Original study

Effect of using different number and type of records from different generations as reference population on the accuracy of genomic evaluation

Azade Boustan[1], Ardeshir Nejati-Javaremi[2], Mohammad Moradi Shahrbabak[2] and Mahdi Saatchi[3]

[1]Department of Animal Science, Moghan college of Agriculture and Natural Resources, University of Mohaghegh Ardabili, Ardabil, Iran, [2]Department of Animal Science, University of Tehran, Karaj, Iran, [3]Department of Animal Science, Iowa State University, Ames, USA

Abstract

One important question about genomic evaluation is how the distance between generations of individuals in reference populations and selection candidates would affect the accuracy of the genomic estimated breeding value of selection candidates. There were two schemes in the present study. In the first scheme, a genome consisting of 30 chromosomes each with 100 equally-spaced single nucleotide polymorphisms (SNPs) for each individual and in the second scheme a genome consisting of 3 chromosomes each with 1 000 equally-spaced SNPs was simulated. To generate enough linkage disequilibrium between loci, random mating for 50 generations was done in a finite population. In generation 51, the population size was expanded to 250 individuals. This structure was continued until generation 55. Individuals in generation 55 were juvenile and did not have phenotypic records and were selection candidates. Heritability was assumed to be 0.3. Our results showed that using information from more distant generations would decrease the accuracy of genomic estimated breeding values of selection candidates but in the scheme in which the marker distance was 1 cM, an increasing generation number between reference population and selection candidates would decrease the accuracy more than in the scheme in which the marker distance was 0.1 cM. According to our results using estimated breeding values of a reference population instead of phenotypic records would increase accuracy extremely.

Archiv Tierzucht 56 (2013) 68, 684-690
doi: 10.7482/0003-9438-56-068

Received: 04 July 2012
Accepted: 20 December 2012
Online: 14 June 2013

Corresponding author:
Azade Boustan; email: boustan_62@yahoo.com
Department of Animal Science, Moghan college of Agriculture and Natural Resources, University of Mohaghegh Ardabili, Ardabil, Iran

Keywords: genomic, simulation, reference population, selection candidate, phenotypic record, breeding value

Abbrevations: EBV: estimated breeding value, GEBV: genomic estimated breeding value, LD: linkage disequilibrium, MAS: marker assisted selection, QTL: quantitative trait locus, SNP: single nucleotide polymorphism, TBV: true breeding value

Introduction

During the past decades, developments in molecular genetics have resulted in identification of several genes or genetic markers associated with genes that affect important traits in livestock (Dekkers 2004). Adoption of marker assisted selection (MAS) by the dairy industry was limited. There were several reasons for the limited use of MAS. Firstly, for many quantitative traits such as production and health traits in dairy cattle, a large number of loci are affecting the trait with any one locus capturing only a limited proportion of the total genetic variance. Consequently, relatively small gains were possible with the limited number of markers available. Secondly, the costs of genotyping these markers were high and thirdly, the complexity of calculating breeding values including marker information was a further limit to the application of MAS (Hayes *et al*. 2009).

Meuwissen *et al*. (2001) suggested a procedure to estimate breeding values using genome-wide dense markers. Because dense markers were used for calculating genomic estimated breeding values (GEBVs), every quantitative trait locus (QTL) was in population-wide linkage disequilibrium (LD) with some markers. Furthermore, because genome-wide markers were used, all QTLs were regarded simultaneously. Selection of animals based on their GEBVs was named genomic selection (de Roos 2011).

Genomic method has recently caused a revolution in the livestock genetic evaluation system. The genomic selection revolution began with advance in sequencing of the bovine genome, which led to the identification of many thousands of DNA markers in the form of single nucleotide polymorphisms (SNP) (Hayes *et al*. 2009). Possibly, genomic selection could double the genetic progress rate through selection and breeding from bulls at 2 years of age compared to 5 years of age or later (Schaeffer 2006). Bull breeding companies could save up to 92 % of their costs by avoiding progeny testing (Schaeffer 2006).

Some factors affect accuracy of GEBVs such as heritability of the trait (Calus & Veerkamp 2007), marker density (Solberg *et al*. 2008), generation distance between individuals in a reference population and selection candidates as well as the number of phenotypic records in the reference dataset (Meuwissen *et al*. 2001). In one study, 1 000 SNPs on 10 chromosomes were simulated for each individual in the population. When the number of records in the reference population was reduced from 2 200 to 500, the correlation between true breeding value (TBV) and estimated breeding value (EBV) of selection candidates was reduced about 14-19.4 % for different methods of estimation of GEBVs (Meuwissen *et al*. 2001).

In practice, one of the restricting factors to increase the number of genotyped individuals in a reference population is the cost of genotyping. Moreover, individuals with phenotypic records may belong to different generations. One important question is whether and how generation distance between individuals in a reference population and selection candidates would affect the accuracy of GEBVs of selection candidates.

The objectives of this study were to explore 1) the effect of generation number between individuals in a reference population and selection candidates, 2) the effect of number of phenotype records or EBVs in a reference population on accuracy of GEBVs of selection candidates and 3) comparison between using phenotype records and EBVs in a reference population for the effect on the accuracy of GEBVs of selection candidates.

Material and methods

In the present study, 3 000 SNPs were simulated for each individual. In one scheme, 30 chromosomes each with 100 equally-spaced SNPs (each cM one SNP) and in another scheme 3 chromosomes each with 1 000 equally-spaced SNPs (each 0.1 cM one SNP) were generated. In two schemes a total number of 50 QTLs (that distributed on chromosomes randomly) was generated. Only the additive genetic effect was considered. Gene effects for each QTL were assigned randomly from a standard normal distribution. Fifty QTLs covered the total genetic variance. True breeding value for each individual was assumed to be the sum of effects of these QTLs. Single nucleotide polymorphisms and QTLs were assumed to be biallelic with equal initial allelic frequencies. A population that consisted of 50 males and 50 females was simulated. To generate enough LD between loci, random mating for 50 generations was done. Our criterion for calculation of LD in generation 50 was r^2. The formula for calculation of r^2 was as follows (Hill & Robertson 1968):

$$r^2 = \frac{D^2}{m} \tag{1}$$

$$m = f(A1).f(A2).f(B1).f(B2) \tag{2}$$

$$D = f(A1B1).f(A2B2) - (A1B2).f(A2B1) \tag{3}$$

In this formula, *f* shows the frequency, for example *f* (A1) is the frequency of *A1* and *f (A1B1)* is the frequency of *A1B1* haplotype.

To generate recombinant haplotypes for each individual, Haldane mapping function was used. After 50 generations, five generations each with 250 individuals (125 females and 125 males) were simulated. Individuals in four generations had marker information, phenotypic records and EBVs and they were used as reference population. Individuals in the 5th generation were selection candidates. They did not have phenotypic records or EBVs. They only had marker information. We used the BLUP method proposed by Meuwissen *et al*. (2001) in this study for calculation of GEBVs of selection candidates (Meuwissen *et al*. 2001). Meuwissen *et al*. (2001) concluded that the BLUP method results in a reasonably high accuracy of predicting TBV (Meuwissen *et al*. 2001). Following formula for calculation of marker effects was used:

$$y = Xb + Zm + e \tag{4}$$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I_{\gamma} \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \tag{5}$$

$$\gamma = \frac{\sigma_e^2}{\sigma_g^2} \tag{6}$$

Where *X* and *Z* are coefficient matrices, *b* is the vector of means, *y* is the vector of observations, *m* is the vector of random marker effects and *e* is the vector of random residual effects.

Marker effects were obtained from data of the reference population. Genomic estimated breeding value for each selection candidate was obtained from the following formula:

$$GEBV_i = Z_i \hat{m}_i \tag{7}$$

The vector of observations was phenotypic records or EBVs of the reference population. Estimated breeding values were simulated with these reliabilities from simulated TBVs using following formula (Saatchi 2009):

$$EBV = (TBV \times \sqrt{REL}) + (Error \times PEV) \tag{8}$$

$$PEV = \sqrt{(1 - REL)} \times STD^2 \tag{9}$$

Where *REL* is the reliability of EBVs, *PEV* is the predicted error variance and *STD* is the genetic standard deviation of trait. Error was a random sample from a normal distribution with mean zero and standard deviation of 1.

The heritability of the trait in the present study was assumed to be 0.3. Reliabilities of EBVs were assumed to be 0.93, 0.86, 0.75 and 0.70 for male animals for generation 1, 2, 3 and 4 and 0.80, 0.75, 0.70 and 0.65 for female animals, respectively (Khansefid 2010). Individuals in generation 5 (young animals) did not have phenotypic records. They were selection candidates and GEBVs were calculated for them.

The accuracy of estimation of GEBVs was the correlation between TBVs and GEBVs. In the present study, single SNP genotypes of animals were used. We replicated each simulation 10 times and an average of 10 replicates was reported.

Results and discussion

The considered LD after 50 generations of random mating between 100 individuals was 0.19 and 0.13 when the marker distance was 0.1 cM and 1 cM, respectively. Accuracies of GEBVs in selection candidates using different numbers of individuals from different generations with 0.1 cM and 1 cM marker distance are shown in Tables 1 and 2.

Our results indicated that using information of generations closer to the generation of selection candidates, results in more accurate GEBVs compared to using information of more distant generations in the reference population. One reason for this result is that LD between markers and QTL affect the accuracy of the genomic method (Habier *et al.* 2007) and an increasing generation between the reference population and selection candidates would decrease LD. The other reason for decreasing GEBV accuracy by using information of more distant generations could be a weaker relationship between individuals of selection candidates and the reference population.

Table 1
Accuracy of GEBVs (± SE) in selection candidates using different number of individuals from different generations with 0.1 cM marker distance

Observation vector	Reference population				
	Generations 51-54 (1 000 individuals)	Generations 51-52 (500 individuals)	Generations 53-54 (500 individuals)	Generation 51 (250 individuals)	Generation 54 (250 individuals)
Breeding value	0.84 ± 0.02	0.73 ± 0.04	0.79 ± 0.02	0.66 ± 0.05	0.79 ± 0.03
Phenotype	0.72 ± 0.03	0.59 ± 0.06	0.66 ± 0.04	0.46 ± 0.05	0.56 ± 0.04

Table 2
Accuracy of GEBVs (± SE) in selection candidates using different number of individuals from different generations with 1 cM marker distance

Observation vector	Reference population				
	Generations 51-54 (1 000 individuals)	Generations 51-52 (500 individuals)	Generations 53-54 (500 individuals)	Generation 51 (250 individuals)	Generation 54 (250 individuals)
Breeding value	0.66 ± 0.04	0.45 ± 0.05	0.66 ± 0.04	0.32 ± 0.04	0.65 ± 0.04
Phenotype	0.52 ± 0.04	0.29 ± 0.06	0.50 ± 0.06	0.20 ± 0.05	0.46 ± 0.05

It is obvious from our results that an increasing marker density would increase the accuracy of GEBVs of selection candidates. This is in agreement with results of Calus *et al.* (2008). Our results also showed that in the scheme, in which the marker distance was 1 cM, an increasing generation number between the reference population and selection candidates would decrease the accuracy of GEBVs of selection candidates more than the scheme in which the marker distance was 0.1 cM. Hayes (2007) stated that if the number of available markers per chromosome is limited, the association between the markers and the QTL will persist only for a limited number of generations due to recombination (Hayes 2007).

Our results also showed that an increasing number of individuals in the reference population would increase the accuracy of GEBVs of selection candidates. Meuwissen *et al*. (2001) used equally-spaced markers (each 1 cM) in a simulation study with different numbers of phenotypic records in the reference population to estimate GEBVs for selection candidates. Based on BLUP evaluation, accuracy of GEBVs was 0.579, 0.659 and 0.732 when the reference population had 500, 1 000 and 2 200 records, respectively. They concluded that increasing individuals in the reference population would result in increased accuracy of GEBVs of selection candidates (Meuwissen *et al*. 2001). One difference between accuracies in the research of Meuwissen *et al*. (2001) and the present study is that the heritability of trait in their research was 0.5 and in the present study it was 0.3. Calus & Veerkamp (2007) also concluded that an increase in individuals in the reference population would result in increased accuracy of GEBVs of selection candidates (Calus & Veerkamp 2007). Saatchi (2009) used equally-spaced markers (each 1 cM) in a simulation study with different numbers of phenotypic records in the reference population. When there were 500 individuals from two former generations in the reference population and the heritability of the trait was 0.5 and 0.1, the accuracy of GEBVs of selection candidates was 0.57 and 0.31, respectively and when there were 1 000 individuals from four former generations in the reference population and the heritability of the trait was 0.5 and 0.1, the accuracy of GEBVs of selection candidates was 0.63 and 0.37, respectively. Therefore, he concluded that an increase in individuals in the reference population would result in increased accuracy of GEBVs of selection candidates (Saatchi 2009). Muir (2007) also showed that increasing individuals in the reference population would increase accuracy of GEBVs of selection candidates (Muir 2007).

Our results also showed that using EBVs of the reference population instead of phenotypic records would greatly increase GEBVs of selection candidates. This is probably due to the fact that some environmental effects are omitted from EBVs but phenotypic records do have these environmental effects. In conclusion, an increasing number of individuals in the reference population would increase the accuracy of GEBVs but the cost of genomic evaluation is high. Therefore, using EBVs of individuals instead of phenotypic records and using information of generations closer to the generation of selection candidates is strongly recommended.

References

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178, 553-561

Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Anim Breed Genet 124, 362-368

Dekkers JCM (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J Anim Sci 82 (Suppl.), E313-E328

De Roos APW (2011) Genomic selection in dairy cattle. Ph.D. thesis, Wageningen University, the Netherlands

Habier D, Fernando RL, Dekkers JCM (2007) The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 177, 2389-2397

Hayes BJ (2007) QTL Mapping, MAS and Genomic Selection. A short course organized by Animal Breeding & Genetics. Department of Animal Science, Iowa State University June 4-8 2007, USA

Hayes BJ, Bowman PJ, Chamberlain AC, Goddard ME (2009) Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92, 433-443

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor App Genet 38, 226-231

Khansefid M (2010) [Genetic evaluation of animals based on sires' genotypes for dense markers using computer simulation]. MSc, Tehran University, Iran [in Persian]

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819-1829

Muir WM (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

Saatchi M (2009) [Estimation of breeding values using dense marker information in dairy cattle population]. PhD, Tehran University, Iran [in Persian]

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123, 218-223

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86, 2447-2454