values on estimated breeding values to calculate the bias. Therefore, we used this statistics bias. A regression coefficient of one denotes an unbiased prediction while any deviation from one indicates bias. Accuracy of evaluation was measured using correlation of TBVs and GEBVs.

To show the effect of selection on variance, the variance of allele frequencies was calculated using the following equation (Falconer, 1996)]:

$$\sigma_q^2 = q(1-q)/2N \quad (4)$$

in which, q is the frequency for the intended allele and N is the number of individuals in the population.

Results and discussion

Distribution of allele frequencies (Figure 1) as well as the average LD (Table 1) for the last generation of historical population showed a mutation-drift equilibrium in all loci and a fairly random assortments of haplotype blocks along genome.

Regression of TBVs on GEBVs for two scenarios and three traits in terms of their h^2 showed an unbiased prediction for GEBVs for all scenarios and traits in the first generation that SG was at random (Figure 2). The values were close to one regardless of SI and h^2. These findings were in agreement with many previous studies as expected from the theory of BLUP that is based on random selection of individuals to contribute to the genetic pool of the next generations (Vitezica et al., 2011; Zhao et al., 2012; Hsu et al., 2017). The regression values deviated rapidly from the unity and consequently the bias appeared when SG started from generation two and beyond based on higher breeding values

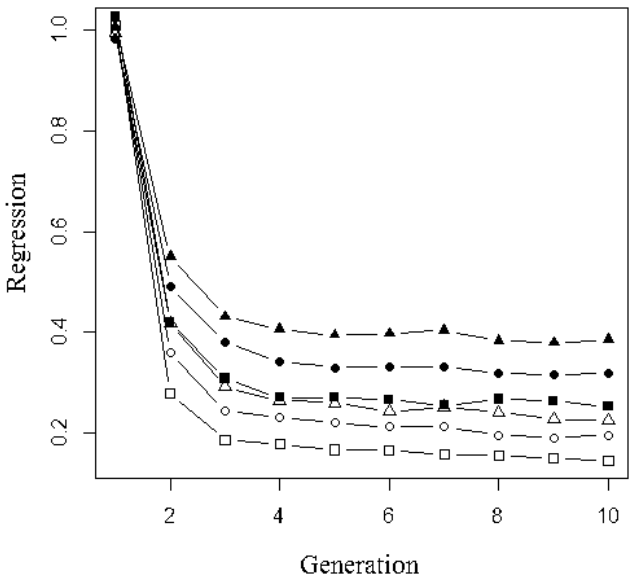

Figure 2. Trend of regression coefficient of true breeding values on genomic estimated breeding values in a population undergoing selection for 10 generations; close characters are for 50 percent selection proportion. Open characters are for 10 present selection proportion; rectangle, circles and triangles are for heritabilities of 0.1, 0.25 and 0.4, respectively.

instead of a random selection. Regressions dropped dramatically by almost a factor of two to three times but differed for every SI or h^2**.** The values became 0.42, 0.49, and 0.55 for SP50 and 0.27, 0.36, and 0.42 for SP10 for traits with h^2 of 0.1, 0.25 and 0.40, respectively (Table 2). It seems that truncation selection not only reduces the variance of population both phenotypically and genetically which is called Bulmer effect (Bulmer, 1971) but also is a key driver for bias in the genomic predictions (Van Grevenhof et al., 2012). It was reported that reduction in response due to the Bulmer effect is always larger for selection based on genetic merits of animals than for selection based directly on phenotypic information (Van Grevenhof et al., 2012). Moreover, the impact of the Bulmer effect on other parameters such as response to selection and accuracy of genomic predictions is the same for GS as for traditional BLUP selection (Van Grevenhof et al., 2012). Reduction of regressions were continued but with a lower rate in generations 3 and 4 (Figure 2). It is clear that after generation 4, the regressions were almost constant up to generation 10. The three times reductions in the regression values after only one generation of truncation selection in generation two, became to almost one tenth reduction for 6 generations of selection after generation 4 (generations 4 to 10). In generation 4, the values were 0.27, 0.34, and 0.41 for SP50 and 0.18, 0.23, and 0.26 for SP10 in the traits with h^2 of 0.1, 0.25 and 0.40, respectively (Table 2). The corresponding values in generation 10 were 0.25, 0.32, 0.38, 0.14, 0.19, and 0.22, respectively (Table 2). The results showed that from generation 1 to 4 the regression values reduced by 73, 66, and 60 percent for SP50 and 82, 77, and 74 percent for SP10 in the traits with h^2 of 0.1, 0.25 and 0.40, respectively. The corresponding reductions from generation 4 to generation 10 were only 2, 3, 2, 4, 4, and 4 percent, respectively. This pattern of reduction in the regression coefficients of TBVs on EBVs showed that the main biasedness happened at the early generations of selection after random population. The results indicated that bias was in agreement with the magnitude of SI which is lower in smaller SI and vice versa. But it was in contrast to h^2 where the bias was higher for lower heritable traits and vice versa. The results indicated that for a trait with a given heritability, for example 0.1, when the intensity of selection increased from 0.798 to 1.755 (50% selection proportion to 10% selection proportion) the bias increased harmoniously, because of the reduction in the genetic and ultimately phenotypic variances. This reduction in the genetic variance ended up with reduction in heritability due to selection. It can be concluded that any sources that change the variance and heritability, also affect the estimations which result in biased predictions. The results also showed the effect of increasing heritability in a given selection intensity which improved the estimations and helped to reduce the bias. The interesting point was that even though the reduction in variance was higher for higher heritable traits due to selection based on genetic merits, biasedness was lower for such traits. Nevertheless, the reduc-

Table 1. Average linkage disequilibrium (LD) for the last generation of historical population according to genetic distance between markers

Genome length (cM)	Average LD (R2)	Standard deviation
[0, 0.05)	0.264	0.0007
[0.05, 0.1)	0.169	0.0006
[0.1, 0.2)	0.098	0.0003
[0.2, 0.3)	0.056	0.0001
[0.3, 0.4)	0.038	0.0001
[0.4, 0.5)	0.029	0.0001
[0.5, 0.6)	0.023	0.0001
[0.6, 0.7)	0.019	0.0001
[0.7, 0.8)	0.017	0.0001
[0.8, 0.9)	0.015	0.0000
[0.9, 1)	0.013	0.0000
[1, 2)	0.009	0.0000
[2, 3)	0.005	0.0000
[3, 4)	0.004	0.0000
[4, 5)	0.003	0.0000

Table 2. The Regression of TBVs on GEBVs over generations 1 to 10

	50% Selection proportion			10% Selection proportion		
Generation	h^2=10	h^2=25	h^2=40	h^2=10	h^2=25	h^2=40
1	1. 03	0. 98	1. 00	1. 00	1. 00	0.99
2	0. 42	0. 49	0. 56	0. 28	0. 36	0. 42
3	0.31	0. 38	0. 43	0. 19	0.00	0. 29
4	0. 27	0. 34	0. 41	0. 18	0. 23	0. 26
5	0. 27	0. 33	0. 40	0. 17	0. 21	0. 26
6	0. 27	0. 33	0. 40	0. 16	0. 21	0. 24
7	0. 257	0. 32	0. 40	0. 16	0. 21	0. 25
8	0. 27	0. 32	0. 38	0. 15	0. 19	0. 24
9	0. 27	0. 32	0. 39	0. 15	0. 19	0. 23
10	0. 25	0. 32	0. 38	0. 14	0. 19	0. 22

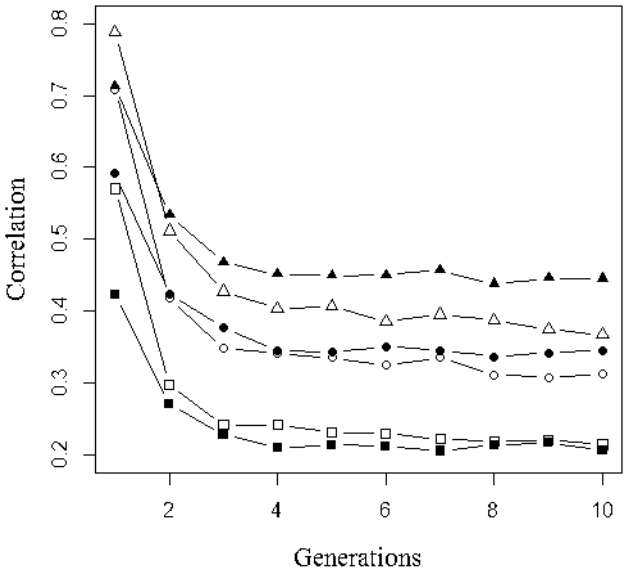

Figure 3. Trend of accuracy of genomic estimated breeding values in a population undergoing selection for 10 generations; Close characters are for 50 percent selection proportion. Open characters are for 10 present selection proportion; Rectangle, circles and triangles are for heritabilities of 0.1, 0.25 and 0.4, respectively.

tion in variance is a key driver for biased estimation of the breeding values, with greater resemblance between parents and offspring generation resulting in more accurate, less biased, predictions.

As shown in Figure 3, the accuracies of prediction were 0.42, 0.59, and 0.71 for SP50, and 0.57, 0.71, and 0.79 for SP10 for traits with h^2 of 0.1, 0.25 and 0.40, respectively (Table 3). The accuracy increased by increasing h^2 as expected, and was less dependent on SI while selection started based on higher EBVs from generation 2 onward. This phenomenon is mainly due to dramatic reduction of genetic variance as shown by calculations of the variance of allelic frequencies (Figure 4). The pattern of reduction in accuracies was the same as for the regression coefficients as well as for the variances of allelic frequencies (Figures 2, 3, and 4).

It is well known that reduction in the accuracy of genomic predictions is a result of selection (Bijma, 2012). It seems the main driver for such patterns is the genetic variance in each generation (Figure 5). Reduction in genetic variance results in the reduction of heritability as a consequence. Since selection impacts on the variance components, heritability estimates are less accurate in the population undergoing selection (Cesarani et al., 2019). It was reported that under different scenarios of genotyping, random selection results in the most accurate heritability estimates. Moreover,

Table 3. Correlations of TBVs and GEBVs over generations 1 to 10

	50% Selection proportion			10% Selection proportion		
Generation	h^2=10	h^2=25	h^2=40	h^2=10	h^2=25	h^2=40
1	0. 42	0. 59	0. 71	0. 570	0. 71	0. 79
2	0. 27	0. 42	0. 53	0. 30	0. 42	0. 5
3	0. 23	0. 38	0. 47	0. 24	0. 35	0. 43
4	0. 21	0. 34	0. 45	0. 24	0. 34	0. 401
5	0. 21	0. 34	0. 454	0. 23	0. 33	0. 41
6	0. 21	0. 35	0. 45	0. 23	0. 33	0. 39
7	0. 20	0. 35	0. 46	0. 23	0. 34	0. 40
8	0. 21	0. 34	0. 43	0. 22	0. 31	0. 39
9	0. 22	0. 34	0. 45	0. 22	0. 31	0. 37
10	0. 21	0. 34	0. 45	0. 21	0. 31	0. 37

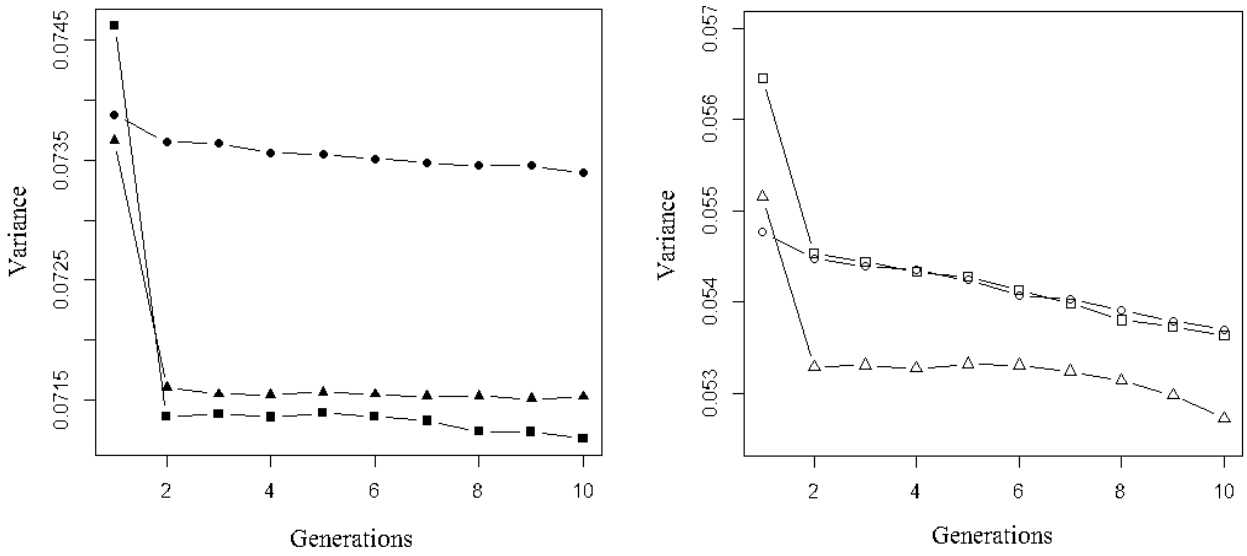

Figure 4. Trend of the change in variances of allele frequencies in a population undergoing selection for 10 generations; close characters are for 50 percent selection proportion (left). Open characters are for 10 present selection proportion (right); Rectangle, circles and triangles are for heritabilities of 0.1, 0.25 and 0.4, respectively. Variance of the change in allele frequencies calculated using the formula below: p×(1-p)/2N (Falconer, 1996).

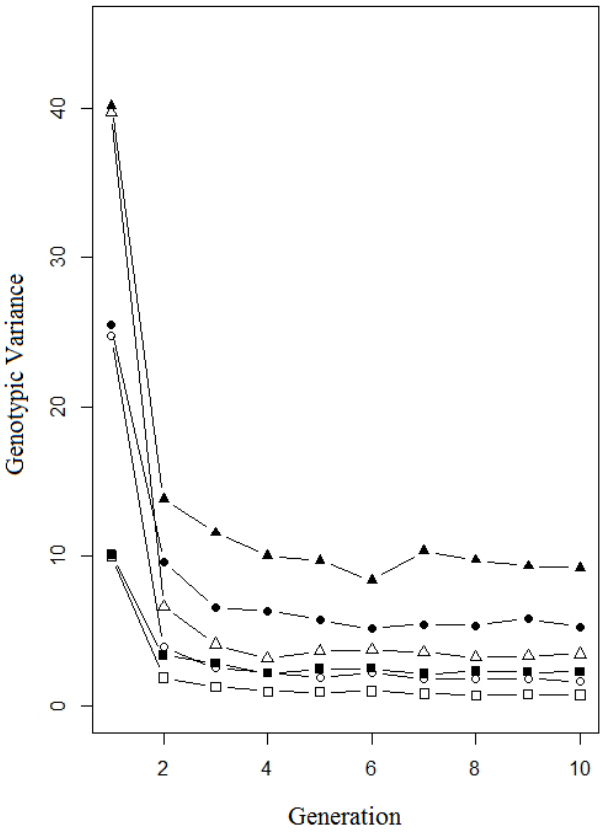

Figure 5. Trend of change in additive genetic variances over generations in the population undergoing selection for 10 generations. Close characters are for 50 percent selection proportion. Open characters are for 10 present selection proportion; Rectangle, circles and triangles are for heritabilities of 0.1, 0.25 and 0.4, respectively.

among statistical methods, the single-step GBLUP is less affected by selection in terms of variance components and heritability estimates (Cesarani et al., 2019). A most recent study showed that accounting for inbreeding and using single-step GBLUP reduces the biasedness and increases the accuracy of prediction (Gowane et al., 2019). Results showed that scaling down the GEBVs using a scale parameter helped removing the biasedness in generations4 onwards.

Conclusions

Trends of regression coefficients showed that the bias was large when selection started from a random population. Despite the fact that SI influenced bias, it was not the case for the accuracies after starting selection. Exploring the trends of genetic variances, shown by the variances of allelic frequencies, indicated that the main driver for accuracy and bias was the change in the genetic variance. Our data showed that the bias would not be a big issue anymore due to more homogeneous populations across generations.

References

Bijma, P., 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *Journal of Animal Breeding and Genetics* 129, 345-358.

Bulmer, M.G., 1976. The effect of selection on genetic variability. *Genetic Research* 28, 101-17.

Cesarani, A., Pocrnic, I., Macciotta, N.P.P., Fragomeni, B.O., Misztal, I., Lourenco D.A.L., 2019. Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies. *Journal of Animal Breeding and Genetics* 136, 40-50.

Ducrocq, V., 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *Journal of Dairy Science* 94, 1011-1020. Ehsani, A., Janss, L., Christensen, O., 2010. Effects of selective genotyping on genomic prediction. Processings of the 9th World Congress on Genetics Applied to Livestock Production. Germany.

Falconer, D.S., 1996. Introduction to Quantitative Genetics. Prentice Hall, Harlow, England.

Gowane, G.R., Lee, S.H., Clark, S., Moghaddar, N., Al-Mamun, H.A., van der Werf, J.H.J., 2019. Effect of selection and selective genotyping for creation of reference on bias and accuracy of genomic prediction. *Journal of Animal Breeding and Genetics* 136, 390-407.

Henderson, C.R., 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31, 423-447.

Hsu, W.L., Garrick, D.J., Fernando, R.L., 2017. The accuracy and bias of single-step genomic prediction for populations under selection. G3: Genes, Genomes, Genetics **7**, 2685-2694.

Kuehn, L.A., Lewis, R.M., Notter, D.R., 2007. Managing the risk of comparing estimated breeding values across flocks or herds through connectedness: a review and application. *Genetics Selection Evolution* 39, 225-247.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157, 1819-1829.

Meuwissen, T.H., Goddard, M.E., 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution* 36, 261-279.

Patry, C., Ducrocq,V., 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *Journal of Dairy Science* 94, 1011-1020.

Powell, J.E., Visscher, P.M., Goddard, M.E., 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics* 11, 800-805.

Reverter, A., Golden, B.L., Bourdon, R.M., Brinks, J.S., 1994. Technical note: detection of bias in genetic predictions. *Journal of Animal Science* 72, 34-37.

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. *Statistical Science* 6, 15-32.

Sargolzaei, M., Schenkel F.S., 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics* 25, 680-681.

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. *Journal of Dairy Science* 92, 16-24.

Van Grevenhof, E.M., Van Arendonk, J.A., Bijma, P., 2012. Response to genomic selection: The Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. *Genetics Selection Evolution* 44, 26.

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. *Genetics Research* 9, 357-366.

Zhao, Y., Gowda, M., Longin, F.H., Würschum, T., Ranc, N., Reif, J.C., 2012. Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theoretical and Applied Genetics* 125, 707-713.

Communicating editor: *Ali K. Esmailizadeh*